

---

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES  
DER FAKULTÄT FÜR CHEMIE UND PHARMAZIE  
DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

DEVELOPMENT OF SENSITIVE AND QUANTITATIVE  
PROTEOMICS STRATEGIES TO STUDY PHOSPHO- AND  
UBIQUITIN-SIGNALING IN HEALTH AND DISEASE

ÖZGE KARAYEL EREN

AUS  
KONAK, IZMIR, TURKEY

2020

---

## **Erklärung**

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Professor Dr. Matthias Mann betreut.

## **Eidesstattliche Versicherung**

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.  
München, den 07.12.2020

---

Özge Karayel Eren

Dissertation eingereicht am 02.09.2020

1. Gutachter: Prof. Dr. Matthias Mann
2. Gutachter: Prof. Dr. Brenda A. Schulman

Mündliche Prüfung am 30.09.2020

## Abstract

Post-translational modifications (PTMs) are vital control mechanisms which govern all major aspects of cellular life, including cell division and protein degradation. They play a key role in increasing functional diversity of the proteome by regulating activity, localization, and interaction of proteins with other cellular molecules. Aberrant PTMs are associated with a variety of human diseases, including cancer and neurodegeneration. To better understand the biological function of PTMs in health and disease, it is crucial to identify the various modifications of proteins. Mass spectrometry (MS)-based proteomics is the method of choice to quantitatively study PTMs on a global scale as well as in a targeted manner. Major advances in enrichment strategies, acquisition methods, instrument performance, and computational analysis tools now enable the in depth analysis of PTMs and transformed the cell signaling field. However, studying PTMs by MS remains challenging mostly due to the complexity of workflows, relatively low sensitivity and data incompleteness in single run analyses. In this thesis, I aimed to develop quantitative methods that allow the analysis of phospho and ubiquitin modified proteomes with very high accuracy and sensitivity. Furthermore, I applied these methods to study various challenging biological and pathophysiological processes including erythropoiesis and neurodegeneration.

Firstly, we developed a highly sensitive workflow for large-scale quantitative phosphoproteomics based on the EasyPhos platform [1]. We simplified the previously published workflow substantially and, in parallel, made it more streamlined, scalable, and applicable for higher sensitivity phosphoproteomics. Our optimized protocol requires only a few hundreds of micrograms protein material and takes about a day. It enables the analysis of phosphoproteomes at a depth of tens of thousands of quantified phosphorylation sites. Encouraged by the high sensitivity and reproducibility of our protocol, we applied it to study human erythropoiesis, so far poorly understood globally with regards to phosphorylation. We combined fluorescence-activated cell sorting (FACS)-based cell enrichment with our EasyPhos protocol and the latest proteomics methods. We quantified and dynamically tracked 7,400 proteins and 27,000 phosphorylation sites of five distinct maturation stages of in vitro reconstituted erythropoiesis of CD34+ HSPCs. Our system-wide analysis shed unprecedented light on the stage-specific dynamic developmental regulation of human erythropoiesis at the protein and post-translational levels.

In addition to large-scale strategies for the analysis of protein phosphorylation, we also developed an ultra-sensitive MS-based targeted assay to accurately quantify known site-specific phosphorylations. I used this assay to uncover potential therapeutic targets of the LRRK2 kinase and to demonstrate their relevance as biological markers for Parkinson's disease (PD) in the clinics. Previously, we had identified several Rab GTPases (Rab8A, Rab10, and Rab12) as physiological targets of the Parkinson's disease kinase LRRK2 [2]. Interestingly, the LRRK2 phosphorylation site on Rab proteins is highly conserved in ~50 Rab proteins. Our systematic

## Abstract

---

proteomics analysis using this targeted assay extended the list of Rab protein family members phosphorylated by LRRK2 at the endogenous level. To monitor LRRK2 activity accurately and assess LRRK2 inhibitor efficacy in clinical samples, we modified the assay to determine the phosphorylation occupancy, which represents the percentage of phosphorylated proteins. Indeed, our sensitive and highly accurate MS-based assay successfully measured the stoichiometry of phosphorylated Rab proteins in neutrophil cells collected from PD patients and healthy controls. We showed that pathogenic mutation carriers had more than two-fold increased pRab10 levels compared to healthy subjects, confirming the relevance of pRab10 as treatment efficacy and stratification marker. It received much attention through the Michael J. Fox foundation network, dedicated to combat PD, and will now be used by the PD community at large.

In the second part of the thesis, I present my work on different aspects of ubiquitin signaling, including the identification of ubiquitinated proteins and of the ubiquitin chain architecture. To improve the identification and quantification of ubiquitination sites we set out to investigate the power of a new state-of-the-art MS acquisition method, so called data-independent acquisition (DIA). This approach increased the reproducibility, quantitative accuracy, and depth of ubiquitinome analysis in single runs compared to the conventional data-dependent acquisition method. The application of our DIA-based workflow to the circadian rhythm for the first time revealed hundreds of cycling ubiquitination sites and dozens of cycling ubiquitin clusters within individual membrane protein receptors and transporters. Secondly, we studied modification by UBL3, which is a novel ubiquitin-like molecule. We used a comprehensive MS-based proteomic analysis to better understand its physiological function and identified proteins that interact with this unusual modification in human cells. In further collaboration with the Schulman group, we established a fast and robust DIA workflow to identify the components and targets of the highly conserved Gid E3 ligase complex - a novel family of multisubunit E3s - that play an essential role in glucose-induced degradation (Gid) of Fbp1 under different metabolic conditions. Lastly, we developed a targeted MS strategy to quantify the distribution of different ubiquitin chain linkage types to investigate the potential roles of the acceptor lysine architecture on ubiquitylation.

Towards the end of my PhD journey early this year, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged and quickly spread, becoming a major global health crisis of our time. This sudden global emergence of SARS-CoV-2 urgently required an in-depth understanding of molecular functions of viral proteins and their mechanisms of host manipulation. To contribute to this endeavor, together with the group of Andreas Pichlmair from the Technical University Munich, we applied a host of diverse sensitive and quantitative proteomics strategies. Thereby we characterized how SARS-CoV-2 and the related coronavirus SARS-CoV manipulate the host proteome and phospho- and ubiquitin-signaling in a systems-wide manner. Our findings are of great utility for the scientific community and will provide an important basis for further translational research.

**Table of Content**

1. Introduction .....	1
1.1. Mass spectrometry (MS)-based proteomics .....	1
1.1.1. Bottom-up or 'shotgun' MS-based proteomics workflow.....	2
1.1.1.1. Sample preparation methods .....	3
1.1.1.2. Liquid chromatography-mass spectrometry (LC-MS) .....	5
1.1.1.3. Tandem MS and fragmentation techniques .....	6
1.1.1.4. Acquisition methods .....	7
1.1.1.5. Protein identification.....	9
1.1.1.6. Protein quantification.....	10
1.1.2. In-depth cellular and bio-fluid proteomes and biomarker discovery .....	12
1.2. Post-translational modifications (PTMs) and their role in signaling .....	13
1.2.1. Phosphorylation .....	14
1.2.1.1. Protein phosphorylation-mediated cellular signaling.....	15
1.2.1.2. Phosphoregulators: kinases and phosphatases .....	17
1.2.1.3. Quantifying phospho-signaling by MS: phosphoproteomics .....	18
1.2.1.4. Phosphoproteomics workflows and enrichment techniques .....	19
1.2.2. Parkinson' disease and phosphorylation .....	22
1.2.2.1. Parkinson' disease .....	22
1.2.2.2. LRRK2 kinase in PD and its cellular targets .....	22
1.2.2.3. LRRK2 inhibitors and clinical trials.....	24
1.2.2.4. Biomarkers of Parkinson's disease .....	25
1.2.3. Ubiquitin and ubiquitin-like modifications .....	27
1.2.3.1. Overview of the ubiquitin system .....	27
1.2.3.2. Cellular ubiquitin-signaling .....	29
1.2.3.3. Quantifying ubiquitin-signaling by MS-based proteomics .....	30
2. Aims of the thesis.....	33
3. Publications .....	35
3.1. Phosphorylation .....	35

## Table of Content

---

3.1.1. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform	35
3.1.2. Integrative proteomics reveals principles of dynamic phospho-signaling networks in human erythropoiesis .....	55
3.1.3. Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis.....	81
3.1.4. Accurate MS-based Rab10 phosphorylation stoichiometry determination for LRRK2 activity in Parkinson's disease .....	103
3.2. Ubiquitin or Ubiquitin-like modification.....	133
3.2.1. Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology .....	133
3.2.2. UBL3 modification influences protein sorting to small extracellular vesicles .....	157
3.2.3. DIA-based systems biology approach unveils novel substrates of the GID E3 ligase in response to a metabolic shift.....	170
3.2.4. Interconversion between Anticipatory and Active GID E3 Ubiquitin Ligase Conformations via Metabolically Driven Substrate Receptor Assembly .....	191
3.2.5. Linkage-specific ubiquitin chain formation depends on a lysine hydrocarbon ruler ..	206
3.2.6. Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV	225
4. Discussion and Outlook .....	249
5. References .....	252
Appendix .....	260
Acknowledgement .....	289

## 1. Introduction

### 1.1. Mass spectrometry (MS)-based proteomics

Complete human genome sequences have revealed the existence of over 20,000 protein-coding genes [3]. However, the number of genes alone does not correlate with the complexity of an organism. The human proteome in particular is diversified by protein variations via alternative splicing (isoforms, multiple protein products from a single gene), and post-translational modifications (PTMs) such as phosphorylation. These modifications together yield millions of different possible ‘proteoforms’, adding orders of magnitude of complexity to the human proteome [4].

During the 1990s, changes in mass spectrometry (MS) instrumentation and techniques revolutionized protein analysis. Especially in the last decades, MS-based methods have gained comparable popularity as next-generation sequencing in genomics and transcriptomics [5]. MS-based strategies have developed to explore complete proteomes of unicellular and higher organisms. Draft maps of the human proteome were generated by proteomic profiling of human tissues and primary cells using high-resolution mass spectrometry [6, 7]. These studies provided spectral evidence for the expression of proteins encoded by over 17,000 genes, accounting for up to 84% of the annotated protein-coding genes in the human genome [6], although reanalysis of the data revised this number down to about 12,000 genes [8].

MS technology determines the mass/charge ( $m/z$ ) ratios of ions. Depending on the charge state of a measured ion, the  $m/z$  ratio is converted into its molecular mass with a standard unit of Dalton (Da). In MS-based proteomics, peptides must first be ionized in their intact form; however, ionization had been unachievable for labile biomolecules. This obstacle was overcome with the development of two soft ionization techniques (electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI)) in the late 1980’s [9, 10]. John Fenn received the Nobel Prize in Chemistry in 2002 for the discovery of ESI. In this technique, analytes (peptides) are directly ionized from a liquid phase using a high voltage via rapid solvent evaporation. A very high potential (kilovolt) is applied at the end of the capillary column, generating a spray of charged droplets. Upon solvent evaporation, the charge density increases, resulting in charged ions that are transferred into the vacuum of the mass spectrometer. ESI was a milestone for the field of MS-based proteomics and has become very popular since it can directly be coupled to a liquid chromatography (LC) system, which is ideally suited to the analysis of complex protein and peptide mixtures.

In principle, there are two MS-based proteomic strategies top-down and bottom-up proteomics dealing with proteins and peptides, respectively. In the analysis of intact proteins from complex

biological systems, top-down proteomics is typically combined with an up-front purification step [11]. This approach in theory allows for almost complete coverage of protein sequence and is used to characterize protein isoforms and complexes, unusual PTM conformations, and especially therapeutic antibodies. However, it remains experimentally and computationally challenging due to poor ionization properties, intricate charge patterns, and MS/MS spectra of high molecular mass organic molecules. In contrast, bottom-up (shotgun) proteomics refers to the proteolytic digestion of proteins before analysis by MS [12]. The term implies that information on proteins is reconstructed from measured and identified peptide fragments. This approach is by far the most widely used for in-depth and tissue-specific large-scale proteome analysis. Unlike the top-down approach, sophisticated instrumentation and expertise to interpret complex MS/MS spectra are not needed. However, the main limitation of this approach is that the limited coverage of the protein sequence is usually obtained since only a small and variable proteome fraction can be recovered. This may cause the loss of significant amount of information about PTMs and alternative splice variants.

### **1.1.1. Bottom-up or ‘shotgun’ MS-based proteomics workflow**

A typical bottom-up MS-based proteomics workflow has three main steps: (i) sample preparation, (ii) LC-MS/MS analysis, and (iii) data analysis (Figure 1). Sample preparation includes the extraction of proteins from biological material and the digestion of intact proteins into peptides. Peptides are first separated by reversed-phase high-performance liquid chromatography (HPLC) based on their hydrophobicity, typically using C<sub>18</sub> modified silica. As the peptides elute from the chromatographic column, they are ionized via ESI and analyzed by mass spectrometry. Proteins are identified from the generated peptide mass spectra that provide information about the abundance, amino acid sequence and PTMs when combined with database searching.

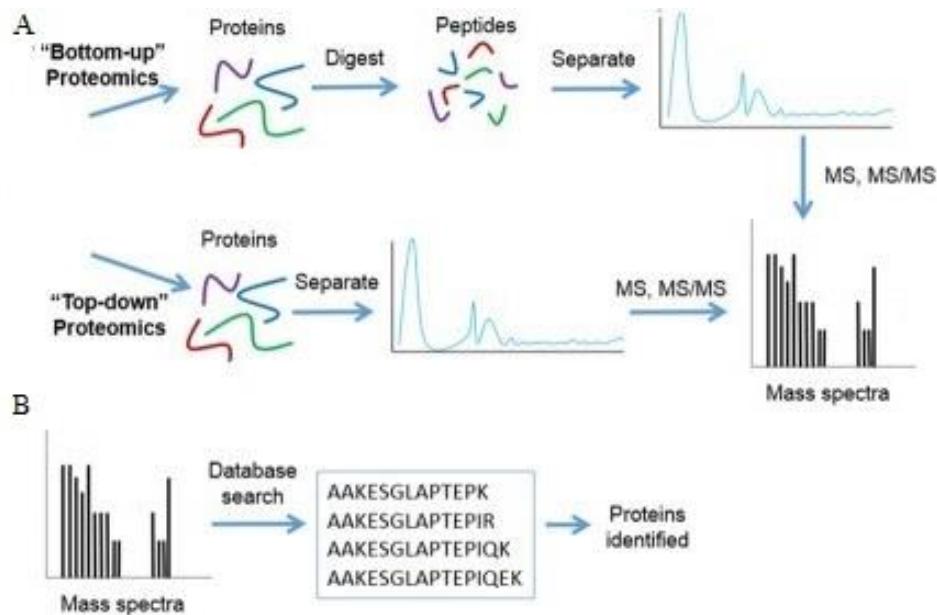


Figure 1. Top-down vs. bottom-up proteomics workflows (adapted from <https://www.creative-proteomics.com/blog/>).

### 1.1.1.1. Sample preparation methods

Sample preparation is a crucial step in shotgun proteomics. To identify thousands of proteins from a complex lysate, we need robust sample preparation methods. These include protein extraction, reduction of disulfide bonds, and selective alkylation of cysteines without non-specific modification of other amino acids, reproducible proteolysis (digestion), and clean-up for complete removal of contaminants including detergents, lipids, and salts before LC-MS/MS analysis. Sample preparation protocols need to be adapted to sample type and amount. Yet, regardless of the protocol used, the quality of MS analysis and reliability and accuracy of results are highly dependent on consistent sample preparation. Any variability associated with this step should be further addressed by the community for this technology to reach its full potential in research and clinical settings.

All protocols first require the lysis of the biological material and efficient extraction of proteins. Depending on the sample type, mechanical breakdown, such as sonication, bead-milling, or heating that increases lysis efficiency can be performed. Next, the cysteines of the extracted proteins are reduced and alkylated to disrupt disulfide bridges and prevent reforming of these reduced disulfide bridges. Reducing agents such as tris(2-carboxyethyl)phosphine (TCEP) and dithiothreitol (DTT) and alkylating agents such as iodoacetamide (IAA) and chloroacetamide (CAA) are typically used in proteomics. For proteolytic digestion of proteins, trypsin is the enzyme of choice. It has high cleavage specificity for the C-terminus of lysines and arginine. It sometimes generates peptides that are not of ideal length (less than seven or longer than 25 amino acids)

for effective and informative MS analysis. In that case, other enzymes such as chymotrypsin, AspN, Lys-N, Lys-C, Arg-C, or Glu-C can be employed to increase overall protein sequence coverage [13].

Protein analysis by SDS polyacrylamide gel electrophoresis is a routine method in biology research. The development of sample preparation workflows utilizing SDS polyacrylamide gel electrophoresis have accelerated MS-based proteomics in its early days [14, 15]. In such methods, proteins present in the gel are directly degraded by trypsin. They also allow fractionation of the proteome through the excision of the entire 1D gel in several pieces. However, in-gel digestion limits throughput, reproducibility and unbiased protein analysis. In-solution digestion largely overcomes these problems. Chaotropic agents, such as urea, are widely employed as they can be directly used for extraction and digestion of proteins in one buffer. However, MS-compatible detergents such as sodium deoxycholate (SDC) perform better compared to urea and they were shown to enhance trypsin activity many-fold at low concentration (<1%). They can be removed by acidification. Another protocol that uses SDS for complete protein solubilization is the 'Filter-Aided Sample Preparation' [16]. In FASP, SDS can be washed away through multiple washing steps with urea because it repurposes spin-filter matrices for the removal of detergents and chaotropic agents, protein digestion, and the isolation of peptide fractions.

The development of a tip-based peptide micro purification system named Stop and Go Extraction tips (StageTips), which consists of a tiny disk of membrane-embedded separation material, was a cornerstone in the field [17]. It allowed convenient sample handling and a universal sample preparation system for proteomics. It is commonly used for multidimensional fractionation as well as desalting, filtration, and concentration before mass spectrometry analysis. Moreover, StageTip protocols employing slightly milder detergents than SDS, such as sodium deoxycholate (SDC), allowed to perform the whole sample preparation steps in the tips (iST, 'in-StageTip') [18]. This protocol addressed some of these challenges for effective, high-throughput, and reproducible workflows as it considerably reduced sample preparation time, contamination, and loss. Further developments in fast and high-throughput sample preparation methods [19], such as MSTern [20], MS3 [21] and PAC (protein aggregation capture) [22] also addressed several problems of long and laborious classical sample preparation methods, enabling in-depth characterization of proteins and revealing the complexity of the human proteome with identification of over a thousand of distinct proteins.

Peptide fractionation before LC-MS/MS analysis can obtain deeper proteome coverage, especially in complex biological samples. Separation helps to decrease sample complexity in each MS run and increase the total amount of material injected onto the analytical column and overall detectability of low abundant peptides. There are several different fractionation principles,

including high pH reversed-phase fractionation or strong cation exchange (SCX) into proteomics workflows. Some of these techniques will be discussed in detail in the ‘phosphorylation’ section. We should here at least mention that off-line high pH reversed-phase fractionation methods are particularly popular in the field as they provide orthogonal separation when combined with the low pH separation of the online LC-MS analysis, yielding overall better peptide identifications [23]. Our group has developed a ‘loss-less nano-spider’ fractionator, which automatically concatenates the collected fractions via a rotating valve [24]. This fractionator enabled the quantification of around 12,000 proteins from very low- $\mu$ g starting peptide material.

### **1.1.1.2. Liquid chromatography-mass spectrometry (LC-MS)**

The LC-MS technique is a dual-selectivity and powerful analytical tool. The LC part separates compounds by their physicochemical properties related to the liquid phase whereas MS measures each component of a complex mixture, specifically their mass-to-charge ratio. The peptide mixtures are subjected to LC separation based on the different hydrophobic interactions with a stationary phase, which is typically octadecyl carbon chain (C<sub>18</sub>)-bonded silica. In reverse-phase chromatography, the sample is separated based on the molecule’s polarity preference to either the polar mobile phase or the non-polar stationary phase. A linear increase in the percentage of organic solvent such as acetonitrile in the aqueous buffers provides elution of peptides from a reversed-phase column. The addition of formic acid to the solutions provides a source of protons, hence improving the ionization efficiency. As peptides elute from the chromatographic column, they are ionized via ESI, and the resultant charged ions are then transported via an ion transfer tube into the vacuum part of the mass spectrometry. Long columns (around 40-50cm) packed with small sized particles require high pressure but are commonly used since they provide better chromatographic resolution and a lower number of co-eluting peptides.

Mass spectrometers can include many types of mass analyzers depending on the method which is chosen to separate and measure ions [25]. Single mass analyzers such as quadrupole and time-of-flight (TOF) were commonly used for measuring peptide species. Ion trap MS systems temporarily accumulate ions of a selected range before separating them by mass. The quadrupole mass analyzer contains four parallel cylindrical metal rods (electrodes) inside a vacuum chamber. These four rods are positioned equidistant from the center axis [25]. Currents are applied to the quadrupole so that ions with specific m/z can pass through and reach the detector. The quantity of ions that reach the detector is converted to a signal. This type of analyzer has high reproducibility and sensitivity but poor resolution and speed. TOF analyzers separate ions with different m/z by measuring the time taken for the ions to travel through a field-free region [25]. Although these analyzers have the highest scanning speed, until recently they suffered from low mass resolution.

Over the last decade, new developments such as Orbitrap devices (a type of ion trap mass analyzer) have enabled LC-MS instruments to rapidly acquire a full range of spectral information with high resolving power. They provide high selectivity and the capability for accurate mass measurement. Currently, they are the most commonly used mass analyzer in proteomics. They consist of a central spindle and two outer electrodes that trap ions in an orbital motion around the spindle [25]. The frequencies of oscillation of ions along the length of the spindle are proportional to the square root of their masses and converted to a mass spectrum using the Fourier transform (FT). Coupling FT to a phased spectrum deconvolution method ( $\Phi$ SDM) has been shown to double the mass resolution without increasing the cycle times; however, it requires extremely high computational power [26].

### **1.1.1.3. Tandem MS and fragmentation techniques**

Mass alone is not sufficient for complete characterization of a peptide. Therefore, MS instruments first record the  $m/z$  values of the precursor ions (intact peptides), which is referred to as MS1 and then do tandem MS (MS/MS, MS2), meaning to conduct multiple rounds of mass spectrometry. Selected ions (precursor ions) are broken down into fragments (product ions) in a collision cell and the masses of the resulting fragment ions are analyzed to reveal the chemical structure of the precursor ion (Figure 2). This is necessary to improve the specificity of the mass spectrometer for peptide identification. In tandem mode, the TopN most abundant peptides from each MS1 scan (full scan) are typically isolated to be fragmented. Therefore, this approach is somewhat biased toward the analysis of the most abundant, higher intensity peptides. Cycle times can be determined by the number of peaks to be fragmented (N) and the transient times for MS1 and MS2 scans in the case of Orbitrap analysis. There are various methods for fragmenting peptides in tandem MS, including collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), and electron-transfer dissociation (ETD) [27]. Each generates different ion series required for peptide sequence identification and unambiguous PTM site assignment.

In CID, fragmentation of precursor ions happens through collisions with an inert gas such as helium, leading to peptide bond dissociation and generation of N-terminal b- and C-terminal y-type ions [27]. It is mostly biased for the more effective fragmentation of small and low-charge state peptides and profoundly affected by the distribution of positive charges along the peptide backbone. HCD is a CID technique but characterized by higher activation energy compared to CID and specific to the Orbitrap mass spectrometers [27]. HCD, in conjunction with the Orbitrap mass analyzer, displays superior resolution and mass accuracy, producing very high-quality spectra. In HCD, ions traverse the C-trap to enter the HCD cell where fragmentation takes place. After fragmentation, the product ions are returned to the C-trap and then injected into the Orbitrap mass analyzer. For tryptic peptides, this technique predominantly results in y-type fragment ions and requires more ions to accumulate to generate a signal in the Orbitrap compared to CID in an ion trap. As a result of this, spectral acquisition time is longer compared to CID. ETD induces

fragmentation of cations by transferring electrons to them [27]. Although it displays lower fragmentation efficiency compared to CID or HCD, it is better at fragmenting longer, multi-charged peptides, or even intact proteins. It creates mostly c- and z-type ions. It has also been employed to detect labile PTMs, which are challenging to characterize using CID because these PTMs break off first, which loses positional information.

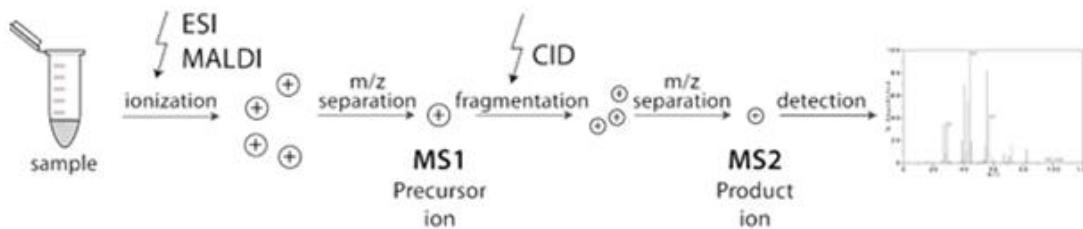


Figure 2. Protein identification by tandem mass spectrometry (adapted from <https://www.creative-proteomics.com/blog/>)

#### 1.1.1.4. Acquisition methods

Shotgun proteomics typically uses a data-dependent selection of precursor ions to generate fragment ion scans (data-dependent acquisition (DDA)), which is the TopN method described in the previous section. In DDA mode, the mass spectrometer selects the most intense peptide ions, and then they are fragmented and analyzed. To prevent the re-fragmentation of peptides, precursors with the same mass are excluded from resequencing for about the time taken for a typical peptide to elute from the LC column. The resultant spectra contain information about the  $m/z$  values, retention times, and ion abundances for detected fragment ions. Although this method has been very successfully used for a long time, the semi-stochastic nature of DDA impedes reproducible quantification of peptides across multiple samples.

To overcome this challenge, the MaxQuant software developed in our group has a ‘match between runs’ feature, which boosts the number of identifications. If a peptide is present in several samples, but not identified via MS/MS in all of them, ‘match between runs’ allows the transfer of this MS/MS information from one sample to others. For this approach, samples must be analyzed using the same chromatographic method (gradient) because it uses the retention time and the masses to match the identifications from one run to another and the false discovery rate (FDR) of match between runs can be difficult to determine. Moreover, our group has developed a method named ‘BoxCar,’ which helps to increase the depth of proteomes [28]. The limited capacity of the C-trap (about a million ions), which allows the analysis of only 1% of generated ions at the MS1 level, can be used more efficiently by dividing the mass range into a number of sequentially filled segments. This approach allocates longer injection times for low abundant ions, thereby maximizing the usage of incoming ions and increasing the dynamic range

at the MS1 level. Consequently, fewer missing values are observed using BoxCar when matching identifications from a library at the MS1 level.

In the bottom-up proteomics, targeted proteomics and data-independent acquisition (DIA) methods are the main alternatives to DDA. Like DDA, DIA is also discovery-oriented but compared to DDA it can generate more complete data sets. In contrast to the semi-stochastic, intensity-based precursor picking of DDA, DIA generates comprehensive fragment-ion maps [29, 30]. In each cycle, the instrument isolates a mass window of precursor ions (for example 25 m/z units wide) and measures fragments of all precursors found within that window. It cycles through mass windows across the entire mass range to generate MS2 scans that cover all detected precursors. The most well-known method that generates DIA data is called SWATH (Sequential Windowed Acquisition of All Theoretical Fragment ions) [30]. As the same precursor isolation window is fragmented again at each cycle during the entire analysis, a time-resolved record of the fragment ions of all the precursors is obtained through the entire mass range. This type of data is composed of highly multiplexed fragment ion maps that display a higher grade of complexity. For this reason, the analysis of such data requires sophisticated deconvolution algorithms for peptide identification, which generally rely on information from pre-existing high-quality spectral libraries. DIA has been also recently propelled by the advances in scan speed and resolution of new mass spectrometers. It now allows us to analyze all co-eluting peptide ions simultaneously, enabling more accurate and precise quantification with less missing values across samples and higher identification rates over a higher dynamic range.

Targeted methods are used for acquiring a predefined set of precursor or fragment transitions with high reproducibility and specificity. There are two main types of targeted acquisition methods: the ones that collect information at peptide ion level such as selected ion monitoring (SIM) or fragment ion level such as single or multiple/parallel reaction monitoring SRM, MRM, and PRM [31] (Figure 3). Briefly, SIM uses a mass analyzer in filtering mode to monitor the signal of a selected precursor ion of interest. SRM is traditionally mostly performed on triple quadrupole instruments (QQQ) as it requires double filtering to sequentially monitor several fragment ions from specific selected precursors. In contrast, PRM can be performed on a quadrupole-Orbitrap mass spectrometer and can therefore be acquired with high resolution and high-throughput capabilities. In this method, precursors are isolated using a mass analyzer, and all fragments are simultaneously analyzed with a second mass analyzer. It can sequentially monitor several peptides within each cycle. Furthermore, parallel monitoring eliminates the need for a prior selection of target peptide transitions. Thus, it requires much less effort than the SRM assay to develop while being much more specific.

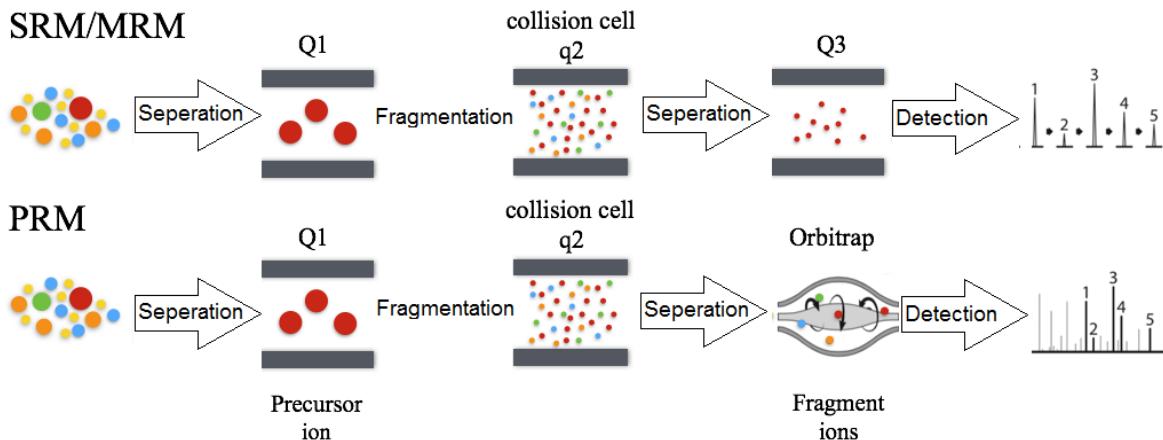


Figure 3. Targeted proteomics strategies (adapted from <https://www.creative-proteomics.com/blog/>)

### 1.1.1.5. Protein identification

The ultimate goal in shotgun proteomics is to identify proteins, which relies on the extraction of information from MS2 spectra. The spectra generated by MS2 contain fragment ions forming N-terminal (b-ions) or C-terminal (y-ions) sequence ladders, and the relationship between the relative abundance (intensity) and  $m/z$  value of each fragment ion of a peptide. The spectral data is first converted into a format readable by downstream software. Next, spectra identification is performed in various ways, including matching into *in-silico* generated peptide sequences or to previously established spectral information such as a spectral library. When using the sequence database search, the experimentally obtained fragment ion spectra are matched to the theoretical ion spectra for a given organism by search engines such as Mascot or Andromeda [32, 33]. In the second method is that sample type-specific spectral libraries can be generated and utilized to fit experimentally observed spectra to it. Another technique for spectrum identification is 'de novo sequencing'. This method is mainly used when there is no or limited database information available.

Furthermore, spectral identification through search engines requires statistical validation of peptide-spectrum matches (PSM) before reporting. For each spectral matching event, scores that represent the resemblance between two spectra are calculated. The scores are then statistically evaluated by calculating probabilities to control for false-positive hits. The most common approach for the false discovery rate (FDR) estimation for MS/MS spectra is the Target-Decoy searching [34]. In this method, the spectra are matched to a database that includes each peptide in the true and reverse amino acid orders. The resultant numbers of hits to the reversed database can be used to define the FDR, which is usually set to less than 1% in global proteomics studies. Furthermore, some peptide sequences are not unique to a specific protein; for instance, they can be matched to several protein isoforms. In such cases, they are usually assigned to the protein for which the most unique peptides (razor peptides) are identified.

### 1.1.1.6. Protein quantification

Protein identification is the first step in gaining insight into a biological system. To detect global changes when studying biological systems, it is crucial to determine differential protein expression or abundance accurately. Quantitative proteomics methods provide a reliable and dynamic analysis of differential expressed proteins in a cell or tissue responding to extrinsic or intrinsic changes. They can be divided into 'relative' or 'absolute' quantification strategies. In the absolute fashion, expression of proteins such as concentrations or copy numbers are determined using spiked-in references. In contrast, in the relative fashion, protein levels are compared amongst different conditions. The relative quantification methods can be categorized into label-based and label-free quantification.

Label-based quantification involves metabolic and chemical labeling methods (Figure 4). SILAC (Stable Isotope Labeling by Amino Acids in Cell Culture) is the most popular in-vivo metabolic labeling method and relies on the metabolic incorporation of a given 'light' or "heavy" form of the amino acid into the proteins. Leucine, arginine or lysine are the amino acids which are commonly used. Cells are usually cultured in medium highly enriched for stable isotopes for more than five passages to ensure complete incorporation. Stable isotopes incorporation does not affect the physical behavior of labeled peptide in terms of elution profiles. After labeling is complete, differentially labeled samples are mixed and analyzed together. Due to the mass difference, the samples can be distinguished at the MS1 level. Relative quantification is performed based on the differences in the precursor areas of differentially labeled peptides in mass spectrometry. Although the most accurate and robust, the main disadvantages of this method are that it is limited mainly to cell line samples, and it requires some time and effort. To overcome these problems, the classical SILAC approach has been extended to methods such as spiking in entire labeled proteomes (mix of cell lines (super-SILAC) [35] and SILAC-labeled protein epitope signature tags (PrESTs) [36]. In addition to the use of SILAC for relative quantification, stable isotope-labeled protein standards (PSAQ) [37] or peptides (AQUA peptides) have been also used as internal standards for absolute standards. Yet, these approaches remain mainly limited to a moderate number of proteins or peptides. However, the SILAC-PrEST method developed in our group successfully quantified over 40 proteins in HeLa cells over a wide abundance range. This method uses a known quantity of recombinant heavy-labeled standards covering unique sequences of the protein of interest that are fused with a purification and solubility tag. There are also other methods for absolute quantification, including the 'proteomic ruler' approach, which allows the copy numbers estimation per cell on a global scale by utilizing the fixed relationship between histones and DNA [38].

Chemical labeling methods are more applicable and suitable for various sample types, including tissues and body fluids, than metabolic labeling since it introduces isotopes *in vitro*. They are mostly performed at the peptide level. One of the early examples is called iTRAQ which utilizes

a multiplexed isobaric chemical tagging reagent [39, 40]. A more popular example for this technique is TMT, tandem mass tags (>10 channels) [41]. Both tags consist of a reporter group for quantification, a balance group so that peptides share the same precursor mass and a peptide reactive group. This reactive group is amine-reactive allowing the attachment of the tag to the N-terminal amine groups of lysine residues of the amino terminus of peptides, labeling all peptides in the sample. These approaches enable multiplexing opportunities, reducing measurement time without increasing spectral complexity. MS2 spectra of isobaric tagged peptides have two types of product ion peaks: the reporter ion that correlates with the abundance in each channel and the peptide fragment ion that allows identification. The disadvantage of TMT is that it suffers from 'ratio compression' due to the relatively low resolution of quadrupole isolation of precursors. The actual ratio between channels is underestimated due to reporter ions from co-isolated and co-fragmented [42]. To overcome this challenge, mass spectrometers that are capable of MS3 are used to fragment the tag-containing fragments further. However, this comes at the expense of sequencing speed and proteome coverage. Our group has developed an alternative method to TMT, which is called EASI-tag, to allow accurate quantification without the ratio compression problem [43]. In this technique, reporter ions are precursor specific and the tag fragments at lower collision energies than the peptide backbone, so the reporters stay attached to the precursors, eliminating interference from other precursors.

Label-free quantification (LFQ) is the most straightforward and most economical approach as it eliminates the need for additional chemical reagents and procedures. Quantitative values can be obtained in two ways. The first approach - known as spectral counting - correlates the number of identified MS/MS spectra from the same protein with its abundance. The second and more recent approach is through the extraction of the area underneath the extrapolated curve of the precursor ion peaks in MS1 scans. The LFQ approach suffers from reduced reproducibility compared to label-based quantification because sample preparation and measurements are performed for each sample separately. Nevertheless, when combining high-resolution instrument measurements with a sophisticated algorithm called MaxLFQ very accurate quantification can be achieved.

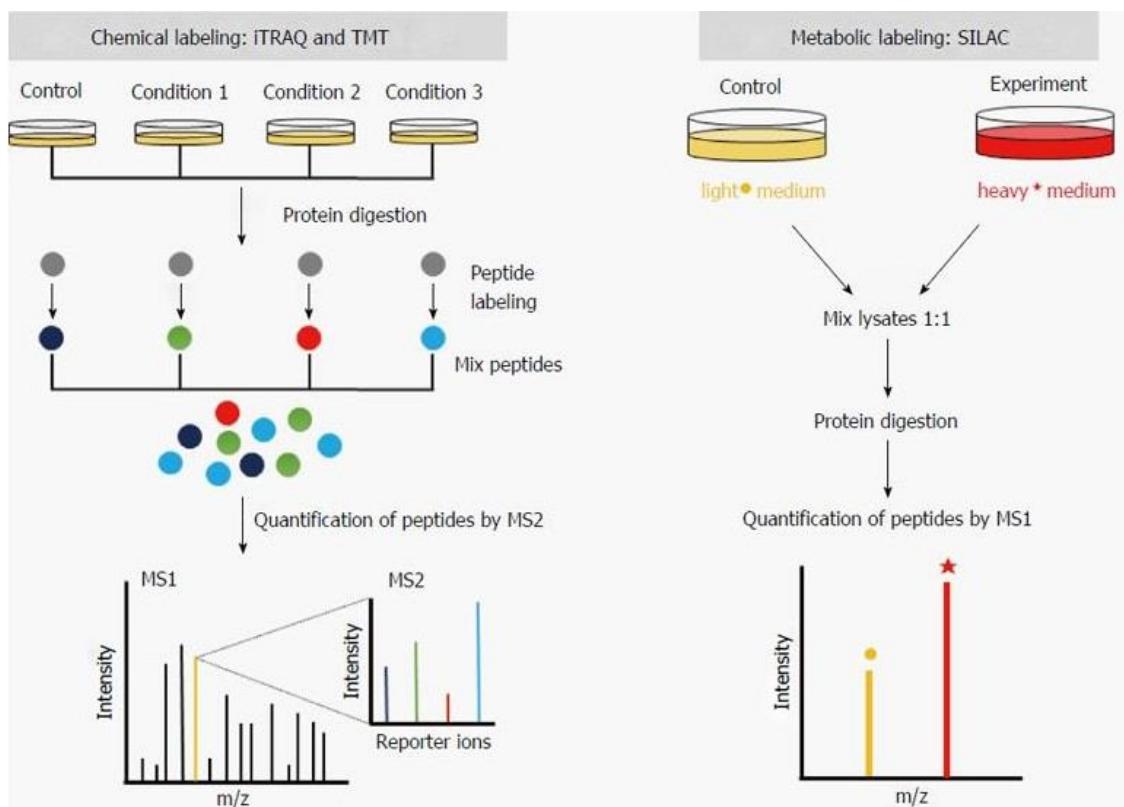


Figure 4. Label-based quantification methods (adapted from [44])

### 1.1.2. In-depth cellular and bio-fluid proteomes and biomarker discovery

Quantitative proteomics holds significant promise for understanding fundamental biologic processes and mechanisms and for discovering diagnostic and prognostic protein markers. The community desires to determine the complete proteomes of various tissues and samples to better understand the complexity of protein-driven mechanisms in the human body. Significant advancements at all levels of today's MS technology, encompassing sample preparation to measurement and subsequent bioinformatics analysis, have enabled the characterization of nearly complete proteomes from cells, tissues, model organisms, and even from very challenging matrices such as biofluids [45, 46]. Thousands of proteins and tens of thousands of modified peptides can be routinely quantified. System-wide proteomic studies explore critical regulatory mechanisms at a large scale and create a link between genomic information, biological function and disease. 'Organellar proteomics' can also be employed to monitor subcellular localization on a system-wide scale [47]. Global interactome studies have likewise become very popular with the development of high resolution instruments, and a recent human interactome studies uncovered interaction partners of thousands of human proteins [48, 49].

Advanced proteomics tools now also provide tremendous opportunities for biomarker-related clinical applications, and there is an enormous interest in profiling body fluids for biomarker

discovery. Biomarkers would be beneficial to define a normal or abnormal condition or disease and monitor a treatment. Numerous studies on human biofluid proteomes that mainly focused on cancers, metabolic and brain diseases have been performed often reporting hundreds of proteins. Recent technological advancements in acquisition methods and software developments enabled in-depth characterization of body fluids such as stool, urine, CSF and plasma, presenting mass spectrometry (MS)-based proteomics as compelling technology for biomarker discovery. Moreover, our group has proposed a “rectangular” biofluid proteome profiling strategy instead of previous “triangular strategies”. The latter involves the discovery of a single or few biomarker candidates in a small cohort, followed by validation typically by immunoassays in a larger cohort. In the “rectangular” strategy, both discovery and validation of biomarkers are performed at maximum possible depth in medium to large size cohorts. Using this strategy in conjunction with the latest MS technologies, our group has published several studies in which proteomes of bodily fluids such as CSF and plasma were profiled to identify biomarkers for various conditions and diseases [50-54].

Clinical proteomics is still a relatively new field but is rapidly growing also due to the development of sensitive and specific methods and clear standards that consistently define and report biomarkers [55, 56]. In particular, issues associated with the collection and storage of the samples are of critical importance in this field [55]. Pre-analytical variation caused by inconsistent sample processing and contaminations during sample collection has a substantial impact on the reported results and may cause the reporting of incorrect biomarkers. For instance, levels of certain proteins, including *a*-synuclein, are much higher in blood than in CSF, suggesting that blood contamination in CSF might affect the reliable quantification of those biomarkers. Standard procedures have to be established and applied when collecting biofluids to ensure that the observed changes are not caused by artifacts related to sample handling and processing.

## **1.2. Post-translational modifications (PTMs) and their role in signaling**

Cells need to respond to external and internal cues in a short time; thus they have a multi-level regulation system capable of sensing, interpreting, and storing the information. Post-translational regulation of proteins via reversible and irreversible modifications is especially crucial for cells to rapidly respond to stimulation and environmental alterations. These modifications can affect structure, stability, and function of proteins and increase the functional diversity of the proteome. Thus far, over 200 PTMs have been discovered regulating proteomes of various organisms. PTMs are very diverse and can be divided into four groups (Figure 5): (i) the modification of the chemical structure of amino acid side chains, (ii) the addition of complex molecules to specific amino acids, (iii) the covalent linkage of polypeptides or small proteins and (iv) the cleavage of the peptide bond (proteolysis). In this thesis, I will primarily focus on two types of PTMs, phosphorylation and ubiquitination.

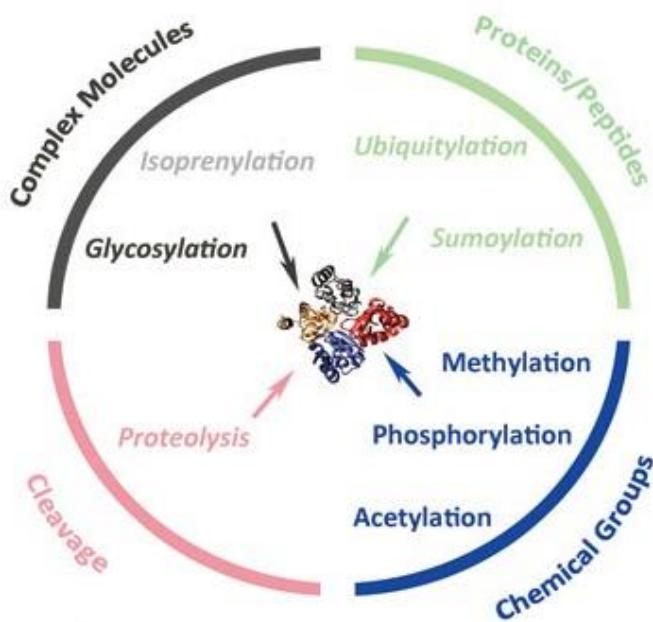


Figure 5. Diversity of post-translational modifications (adapted from [57])

### 1.2.1. Phosphorylation

Phosphorylation is one of the most common and best-studied PTM. It defines the covalent addition of phosphoric acid to the target protein and occurs through the transfer of the  $\gamma$ -phosphate group from adenosine 5'-triphosphate (ATP) to the hydroxyl group of specific amino acids. It is mainly attached to serine, threonine, or tyrosine residues, representing approximately 90%, 9.9%, and 0.1% of protein phosphorylation, respectively. Tyrosine phosphorylation is thus relatively rare and is typically mediated by receptors harboring a tyrosine kinase domain. Less stable phosphorylations may also occur naturally on aspartic acid and histidine residues [58, 59]. The addition of a phosphate group in general modifies the protein from a more hydrophobic apolar state to a more hydrophilic polar state. This can lead to the activation or inhibition of the protein and can induce conformational changes also upon interaction with other molecules.

Phosphorylation is a reversible reaction with a constant balance between phosphorylation and dephosphorylation events mediated by kinases and phosphatase ( $\text{protein} + \text{ATP} \rightleftharpoons \text{phosphoprotein} + \text{ADP}$ ) (Figure 6). The addition of the phosphate group is catalyzed by the enzymatic activities of protein kinases, whereas another type of enzymes, phosphatases, are responsible for its removal [60]. This highly dynamic process is at the heart of signaling pathways, serving as a signal in a plethora of cellular pathways and mechanisms, including the cell cycle and cancer [61]. More than 30% of all encoded proteins are phosphorylated at any given time [62]. Over 250,000 phosphorylation sites have been identified and curated so far in one of the well-known and widely used databases (<http://www.phosphosite.org>).

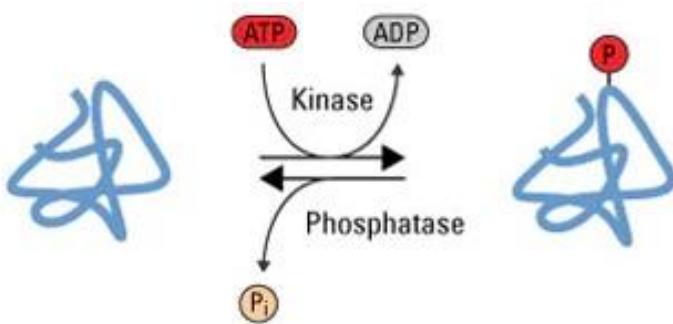


Figure 6. Phosphorylation is a reversible PTM that regulates protein function. Kinases mediate phosphorylation at serine, threonine and tyrosine side chains and phosphatases remove the phosphate group (adapted from <https://www.thermofisher.com/de/de/home/life-science/protein-biology-learning-center>)

### 1.2.1.1. Protein phosphorylation–mediated cellular signaling

Cells can sense and respond to extracellular changes through cell surface proteins. Upon binding of a ligand to a cell-surface receptor, the receptor gets activated, and this sets off a series of signaling events. The chains of molecules known as intracellular signal transduction pathways transfer information from the surface to the interior of the cell. For instance, the receptor can bind another molecule within the cell, which in turn activates its target. Non-protein molecules like phospholipids can also play essential roles in signaling.

Most signal transduction systems in the cell are regulated by protein phosphorylation, which acts as a switch altering protein activity. This serves a variety of functions, including increase in catalysis, providing recognition by interacting protein, conformational changes, and formation of protein complexes or alteration of the cellular protein localization. It can also inactivate the protein or cause its degradation. These features make protein phosphorylation a critical regulatory mechanism of the most crucial cellular processes, including proliferation, apoptosis, differentiation, and development.

Phospho-signaling pathways are dynamically regulated within seconds at multiple levels of positive and negative feedback. One kinase can target hundreds of potential substrates, enabling signal diffusion across many network nodes. Depending on the type and the intensity of the stimulus, signals can be transient or prolonged. Each stimulation leads to a unique signaling network typically inducing a unique biological response. The complexity of signaling networks has recently been more acknowledged based on results of system-wide studies using MS-based proteomics.

To better explain phosphorylation–mediated cellular signaling concepts, I use one of the most studied signaling pathways called the epidermal growth factor (EGF) pathway as an example (Figure 7). This pathway acts through a series of kinases to produce various cellular responses. EGF binds to the receptor (EGFR) and induces its activation by dimerization. The activated receptor recruits various adaptor proteins such as GBR2 that bind to the phosphotyrosine residue in the cytoplasmic part of the receptor. GBR2 recruits SOS to the membrane. Next, SOS activates GDP/GTP exchange, which then recruits Raf to the membrane. Active Raf phosphorylates and activates MEK, which phosphorylates and activates the ERKs. ERKs act on a variety of target molecules, including transcriptional regulators like c-Myc that induces growth and proliferation. These three kinases, Raf, MEK, and the ERKs, together make up a kinase signaling pathway called the mitogen-activated protein kinase (MAPK) cascade. They are found in many organisms, from humans to yeast. The overactive forms of the proteins in this pathway, such as the growth factor receptor, Raf and c-Myc have been associated with cancer as they play a central role in promoting cell proliferation. GRB2 can also recruit another major mediator of EGFR signaling, PI3Ks. They convert phosphatidylinositol-4,5-bisphosphate (PIP2) to phosphatidylinositol-3,4,5-trisphosphate (PIP3). PIP3 recruits PDK1 which phosphorylates AKT. This pathway, also known as PI3K-Akt pathway, plays a role in prosurvival signaling.

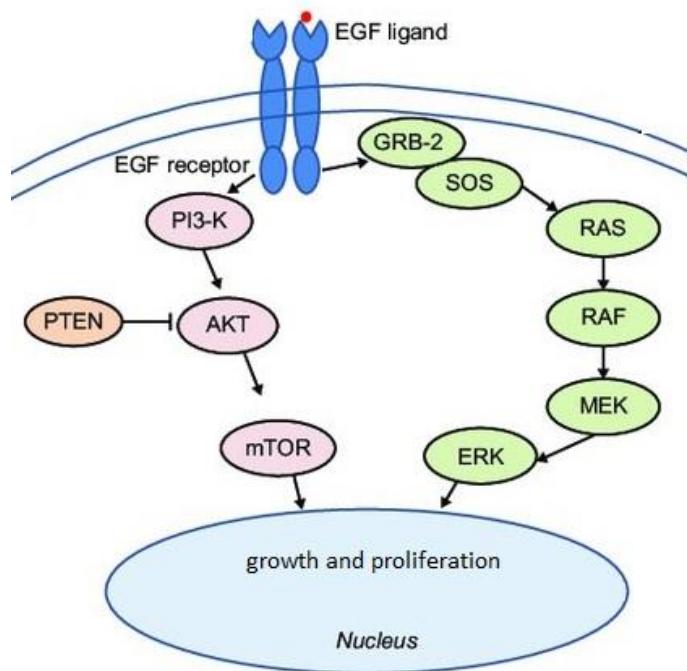


Figure 7. Epidermal growth factor receptor (EGFR) and its downstream signaling proteins (adapted from [63])

### 1.2.1.2. Phosphoregulators: kinases and phosphatases

Kinases and phosphatases are central players of the intracellular signaling cascade in which a signal is passed on to downstream proteins by sequential activation and deactivation of proteins via phosphorylation/dephosphorylation events. Therefore, these enzymes are also crucial for the transduction of the signal by controlling its amplitude. Over 500 kinases are found in the human genome, encompassing 1.7% of the total genome [62]. Kinase activity is controlled in three ways: (i) kinase phosphorylation/autophosphorylation, (ii) binding with activator or inhibitor proteins, or (iii) through localization in relation to their substrate [64].

The conserved protein kinase core has two domains: the N- and C-terminal core [65]. Between these two domains, there is an active-site cleft with the ATP binding site. The activation of the kinase through activation loop phosphorylation or an allosteric mechanism induces a deep cleft where the transfer of the phosphoryl group occurs [65]. Non-catalytic domains of kinases help with the attachment of substrates and allow the recruitment of other signaling components.

Most kinases are serine and threonine (STKs) kinases that phosphorylate the OH group on these amino acids. Some kinase can only act on tyrosine (TKs), and there are also kinases targeting three amino acids (dual-specificity kinases; DSKs) [66]. Protein kinases subfamilies include: AGC (Protein Kinase A, G, and C families (PKA, PKC, PKG)), CaMK (Calmodulin/Calcium regulated kinases), CK1 (Casein Kinase 1), CMGC (CDK, MAPK, GSK3, and CLK), TK (Tyrosine Kinase), TKL (Tyrosine Kinase-Like), RGC (Receptor Guanylate Cyclases), PKL (Protein Kinase-Like) and STE [67]. Many kinase families are highly conserved in yeast to human.

Protein phosphatases act opposite of kinases by removing the phosphate group from proteins. They do not have a common structure, and unlike kinases, their activity is associated with different protein folds and catalytic mechanisms and they typically gain their specificity only as members of multi-protein complexes [60]. Over 200 protein phosphatases have been identified so far and classified into several families, including the phosphoprotein phosphatase (PPP) family, the metallo-dependent protein phosphatase (PPM) family and the protein-tyrosine phosphatase (PTP) family [60]. Although PPP and PMP families mostly act on phosphoserine and phosphothreonine, they can also dephosphorylate phosphotyrosine. PTPs can also dephosphorylate non-protein targets, such as mRNA and phosphoinositides [68].

Mutations in genes encoding for kinases and phosphatases have often lead to decreased cell viability, disrupted cellular signaling pathways and contribute to several human diseases including cancer, autoinflammatory diseases, diabetes, and neurodegenerative diseases [69]. Thus, having a better understanding of impaired phospho-signaling pathways associated with diseases would greatly help with the identification of potential drug targets [69].

### 1.2.1.3. Quantifying phospho-signaling by MS: phosphoproteomics

Many different analytical strategies have been developed and employed to quantify activated signaling networks. Classical biochemical techniques such as phospho-specific antibodies are typically low throughput. For a systems-wide understanding of cellular signaling networks, identification and quantification of a sufficient number of specific phosphorylation sites and their coordinated and temporal response to a perturbation are necessary. Moreover, multiple phosphorylation sites often occur on a single protein and their abundance range spans orders of magnitude, with varying stoichiometry [61]. All this makes detection and quantification of all sites affected by a given stimulus extremely challenging.

Mass spectrometry has transformed the field of cellular phospho-signaling and become the method of choice for studying multiple phosphorylation events simultaneously ('phosphoproteomics') [61, 62]. Global strategies aiming to detect and quantify thousands of phosphorylation sites can identify altered phospho-signaling networks. Recent developments in the MS field have also enabled increasingly sophisticated phosphoproteomics studies. However, detection and quantitation of protein phosphorylation remain analytically challenging in many aspects [70]. For instance, phosphorylation is a dynamic and transient process. Its short life together with its low abundance makes accurate and precise quantification difficult, especially in heterogeneous mixtures. Most quantification methods, including chemical and metabolic labeling strategies, have also been utilized for quantitative phosphoproteomic. Label-free strategies have also gained acceptance and popularity in signaling research as they have the advantage of being very adaptable [71].

Other issues that complicate phosphopeptide identification are the potential suppression of their ionization by their non-phosphorylated counterparts [70] and the labile character of the phosphoester bond. In collision induced dissociation (CID) type of fragmentation, the phosphoester bond tends to break first, leading to a neutral loss of phosphoric acid, creating a - 98 Da ( $H_3PO_4$ ) mass shift [72]. In contrast, the high-energy collision dissociation (beam-type CID, HCD), in which an electrical potential is applied to accelerate peptide ions towards the inert gas, substantially reduces the presence of the neutral loss precursor [73]. Compared to CID and HCD, electron transfer dissociation (ETD) is more suitable for the analysis of peptides with labile modifications as these modifications can be preserved. However, this technique results in a low number of total identification due to its slower scan rate and lower efficiency [73]. Methods combining ETD and HCD fragmentation techniques for complementary ion production (c/z ions) have also been developed to provide data-rich MS/MS spectra, which yield higher phosphoproteome coverage and more confident site localization [74].

Developments of more sensitive and fast mass spectrometers along with improved data analysis software have increased the number of phosphorylation sites identified by DDA. It has become very common to report and quantify tens of thousands of phosphorylation sites, especially when coupled to a prior fractionation step [75, 76]. However, as mentioned before, DDA and MS in general, have some bias toward the accurate and precise analysis of the more abundant, higher intensity peptides. Since phosphorylated peptides are less abundant than their non-phosphorylated counterparts, it remains challenging to accurately detect and quantify low abundant phosphopeptides with DDA. Moreover, the semi stochastic nature of DDA, favoring the most intense peaks, results in poor reproducibility even among replicates. This is particularly a problem for phosphopeptides due to their low abundance. Most phosphorylation events typically present in a cell are not induced by any given stimulation. These unspecific phosphorylations can interfere with the identification of stimulus-specific phosphorylation events as most of the instrument analysis time is spent on these uninformative ones. DIA, propelled by the advances in scan speed and resolution of new mass spectrometer generations, is becoming a compelling alternative to DDA. The superior performance of DIA for sensitive and reproducible MS measurements has recently also been demonstrated for global protein phosphorylation analysis. Notably, a recent study quantified > 20,000 phosphopeptides in 15 min single-run LC-MS analysis per condition without the need for spectral libraries [46].

Targeted MS methodologies appear to be a promising way to detect and accurately quantify low abundant phosphorylated species as they promise ultra-high specificity and sensitivity. Such methods rely on non-data dependent but targeted MS acquisition techniques and usually work with a small number of proteins. Therefore, a priori knowledge of the phosphorylated proteins and signaling network is required. The most common MS-based targeted approaches used for phosphoproteomics are selected/multiple reaction monitoring (SRM, MRM) and parallel reaction monitoring (PRM). These methods rely on tandem mass spectrometry and offer high specificity along with sensitivity as they monitor fragment ions. However, sensitivity is inherently lower compared to single-stage mass spectrometry analysis, such as selected ion monitoring (SIM), as they distribute the precursor ion signal across multiple fragments. These methods have frequently been applied to quantify cellular signaling networks, protein complexes [61] and protein phosphorylation stoichiometry, as well as kinome abundance and phosphorylation states [77]. They can also determine the absolute amount of a given peptide or phosphorylation site in a sample using SIL peptide standards spiked in known quantities into the samples.

#### **1.2.1.4. Phosphoproteomics workflows and enrichment techniques**

In a typical phosphoproteomics experiment, proteins are proteolytically digested, mostly using trypsin, and the complex mixture of digested peptides can be either analyzed directly or fractionated before MS analysis of each fraction. Regardless of the preferred fractionation

method, phosphopeptide enrichment is required before LC-MS/MS analysis. The common methods for the enrichment of phosphopeptides, such as IMAC and MOAC, are typically combined with prior fractionation to increase the number of identified phosphopeptides (Figure 8). There are several factors that drive the need for pre-fractionation and enrichment. First, proteolysis of proteins generates many peptides, of which only a small fraction will carry phosphorylations. Second, the phosphopeptides are usually present at substoichiometric levels, meaning they are less abundant than their nonphosphorylated counterparts. Each fraction enriched for phosphopeptides is analyzed separately. This increases the chance to observe low abundant phosphopeptide ions and to reach the depth required to cover key signaling components. Third, as mentioned above, phosphorylated peptides may have lower ionization potentials than their nonphosphorylated counterparts when measured in positive ion mode.

Common pre-fractionation methods in phosphoproteomics studies are chromatography techniques that provide separation based on charge (SCX and SAX) isoelectric point (IEF), polarity (HILIC) and hydrophobicity (ERLIC) [78]. While more peptides can be identified by fractionation since the complexity in each fraction decreases, the total time required for the analysis increases proportionally with the number of fractions. It also comes at the expense of throughput, robustness, and reproducibility, and often requires specialized equipment and large sample amounts.

One of the earliest enrichment techniques is affinity-based chromatography. Positively charged metal ions, such as Fe (III), immobilized with a solid phase are presented for interaction with negatively charged phosphate groups (immobilized metal affinity chromatography (IMAC)). The use of metal oxides such as titania or zirconia has emerged as the alternative ways for phosphopeptide enrichment (metal oxide affinity chromatography (MOAC)) [79-81]. However, the latter method may suffer from nonspecific binding. The addition of an organic acid, such as 2,3-Dihydroxybenzoic acid (DHB) and glycolic acid into the enrichment buffer helps to increase phosphopeptide specificity [82]. In both chromatography techniques, binding and washing conditions are performed in a very low pH environment. This overcomes the competition between carboxyl acid and phosphoryl groups in binding to the material since carboxylic acid groups are also negatively charged at moderate pH. Finally, elution is performed at a more basic pH [78].

Both materials are typically used to enrich phosphopeptides rather than full length phosphorylated proteins. They both are especially suited for the enrichment of serine, threonine and tyrosine phosphorylations as these residues are stable under acidic pH conditions. Furthermore, IMAC is well known for its affinity for multiple phosphorylated peptides, whereas MOAC, especially titanium dioxide (TiO<sub>2</sub>), preferentially detects singly phosphorylated ones. Although they are thought to be complementary, a recent study showed that they bind the same phosphopeptides but differ in binding capacity and elution efficiency [83].

Ion exchange, especially strong cation exchange chromatography, can be also employed for both the separation of peptides and the enrichment of phosphopeptides. Separation is provided by charge interactions between the analyte and the stationary phase [95]. The stationary phase containing negatively charged strong acidic groups, such as sulfonic acid derivatives, which have affinity for positively charged groups. Tryptic peptides at a low pH typically have a +2 or +3 net charge due to positively charged terminal amino groups and basic amino acid residues (lysine, histidine, and arginine) contributing to the net charge of peptides. However, the presence of a negatively charged phosphate group ( $H_2PO_4^-$ ) decreases the net charge. As peptides elute from the column along increasing salt concentrations in SCX, peptides with higher charge state have a stronger interaction with the stationary phase. Therefore, phosphorylated peptides with a net charge of less than +1 elute earlier compare to nonphosphorylated peptides. The first fractions will therefore be more enriched for phosphopeptides. Given the complexity of the proteolytic digest of the proteome, phosphopeptide enrichment by SCX chromatography is not very specific. Still, it has been frequently used as a complementary technique to affinity-based phosphoenrichment.

More targeted approaches requiring a prior knowledge of cellular signaling networks to be analyzed also exist. For instance, there are strategies utilizing phosphorylation motif antibodies and naturally occurring phospho-binding domains such as SH2 domains, which recognize and bind to the tyrosine phosphorylation site to study signaling networks. Perhaps the most common targeted approach is tyrosine phosphorylation analysis after phosphotyrosine peptide immunoprecipitation. This technique has been coupled with different labeling approaches to quantify the temporal response to growth factor stimulation [84, 85] and the effects of oncogenic kinases on cellular signaling networks [86, 87].

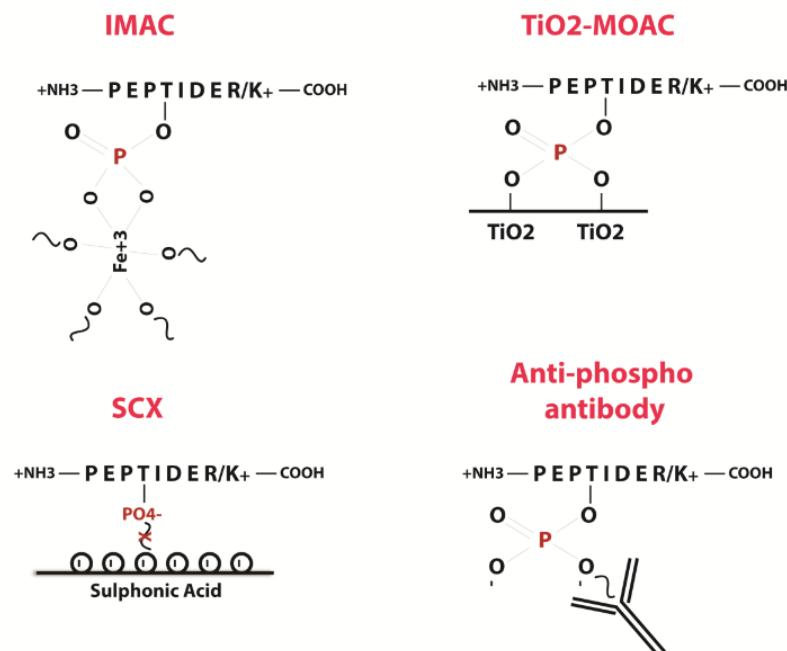


Figure 8. Strategies for phospho-specific enrichment and pre-fractionation techniques

## 1.2.2. Parkinson' disease and phosphorylation

### 1.2.2.1. Parkinson' disease

Parkinson's disease (PD) is the second common neurologic disorders after Alzheimer's disease. It is mainly caused by the premature death of dopaminergic neurons in the substantia nigra pars compacta (SNc) brain region and characterized by cytoplasmic  $\alpha$ -synuclein (SNCA)-containing protein aggregates (Lewy bodies). It mainly affects the motor system, resulting in muscle stiffness, resting tremor, and slowness of movement, which are critical features for the diagnosis. Many PD patients also suffer from postural instability and other non-motor symptoms.

PD is diagnosed in up to 10 million people worldwide (<http://www.pdf.org/>) and its prevalence is known to increase with age. Less than 5% of PD cases have an onset before the age of 40, about one percent of the population above 60 and up to 5 percent above 85 are diagnosed with PD [88]. The majority of PD cases are idiopathic; the cause of the disease remains unknown. However, about 10% have been linked to a genetic cause. Several genes are known to segregate with familial forms of PD, such as LRRK2, GBA, PARK8, or SNCA. The individuals with monogenic forms of the disorder -caused by a single mutation in a dominantly or recessively inherited gene- or carrying genetic risk factors have a higher risk of developing the disease (2 to 5%) [88]. Defined genetic causes collectively account for around 30% of the familial and up to 5% of the sporadic cases [89]. Recently, a meta-analysis of genome-wide association studies (GWAS) of PD identified about 40 genetic susceptibility loci associated with late-onset PD [90, 91]. A few of them have been identified as causal, yet the underlying mechanisms remain mostly elusive for the majority of them [91].

PD is also a progressive disease. Patients typically experience increased severity of symptoms when growing older, which is associated with increased treatment costs. The most effective PD medication is currently levodopa (L-dopa), a central nervous system agent, which is inexpensive and effective. Levodopa-based treatment does not cure the disease; it mainly influences apparent progression and alleviates symptoms. The patients are usually diagnosed after the disease manifests and severe motor impairment is evident, however, by this time irreversible brain damage has already occurred. This highlights the urgent need for disease-modifying therapeutics. Unfortunately, there is no specific and sensitive tests to detect PD early to aid the development of new therapeutic strategies.

### 1.2.2.2. LRRK2 kinase in PD and its cellular targets

Leucine-rich repeat kinase 2 (LRRK2/Park8 loci) was identified in 2004 and it is one of the key genes that are linked to PD pathogenesis. Mutations in the LRRK2 gene are the most common

cause of autosomal dominant PD, representing around 2% of total Parkinson's cases (<https://www.michaeljfox.org/>). Ubiquitously expressed LRRK2 is a large and multi-functional and multi-domain protein. It has four distinct domains that are implicated in protein-protein interaction (ARM, ANK, LRR, and WD40), two domains involved in GTPase function (Roc and COR), and the kinase domain. Pathogenic LRRK2 mutations are frequently found in both the GTPase (R1441C/G/H, Y1699C) and kinase (G2019S, I2020T) domains, suggesting a link between the enzymatic activity of LRRK2 and PD pathogenesis [92] (Figure 9). The most common mutation is the S→G substitution of serine 2019 (G2019S), which accounts for 4% of familial and 1% of sporadic PD cases worldwide [93]. This mutation resides in the activation loop of the kinase domain and increases its activity of about 3-fold. Strikingly, mutations in LRRK2 are also found in non-familial forms of PD, and the pathology of sporadic and LRRK2-linked PD is almost indistinguishable. This suggests that a shared LRRK2-driven molecular pathway is driving pathogenesis. Thus far, LRRK2 has been associated with various cellular processes, including mitochondrial disease, vesicular trafficking, and autophagy. Several studies reported pathogenic LRRK2 driven functional alterations in the same pathways as well as in lysosomal degradation and immune responses [93] (Figure 9).

Although some potential LRRK2 substrates, including ezrin/radixin/moesin proteins, have been described in the last decades [94-96], true and reproducible physiological targets of LRRK2 had proven elusive for a long time. Martin Steger from our group previously identified several Rab GTPases (Rab8A, Rab10, and Rab12) as physiological targets of LRRK2 [2] (Figure 9). The phosphorylated Thr/Ser residue on Rab proteins is highly conserved in ~50 Rab proteins and lies within their effector-binding switch-II motif. The Rab GTPase family is involved in the regulation of intracellular membrane trafficking pathways. They cycle between cytosol, where they are bound to a protein called GDI (GDP Dissociation Inhibitor) and membrane compartments where they are found in GTP-bound active form and recruit specific effector proteins [97]. They play important roles in the formation of transport vesicles, docking and fusion of those vesicles with target membranes and binding of motor proteins for vesicle transport [97]. Our group showed further that the PD-associated mutant LRRK2 (in the hyperactive form) deactivates Rab proteins by phosphorylation, which leads to reduced affinities for GDIs and results in their accumulation at the cell membrane.

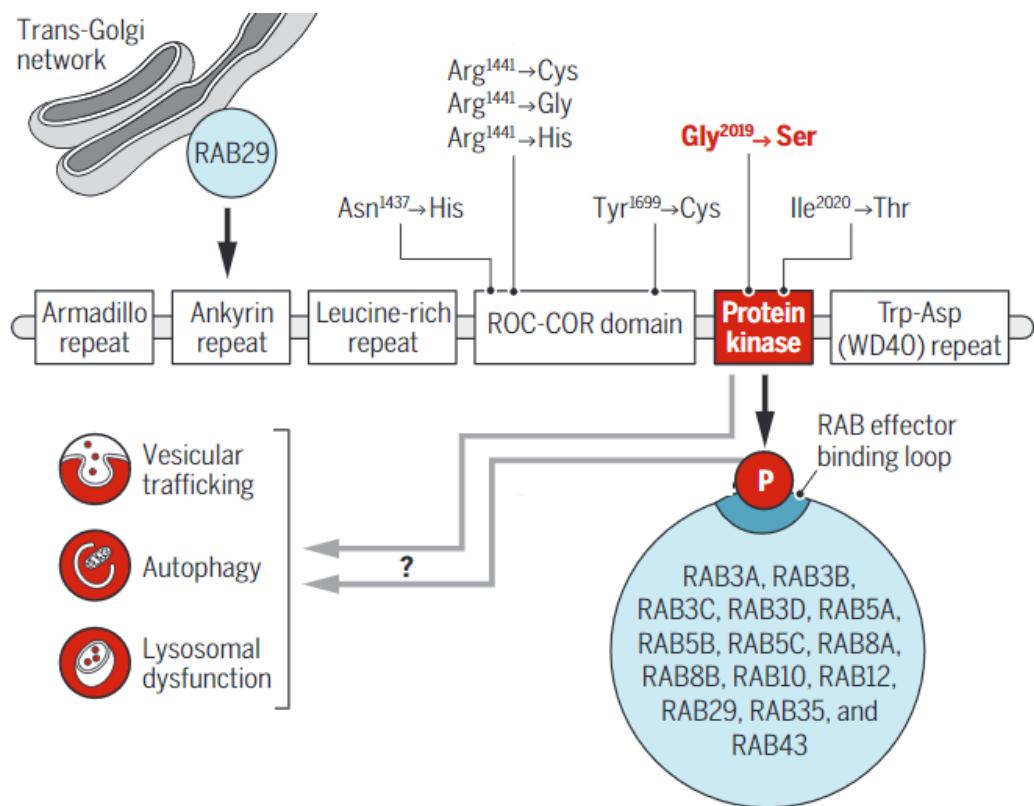


Figure 9. Parkinson-associated mutations that activate LRRK2 kinase activity a subgroup of RAB GTPases phosphorylated by LRRK2 and that regulate various cellular processes.

### 1.2.2.3. LRRK2 inhibitors and clinical trials

Pathogenic mutations in LRRK2 are found in both sporadic and familial PD patients and increase its kinase activity. LRRK2-associated PD cases, whether familial or not, share similar pathophysiological processes. As LRRK2 associated PD closely resembles idiopathic PD (iPD) LRRK2-targeted therapies may be useful for the treatment of a large group of patients. Together, these make LRRK2 a promising target for the treatment of PD. More than 100 compounds targeting LRRK2 kinase activity have been reported in the last decade, evidencing the massive interest in therapy of LRRK2-dependent PD [98]. Some inhibitors, including MLi-2, are in active preclinical development [99]. Recently, Denali Therapeutics has undertaken pre-clinical research with selective LRRK2 inhibitors, known as DNL201 and DNL151. Safety and target engagement of DNL201 have already been assessed, and it has passed the phase I in a study of healthy volunteers [93].

Insufficient measures of drug target engagement are one of the largest challenges in clinical trials. There is a great need to monitor LRRK2 kinase activity and reliably quantify these biomarkers in biofluids and human-derived cells. Such tools would be highly valuable to (i) assess target engagement and monitor compliance of LRRK2 kinase inhibitors and treatment efficacy,

(ii) detect non-manifesting carriers early to prevent disease onset, and (iii) stratify idiopathic PD patients who could benefit from LRRK2 based therapies.

To assess LRRK2 activity, phosphorylations at Ser910, Ser935, Ser955, and Ser973 had been the only option as they are highly responsive to inhibition of LRRK2 activity although the upstream kinase remains unknown [100-103]. However, these phosphorylations do not correlate with intrinsic cellular LRRK2 kinase activity as many knock-in pathogenic mutations were shown not to affect them, and S935, in particular, is still phosphorylated in kinase-inactive LRRK2 [104, 105]. Another conserved LRRK2 fingerprint is the autophosphorylation at the Ser1292, which correlates well with LRRK2 kinase activity [106, 107]. However, pS1292 has been mostly detected in models and systems where LRRK2 was overexpressed or enriched because it has very low occupancy in tissue [106, 108, 109]. There is also no sensitive phospho-specific antibody available to reliably detect and quantify this phosphorylation. Furthermore, the quantification of this site by an MS-based assay is difficult as the proteolytic sites around Ser1292 phosphorylation are either too close or too far to produce a peptide that can be detected by MS. Additionally, miscleaved tryptic peptides covering this site are produced with low and variable yield thereby complicating absolute quantification.

Among the substrates of LRRK2, Rab phosphorylations have emerged as promising readouts to monitor LRRK2 activity as all pathogenic mutations in LRRK2 increase Rab phosphorylation *in vivo* [110-112]. In this thesis, one of our aims was to quantitatively assess LRRK2-mediated low levels of Rab phosphorylation by a sensitive and robust mass spectrometry (MS)-based assay in bodily fluids and/or PD patient-derived cells or tissues.

#### **1.2.2.4. Biomarkers of Parkinson's disease**

Modern medical science is driven by the idea that inter-individual biological variances can determine the differences in disease presentation and response to treatment [113]. Investigation of diseased tissues or body fluids of individuals might reveal inter-individual differences, provide biological markers diagnosing diseases in early stages and predict disease progression. In particular, early diagnosis of PD, before substantial neurodegeneration occurs, could be key to slow progression of the disease and prevent dopamine loss. Effective disease-modifying therapy needs to be initiated before or with the disease onset. However, there are currently no reliable and sensitive biomarkers to detect PD early. Such biological markers could assist in early and differential diagnosis (diagnostic biomarkers) of the disease and in tracking its progression (prognostic biomarkers), and monitoring whether and how patients respond to a therapy (predictive and treatment response biomarkers).

Although there is still no PD-specific biological marker available, several potential candidates have been reported in the past several decades [114]. Among them,  $\alpha$ -Synuclein has attracted substantial attention and has become one of the most investigated potential molecular biomarkers of PD [115]. Human  $\alpha$ -syn - encoded by the SNCA gene - is expressed in the brain and found in an aggregated form (Lewy bodies) in the brain tissue of PD patients, which is a PD hallmark [116]. A mutation in the SNCA gene causes one of the monogenetic forms of PD. It has been suggested that  $\alpha$ -syn can be secreted into the extracellular space of the brain and therefore it should be detectable in biofluids [117]. Indeed, some biomarker studies have reported changes in the levels of  $\alpha$ -syn in CSF and plasma of PD patients compared to control individuals [118-120]. When secreted at high concentrations, extracellular  $\alpha$ -syn can potentially damage healthy functioning neurons and spread the PD pathology by propagating in a prion-like fashion [121].

Apolipoproteins are also involved in the development of many neurodegenerative disorders, and they could potentially be used as PD biomarkers [122]. Especially, two of them - apolipoprotein A1 (ApoA1), the main constituent of high-density lipoprotein (HDL) particles, and apolipoprotein E (ApoE), responsible for lipid transportation in the brain- have been suggested as PD biomarkers. Lower levels of apoA1 have been reported in the CSF of PD patients [123, 124], potentially meaning that less efficient lipid transport leads to reduced brain cholesterol homeostasis [125]. Moreover, orexin, a neuropeptide hormone expressed by a small number of neurons in the hypothalamus, was found to be lower in PD patients compared to no-PD subjects and thereby correlate with the severity of disease. As this hormone regulates many physiological functions, including the sleep-wake cycle [126], low levels of orexin could explain excessive daytime sleepiness from which PD patients frequently suffer in the late stages [127].

The development of useful biomarkers generally depends on the availability and the quality of collections of diseased tissues or body fluids of individuals. Tissues like the brain are usually not easily accessible as they can only be collected after surgery or autopsy. Arguably, the analysis of cerebrospinal fluid (CSF) would be the ideal strategy for PD as it primarily manifests in the central nervous system and has no barrier to the brain fluid. However, CSF is highly challenging to obtain since it requires invasive techniques for sample collection. In contrast, other body fluids are easily accessible in living patient; especially urine has the advantage that it can be obtained in non-invasively. It also contains proteins from distal organs (potentially also from the brain). However, it remained to be investigated if and to what extent PD and other neurodegenerative disorders affect the urinary proteome. We have developed a pipeline combining the streamlined and highly reproducible MS<sup>n</sup> workflow [20] with our state-of-the-art DIA proteomics workflow for urinary proteome profiling in PD, as discussed later in this thesis.

### 1.2.3. Ubiquitin and ubiquitin-like modifications

The ubiquitin fold is a versatile module that comprises a structurally conserved protein family that regulates a plethora of processes in eukaryotic cells [128]. The high interspecies sequence conservation indicated that their biological function is also highly conserved. Unlike small chemical modifications of amino acids, which regulate protein function, such as phosphorylation, ubiquitin (Ub) itself is a small protein that is coded by four genes in the human genome. It has remarkable conservation in eukaryotic organisms, differing at only three out of 76 positions between yeast and human. Most ubiquitin-like modifiers (Ubl) such as Nedd8 and Sumo are also found in almost all eukaryotes. These proteins act as signals and share a common conjugation mechanism. They are attached to another protein through an isopeptide bond between the modifier's terminal glycine and an amine group of the target protein, provided by either a lysine residue within sequence or by the N-terminus [128]. Furthermore, there are type II ubiquitin-like domains which also show significant sequence similarity to ubiquitin [128]. However, unlike ubiquitin that requires C-terminal chemistry, they lack the terminal diglycine signature, hence cannot be conjugated.

#### 1.2.3.1. Overview of the ubiquitin system

The attachment of ubiquitin to a substrate protein is called ubiquitination. This process involves three main steps performed by different sets of enzymes: (i) activation by ubiquitin-activating enzymes (E1s), (ii) conjugation by ubiquitin-conjugating enzymes (E2s) and (iii) ligation by ubiquitin ligases (E3s). So far, two E1s, a limited number of E2s (~30), and hundreds of E3s have been identified in humans [129] (Figure 10). This three-step hierarchical mechanism occurs in all ubiquitination reactions, independent of the fate of the substrate.

The highly conserved activation of the C-terminus of the ubiquitin (Ub) protein is an ATP-dependent two-step process in eukaryotes. The formation of Ub-adenylate intermediate is followed by the reaction of this intermediate with an E1 cysteine residue to form an E1~Ub thiol ester. A fully loaded E1 carries two molecules of activated ubiquitin, one as an adenylate and the other as a thiol ester, which is transferred to a cysteine residue of the next enzyme in the cascade, the E2. E3 ligases mediate the final step, which is the transfer of Ub to the substrate. Depending on the E3 type, this can either occur by direct transfer of ubiquitin or after thioester formation of the ubiquitin with the E3. Subsequently, an isopeptide bond is formed between the lysine  $\epsilon$ -amino group of the substrate and the C-terminal carboxyl group of Ub. In mammalian cells, ubiquitination ligation to the N-terminal amino group and non-canonical residues such as serine, threonine, or cysteine forming peptide, ester, or thioester linkages, have also been described.

The principles outlined above are well established. However, most of the time, the substrate, the ubiquitin, and the ubiquitin loaded E2 must all be present together at the reaction center. This indicates that the E3 ligases must play an important role in coordinating the alignments of multiple proteins involved in the reaction. Strikingly, the human genome encodes for about 500-600 ubiquitin ligases, which is comparable to the over 500 predicted kinases. They are highly important to control the efficiency of the ubiquitination reaction and to confer substrate specificity. Given that each ubiquitin ligase targets several substrates, dissection of the diverse mechanisms underlying substrate recognition by E3s would help to better understand how ubiquitination drives various cellular functions. In general, E3 ligases contain different domains and can be mechanistically divided into three main classes: RING (Really Interesting New Gene)/U-box, HECT (Homologous to E6AP C-terminus) and RBR (RING-Between-RING) type ligases [130]. The RING and U box types constitute the largest E3 subfamilies. They interact noncovalently with ubiquitin-loaded E2 ligases and transfer ubiquitin from the E2 to the substrate in a single step [130]. In contrast, the HECT-type and RBR type E3s require a catalytic cysteine residue and form a thioester-linked ubiquitin intermediate before passing ubiquitin on to a target protein via the transthioesterification reaction [130].

Ubiquitination is a reversible event, and its removal is mediated by several specialized families of proteases, the deubiquitinases (DUBs) [131]. More than 100 deubiquitinating enzymes hydrolyzing peptide bonds play crucial roles in recycling ubiquitin from proteasome substrates and in stabilizing proteins by counteracting their poly-ubiquitination.

Ubiquitin can be linked to substrates as a monomer, or in the form of isopeptide-linked polymers called polyubiquitin chains. Eight distinctive linkages can be formed either through Met1 (linear Ub chain) or seven lysine residues (Lys6, Lys11, Lys27, Lys29, Lys33, Lys48- and Lys63). The polymerization state is important as its structure can influence the substrate's fate. Ubiquitin can also be modified by other modifications, including phosphorylation, acetylation and ubiquitin-like modifications such as sumoylation [132]. This complexity was previously has called the 'ubiquitin code' [133], describing that various ubiquitin modifications create a plethora of ubiquitin chain topologies, which can affect the fate of a substrate protein in multiple ways.

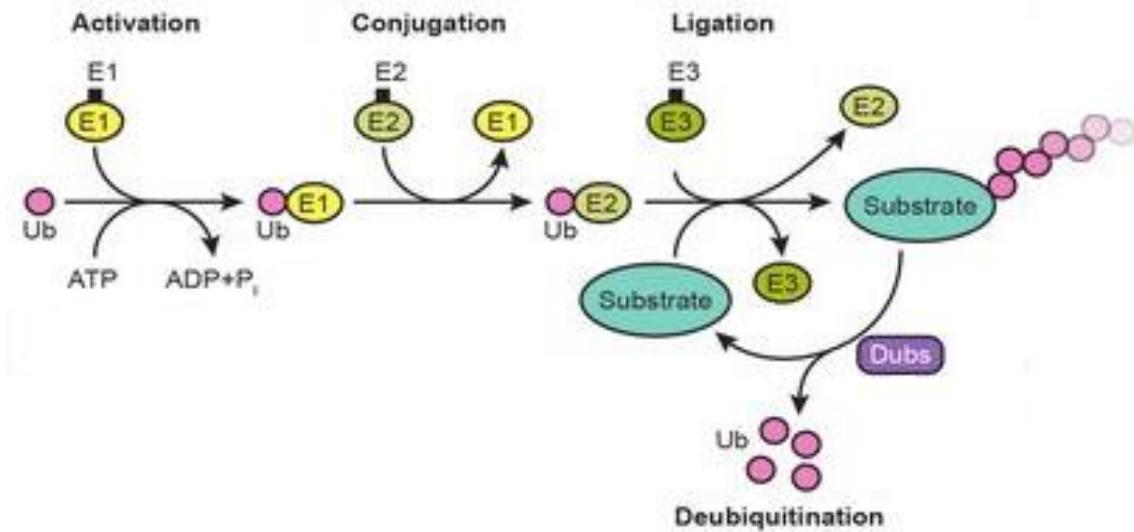


Figure 10. Summary of the ubiquitin system (adapted from [134])

### 1.2.3.2. Cellular ubiquitin-signaling

The importance of this protein-based modification first emerged 40 years ago when its well-known role to mark proteins for degradation and functionally distinct signals in proteasomal and lysosomal proteolysis was discovered [135]. Ubiquitination affects proteins in many ways and is also known to have non-proteolytic roles such as changing cellular localization of targets, modulating their activities, and altering protein-protein interactions. Large-scale proteomics studies have identified thousands of proteins targeted for ubiquitination [136-139], indicating that most proteins experience ubiquitination at some point in their cellular lifetime.

The mono-ubiquitination plays an important role in many other cellular processes such as protein and membrane trafficking, protein kinase activation, DNA damage repair, and DNA replication [140]. With the discovery of different linkage types and linkage-specific enzymes, researchers became very interested in studying ubiquitin chain signaling to identify distinct roles of various linkage types. As mentioned above, besides the heterotypic chains branched from eight possible linkage points, Ubiquitin can harbor Ubl modifications or acetylation on Lys residues and eleven potential phosphorylation sites, generating a complex code of a nearly unlimited number of potential combinations [132].

Poly-ubiquitin chains, in particular the most abundant K48-linked chain, serve as recognition signals for the 26S proteasome and initiate proteolysis of the substrate, which is the primary regulator system for protein abundance in cells [141]. In contrast, the second most abundant chain type linked via K63 has various non-degradative roles, such as endosomal sorting [141].

K33 linkages are implicated in post-Golgi protein trafficking [142]. Although the cellular function of K6-linked ubiquitin chains remains unclear, several studies have connected it to mitophagy and mitochondrial quality control. The anaphase-promoting complex (APC/C), together with the E2 enzyme UBE2C forms K11 linkage, which is a powerful signal for proteasomal degradation that regulates mitotic exit [143]. K27 chain plays a role in innate immunity and DNA damage response which is triggered by the ATM pathway [144]. The K27-linked polyubiquitination of histone 2A (H2A) proteins represents the major chain type on chromatin upon DNA damage [145]. While K29-linked chains are known to inhibit the Wnt signaling pathway, linear ubiquitin chains (M1-linked) play pivotal roles in inflammatory and immune responses by regulating the activation of NF- $\kappa$ B [144]. Moreover, ubiquitin phosphorylation is involved in mitochondrial quality control through mitophagy, which is an important mechanism linked to neurological disorders such as Parkinson's disease. The phosphorylation of Ser65 on ubiquitin by PINK1 is required for activation of the ubiquitin ligase Parkin. This activation is essential for the recruitment of autophagy receptors and promotes the formation of K6-, K11-, K48- and K63-linked chains on MOM proteins by Parkin (reviewed in [144]). Although we still lack a clear and complete picture of the biochemical mechanisms carried out by eukaryotic ubiquitin-like modifiers, many studies have also elucidated their essential functions. For instance, sumoylation regulates nucleocytoplasmic transport and cell cycle progression [146] while ISGylation plays an important and specific role in interferon-mediated responses to viral infection [147]. Finally, dysregulation of the components of the ubiquitin system contributes to many diseases, such as cancer, neurodegenerative diseases, cardiovascular diseases, immunological disorders, and inflammatory diseases [5]. Unsurprisingly, the biotechnology and pharmaceutical industries has started to search for modes of regulations and potential chemical inhibition of the ubiquitin system.

#### **1.2.3.3. Quantifying ubiquitin-signaling by MS-based proteomics**

Ubiquitination is one of the most challenging PTMs to study due to its size, low abundance, and dynamic regulation. Furthermore, the large number of proteins that constitute the ubiquitin system and the enormous number of ubiquitin substrates require sophisticated approaches to study different aspects of ubiquitin signaling, including substrates and chain architecture (types, length, quantity, PTMs, etc.). Mass spectrometry has become a very powerful tool in quantifying ubiquitin-signaling in cells and tissues, and recent advancements in MS techniques and biochemical methods have greatly improved our understanding of the physiological significance of the ubiquitin system.

The pioneering method to identify ubiquitin targets used transient or ectopic expression of N-terminal epitope-tagged ubiquitin variants, such as 6xHIS-Ub and STREP-Ubi [136, 148]. These variants can still be conjugated to the substrate proteins and the conjugates are then enriched by affinity purification. Another strategy employed biotinylation under a very stringent denaturing

condition to identify ubiquitinated substrates [149]. Besides the identification of target proteins, MS also allows the precise mapping of ubiquitinated sites on targeted proteins (Figure 11). The most common method takes advantage of an immunoprecipitation approach, in which monoclonal antibodies specifically enrich for peptides derived from the ubiquitinated region of proteins [137, 138, 150]. The C-terminus of the mature ubiquitin has the amino acid sequence KESTLHLVLRLRG, in which the last glycine (Gly, G) is conjugated to lysine residues on target proteins. After proteolytic processing using trypsin the conjugated ubiquitin is released. This leaves two Gly residues on the modified lysine residues, generating a new type of peptide. These peptides are termed “ubiquitin remnant-containing peptides”. The di-glycine (di-Gly) adduct creates a 114.04 Da monoisotopic mass difference. Therefore, detection of this unique mass difference on lysine residues by MS allows the identification of both the ubiquitination site as well the ubiquitinated protein. The commercialization of di-Gly antibodies has significantly accelerated MS-based ubiquitinome analysis and enabled a variety of quantitative and systems-wide studies [151-155]. However, there are two major limitations of this approach. The first one is its inability to recognize linear ubiquitin signature peptides and the second one is that two ubiquitin-like modifiers, NEDD8 and ISG15, can also generate an identical di-Gly remnant on modified lysines when cleaved by trypsin. Therefore, it is impossible to distinguish them among all di-Gly based modifications detected by mass spectrometry. Although these ubiquitin-like modifications only account for a very small percentage of di-Gly modified peptides (<6%), the “ubiquitin remnant-containing peptides” do not represent the pure population of ubiquitinated peptides [156]. A recently described antibody targeting a longer remnant generated by the LysC digestion can exclude ubiquitin-like modifications and allow detection of N-terminal ubiquitination [139, 157].

To assess ubiquitin linkage-type, chain size and architecture, several methods have been developed in the last decade. Monoclonal antibodies that can recognize all ubiquitinated proteins or chains are important reagents in this field [158] and they have also been utilized in proteomics studies [143, 159]. However, these antibodies often exhibit high cross-reactivity. Ubiquitin traps, such as tandem ubiquitin-binding entities (TUBEs) [160], differing in number and types of UBD (ubiquitin-binding domain), linker length and type of epitope tag have also been engineered for affinity purification of ubiquitinated proteins. These tools have the advantage of protecting ubiquitinated proteins from DUBs during purification, and purified proteins can be further analyzed by western blotting or mass spectrometry. However, neither these tools nor antibodies can selectively recognize atypical linkages.

Recently, screening for non-antibody protein scaffolds resulted in the linkage specific ubiquitin affinity reagents for the detection of K6 and K33-linked polyubiquitin chains [161]. Although these linkage specific reagents are very useful, they cannot distinguish complex topology of ubiquitin branches. The ubiquitin chain enrichment middle-down MS (UbiChEM-MS) method allowed the characterization of branched conjugates by combining chain enrichment using linkage specific

UBDs, minimal trypsin digestion and middle-down MS [162]. Furthermore, the recently published Ub-clipping approach uses an engineered viral protease that removes ubiquitin but keeps the branching information on the modified lysine [163]. The remaining ubiquitin and substrate polypeptides are kept intact. This method enables quantification of multiple diGly-modified branch points and can also determine co-existing PTMs on ubiquitin modification. There are also several antibodies suitable for Western blotting that detect small chemical modifications such as Ser65 phosphorylation and Lys48 acetylation.

Quantitative proteomic methods, including SILAC and TMT labeling, can be coupled with the enrichment strategies to analyze relative changes in ubiquitination and identify perturbation-relevant regulatory sites in complex signaling networks. Absolute quantification was also adopted to quantify the abundance of ubiquitin chain linkage peptides (Ub-AQUA) by employing special labeled ubiquitin peptide standards [164]. An alternative approach called Ub-PSAQ has the advantage to account for experimental loss of protein by using ubiquitin protein standards. Several studies that developed or utilized the tools described above have been published in the last decades. They improved our understanding of the complex nature of ubiquitination-dependent regulation, the ubiquitin system and signaling.

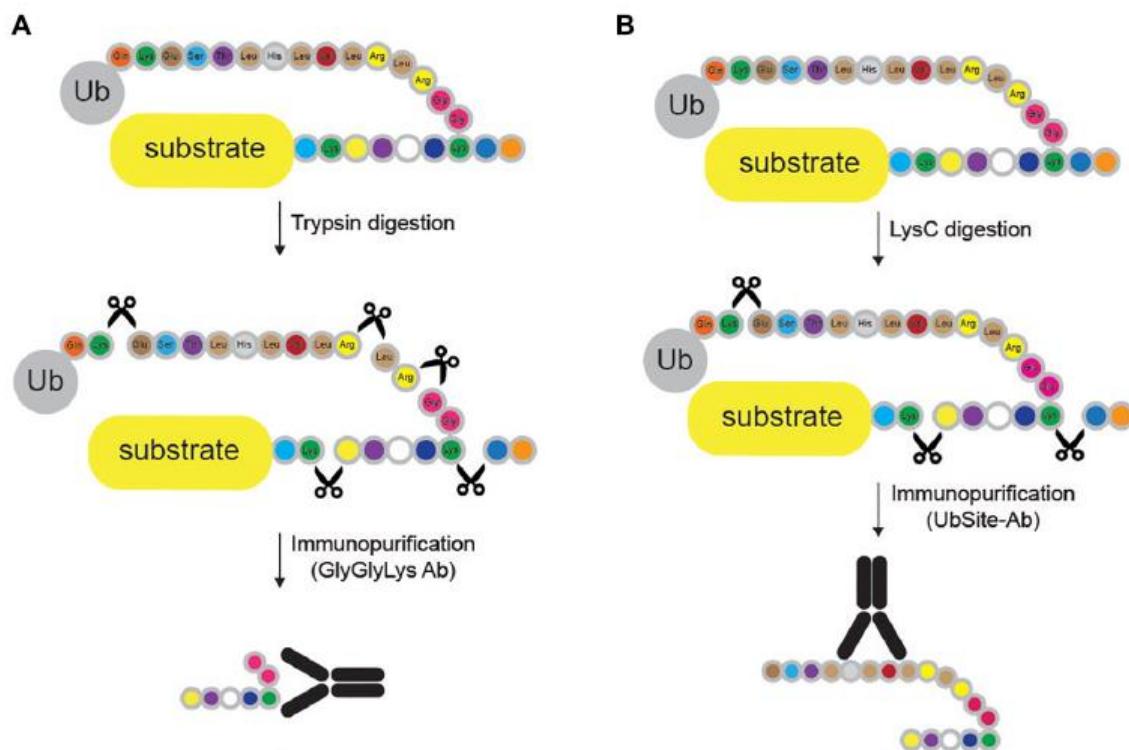


Figure 11. Methods for mapping ubiquitination sites in proteins (adapted from [129])

## 2. Aims of the thesis

In this thesis, I developed and applied different MS-based quantitative proteomics methodologies with high accuracy and sensitivity to investigate the system-level dynamics of post-translational modifications, which diversify and extend protein function beyond what is directly dictated by gene transcripts. I primarily focused on phospho- and ubiquitin-signaling networks.

In the first part of my thesis, we aimed to develop a streamlined and scalable sample preparation method for high sensitivity phosphoproteomics. We reasoned that such a protocol would substantially minimize sample loss and variability, thereby allowing high reproducibility and in-depth coverage of small input material. Our single-run, label-free workflow, which was built on the EasyPhos platform could quantify over tens of thousands of phosphorylation sites and is now routinely and successfully used in our laboratory. Next, we applied our protocol to obtain high coverage and accurate proteome and phosphoproteome data from *in vitro* reconstituted CD34+ HSPCs undergoing different stages of erythropoiesis. In this study, our major goal was to gain a system-wide understanding of how post-transcriptional and translational mechanisms are driving terminal maturation. This project was carried out in collaboration with the group of Mitchell J. Weiss at the St. Jude Children's Research Hospital and of Brenda Schulman at the Max-Planck Institute of Biochemistry.

Furthermore, we developed sensitive and accurate targeted MS-based phosphoproteomics methods. Due to the low stoichiometry of endogenous phosphorylated peptides, accurate measurements of phosphorylation levels upon kinase inhibitor treatment in clinical studies are challenging. We applied our targeted assay to detect PD kinase LRRK2 targets by systematically analyzing Rab protein family members that harbor the conserved LRRK2 phosphorylation site. Moreover, we modified the assay to robustly detect and quantify occupancies of phosphorylated Rab protein in PD patient cells. This assay is now a robust, highly sensitive and specific tool to monitor LRRK2 activity and assess LRRK2 inhibitor efficacy in clinical studies. With the assay in hand, we demonstrated that compared to healthy controls, the pathogenic mutation carriers have more than two-fold increased pRab10 levels. Our results establish pRab10 as a treatment efficacy- and stratification marker. These projects were conducted in an exceptionally fruitful collaboration with the Alessi Lab from the University of Dundee, Suzanne Pfeffer from Stanford University and funded by the Michael J Fox Foundation.

In the second part of my thesis, I embarked on the mass spectrometry-based analysis of the ubiquitinome by studying different aspects of ubiquitin signaling. In a team effort with a PhD student from our group, Fynn Hansen, we developed a sensitive and reproducible workflow for the study of the ubiquitin-modified proteome. Firstly, we aimed to investigate the power of Data Independent Analysis (DIA), which substantially improved data completeness and sensitivity in

single run of ubiquitinome analysis compared to the DDA method. To demonstrate the capabilities of our workflow, we also applied it to the well-studied TNF-signaling pathway and performed global ubiquitinome analysis of circadian rhythm. I also identified proteins modified by UBL3 -a novel ubiquitin-like molecule- in human cells. Furthermore, we aimed to establish a fast DIA workflow for yeast proteome profiling in a collaboration with the group of Brenda Schulman, where we subsequently used this workflow to find components and targets of the Gid E3 ligase complex under different metabolic conditions in yeast. Lastly, together with Brenda Shulman's group we investigated how the location of the ubiquitin lysine mutation determines the distribution of di-ubiquitin linkage types. For this goal, we established a targeted MS strategy to quantify the levels of Ub chain linkage types.

Since the beginning of the COVID-19 pandemic many research groups focus on the investigation of SARS-CoV2 and its mechanism of action to develop useful therapies. However, there were only a few systems-wide studies of the cellular correlates of viral infection at the proteomic level and they did not use the latest MS technology. Therefore, in the last months of my PhD study, my main objective was to use my extensive experience in MS-based proteomics to study host signaling processes affected by the SARS-CoV-2 at the levels of phosphorylation and ubiquitination. Together with the Pichlmair group at the TUM, we succeeded in obtaining a large picture of the infection process in lung cells at the molecular level.

### 3. Publications

#### 3.1. Phosphorylation

##### 3.1.1. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform

Sean J Humphrey<sup>1,2\*</sup>, Ozge Karayel<sup>3</sup>, David E James<sup>1,2,4</sup>, Matthias Mann<sup>3,5\*</sup>

<sup>1</sup>School of Life and Environmental Sciences, The University of Sydney, Australia <sup>2</sup>The Charles Perkins Centre, University of Sydney, Australia <sup>3</sup> Max Planck Institute of Biochemistry, Martinsried, Germany <sup>4</sup> Sydney Medical School, The University of Sydney, Australia <sup>5</sup>NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Denmark \* Corresponding authors

Published in *Nature Protocols* (2018)

Advances in enrichment strategies, instrument performance, and computational analysis tools have enabled global studies of dynamic protein phosphorylation. However, conventional phosphoproteomics generally entails complex workflows that require specialized equipment and relatively high amounts of sample lysate (~10 mg). Typically, urea-based protein digestion is followed by peptide desalting and lyophilization. These steps have the potential to cause sample loss. To reach depths required to cover key signaling components, many phosphoproteomic workflows also comprise fractionation methods, resulting in numerous LC-MS/MS measurements per biological sample to be analyzed. Finally, phosphopeptides are enriched using IMAC or TiO<sub>2</sub> materials. The streamlined phosphopeptide enrichment protocol called EasyPhos was published by Sean J. Humphrey in 2015 to address key challenges in this sample preparation process. This protocol enabled the study of global phosphoproteome dynamics with minimal sample amounts and reduced measurement times [1]. Although the protocol was being routinely and successfully applied in our laboratory and allowed the analysis of phosphoproteomes at a depth of >10,000 quantified phosphorylation sites, we still aimed for further improvements on the EasyPhos protocol.

In this project, our objective was to make the upstream (lysis) steps of EasyPhos protocol more streamlined and scalable, and to obtain even higher sensitivity in phosphoproteomics. We eliminated protein precipitation and peptide clean-up and performed all steps in 96-well plates. Our protocol significantly minimized opportunities for sample loss and variability, thereby allowing high reproducibility and small input size ( $\leq 200$   $\mu$ g of protein starting material). The entire protocol only takes about one day, whereas MS measurements require less than an hour per sample. We published the updated protocol in detail in *Nature Protocols* in 2018, where it has been cited 50 times so far.

# High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform

Sean J. Humphrey<sup>1,2\*</sup>, Ozge Karayel<sup>3</sup>, David E. James<sup>1,2,4</sup> and Matthias Mann<sup>1,5\*</sup>

Mass spectrometry has transformed the field of cell signaling by enabling global studies of dynamic protein phosphorylation ('phosphoproteomics'). Recent developments are enabling increasingly sophisticated phosphoproteomics studies, but practical challenges remain. The EasyPhos workflow addresses these and is sufficiently streamlined to enable the analysis of hundreds of phosphoproteomes at a depth of >10,000 quantified phosphorylation sites. Here we present a detailed and updated workflow that further ensures high performance in sample-limited conditions while also reducing sample preparation time. By eliminating protein precipitation steps and performing the entire protocol, including digestion, in a single 96-well plate, we now greatly minimize opportunities for sample loss and variability. This results in very high reproducibility and a small sample size requirement ( $\leq 200$   $\mu$ g of protein starting material). After cell culture or tissue collection, the protocol takes 1 d, whereas mass spectrometry measurements require ~1 h per sample. Applied to glioblastoma cells acutely treated with epidermal growth factor (EGF), EasyPhos quantified 20,132 distinct phosphopeptides from 200  $\mu$ g of protein in less than 1 d of measurement time, revealing thousands of EGF-regulated phosphorylation events.

## Introduction

Phosphorylation is a widespread post-translational modification that regulates essentially all cellular processes<sup>1</sup>. With at least two-thirds of cellular proteins harboring one or more phosphorylation sites<sup>2</sup>, elucidating complex phosphorylation events on a proteome-wide scale is a major endeavor. Mass spectrometry (MS)-based proteomics has been used to study dynamic cell signaling for more than a decade<sup>3,4</sup>, and phosphoproteomics has now become a burgeoning field<sup>5–9</sup>.

We recently developed the EasyPhos technology to facilitate rapid and reproducible phosphopeptide enrichment by optimizing and minimizing workflow steps and performing the entire protocol from digestion to enrichment in parallel 96-well format<sup>10</sup>. EasyPhos maximizes phosphoproteome coverage and reproducibility of phosphopeptide quantification in single liquid chromatography–MS (LC–MS) runs, without requiring fractionation. The single-run format greatly increases sample throughput, enabling large experiments comprising hundreds of samples to be analyzed with minimal measurement time, in turn enabling sophisticated phosphoproteomics experiments to be performed. We have recently applied this workflow to several studies of global phosphorylation dynamics, including those examining time-resolved insulin signaling<sup>10</sup> and circadian rhythms<sup>11</sup> in the liver, kappa opioid receptor signaling in the brain<sup>12</sup> and glucose- and drug-enhanced insulin secretion in pancreatic insulinoma cells<sup>13</sup>, as well as using it to identify bona fide substrates of LRRK2 (ref. <sup>14</sup>).

Here we provide an in-depth and updated workflow for the enrichment of phosphopeptides from complex samples in single-run format using the EasyPhos platform. We highlight critical steps in the preparation of samples for phosphoproteome analysis and evaluate instrument parameters for efficient sequencing of phosphopeptides on two versions of the benchtop Orbitrap mass spectrometer, which should serve as a starting point for phosphoproteome measurements on other LC–MS platforms. We also briefly describe the analysis of phosphoproteomics data using the MaxQuant computational platform and demonstrate how the protocol is appropriate for rapidly measuring global phosphoproteomes by analyzing EGF-treated glioblastoma cells.

<sup>1</sup>School of Life and Environmental Sciences, The University of Sydney, Sydney New South Wales, Australia. <sup>2</sup>The Charles Perkins Centre, The University of Sydney, Sydney New South Wales, Australia. <sup>3</sup>Max Planck Institute of Biochemistry, Martinsried, Germany. <sup>4</sup>Sydney Medical School, The University of Sydney, Sydney New South Wales, Australia. <sup>5</sup>NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. \*e-mail: sean.humphrey@sydney.edu.au; mman@biochem.mpg.de

## PROTOCOL

## NATURE PROTOCOLS

## Approaches for analyzing global phosphorylation by mass spectrometry

The low abundance, high complexity, and highly dynamic nature of phosphorylation necessitates the enrichment of peptides containing only phosphorylated serine, threonine, and tyrosine residues before measurement by MS. To this end, several methods for enriching phosphopeptides from a complex background of nonphosphorylated peptides have been developed. Most strategies have used a dimension of peptide fractionation, either after<sup>15,16</sup>, or more often before, a second phosphopeptide enrichment stage using immobilized-metal affinity chromatography (IMAC)<sup>3,17–20</sup> or metal-oxide affinity chromatography (MOAC; typically using TiO<sub>2</sub> (refs<sup>21–24</sup>)). The peptide-fractionation approach of choice has most commonly been strong cation exchange (SCX) chromatography<sup>4,19,25–27</sup>; however, more recently, high-pH reversed-phase chromatography combined with fraction concatenation<sup>28</sup> has also been shown to be particularly powerful<sup>29</sup>. Such multistage approaches have enabled deep phosphoproteome coverage<sup>2,30</sup>; however, this often requires specialized fractionation equipment and can be difficult to make sufficiently robust, especially for label-free quantification.

Single-run phosphoproteomics workflows, in which one or more biological samples are measured in each LC–MS run, are particularly powerful because they greatly improve sample throughput, accelerating the measurement of studies comprising many samples, such as time-series experiments<sup>10,31</sup>. Single-run phosphoproteomics methods also have an inherent sensitivity advantage over multistage approaches involving peptide fractionation because peptide species are not diluted between multiple fractions. For example, a previous single-run approach quantified ~1,000 phosphorylation sites from only 10,000 cells<sup>32</sup>. We recently developed a single-run phosphoproteomics workflow, called EasyPhos, with the aim of simplifying the measurement of very large numbers of phosphoproteomes without compromised coverage<sup>10</sup>. A key feature of EasyPhos is the elimination of peptide desalting steps before phosphopeptide enrichment. In other protocols, peptide desalting is typically performed by solid-phase extraction (SPE), using C18 or a mixed-mode material that enables the removal of interfering compounds that are not compatible with the subsequent phosphopeptide enrichment steps. In the original EasyPhos workflow, these peptide desalting steps are eliminated by performing protein precipitation, followed by protein digestion in a buffer that is directly compatible with phosphopeptide enrichment, 2,2,2-trifluoroethanol (TFE)<sup>10</sup>. This offers several advantages: (i) considerable costs are saved by eliminating materials required for peptide desalting by SPE (often several hundred dollars for 100 samples); (ii) opportunities for peptide losses during desalting are minimized (because of incomplete binding to the solid phase, incomplete peptide reconstitution, or precipitation); (iii) variability both within and between experiments is reduced; and (iv) it requires fewer time-consuming steps. Collectively, this facilitates the enrichment of many samples in parallel.

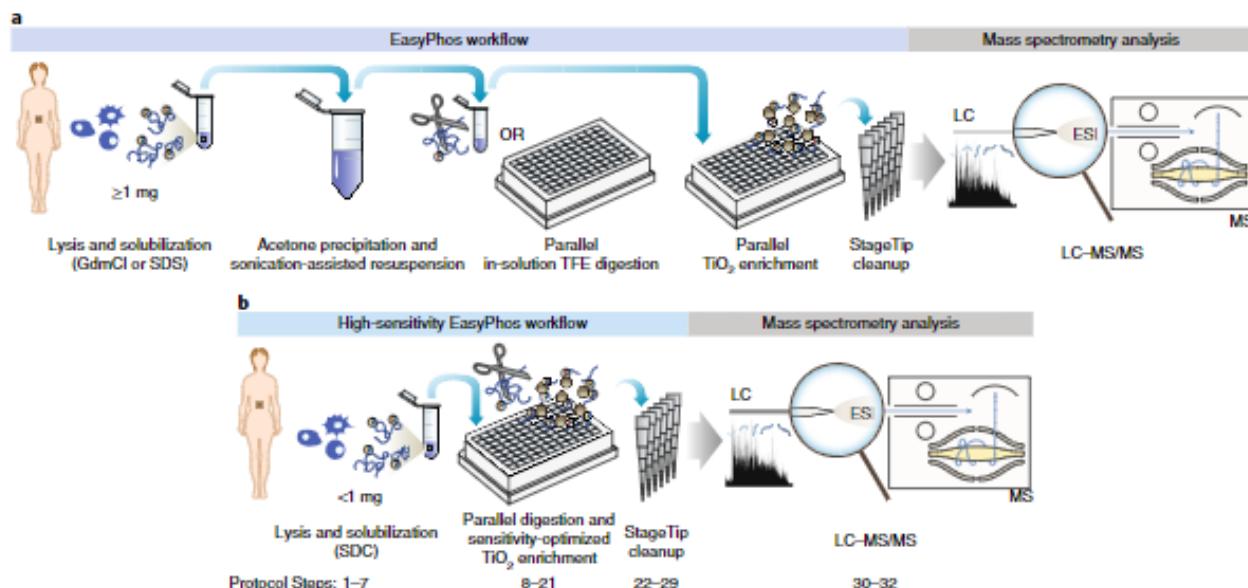
Although the published EasyPhos method performs well in diverse biological contexts, it was originally optimized for ~1 mg of protein starting material. To enable its use in situations in which only minuscule amounts of material are available, we have made several key improvements, described here. First, we now eliminate the requirement for protein precipitation, enabling the entire workflow to be performed in a single tube or a 96-well plate. This limits sample losses during precipitation and during contact with multiple types of plasticware, reduces processing time by a whole day, and facilitates rapid and reproducible sample handling. Second, we extensively re-evaluated all buffers used throughout the workflow to maximize performance with reduced amounts of starting material. To balance sensitivity with deep phosphoproteome coverage, we aimed for a five-fold reduction in protein material used in a typical experiment (to 200 µg). Together, these changes markedly improve the performance of EasyPhos in high-sensitivity applications in which <1 mg of protein material is available. The refined workflow is therefore now applicable to a greater range of biological and clinical contexts. Moreover, the method can now be completed in less than half the time of the original workflow, with fewer manual sample-handling steps, and substantially reduced material costs.

## Improvements to the EasyPhos workflow

In the originally described method, proteins were solubilized in buffers containing powerful chaotropic agents such as guanidine hydrochloride (GdmCl) (refs<sup>10,13</sup>) or sodium dodecyl sulfate (SDS; ref.<sup>11</sup>). These reagents must thereafter be removed because they are not directly compatible with phosphopeptide enrichment, they strongly inhibit the activity of trypsin (GdmCl)<sup>33</sup>, and they interfere with reversed-phase separation of peptides and electrospray ionization (SDS). This necessitates protein precipitation using acetone or chloroform–methanol, and protein pellets must be resuspended as microparticles using ultrasound in a mild TFE-based buffer for proteolytic digestion

## NATURE PROTOCOLS

## PROTOCOL



**Fig. 1 | Schematic of the EasyPhos platform for high-throughput phosphoproteomics.** **a,b,** Steps involved in preparing samples using the originally published EasyPhos workflow (**a**) and the updated high-sensitivity workflow described in this protocol (**b**). Samples are lysed in sodium deoxycholate (SDC) buffer, digested with Lys-C and trypsin, and enriched directly in 96-well-plate format, followed by StageTip cleanup and LC-MS/MS measurement. Steps associated with the protocol described here are indicated.

(Fig. 1a). This approach is highly effective at removing interfering buffer components and works particularly well with large amounts of protein at high concentrations. However, protein precipitation is often not reliable when minimal material is available, and it is also not easily applied to large numbers of samples. To resolve these limitations, here we present an updated version of the EasyPhos workflow in which lysis, digestion, and phosphopeptide enrichment steps can be performed in a single tube, without the need for intermediate protein precipitation or peptide cleanup steps (Fig. 1b). We reasoned that this could be achieved by identifying a lysis buffer that simultaneously (i) is strong enough to efficiently lyse cells and disrupt intracellular membranes to enable effective protein solubilization; (ii) does not substantially interfere with the activity of trypsin; and (iii) is completely compatible with phosphopeptide enrichment steps and subsequent LC-MS measurement. Sodium deoxycholate (SDC) satisfies the first two criteria, efficiently solubilizing proteins at high concentrations without inhibiting trypsin, and producing high-quality proteome digests<sup>34,35</sup>. However, phosphopeptide enrichment is performed at low pH (e.g., using buffers containing 6% (vol/vol) trifluoroacetic acid (TFA) to protonate acidic residues (glutamic and aspartic acid), thereby limiting their binding to MOAC or IMAC beads. This in turn reduces the recovery of unwanted nonphosphorylated acidic peptides. Under these low-pH conditions, however, SDC precipitates due to the extremely low aqueous solubility of deoxycholic acid (the conjugate acid of SDC), and risks co-precipitating peptides, complicating its use as a lysis buffer for phosphoproteomics workflows.

To resolve the acid-induced precipitation of SDC, we investigated the solubility of deoxycholic acid in various solvents. Although it is essentially insoluble in water (<0.05% (wt/vol)), we found that it is around fivefold more soluble (~0.25% (wt/vol)) in 50% (vol/vol) acetonitrile (ACN), and a striking >100-fold more soluble (>5% (wt/vol)) in 50% (vol/vol) isopropanol (ISO). ACN is a polar aprotic solvent, whereas ISO is a polar protic solvent and can therefore better solubilize bile acids such as deoxycholic acid via hydrogen bonding. The presence of at least 50% (vol/vol) ACN has been a mainstay feature of almost all phosphopeptide enrichment protocols because it very effectively reduces nonspecific binding of nonphosphorylated peptides<sup>36</sup>. It was previously shown that high concentrations of ACN (70% (vol/vol)) enable the use of low concentrations of SDC (12 mM, i.e., 0.5% during lysis and 0.1% during proteolysis) in single-run phosphoproteomics workflows<sup>32</sup>. Here we show for the first time that by substituting ACN for ISO in the phosphopeptide enrichment buffer, we can exploit this ability of ISO to eliminate the precipitation of even very high concentrations of SDC (>1% (wt/vol)) under acidic conditions. This in turn provides a buffer that satisfies all three conditions presented above and eliminates the need to perform any protein precipitation or peptide

## PROTOCOL

## NATURE PROTOCOLS

desalting steps before phosphopeptide enrichment. The resulting phosphoproteomics workflow is therefore ideally suited to high-sensitivity applications. It also increases throughput and applicability to robotic automation. ISO is also an attractive wash-buffer solvent, as it is both relatively inexpensive (costing around half as much as ACN) and, as we found in practice, particularly effective at solubilizing and removing nonpeptide contaminants from beads and plasticware without interfering with phosphopeptide enrichment. Therefore, ISO replaces ACN in both enrichment and wash buffers in the updated workflow.

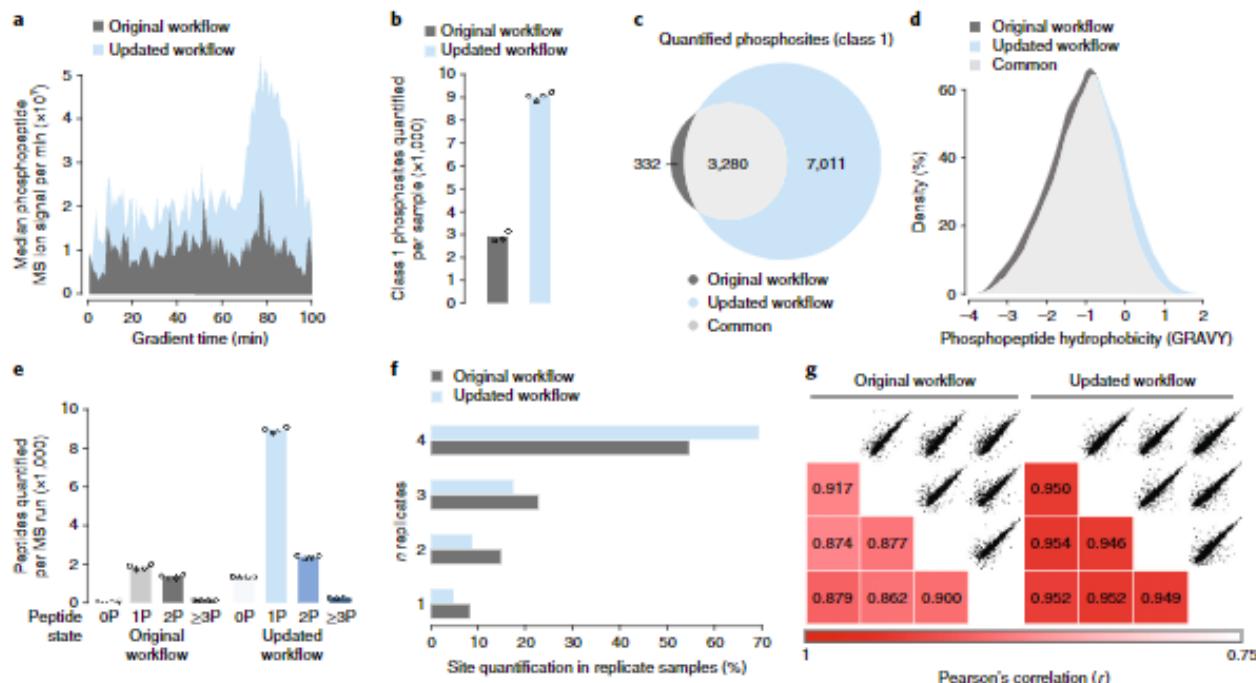
Because EasyPhos was originally designed to handle ~1 mg of starting material, we also systematically investigated experimental parameters with a view toward improving performance by using smaller amounts of starting materials. These included the specific composition of enrichment, wash and elution buffers and conditions, as well as steps for post-enrichment phosphopeptide desalting. Many parameters had little effect on the performance of the workflow at higher sample loads, but when smaller amounts of starting materials were used, opportunities for maximizing phosphopeptide identification became apparent. For example, using limited amounts of starting materials, we found that reducing the concentration of monopotassium phosphate ( $\text{KH}_2\text{PO}_4$ ), used in the enrichment buffer to enhance the specificity of phosphopeptide enrichment, from 5 to 1 mM resulted in consistently higher phosphopeptide yields and substantial gains in sensitivity. Through these optimizations, we ultimately improved the performance of the protocol under sample-limited conditions.

#### Evaluating the sensitivity and reproducibility of the method

To quantitatively compare the performance of the updated method when used with low amounts of starting materials, we directly compared it with that of the original workflow<sup>10</sup>, including the acetone-precipitation steps. With 200  $\mu\text{g}$  of protein starting material, and using 90-min LC gradients on a Q Exactive HF mass spectrometer, the median MS1 ion signal contributed by phosphopeptides is around twofold greater across the entire LC gradient when using the workflow described in this protocol (Fig. 2a), leading to the identification of approximately threefold more phosphorylation sites under otherwise identical conditions (starting material, mass spectrometer, and LC) (Fig. 2b). The updated workflow also identifies >90% of the phosphorylation sites found by the original workflow, in addition to a large pool of newly identified sites (Fig. 2c). These tend to be more hydrophobic (Fig. 2d), despite being shorter in length on average, and have fewer missed cleavages (Supplementary Fig. 1a,b). This suggests that SDC digestion may be better able to access cleavage sites that were previously inaccessible. The majority of newly identified phosphopeptides are singly phosphorylated, because of optimized enrichment buffers, and notably, this comes with a gain rather than loss of higher-phosphorylated (2P and 3P) species (Fig. 2e). A larger proportion of phosphopeptides were quantified in all replicate samples measured (Fig. 2f, Supplementary Fig. 1c). This is likely due to improved phosphopeptide MS1 signal intensity, again due to optimized enrichment conditions. Elimination of protein precipitation and reduced sample-handling steps have a beneficial effect on reproducibility, with correlation coefficients of workflow replicates as high as 0.95 (average) as compared with 0.88 (average) under conditions identical to those of the original workflow (Fig. 2g).

To further evaluate performance across a range of starting materials, we performed a dose-titration experiment in which the protein amount was varied from 25 to 750  $\mu\text{g}$ , and phosphopeptides were measured on a Q Exactive HF benchtop instrument with 90-min LC gradients (Supplementary Fig. 2a). We performed all experiments in quadruplicate, with replicate samples enriched from identical starting material divided according to protein content (bicinchoninic acid (BCA) assay) and subsequently kept entirely separate for the complete workflow (i.e., the 25- $\mu\text{g}$  condition corresponds to 25  $\mu\text{g}$  of protein processed four times separately, not 100  $\mu\text{g}$  of starting material processed once and diluted for four replicate injections). This is an important distinction, as the adsorption of small amounts of protein and peptide to plasticware probably has a major effect on sensitivity, as well as variability. The updated workflow is capable of quantifying almost 4,000 phosphopeptides in each sample from as little as 25  $\mu\text{g}$  of cell lysate and >8,000 phosphopeptides using the 'match between runs' algorithm in MaxQuant and a library comprising the other measurements in the same experiment (Supplementary Fig. 2a). Under these sample-type, MS and LC conditions, we found that the method effectively saturates at ~500  $\mu\text{g}$ , with ~16,000 phosphopeptides quantified in each sample.

Recent improvements in benchtop Orbitrap mass spectrometers have increased scanning speed and ion isolation performance (Q Exactive HF)<sup>37,38</sup>, and more recently, the manufacturer has increased ion current by incorporating an ion funnel<sup>39</sup> and a larger ion transfer capillary (Q Exactive



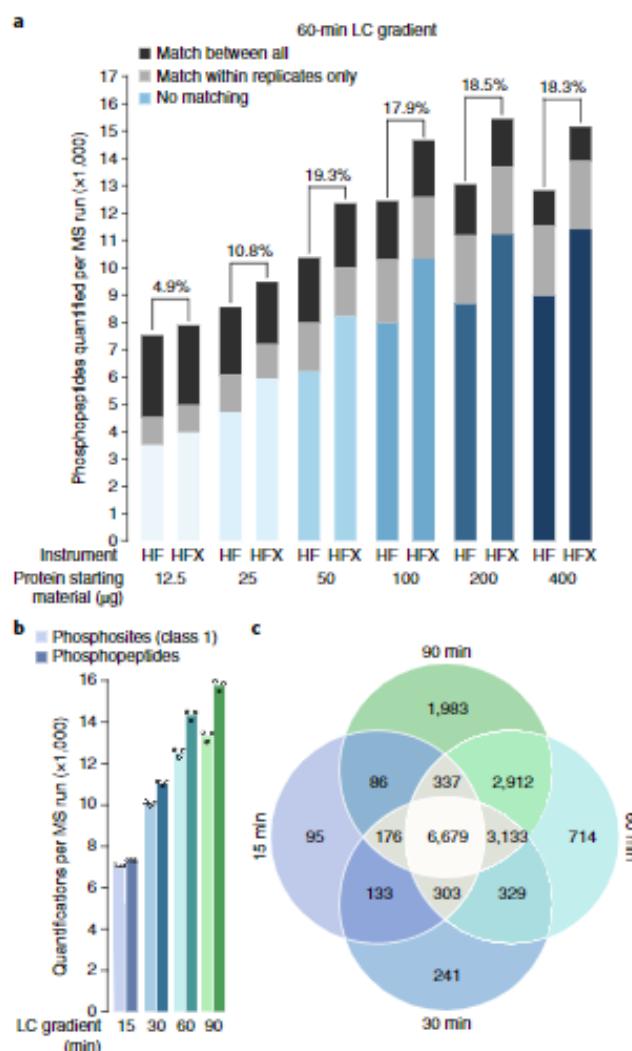
**Fig. 2 | Performance evaluation of the EasyPhos method using limited starting material.** All data were measured on a Q Exactive HF, and samples were prepared from protein equally divided into four workflow replicates ( $n = 4$ ) for the two workflows. For each replicate, 200  $\mu$ g of protein (determined by BCA assay) from untreated U-87 cells was processed using either the EasyPhos workflow as originally described<sup>10</sup> ('Original workflow') or the updated protocol described here ('Updated workflow'). All samples were measured with identical conditions on a Q Exactive HF mass spectrometer with a 90-min LC gradient (3–41% buffer B; 80% (vol/vol) ACN/0.1% (vol/vol) formic acid). Only the amount of protein for a single replicate (200  $\mu$ g) was prepared for each replicate sample. **a**, Median phosphopeptide MS1 intensity per 1-min window across the acquisitions. **b**, Average number of quantified phosphorylation sites (class 1; localization probability  $> 0.75$ ) in each sample. **c**, Overlap between all class 1 phosphorylation sites quantified by each method. **d**, Distribution of peptide hydrophobicity (GRAVY<sup>51</sup>) for quantified phosphopeptides. **e**, Proportion of phosphopeptides containing one (1P), two (2P), or three (3P) phosphate groups. **f**, Distribution of phosphorylation sites quantified in  $n$  replicate samples. **g**, Pearson correlation coefficients and scatterplots depicting the log<sub>2</sub>-transformed intensities of phosphopeptides from the replicate phosphopeptide enrichment experiments.

HF-X)<sup>40</sup>. This development is particularly relevant to phosphoproteomics experiments because these are constrained by the weak ion signal of low-abundance phosphopeptides. This results in relatively long MS/MS accumulation times to achieve the high fragmentation spectral quality needed to accurately localize phosphorylation sites, which in turn impairs the ability to accumulate precursor ions and to detect fragment ions in parallel. The increased ion current provided by the latest generation of benchtop Orbitrap instruments therefore promises to be particularly beneficial for phosphoproteomics experiments. We optimized key parameters to maximize phosphopeptide identification on this new instrument, including ion funnel radio frequency (RF), collision energy, and MS2 resolution and injection time (Supplementary Fig. 3), and compared the performance of the EasyPhos platform under high-sensitivity scenarios. Beginning with only 12.5  $\mu$ g of material, we measured phosphopeptides with shorter LC gradients of 1 h on both the Q Exactive HF ('HF') and the Q Exactive HF-X ('HF-X') instruments (Fig. 3a and Supplementary Fig. 2b). As with the previous dose-titration series, we performed quadruplicate experiments with replicate samples representing separate workflow repeats using the same protein starting material. The Q Exactive HF-X instrument provided an ~20% improvement in phosphopeptide identification from 50  $\mu$ g of material, on a par with the performance of the Q Exactive HF instrument at twice the sample load (100  $\mu$ g). The Q Exactive HF-X saturated at similar sample load; however, the maximum phosphopeptide identification achieved per sample was considerably higher, reaching >15,000 phosphopeptides in each sample measured.

Finally, we assessed the impact of different LC gradient durations on phosphopeptide identification using the EasyPhos platform. For a gradient time denoted  $x$ , all gradients went from 2.4 to 15.2% (vol/vol) ACN in  $0.67 \times x$  min, followed by 15.2 to 32.8% (vol/vol) ACN in  $0.33 \times x$  min (Equipment setup). All gradients were followed by a washout of 72% (vol/vol) ACN for 10 min. This

## PROTOCOL

## NATURE PROTOCOLS



**Fig. 3 | Evaluation of the performance of the updated EasyPhos workflow with respect to starting material, MS instruments, and LC gradient duration. a**, Average number of phosphopeptides quantified in each MS run of a dose-titration experiment spanning 12.5–400 µg of total protein material (determined by BCA assay) from untreated U-87 cells. Protein was pooled immediately after cell lysis and divided into four workflow replicates ( $n = 4$ ) for each condition. Only the amount of protein for a single replicate was prepared for each independent workflow replicate. Samples were measured on either a Q Exactive HF ('HF') or a Q Exactive HF-X ('HF-X') mass spectrometer as indicated, using 60-min LC gradients. Data were processed with or without the 'Match between runs' algorithm in MaxQuant enabled as indicated, to transfer MS2 identifications between samples. **b**, Number of class 1 phosphorylation sites (lighter bars) and phosphopeptides (darker bars) quantified in each MS run on a Q Exactive HF-X instrument from 200 µg of protein (determined by BCA assay) from untreated U-87 cells, with different LC gradient durations ( $n = 3$ ): for a gradient time denoted  $x$ , all gradients went from 2.4 to 15.2% (vol/vol) ACN in  $0.67 \times x$  min, followed by 15.2 to 32.8% (vol/vol) ACN in  $0.33 \times x$  min. All gradients were followed by a washout of 72% (vol/vol) ACN for 10 min. **c**, Overlap of total quantified phosphorylation sites (class 1; localization probability  $> 0.75$ ) measured using the different LC gradient lengths depicted in **b**.

approximately matches the solvent transit time from our pump to the emitter with the 40-cm column and  $350 \text{ nl min}^{-1}$  flow rate used, and therefore MS phosphopeptide data acquisition duration (i.e., the time during which the vast majority of peptides are eluting) closely approximates the gradient time indicated. With enrichments from 200 µg of unstimulated U-87 cell lysate measured on the Q Exactive HF-X system, we quantified over 13,000 class 1 phosphorylation sites and  $>15,000$  unique phosphopeptides in each run with 90-min gradients, and more than 7,000 class 1 sites and phosphopeptides in each run with 15-min gradients (Fig. 3b). From three replicate measurements, there

were in total 15,306 class 1 phosphorylation sites quantified with 90-min gradients and 7,809 class 1 sites with 15-min gradients (Fig. 3c). This represents a per-run phosphoproteome depth of ~75% greater than that of another recent single-shot phosphoproteomics study measured from 200 µg of HeLa cells, also using fast 15-min gradients on the HF-X instrument<sup>40</sup>.

## Materials

### Reagents

- Cells or tissue samples. We demonstrate our protocol by using the human glioblastoma cell line U-87 MG (ATCC, cat. no. HTB-14), which tested negative for mycoplasma contamination. This protocol can be readily adapted for use with various cell lines, primary cells, and tissues (see Reagent setup for details on cell/tissue preparation). **!CAUTION** All experiments involving animal or human samples should adhere to the relevant governmental and institutional ethics guidelines and regulations.
- **!CAUTION** The cell lines used in your research should be regularly checked to ensure that they are authentic and are not infected with mycoplasma.
- Sodium deoxycholate (SDC; Sigma-Aldrich, cat. no. D6750)
- Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl; Pierce, cat. no. 20490)
- 2-Chloroacetamide (CAM; Sigma-Aldrich, cat. no. C0267)
- Bicinchoninic acid (BCA) Protein Assay Kit (Pierce, cat. no. 23225)
- Potassium hydroxide (KOH; Sigma-Aldrich, cat. no. P5958)
- Calcium chloride (CaCl<sub>2</sub>; Sigma-Aldrich, cat. no. 793639)
- Potassium dihydrogen phosphate (KH<sub>2</sub>PO<sub>4</sub>; Merck, cat. no. 5438410100)
- Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl)
- Endopeptidase Lys-C (Wako Chemicals, cat. no. 129-02541)
- Proteomics-grade modified trypsin (Sigma-Aldrich, cat. no. T6567). We use Sigma proteomics-grade trypsin because we find it provides high-quality digests under the conditions described and is available in various sizes.
- Acetonitrile (ACN; Fisher Scientific, cat. no. A955-4)
- Methanol (MeOH; Fisher Scientific, cat. no. A456-1)
- Isopropanol (ISO; Fisher Scientific, cat. no. A461-1)
- Acetic acid (AcOH; Fisher Scientific, cat. no. A11350)
- Trifluoroacetic acid (TFA; Merck, cat. no. 8082600100)
- Ammonia solution (25% wt/vol; NH<sub>4</sub>OH; Merck, cat. no. 5330030050)

### Equipment

- Deep-well plates, 96 wells (2-ml square wells; Eppendorf, cat. no. 0030504305). We use Eppendorf 2-ml deep-well plates because the design of these deep-well plates (square edges with round bottom) simplifies rapid supernatant aspiration, allowing aspiration without disturbing the bead pellets (Supplementary Fig. 4).
- Silicone sealing mats for 2-ml deep-well plates (Eppendorf, cat. no. 0030127960). We use Eppendorf sealing mats because they have high chemical resistance, do not leach polymers into the samples, and seal tightly, preventing leakage during vigorous mixing.
- Temperature-controlled high-speed orbital shaker. We use an Eppendorf ThermoMixer C (Eppendorf, cat. no. 5382000015) because adapters are available for both 2-ml tubes and deep-well plates, it supports high mixing velocities (2,000 r.p.m.), and a 'ThermoTop' accessory is available that prevents lid condensation during prolonged incubations. If an alternative shaker is used, ensure that it can be heated to 95 °C, can accommodate both 2-ml tubes and deep-well plates, and can operate at mixing velocities of 2,000 r.p.m.
- Microtip-probe sonicator (Branson, cat. no. SFX250 or SFX450) and microtip probe (Branson, cat. no. 9655V30) or a powerful sonicating water bath. We use the Diagenode Bioruptor Plus (Diagenode, cat. no. B01020001) or Q-Sonica Q800R (Q-Sonica, cat. no. Q800R) systems because they provide a very high output power, sufficient to shear DNA, and enable processing of 12 samples in parallel without sample-instrument contact.
- 2-ml tube adapter for orbital shaker (Eppendorf, cat. no. 5362000035)
- Deep-well plate adapter for orbital shaker (Eppendorf, cat. no. 5363000039)
- ThermoTop accessory for orbital shaker (Eppendorf, cat. no. 5308000003)
- Solid-phase extraction disks for SDB-RPS StageTips: Empore SDP-RPS (Sigma-Aldrich, cat. no. 66886-U). We use self-fabricated plugged StageTips because they are inexpensive and enable the use of

## PROTOCOL

## NATURE PROTOCOLS

mixed-mode SDB-RPS (strong cation exchange, reversed-phase) material. Commercial alternatives are available; however, buffers may need to be adapted as appropriate.

- Solid-phase extraction disks for C8 Stage'ips: Empore C8 (Sigma-Aldrich, cat. no. 66882-U)
- Titanium dioxide ( $\text{TiO}_2$ ) beads: Titansphere Phos- $\text{TiO}$  bulk, 10  $\mu\text{m}$  (GL Sciences, cat. no. 5010-21315). Alternative beads and bead chemistries such as IMAC may be used; however, the optimal composition of enrichment, wash, and elution buffers may vary from those defined here.
- In-house 96-well Stage'ip centrifuge device. 3D printers using fused-deposition modeling such as the Zortrax M200 or Ultimaker 3 instruments enable the rapid (~12-h) and inexpensive fabrication of devices capable of holding Stage'ip devices and collection tubes for centrifugal-based processing of up to 96 samples in parallel. If such a device is not available, a pipette-tip box may serve as a suitable Stage'ip holder. **!CAUTION** Care should be taken when centrifuging to ensure items are fully contained within centrifuge buckets, that they are equally balanced, and that centrifugation speeds used do not result in damage.
- Electronic positive-displacement pipette or eight-channel electronic pipette: Eppendorf Multipette E3 (Eppendorf, cat. no. 4987000010) and Combitips (5 ml, 10 ml; Eppendorf, cat. no. 0030089669, 0030089677) or Eppendorf Xplorer (Eppendorf, cat. no. 4861000163) and pipette tips, for rapid liquid dispensing (50–1,250  $\mu\text{l}$ ; Eppendorf, cat. no. 0030073320)
- Eight-channel 200- $\mu\text{l}$  pipette: Eppendorf Research plus (Eppendorf, cat. no. 3122000051) or Thermo Fisher Finnpipette F1 (Thermo Fisher Scientific, cat. no. 4661030N) and 200- $\mu\text{l}$  pipette tips (Eppendorf, cat. no. 0030073266)
- PCR strip tubes (Thermo Fisher Scientific, cat. no. AB-0452)
- Sealing mats for PCR tubes: Axymat silicone sealing mat (Axygen, cat. no. AM-96-PCR-RD)
- Vacuum liquid aspiration line and disposable borosilicate Pasteur pipettes, 5½ inches (Corning, cat. no. 7095D-5X)
- Evaporative concentrator (Eppendorf Vacuum Concentrator Plus or Genevac EZ-2), with 96-well plate rotor **▲CRITICAL** Timings provided are based on the Eppendorf instrument and may vary with other instruments, depending on evaporative rate.
- Nanospray columns for online ultra-high-performance liquid chromatography (UHPLC)–MS/MS analysis. We use self-pulled fused-silica (Polymicro, cat. no. 1068150019), 360- $\mu\text{m}$  o.d.  $\times$  75- $\mu\text{m}$  i.d., 8- to 10- $\mu\text{m}$  i.d. tip, 40-cm-length columns packed in house with C18 material (ReproSil-Pur, 120  $\text{\AA}$ , C18-AQ, 1.9  $\mu\text{m}$  (Dr. Maisch, cat. no. r119aq)). Commercial alternatives may also be used, such as EASY-Spray (Thermo Fisher Scientific, cat. no. ES803) or Aurora (IonOpticks, cat. no. AUR-25075C18A) columns.
- UHPLC system for online LC–MS/MS analysis (Thermo Easy-nLC 1000 or 1200 or Dionex Ultimate 3000 systems (Thermo Fisher Scientific)). We use the Easy-nLC 1000 and 1200 UHPLC systems, which can operate at nanoflow rates at pressures up to 1,000 and 1,200 bar, respectively. Non-UHPLC systems can be used, particularly with shorter columns, but we find that phosphopeptide identifications typically decrease under these conditions.
- Column oven enabling heating of nanospray columns to 50 °C (Sonation, cat. no. PRSO-V1)
- Mass spectrometer. Any modern mass spectrometer capable of MS1 and MS2 scans with sufficiently high resolution and scanning speed, including instruments from the Q Exactive, Orbitrap Fusion, LTQ-Orbitrap Elite, and LTQ-Orbitrap Velos series, as well as several TOF instruments may be used.
- Current MaxQuant release (<http://maxquant.org>) or other software for analyzing raw proteomics data, installed on a suitable workstation or server. For in-depth details on MaxQuant setup and minimum system requirements, see Tyanova *et al.*<sup>41</sup>.

### Reagent setup

#### Stock solutions

Stock solutions consist of 1 M Tris-HCl (pH 8.5), 5 M potassium hydroxide (KOH), 100 mM  $\text{KH}_2\text{PO}_4$ , and 2 M calcium chloride ( $\text{CaCl}_2$ ). All stock solutions can be stored at room temperature (RT; 20–25 °C) for >1 year.

#### SDS lysis buffer

SDC lysis buffer contains 4% (wt/vol) SDC and 100 mM Tris-HCl (pH 8.5). **▲CRITICAL** This buffer should be prepared fresh because SDC gels and crystallizes upon storage in solution.

**Reduction/alkylation buffer**

Reduction/alkylation buffer contains 100 mM TCEP and 400 mM 2-chloroacetamide (CAM), brought to pH 7–8 with KOH. Empirically, we find that 80  $\mu$ l of 5 M KOH added to 920  $\mu$ l of 100 mM TCEP/400 mM CAM results in a solution of pH ~7. **! CAUTION** CAM is toxic. Prepare this solution in a fume hood and handle it with gloves. **▲ CRITICAL** Adjust the pH of reduction/alkylation buffer by pipetting a small amount onto a pH indicator strip to ensure that the pH is neutral (pH ~7–8). This buffer should be prepared immediately before use to ensure full activity of CAM, and excess must be discarded safely after use.

**Trypsin buffer**

Trypsin buffer consists of 0.05% (vol/vol) AcOH and 2 mM CaCl<sub>2</sub>. This buffer can be prepared and stored at –20 °C for >1 year.

**Enzyme reconstitution**

Reconstitute a 1-mg vial of lyophilized trypsin in 1 ml of trypsin buffer (1 mg ml<sup>-1</sup>) and reconstitute a 10-AU vial of Lys-C in 3 ml of Milli-Q water (~1 mg ml<sup>-1</sup>). Resuspend the enzymes by vortexing, and then centrifuge each vial (1,000g for 1 min at RT). **▲ CRITICAL** After reconstitution, enzymes should be divided into aliquots, which can be stored for at least 6 months at –80 °C. Avoid multiple freeze–thaw cycles.

**EasyPhos loading buffer**

EasyPhos loading buffer consists of 6% (vol/vol) TFA/80% (vol/vol) ACN. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**EP enrichment buffer**

EP enrichment buffer consists of 48% (vol/vol) TFA and 8 mM KH<sub>2</sub>PO<sub>4</sub>. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**EP wash buffer**

EP wash buffer consists of 5% (vol/vol) TFA/60% (vol/vol) ISO. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**EP transfer buffer**

EP transfer buffer consists of 0.1% (vol/vol) TFA/60% (vol/vol) ISO. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**EP elution buffer**

Add 200  $\mu$ l of ammonia solution (NH<sub>4</sub>OH) to 800  $\mu$ l of 40% (vol/vol) ACN. **▲ CRITICAL** This buffer must be prepared fresh (within 1 h of use) because the pH will begin to increase due to its high volatility, thereby reducing its elution strength.

**SDB-RPS StageTip loading buffer**

SDB-RPS StageTip loading buffer consists of 1% (vol/vol) TFA in ISO. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**SDB-RPS StageTip wash buffer 1**

SDB-RPS StageTip wash buffer 1 consists of 1% (vol/vol) TFA in ISO. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

## PROTOCOL

## NATURE PROTOCOLS

**SDB-RPS StageTip wash buffer 2**

SDB-RPS StageTip wash buffer 2 consists of 0.2% (vol/vol) TFA/5% (vol/vol) ACN. **! CAUTION** TFA solutions are corrosive. Prepare the solutions in a fume hood and handle with gloves. This buffer is stable for >3 months at RT.

**SDB-RPS StageTip elution buffer**

Add 20  $\mu$ l of NH<sub>4</sub>OH to 4 ml of 60% (vol/vol) ACN. **▲ CRITICAL** This buffer must be prepared fresh (within 1 h of use) because the pH will begin to increase due to its high volatility, thereby reducing its elution strength.

**MS loading buffer**

MS loading buffer consists of 0.3% (vol/vol) TFA/2% (vol/vol) ACN. This buffer is stable for >6 months at RT.

**Cell or tissue preparation**

Protein material derived from cells or tissues can be used with the protocol. Tissue samples should be snap-frozen and can be stored under liquid nitrogen (LN<sub>2</sub>) for >6 months. Before use, tissue should be ground to a fine powder using a mortar and pestle under LN<sub>2</sub>. Powdered tissue can be easily handled by ensuring all surfaces contacting the sample, including tubes and metal spatulas, are prefrozen in LN<sub>2</sub>. **▲ CRITICAL** Before lysis, cell and tissue samples should be free of phosphate salts, including disodium phosphate and monopotassium phosphate, such as is found in phosphate-buffered saline, as these can interfere with phosphopeptide enrichment efficiency. If phosphate-containing buffers are used to wash cells or tissues, a final wash should be performed with a phosphate-free buffer such as Tris-buffered saline.

**Equipment setup****StageTip setup**

C8 and SDB-RPS StageTips are prepared as described<sup>42</sup>, using Empore solid-phase extraction disks. We use two plugs of C8 or SDB-RPS material for the respective StageTips, punched out by using a blunt-end 14-gauge syringe to pierce the ends of 200- $\mu$ l pipette tips.

**Ultra-high-performance liquid chromatography (UHPLC) setup**

The following UHPLC program is used at a flow rate of 350 nl min<sup>-1</sup>, where buffer B is composed of 80% (vol/vol) ACN/0.1% (vol/vol) formic acid and buffer A is composed of 0.1% (vol/vol) formic acid.

Time interval (min)	Gradient (% buffer B)
0	3
40	20
60	40
65	90
70	90

**Mass spectrometer setup**

The MS parameters used for the LC-MS/MS analysis are as follows.

Setting	Value (Q Exactive HF)	Value (Q Exactive HF-X)
<b>Instrument</b>		
Polarity	Positive	Positive
S-lens/ion-funnel RF level	60	45
Capillary temperature	300	300
<b>Full MS</b>		
Microscans	1	1
Resolution	60,000	60,000
Automatic gain control target	$3 \times 10^6$ ion counts	$3 \times 10^6$ ion counts

Table continued

## (continued)

Setting	Value (Q Exactive HF)	Value (Q Exactive HF-X)
Maximum ion time	120 ms	120 ms
Scan range	300–1600 $m/z$	300–1600 $m/z$
<b>dd-MS<sup>2</sup></b>		
Microscans	1	1
Resolution	15,000	15,000
Automatic gain control target	$1 \times 10^5$ ion counts	$1 \times 10^5$ ion counts
Maximum ion time	100 ms	50 ms
Loop count	10	10
Isolation window	1.6 $m/z$	1.6 $m/z$
Isolation offset	0	0
Fixed first mass	100 $m/z$	100 $m/z$
Normalized collision energy	27	27
<b>DD settings</b>		
Minimum AGC target	$2 \times 10^4$ ion counts	$2 \times 10^4$ ion counts
Apex trigger	2–4 s	2–4 s
Charge exclusion	Unassigned, 1, $\geq 5$	Unassigned, 1, $\geq 5$
Peptide match	Preferred	Preferred
Exclude isotopes	On	On
Dynamic exclusion	30 s	30 s

AGC, automatic gain control; DD, data dependent.

**▲ CRITICAL** The LC parameters described above are specific to peptide separation using a nanospray column packed with ~40 cm of ReproSil-Pur 1.9- $\mu$ m C18 resin, and the MS parameters are specific for the analysis of phosphopeptide samples using a Q Exactive HF or Q Exactive HF-X mass spectrometer. Settings are intended to serve as a guide and may need to be adjusted for measurements using alternative LC and MS instruments. For this purpose, we suggest performing several enrichments of a typical sample type and pooling phosphopeptide elutions to provide multiple sample injections for identifying optimal instrument settings.

## Procedure

**▲ CRITICAL** See Fig. 1b for an overview of the workflow.

Lysate preparation ● Timing 1 h

1 Chill SDC lysis buffer to 4 °C and add lysis buffer to the cells or powdered tissues (Reagent setup) to achieve a protein concentration of ~2–5 mg ml<sup>-1</sup>.

**▲ CRITICAL STEP** The use of phosphatase and protease inhibitors in lysis buffers is discouraged because chemical inhibitors differentially inhibit different enzymes and could therefore affect distinct phosphopeptide populations<sup>43,44</sup>; they may also interfere with proteome digestion. Heat treatment immediately upon lysis of cell and tissue samples (Step 2) is therefore a preferable nonchemical enzyme-inactivation method for phosphoproteome analysis<sup>45,46</sup>.

**▲ CRITICAL STEP** Because of the absence of any protein-concentration steps, volumes should be kept to a minimum during lysis. For cell lines, we recommend that ~200  $\mu$ g of total protein starting material be used. For tissues, more material may be required to maximize phosphoproteome coverage; however, optimal starting material should be determined for each sample type.

**▲ CRITICAL STEP** In our experience, for cell-based label-free experiments, three biological replicates represent a minimum for identifying a large proportion of regulated phosphorylation events, with four biological replicates serving as a typical optimal balance between MS measurement time and regulated phosphorylation site identification. For tissue samples, in which biological variability may be greater, more replicates may be beneficial, and this should be investigated for different sample types.

2 Immediately heat-treat lysates for 5 min at 95 °C to facilitate lysis and to inactivate endogenous proteases and phosphatases.

## PROTOCOL

## NATURE PROTOCOLS

- 3 Homogenize the lysates with sonication at 4 °C (three 30-s cycles of 1 s on and 1 s off at 80% output power for tip–probe sonicator, or two cycles of 10 min at maximum output power for a Bioruptor).  
? TROUBLESHOOTING
- 4 Perform a BCA assay to determine the protein concentration. If you are not working with stable isotope labeling with amino acids in cell culture (SILAC)-labeled samples continue to Step 6.
- 5 (Optional) For SILAC-labeled samples, prepare a single sample by mixing equal amounts (by total protein content) from each SILAC-encoded sample.
- 6 Dilute all samples to equal protein concentrations based on BCA assay results to achieve the desired starting material in a final volume of 270  $\mu$ l of lysis buffer in a deep-well, 96-well plate.

## ? TROUBLESHOOTING

## Reduction and alkylation ● Timing 5 min

- 7 Reduce disulfide bonds and carbamidomethylate cysteine residues by adding a 1:10 volume (30  $\mu$ l) of reduction/alkylation buffer to the samples. Seal the deep-well plate with a silicone sealing mat and incubate for 5 min at 45 °C with shaking (1,500 r.p.m.).

## Enzymatic digestion ● Timing 16–18 h

- 8 Remove the samples from heat and allow them to cool to RT. Add Lys-C and trypsin at an enzyme-to-substrate ratio of 1:100 (wt/wt) (e.g., 2  $\mu$ g of each enzyme for 200  $\mu$ g of protein), reseal, and digest overnight at 37 °C with shaking (1,500 r.p.m.).

## ? TROUBLESHOOTING

■ PAUSE POINT Digested samples can be stored for weeks at –20 °C or for months at –80 °C.

## Phosphopeptide enrichment ● Timing 3 h

▲ CRITICAL For all steps involving StageTips, centrifugation durations indicated are a guide only—longer or shorter spin speeds may be required to pass buffers completely through, depending on the compression of the StageTip bed.

- 9 Prepare one C8 StageTip for each sample, as described in the Equipment setup and in the protocol by Rappaport et al.<sup>42</sup>.

■ PAUSE POINT StageTips can be prepared in advance and stored in a covered pipette-tip box for several months at RT.

- 10 Add 400  $\mu$ l of ISO to each sample and mix thoroughly (1,500 r.p.m., 30 s).

▲ CRITICAL STEP Samples must be adequately mixed with ISO before continuing with the addition of EP enrichment buffer (Step 11) in order to prevent formation of precipitates.

- 11 Add 100  $\mu$ l of EP enrichment buffer (Reagent setup) to the samples and mix thoroughly (1,500 r.p.m., 30 s).

▲ CRITICAL STEP Samples should be carefully inspected for the presence of precipitate or cloudiness at this stage. If precipitate is present, samples should be cleared by centrifugation (2,000g for 15 min at RT) in the deep-well plate, and supernatants should be carefully transferred to clean wells of a 96-well plate before addition of TiO<sub>2</sub> beads.

## ? TROUBLESHOOTING

- 12 Weigh the deep-well plate (including silicone sealing lid) containing the samples, and prepare a balance plate for subsequent centrifugation steps by filling an unused plate with water to match the mass of the sample plate.

- 13 Resuspend the TiO<sub>2</sub> beads in EP loading buffer at a concentration of 1 mg  $\mu$ l<sup>–1</sup>, and pipette an aliquot of suspended beads into each sample, using whichever of the following is greater: (i) a 12:1 (wt/wt) bead to protein ratio (e.g., 6 mg of beads per 0.5 mg of protein) or (ii) 5 mg per sample (e.g., for 500  $\mu$ g of sample, prepare a 6- $\mu$ l aliquot of EP loading buffer containing 6 mg of beads). Seal the deep-well plate with silicone sealing mat, taking care to ensure that a tight seal is achieved, and incubate the plate at 40 °C with shaking (2,000 r.p.m.) for 5 min.

▲ CRITICAL STEP We recommend weighing out sufficient beads for several additional samples to account for losses and to avoid running out mid-experiment.

▲ CRITICAL STEP TiO<sub>2</sub> beads settle very quickly and must be mixed thoroughly immediately before pipetting to ensure that the beads are evenly distributed.

▲ CRITICAL STEP Leaks will occur if the sealing mat is not securely attached. A flat solid object such as a tube rack can be used to apply force to help firmly press the sealing mat onto the deep-well plate.

## ? TROUBLESHOOTING

- 14 Pellet the beads by centrifugation (2,000g for 1 min at RT) and aspirate (discard) the nonphosphopeptide supernatant using a glass aspirator tip attached to a vacuum hose.  
**▲CRITICAL STEP** The supernatant can be almost completely removed by aspirating down the corner of the sample well, stopping only upon touching the base of the well, as shown in Supplementary Fig. 4.
- 15 Add 1 ml of EP wash buffer to the samples, seal the plate with a silicone sealing mat, and incubate it at RT with shaking (2,000 r.p.m.) for 30 s. Adjust the weight of the balance plate as required, pellet the beads by centrifugation (2,000g for 1 min at RT), and aspirate (discard) the supernatant.
- 16 Repeat Step 15 a further four times.
- 17 After the final wash, resuspend the beads in 75  $\mu$ l of EP transfer buffer (Reagent setup) and transfer to the top of a C8 StageTip (Equipment setup). Add another 75  $\mu$ l of EP transfer buffer to each well to capture any remaining beads and transfer these to the same C8 StageTip.
- 18 Place the StageTips into centrifuge adapters and spin through to dryness (1,500g for ~8 min at RT), capturing the  $\text{TiO}_2$  beads on top of the C8 material.  
**▲CRITICAL STEP** Ensure that all transfer buffer has passed through the StageTips before proceeding to elution. In the case that not all buffer has passed through the StageTips, centrifuge for another 3–5 min.
- ? TROUBLESHOOTING**
- 19 Elute the phosphopeptides with 30  $\mu$ l of EP elution buffer (Reagent setup), centrifuge (1,500g for ~4 min at RT) to dryness, and collect the eluates into clean PCR strip tubes.
- 20 Repeat Step 19 and collect the two eluates into a single tube.
- 21 Immediately place the tubes into an evaporative concentrator and concentrate under vacuum at 45 °C until  $\leq$ 15  $\mu$ l of sample remains, without drying the samples completely.  
**▲CRITICAL STEP** Do not allow the samples to concentrate to dryness here.
- ? TROUBLESHOOTING**

#### StageTip desalting of phosphopeptides ● **Timing 1 h**

- 22 Prepare one SDB-RPS StageTip for each sample as described in the Equipment setup and in the protocol by Rappaport et al.<sup>42</sup>.  
**■PAUSE POINT** StageTips can be prepared in advance and stored in a covered pipette-tip box for several months at RT.
- 23 Add 100  $\mu$ l of SDB-RPS loading buffer to each sample (Reagent setup), and then transfer each sample to the top of an SDB-RPS StageTip (Equipment setup).
- 24 Place the StageTips into centrifuge adapters and centrifuge the sample through to dryness (1,500g for ~8 min at RT).
- ? TROUBLESHOOTING**
- 25 Wash the StageTips once with 100  $\mu$ l of SDB-RPS wash buffer 1 and centrifuge the sample through to dryness (1,500g for ~8 min at RT).
- 26 Repeat Step 25 with SDB-RPS wash buffer 2 (Reagent setup).  
**■PAUSE POINT** Phosphopeptides can be stored bound to StageTip material immediately before elution, for several weeks at 4 °C.
- 27 Elute the phosphopeptides by adding 60  $\mu$ l of SDB-RPS elution buffer (see Reagent setup) to the top of the StageTip, and centrifuge through to dryness (1,500g for ~5 min at RT). Collect the eluates into clean PCR strip tubes.
- 28 Immediately place the tubes into an evaporative concentrator and concentrate under vacuum at 45 °C to dryness.  
**▲CRITICAL STEP** The time to dryness will depend on the number of samples being processed. In a correctly working Eppendorf Vacuum Concentrator Plus, this step takes 40 min; however, after the first 30 min have elapsed, samples should be checked at intervals of 10 min and concentrated until no liquid remains.
- 29 Reconstitute the phosphopeptides by adding 7  $\mu$ l of MS loading buffer (Reagent setup) and either shaking for 2 min at 2,000 r.p.m. or incubating for 5 min in a sonicating water bath on low power, then centrifuge the samples at 2,000g for 1 min at RT.  
**■PAUSE POINT** Phosphopeptide samples can be stored in MS loading buffer for several days at 4 °C, several weeks at –20 °C, or several months at –80 °C, provided an airtight seal is maintained on the sample tubes.

## PROTOCOL

## NATURE PROTOCOLS

## LC-MS/MS ● Timing 1 d

30 Place the samples in an LC autosampler cooled to 4–8 °C. If the HPLC autosampler cannot accommodate PCR strip tubes, transfer the phosphopeptide samples to appropriate autosampler vials. We typically inject samples directly from the PCR tubes in which they were dried in Step 28.

31 Set up the nanospray column with a column heater set to 50 °C.

32 Analyze each phosphopeptide sample using the UHPLC and MS parameters outlined in the Equipment setup. For each analysis, we typically load 6  $\mu$ l of sample at a constant pressure of 950 bar ( $\sim$ 500 nl min $^{-1}$  with our 40-cm packed-tip emitters).

## ? TROUBLESHOOTING

## Data analysis ● Timing &gt;1 d

▲ CRITICAL We typically analyze raw phosphoproteomics data using MaxQuant<sup>41,47</sup> and perform downstream bioinformatics analysis using the Perseus platform<sup>48</sup> (described below). However, several other suitable programs are freely or commercially available that can be used to process phosphoproteomics data.

33 Specify the parameters used for searching label-free phosphoproteomics data using MaxQuant. Typical parameters are outlined in the table below.

Setting	Value
<b>Raw data</b>	
Parameter group	All samples: Group 0
Experiment	Enter unique name, suffixed by biological replicate (e.g., 'Control_1')
Fraction	All samples: 1
PTM	True
<b>Group-specific parameters</b>	
Type	Standard
Multiplicity	1
Labels	None
Variable modifications	Oxidation (M), Acetyl (protein N-term), and Phospho (STY)
Digestion mode	Specific
Enzyme	Trypsin/P
Maximum missed (cleavages)	2
<b>Global parameters</b>	
FASTA files	Specify protein sequence databases
Fixed modifications	Carbamidomethyl (C)
Second peptide	True
Match between runs	True
Match time window	0.7 min
Alignment time window	20 min
Advanced site intensities	Yes

▲ CRITICAL STEP Default settings in MaxQuant are typically optimal and can normally be left unchanged, with a few exceptions such as defining variable modifications as described above. See Tyanova et al.<sup>41</sup> for in-depth information about MaxQuant settings.

## Troubleshooting

## Samples are highly viscous after lysis (Step 3)

This is caused by the release of genomic DNA and is completely normal. It is essential to ensure sufficient sonication of samples has been performed to effectively shear DNA and reduce viscosity, before proceeding with subsequent steps. If samples appear 'sticky' upon pipetting, additional sonication is required.

## Sample volume exceeds the specified amount for desired protein yield (Step 6)

Lysis volumes should be kept to a minimum. However, if (after the BCA assay) it is not possible to achieve the desired protein yield within the specified volume (Step 6), the subsequent

volumes—reduction/alkylation buffer (Step 7), ISO (Step 10), and EP enrichment buffer (Step 11)—should be scaled proportionally. In the case that larger volumes are used, care should be taken during digestion to ensure that mixing velocity is not so high that samples come in contact with the lid for prolonged periods (Step 8).

#### Samples leaking from lid of the 96-well plate (Step 8)

The silicone sealing mats recommended in this protocol are effective at sealing at high mixing velocities (2,000 r.p.m.); however, minor leaks may occur if liquid enters the plate-mat interface during sealing. If samples contact the underside of the lid, a brief centrifugation (500g, 1 min, RT) before opening will prevent leaks upon subsequent sealing. It is also recommended to limit the mixing velocity during overnight digestion to a speed that prevents the samples from touching the lid (~1,500 r.p.m. with the volumes shown here). We also use and recommend use of the ThermoTop lid (Equipment) for extended incubation periods, such as overnight digestion, to prevent condensation on the lid.

#### Cloudy peptide solution after addition of EP enrichment buffer (Step 11)

A slight cloudiness after addition of EP enrichment buffer (Step 11) should not interfere with phosphopeptide enrichment. However, the formation of larger amounts of insoluble material at this stage indicates the presence of lipids, which can occur with high-lipid-content samples or when working with large amounts of protein starting material. This can be resolved by centrifugation in the deep-well plate (2,000g, 15 min, RT) and careful transfer of supernatants (without disrupting any insoluble pellets) to a clean 96-well plate, before addition of TiO<sub>2</sub> beads.

#### Uneven distribution of TiO<sub>2</sub> beads between samples (Step 13)

TiO<sub>2</sub> beads settle very quickly, and care must be taken to ensure that equal volumes of beads are dispensed into each sample. Ensuring that the volume of EP loading buffer does not exceed the recommended volume (1  $\mu$ l of buffer per 1 mg of beads) will help mitigate bead settling. Vortexing the beads immediately before pipetting each sample, or by placing the beads in a flat-bottom glass vial with a small magnetic 'flea' on a magnetic stir plate, will ensure that the beads are evenly mixed.

#### Samples do not concentrate in the vacuum concentrator (Step 21)

In the case that samples do not concentrate evenly or quickly in the vacuum concentrator (Steps 21 and 28), this may be the result of insufficient vacuum. This can occur due to worn seals on vacuum concentrator pumps or poor lid sealing, and should be rectified by instrument servicing. In a correctly working Eppendorf Vacuum Concentrator Plus at Step 21, ~10  $\mu$ l of sample should remain after 30 min of concentration at 45 °C using the 'V-AQ' program. When using other instruments, timing should be empirically determined once and used for subsequent experiments.

#### StageTips clog (Steps 18 and 24)

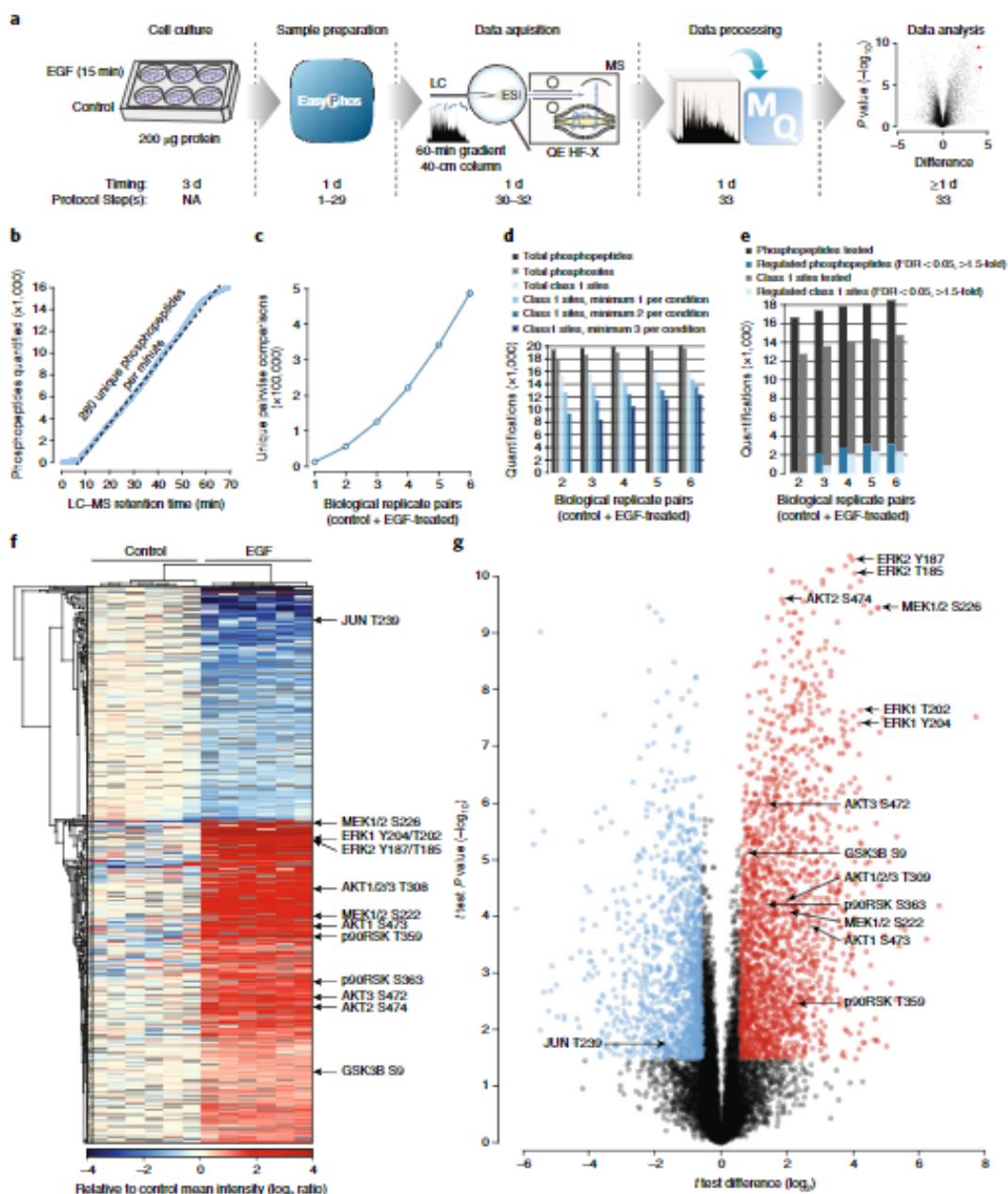
If a StageTip flow is slow or stops entirely during TiO<sub>2</sub> bead capture (Steps 17–19), lipid or other insoluble material may have been present before enrichment. This should be avoided as described under 'Cloudy peptide solution after addition of EP enrichment buffer'. If a StageTip flow is slow during phosphopeptide desalting (Step 24), too much force may have been used to compress StageTip material into the pipette tips when they were made. Slow-flowing StageTips can be used, but may require higher centrifugation speeds for longer durations for solvents to transit.

#### Low phosphopeptide yield (Step 32)

If the yield of phosphorylation sites is lower than expected following data analysis, several parameters in the workflow should be checked, including the digestion and phosphopeptide enrichment specificity (ratio of phosphopeptides to total peptides identified). Issues that can lead to poor recovery of phosphopeptides include inefficient proteome digestion (Step 8), high sample contamination with non-protein constituents during enrichment (Steps 11–13), use of old EP elution buffer (Steps 19 and 20), and insufficient evaporation of EP elution buffer before loading onto SDB-RPS StageTips (Step 21). If large numbers of nonphosphorylated peptides are observed (>30% of the total number of peptides identified), inefficient washing of the beads may have occurred. This could potentially be caused by the presence of particulate or nonpeptide material during enrichment, which can be

## PROTOCOL

## NATURE PROTOCOLS



**Fig. 4 | Demonstration of the phosphoproteomics workflow.** As an example, we investigated the response of U-87 human glioblastoma cells to EGF treatment (100 ng/ml, 15 min) compared with control cells (serum-starved, 18 h). Six independent biological replicates ( $n = 6$ ) were performed per condition, using 200 µg of protein (determined by BCA assay). **a**, Experimental stages and timing for the phosphoproteomics experiment, using the EasyPhos protocol described here. Approximate timing and steps associated with the protocol described here are indicated. **b**, Rate of phosphopeptide quantification across a chromatographic time scale in a single LC-MS measurement. **c**, Actual quantified pairwise comparisons between biological replicate pairs (control and EGF-treated samples) with respect to the number of samples measured. **d**, Number of quantified phosphopeptides, phosphorylation sites (total), and localized phosphorylation sites (class 1; localization probability  $> 0.75$ ) as a function of  $N$  replicate pairs (control + EGF treated) analyzed. The numbers of phosphorylation sites quantified in at least one, two, or three replicates of both control and EGF-treated samples are also shown. **e**, Identification of significantly regulated (permutation-based FDR-controlled  $t$  test; FDR  $< 0.05$  and  $\geq 1.5$ -fold change as compared with control samples) phosphopeptides and phosphorylation sites (class 1; localization probability  $> 0.75$ ) as a function of the number of biological replicate pairs (control and EGF-treated samples) analyzed, as compared with the number of phosphopeptides and phosphorylation sites tested for regulation. **f,g**, Hierarchical clustering (**f**) and volcano plot (**g**) of EGF-treated U-87 phosphoproteomes depicting key components of the EGF-MAPK and EGF-PI3K-AKT networks, produced using the Perseus computational framework. FDR, false-discovery rate; NA, not applicable.

rectified by centrifugation of samples (16,000g, 15 min, RT) before enrichment (Step 11). Poorly operating LC or MS systems because of leaks or contamination can also lead to low numbers of identified phosphopeptides.

### Timing

Steps 1–6, lysate preparation: 1 h  
 Step 7, reduction and alkylation: 5 min  
 Step 8, enzymatic digestion: 16–18 h  
 Steps 9–21, phosphopeptide enrichment: 3 h  
 Steps 22–29, StageTip desalting of phosphopeptides: 1 h  
 Steps 30–32, LC–MS/MS: 1 d  
 Step 33, data analysis: >1 d

### Anticipated results

Using the EasyPhos protocol, we previously showed that >10,000 phosphopeptides can be identified and quantified using 1 mg of starting protein per sample<sup>10</sup>. Here we have described an updated protocol that improves phosphoproteome coverage, especially under conditions in which the amounts of starting materials are limited. With this protocol, 200 µg of human U-87 glioblastoma cells were either treated with EGF for 15 min or left untreated with six full biological replicates (Fig. 4a), leading to high phosphopeptide identification and quantification rates (~280 unique phosphopeptides per min) over a sustained period (Fig. 4b). Together, these data were measured in <1 d and represent around half a million biologically meaningful pairwise comparisons (control versus treatment) (Fig. 4c). In total, this resulted in the quantification of ~20,000 phosphopeptides, comprising 16,021 accurately localized phosphorylation sites (Fig. 4d). When six biological replicates were performed, 85% and 77% of the localized phosphorylation sites were quantified in at least two or three replicates, respectively, for both conditions (control and EGF-treated). When four biological replicates were performed, quantification in at least two or three replicates was 79% and 66%, respectively, enabling the quantitative assessment of phosphorylation site regulation across more than two-thirds of the phosphoproteome (Fig. 4d). We advise that for cell-based label-free experiments using the EasyPhos workflow, three biological replicates represent a minimum for identifying a large proportion of regulated phosphorylation events, with four biological replicates serving as a typical optimal balance between MS measurement time and regulated phosphorylation site identification (Fig. 4c). Downstream bioinformatics analysis is facilitated by the Perseus computational framework<sup>48</sup>, enabling rapid visualization and interactive interrogation of phosphoproteomics data. For example, using this software, we highlighted several major regulatory components of the EGF–MAPK and EGF–PI3K–AKT signaling pathways, superimposed on graphical representations of the underlying quantitative phosphoproteomics data (Fig. 4f,g).

Taken together, the protocol described here enables the rapid identification and quantification of well over 15,000 phosphopeptides in a single-run format, without requiring metabolic or chemical labeling, and from minimal starting materials. Although these experiments were performed using a data-dependent acquisition MS approach, there is nothing in the EasyPhos protocol that is specific to data-dependent or data-independent acquisition MS methods. Indeed, we anticipate that the combination of the updated high-sensitivity EasyPhos workflow with advances in instrument and acquisition methods will result in even further improvements to data completeness, throughput, reproducibility, and quantification accuracy.

### Data availability

The raw MS data from this study have been deposited into the ProteomeXchange Consortium via the PRIDE partner repository<sup>49,50</sup> with the identifier PXD009227 and can be accessed at <https://www.ebi.ac.uk/pride/archive/projects/PXD009227>.

### References

1. Humphrey, S. J., James, D. E. & Mann, M. Protein phosphorylation: a major switch mechanism for metabolic regulation. *Trends Endocrinol. Metab.* **26**, 676–687 (2015).
2. Sharma, K. et al. Ultradep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8**, 1583–1594 (2014).

## PROTOCOL

## NATURE PROTOCOLS

3. Ficarro, S. B. et al. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20**, 301–305 (2002).
4. Olsen, J. V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
5. Lemeer, S. & Heck, A. J. The phosphoproteomics data explosion. *Curr. Opin. Chem. Biol.* **13**, 414–420 (2009).
6. Riley, N. M. & Coon, J. J. Phosphoproteomics in the age of rapid and deep proteome profiling. *Anal. Chem.* **88**, 74–94 (2016).
7. Rigbolt, K. T. G. & Blagoev, B. Quantitative phosphoproteomics to characterize signaling networks. *Semin. Cell Dev. Biol.* **23**, 863–871 (2012).
8. Olsen, J. V. & Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics* **12**, 3444–3452 (2013).
9. Jünger, M. A. & Aebersold, R. Mass spectrometry-driven phosphoproteomics: patterning the systems biology mosaic. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 83–112 (2014).
10. Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat. Biotechnol.* **33**, 990–995 (2015).
11. Robles, M. S., Humphrey, S. J. & Mann, M. Phosphorylation is a central mechanism for circadian control of metabolism and physiology. *Cell Metab.* **25**, 118–127 (2017).
12. Liu, J. J. et al. In vivo brain GPCR signaling elucidated by phosphoproteomics. *Science* **360**, eaao4927 (2018).
13. Sacco, F. et al. Glucose-regulated and drug-perturbed phosphoproteome reveals molecular mechanisms controlling insulin secretion. *Nat. Commun.* **7**, 13250 (2016).
14. Steger, M. et al. Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *eLife* **5**, e12813 (2016).
15. Courcelles, M. et al. Phosphoproteome dynamics reveal novel ERK1/2 MAP kinase substrates with broad spectrum of functions. *Mol. Syst. Biol.* **9**, 129–57 (2013).
16. Kanshin, E., Bergeron-Sandoval, L. P., Isik, S. S., Thibault, P. & Michnick, S. W. A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep.* **10**, 1202–1214 (2015).
17. Posewitz, M. C. & Tempst, P. Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal. Chem.* **71**, 2883–2892 (1999).
18. Andersson, L. & Porath, J. Isolation of phosphoproteins by immobilized metal ( $Fe^{3+}$ ) affinity chromatography. *Anal. Biochem.* **154**, 250–254 (1986).
19. Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1630–1638 (2008).
20. Zhou, H. et al. Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography. *Nat. Protoc.* **8**, 461–480 (2013).
21. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P. & Jørgensen, T. J. D. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* **4**, 873–886 (2005).
22. Liang, S.-S., Makamba, H., Huang, S.-Y. & Chen, S.-H. Nano-titanium dioxide composites for the enrichment of phosphopeptides. *J. Chromatogr. A* **1116**, 38–45 (2006).
23. Sugiyama, N. et al. Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. *Mol. Cell. Proteomics* **6**, 1103–1109 (2007).
24. Thingholm, T. E., Jørgensen, T. J. D., Jensen, O. N. & Larsen, M. R. Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nat. Protoc.* **1**, 1929–1935 (2006).
25. Macek, B., Mann, M. & Olsen, J. V. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.* **49**, 199–221 (2009).
26. Pinkse, M. W. H. H. et al. Highly robust, automated, and sensitive online  $TiO_2$ -based phosphoproteomics applied to study endogenous phosphorylation in *Drosophila melanogaster*. *J. Proteome Res.* **7**, 687–697 (2008).
27. Humphrey, S. J. et al. Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. *Cell Metab.* **17**, 1009–1020 (2013).
28. Wang, Y. et al. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**, 2019–2026 (2011).
29. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).
30. Batth, T. S. et al. Large-scale phosphoproteomics reveals Shp-2 phosphatase-dependent regulators of Pdgf receptor signaling. *Cell Rep.* **22**, 2601–2614 (2018).
31. de Graaf, E. I., Giansanti, P., Altelaar, A. F. M. & Heck, A. J. R. Single-step enrichment by  $Ti^{4+}$ -IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. *Mol. Cell. Proteomics* **13**, 2426–2434 (2014).
32. Masuda, T., Sugiyama, N., Tomita, M. & Ishihama, Y. Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. *Anal. Chem.* **83**, 7698–7703 (2011).
33. Ren, D. et al. An improved trypsin digestion method minimizes digestion-induced modifications on proteins. *Anal. Biochem.* **392**, 12–21 (2009).

34. Masuda, T., Tomita, M. & Ishihama, Y. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J. Proteome Res.* **7**, 731–740 (2008).
35. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
36. Kokubu, M., Ishihama, Y., Sato, T., Nagasu, T. & Oda, Y. Specificity of immobilized metal affinity-based IMAC/C18 tip enrichment of phosphopeptides for protein phosphorylation analysis. *Anal. Chem.* **77**, 5144–5154 (2005).
37. Scheltema, R. A. et al. The Q Exactive HF, a benchtop mass spectrometer with a pre-filter, high performance quadrupole and an ultra-high field Orbitrap analyzer. *Mol. Cell. Proteomics* **13**, 3698–3708 (2014).
38. Kelstrup, C. D. et al. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.* **13**, 6187–6195 (2014).
39. Shaffer, S. A., Prior, D. C., Anderson, G. A., Udseth, H. R. & Smith, R. D. An ion funnel interface for improved ion focusing and sensitivity using electrospray ionization mass spectrometry. *Anal. Chem.* **70**, 4111–4119 (1998).
40. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
41. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
42. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
43. Thingholm, T. E., Larsen, M. R., Ingrell, C. R., Kassem, M. & Jensen, O. N. TiO<sub>2</sub>-based phosphoproteomic analysis of the plasma membrane and the effects of phosphatase inhibitor treatment. *J. Proteome Res.* **7**, 3301–3313 (2008).
44. Pan, C., Gnad, F., Olsen, J. V. & Mann, M. Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* **8**, 4534–4546 (2008).
45. Svensson, M. et al. Heat stabilization of the tissue proteome: a new technology for improved proteomics. *J. Proteome Res.* **8**, 974–981 (2009).
46. Lundby, A. et al. Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat. Commun.* **3**, 876 (2012).
47. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
48. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
49. Vizcaíno, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
50. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).
51. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

### Acknowledgements

We thank members of the Proteomics and Signal Transduction Group, the Metabolic Systems Biology Group, and the staff of SydneyMS for support and discussions; and, in particular, I. Paron, K. Mayr, and G. Sowa for MS technical assistance; J. Cox for bioinformatic tools; and N. Nagaraj, E. Humphrey, B. Parker, and M. Larance for technical discussions. This work was funded in part by the Max Planck Society for the Advancement of Science, the Novo Nordisk Foundation (grant NNF15CC0001) (M.M.), and the National Health and Medical Research Council (NHMRC) (grants GNT1061122 and GNT1086850) (D.E.J.). Thanks to J. Cobcroft for her generous funding support (S.J.H.).

### Author contributions

S.J.H. designed the studies and developed methods. S.J.H. and O.K. performed the phosphoproteomic experiments and analyzed the data. M.M. and D.E.J. supervised the project and wrote the manuscript with S.J.H. and O.K.

### Competing interests

M.M., O.K., and D.E.J. declare that they have no competing interests. S.J.H. is an inventor on a provisional patent (no. 2018901179) relating to parts of the described protocol.

### Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41596-018-0014-9>.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

Correspondence and requests for materials should be addressed to S.J.H. and M.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 6 September 2018

### 3.1.2. Integrative proteomics reveals principles of dynamic phospho-signaling networks in human erythropoiesis

Ozge Karayel<sup>1,a</sup>, Peng Xu<sup>2,a</sup>, Isabell Bludau<sup>1</sup>, Senthil Velan Bhoopalan<sup>2</sup>, Yu Yao<sup>2</sup>, Ana Rita Freitas Colaco<sup>3</sup>, Alberto Santos<sup>3</sup>, Brenda A. Schulman<sup>4,b</sup>, Arno F. Alpi<sup>4,b</sup>, Mitchell J. Weiss<sup>2,b</sup>, Matthias Mann<sup>1,3,b</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany <sup>2</sup> Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN, USA <sup>3</sup> Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark <sup>4</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, Germany <sup>a</sup> These authors contributed equally <sup>b</sup> Corresponding authors

Under revision in *Molecular Systems Biology* (2020)

Human erythropoiesis is a finely controlled multistep developmental process that results in a vastly remodeled proteome to form mature erythrocytes equipped with highly specialized functions. Over the years, focused studies have provided crucial functional insights in erythropoiesis; however, a system-wide understanding of how post-transcriptional and translational mechanisms are driving terminal maturation is still lacking. It was crucial to obtain high coverage proteome and phosphoproteome data to understand mechanisms underlying regulation in human erythropoiesis.

In collaboration with the group of Mitchell J. Weiss at St. Jude Children's Research Hospital and the Schulman group at our Max-Planck Institute of Biochemistry, we established a pipeline combining fluorescence activated cell sorting (FACS)-based cell enrichment procedures with our state-of-the-art proteomics and phosphoproteomics workflows. Applying the EasyPhos protocol, we dynamically monitored 7,400 proteins and 27,000 phosphorylation sites of five distinct maturation stages of in vitro reconstituted erythropoiesis of CD34+ HSPCs. This resulted in a system-wide view of the functional dynamic regulation of erythropoiesis through proteome remodeling. Interestingly and unexpectedly, we found numerous orchestrated changes in solute carriers which provided new state-specific markers. This is the first phosphoproteome study to demonstrate direct evidence for intricate stage-specific regulation of this process by post-translational modification. Proteomics, in conjunction with a kinome-targeting CRISPR-Cas9 screen, we defined the 'erythropoietic kinome' and revealed its critical regulations, including a requirement for PIM1 kinase functions, in erythroid maturation. Tracking multiple signaling cascades along maturation stages allowed us to dissect the sequential attenuation of c-Kit and EPOR/JAK2 signaling pinpointing downregulation of Ras/MAPK signaling, which promotes terminal maturation. After initial suggestion by Brenda Schulman and Arno Alpi, I conceived this study and organized this multi-group collaboration from initial stages to final publication in *Molecular Systems Biology*.

## Integrative proteomics reveals principles of dynamic phospho-signaling networks in human erythropoiesis

Özge Karayel<sup>1,a</sup>, Peng Xu<sup>2,a</sup>, Isabell Bludau<sup>1</sup>, Senthil Velan Bhoopalan<sup>2</sup>, Yu Yao<sup>2</sup>, Ana Rita Freitas Colaco<sup>3</sup>, Alberto Santos<sup>3</sup>, Brenda A. Schulman<sup>4,b</sup>, Arno F. Alpi<sup>4,b</sup>, Mitchell J. Weiss<sup>2,b</sup>, Matthias Mann<sup>1,3,b,c</sup>

<sup>1</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

<sup>2</sup>Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN, USA

<sup>3</sup>Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

<sup>4</sup>Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, Germany

<sup>a</sup> These authors contributed equally

<sup>b</sup> Corresponding authors

<sup>c</sup> Lead contact

Correspondence: schulman@biochem.mpg.de (BAS), aalpi@biochem.mpg.de (AFA), Mitch.Weiss@stjude.org (MW), mmann@biochem.mpg.de (MM)

**Running title:** Integrative proteomics of human erythropoiesis

### SUMMARY

Human erythropoiesis is an exquisitely controlled multistep developmental process and its dysregulation leads to numerous human diseases. Transcriptome and epigenome studies provided insight into system-wide regulation, but we currently lack a global mechanistic view on the dynamics of proteome and post-translational regulation coordinating erythroid maturation. We established a mass spectrometry (MS)-based proteomics to quantify and dynamically track 7,400 proteins and 27,000 phosphorylation sites of five distinct maturation stages of *in vitro* reconstituted erythropoiesis of CD34<sup>+</sup> HSPCs. Our data reveals developmental regulation through drastic proteome remodeling across stages of erythroid maturation encompassing most protein classes. This includes various orchestrated changes in solute carriers indicating adjustments to altered metabolic requirements. To define the distinct proteome of each maturation stage we developed a computational deconvolution approach which revealed stage-specific marker proteins. The dynamic phosphoproteomes combined with a kinase-targeted CRISPR/Cas9 screen uncovered coordinated networks of erythropoietic kinases and pinpoints downregulation of c-Kit/MAPK signaling axis as key driver of maturation. Our system-wide view establishes the functional dynamic of complex phosphosignaling networks and regulation through proteome remodeling in erythropoiesis.

### KEYWORDS

Human erythropoiesis, proteomics, phosphoproteomics, signaling, EPOR, c-Kit, PIM1, kinase, SLC, red blood cells, hematopoietic stem and progenitor cells, systems biology, CRISPR/Cas9 screen

## INTRODUCTION

Human erythropoiesis is a multistep developmental process that maintains stable erythroid homeostasis throughout life and replenishes more than 200 billion erythrocytes lost by senescence in healthy humans [1]. Lineage-committed erythroid progenitors, including burst-forming unit-erythroid (BFU-E) and their colony-forming unit-erythroid, (CFU-E) progeny, undergo enormous expansion, followed by morphological signs of terminal maturation. The first recognizable erythroid precursors are proerythroblasts (ProE), which mature progressively into early basophilic (EBaso) and late basophilic (LBaso) erythroblasts, polychromatic erythroblasts (Poly), orthochromatic (Ortho) erythroblasts and reticulocytes. Terminal erythroid maturation is distinguished by progressive reductions in proliferative capacity and cell size, chromatin condensation, loss of most organelles including the nucleus, and remarkable streamlining of the proteome with expression of specialized cytoskeletal and plasma membrane proteins and finally massive accumulation of hemoglobin [2-4]. This finely tuned developmental process generates mature erythrocytes with the highly specialized function of circulatory oxygen/carbon dioxide transport.

Our knowledge of human erythropoiesis has been greatly advanced by *in vitro* differentiation systems in which primary multipotent CD34<sup>+</sup> hematopoietic stem cell progenitors (HSPCs) are cultured with defined cytokines and other bioactive components to generate reticulocytes [5]. Erythropoiesis is controlled by the essential cytokines stem cell factor (SCF) and erythropoietin (EPO), and their cognate receptors c-Kit and EPOR [6-11]. In general, c-Kit acts to promote progenitor proliferation during early erythropoiesis, while EPOR fosters survival and maturation at later stages, although there is substantial overlap in their activities and some evidence for cross-regulation [9, 12, 13]. Moreover, c-Kit and EPOR trigger remarkably similar signaling pathways including Ras/Raf/MAPK, PI3K/Akt, and JAK2/STAT5 [14-20]. In concert with cytokine signaling, several key erythroid-restricted transcription factors (including GATA-1, FOG-1, SCL/TAL-1, EKLF/KLF1) associate with generalized cofactors to activate the transcription of erythroid-specific genes and suppress those of alternate lineages [21-28].

While focused studies on erythroid cytokine signaling and transcription factors have generated tremendous functional insights into erythropoiesis, they do not provide a systems-wide view. A comprehensive view of erythroid gene expression has been provided by global analysis of erythroid transcriptomes and the epigenome in purified bulk populations and single cells ([29] and reviewed in [30]). These approaches necessarily use global mRNA levels as proxies of protein abundance and infer signaling activity indirectly. A truly system-wide understanding of post transcriptional and translational mechanisms that drive and coordinate terminal maturation is clearly still lacking. Such a dynamic map would complement transcriptome studies to broadly describe the molecular basis of the pathways involved and to understand how cytokine receptor signaling and transcription factors together shape the erythroid proteome.

In contrast to transcriptome and epigenetic studies of erythropoiesis, relatively few proteome studies to date provide a global analysis of the protein landscape. Due to technical limitations, these studies examined only selected maturation stages in limited depth or focused on defined protein families [31-39]. A recent analysis described dynamic changes in protein expression during *in vitro* erythroid differentiation of CD34<sup>+</sup> HSPCs [40]. However, because relatively large numbers of cells were required for proteomic characterization, this study examined semisynchronous erythroid cultures consisting of cells at different stages of maturation. Given the recent dramatic advances in mass spectrometry and label-free quantitative proteomics [41, 42], we reasoned that it may now be possible to obtain accurate high coverage proteome and phosphoproteome quantification from relatively low numbers of purified erythroid precursors at distinct developmental stages.

We developed a pipeline combining fluorescence activated cell sorting (FACS) enrichment procedures with our state-of-the-art proteomics workflow. We uncovered the temporal staging of developmental regulation through proteome remodeling. To identify the distinct proteome defining each maturation stage from proerythroblast to orthochromatic erythroblast, we developed a bioinformatic deconvolution approach which revealed stage-specific proteins and protein families. Importantly, our proteomics workflow enabled detection of more than one thousand membrane

proteins, and identified distinct combinations of solute carrier (SLC) family proteins as stage-specific maturation markers. Pursuing post-translational regulation further, in-depth sensitive quantitation of the global phosphoproteome with our EasyPhos platform [43, 44] provided direct evidence for intricate developmental stage-specific regulation by post-translational modification. To functionally explore the identified signaling modules, we performed a kinome-targeting CRISPR/Cas9 screen, which in combination with our proteomic studies, identified distinct signaling requirements for erythroid maturation. Focusing on networks amongst over 27,000 phosphosites and kinase functions uncovered the sequential attenuation of c-Kit and EPOR/JAK2 signaling, pinpointing downregulation of Ras/MAPK signaling in promoting terminal maturation. Our system-wide data provide a wealth of molecular information regarding the functional dynamics of complex phosphosignaling networks in erythropoiesis, expanding our knowledge and data for cellular principles of regulation through proteome remodeling.

## RESULTS

### Establishing stage-specific proteomes of human erythropoiesis

To investigate the remodeling of the proteomics landscape during human erythropoiesis, we cultured human peripheral blood-derived CD34<sup>+</sup> HSPCs under conditions to support erythroid differentiation (Methods). We obtained highly enriched populations of erythroid precursors at specific developmental stages by FACS using CD235a (GYPA), CD49d (ITGA4), and Band 3 (SLC4A1) markers (Figure 1A-B and Figure S1A) [45]. We isolated early maturation stages (progenitors, ProE, EBaso, LBaso) after 7 days of culture, while later maturation stages (LBaso, Poly and Ortho) were purified at day 14. LBaso stage precursors were isolated after both 7 days and 14 days of culture using the same markers (Figure 1B). Note that SCF was present at 7 days but not at 14 days. Purified cell populations were morphologically homogeneous as judged by May-Grünwald-Giemsa staining (Figure S1B). Due to relatively low cell yields, ProE and EBaso populations were combined in equal cell numbers prior to subsequent analysis. The resulting five populations/stages are henceforth color-coded as follows: progenitors (mostly CFU-E) [45-47], yellow; ProE/EBaso, blue; LBaso day7, light pink;

LBaso day14, dark pink; Poly, dark blue, and Ortho, orange (Figure 1A-B).

Each population was processed in four biological replicates and their tryptic peptides were analyzed in single shots in Data Independent Acquisition (DIA) mode (Methods, Figure 1A). To generate a project-specific library, we separated peptides by high pH reversed-phase chromatography into fractions, followed by data-dependent acquisition (DDA) and analysis with Spectronaut. The resultant library contained more than 9,000 protein groups, 7,479 of which could be matched into the DIA runs of at least one maturation stage (q-value less than 1% at protein and precursor levels, Figure 1C). In the DIA method, small m/z precursor windows are fragmented in a cyclical manner, which turned out to be crucial for preserving the dynamic range of peptide detection in the presence of the very large hemoglobin peptide peaks that would otherwise complicate analyses at later developmental stages. Remarkably, 84% of all detected proteins were consistently quantified at varying levels across all maturation stages and a relatively small percentage was only matched in a single stage. Quantitative accuracy was high, with Pearson correlations > 0.95 and CVs < 20% for 72% of all proteins between the four biological replicates (Figure 1D and Figure S1C). MS signals spanned abundance ranges of five (progenitors) to seven (Ortho) orders of magnitude. As expected, globin proteins increased by approximately one thousand-fold from progenitor to Ortho stage (Figure S1D).

As biological interpretation is facilitated by absolute rather than relative concentration measurements, we employed the 'proteomic ruler' method, which uses the fixed relationship between histones and DNA to estimate proteome-wide copy numbers per cell [48]. Considering that chromatin condensation during erythropoiesis is associated with partial release of major histones from the nucleus and subsequent degradation in the cytoplasm [49], we first assessed the overall histone content of cells in our system, which indeed declined with progressive differentiation (Figure S1E). Taking this into account for the proteomic ruler calculations, we measured an almost four-fold reduction in total protein copy numbers per cell during differentiation with median copy numbers dropping from  $23,380 \pm 371$  in progenitors to  $12,395 \pm 1342$  at the LBaso stage (Figure 1E, Figure S1E and Table S1). In contrast, the average copy numbers of proteins

annotated as “erythrocyte maturation” and “heme biosynthesis” by Gene Ontology (GO) increased by approximately 50-fold from progenitor to Ortho stage (Figure 1E and Table S1). Quantitative comparison and copy number estimation of LBaso stages isolated at either day7 or day14 confirmed their close resemblance at the global proteome level, including marker proteins, such as GYPA, CD49d, Band 3, c-Kit, and several hemoglobin subunits that did not significantly change (Figure 1E and Figure S2A-C). Thus, they were combined for further proteomic analysis unless otherwise noted.

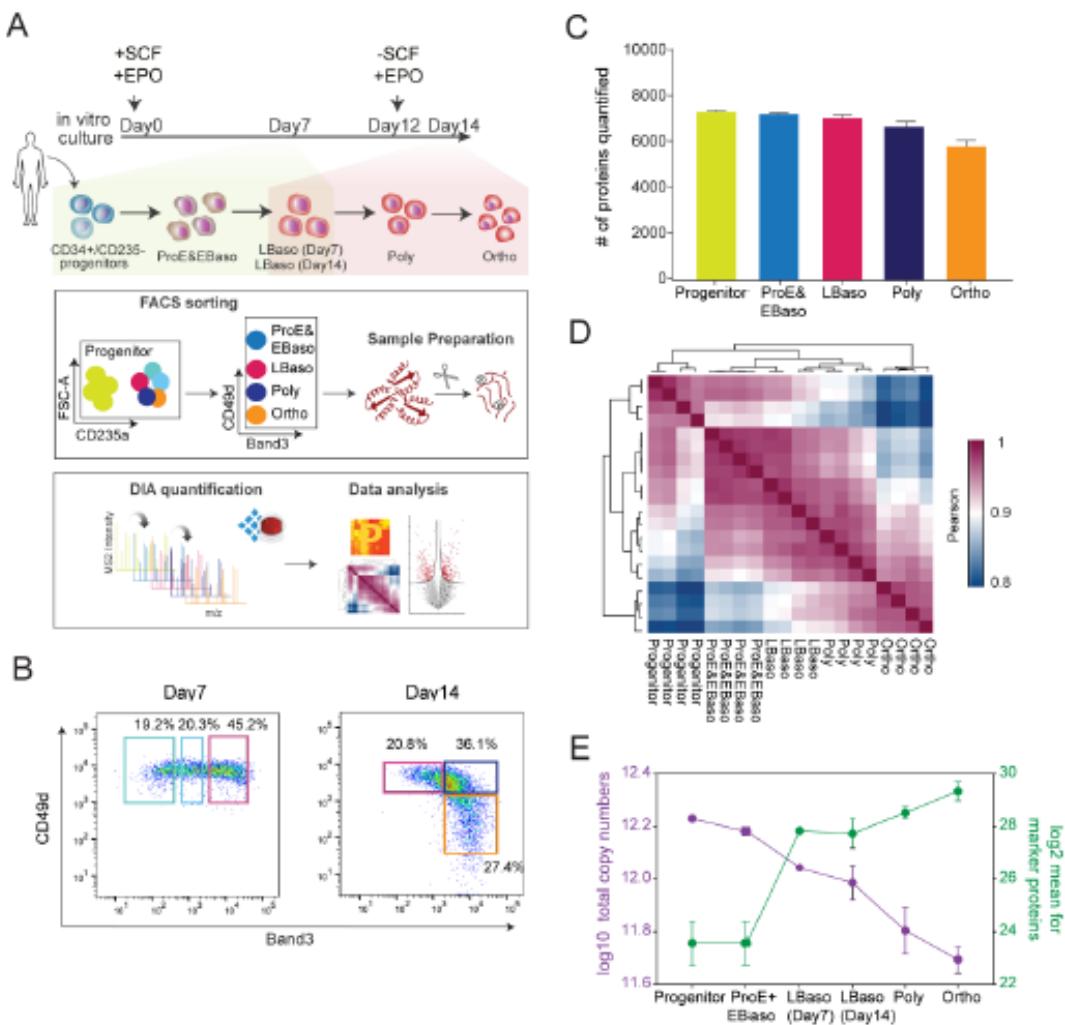
#### Dynamic and stage-specific proteome remodeling in erythropoiesis

The five stages of human erythropoiesis clustered separately by principal component analysis (PCA) with very high concordance between replicates (Figure 2A). Hierarchical clustering of 4,316 proteins with statistically different expressions (ANOVA,  $FDR < 0.01$ ), revealed drastic differences in the stage-specific proteomes. Rather than straightforward increase or decrease in protein levels across differentiation, proteins cluster into one of six distinct profiles of temporal co-expression dynamics (Figure 2B and Table S3). In addition to known developmental themes in each cluster, GO enriched terms point to novel state-specific regulation (summarized in Figure S3). In pairwise comparisons between successive stages, 2,157 proteins (29%) changed significantly at the first transition (Figure 2C). The overall proteome was more stable from ProE/EBaso to Poly stages, with 8.5% proteins up- or down-regulated. In contrast, almost 20% of the

proteome significantly changed in the last investigated transition, reflecting the specialization towards mature erythrocytes (Figure 2C).

To discover unique stage-specific marker proteins we compared all stages against each other (Figure S4A). Interestingly, the Poly stage can be distinguished by the centalspindlin and chromosomal passenger complexes (Benjamini-Hochberg,  $FDR < 0.01$ ). These proteins regulate cytokinesis in the late stages of cell division and also likely participate in erythroblast enucleation. Indeed, mutations in the kinesin KIF23B cause congenital dyserythropoietic anemia associated with erythroid multinuclearity and impaired erythropoiesis [50]. This analysis, like the ANOVA results, revealed the most drastic proteome changes occurring at the transition from progenitors to ProE/EBaso and from Poly to Ortho (Figure S4B). The cumulative proteome remodeling from the progenitors was reflected in a very large fraction of differentially represented proteins at the later maturation stages, Poly and Ortho (44% and 57%, respectively, two sample test,  $FDR < 0.01$  and  $S0 = 0.1$ ) (Figure 2D).

Taken together, our stage-specific proteomic data enable accurate, quantitative and in-depth monitoring of global protein expression during human erythropoiesis. The identified proteins are potentially important for the functional specialization of erythroid cells towards mature erythrocytes and represent excellent starting points for more detailed mechanistic studies.



**Figure 1. Establishing differentiation stage-specific proteomes of human erythropoiesis**

(A) Top panel depicts culture conditions for *in vitro* erythroid differentiation of CD34<sup>+</sup> cells. Shading indicates the presence of SCF and EPO (yellow), or EPO alone (pink). Note that SCF was present at 7 days but not at 14 days. The lower panels indicate the workflow of our study, including FACS gating/sorting strategy of erythroid precursors and single shot DIA analysis.

(B) FACS gating regime to enrich for ProE, EBaso, LBaso, Poly, and Ortho erythroblasts.

(C) Number of different proteins quantified in each differentiation stage.

(D) Correlation based clustering illustrating the reproducibility between biological replicates. High (1.0) and lower (0.8) Pearson correlations are denoted in pink and blue, respectively.

(E) Estimated copy numbers of total molecules (purple) and mean copy numbers of the proteins with GO annotations “erythrocyte maturation” and “heme biosynthesis” (green) per cell across maturation stages.

### Dramatic remodeling of the transmembrane proteome in erythropoiesis

Our data captures distinct regulation of proteins that contribute to the highly specialized erythroblast membrane at later developmental stages. Despite identification of several transmembrane proteins as markers of erythropoiesis over the years [51], there is still limited systems-wide information on them. Our optimized lysis and digestion protocol enabled unbiased access to the membrane proteome and provided a comprehensive view of membrane proteins during differentiation. Across the differentiation stages, we quantified 1,033 plasma membrane proteins (~21% of the total genome-encoded plasma membrane proteome in humans and ~14% in our study), of which 692 changed significantly (Table S1). Our data identifies a plethora of new examples that will aid in pinpointing maturation stages and in better understanding of erythroid biology.

Of the significantly changing membrane proteins, we could map 86% to pathways, with transport of small molecules across plasma membranes among the most represented ( $p = 8.7 \times 10^{-9}$ ). Further functional classification showed markedly strong enrichment of ‘SLC (solute carrier)-mediated transmembrane transport’ (Figure 3A). The roles of SLCs in biology has arguably been understudied, but now there are systematic efforts characterizing their roles [52]. Notably, since identification of “Band 3” as a solute carrier protein (SLC4A1) 35 years ago [53], it has become clear that SLC proteins must have widespread roles in erythropoiesis. Remarkably, our data quantified 101 SLCs, 68 of which significantly change in at least one transition (Figure 3B), likely reflecting remarkable changes in metabolic requirements along the stages of maturation. As summarized in Table S2, 62 of these have known or purported substrates associated with them.

Only 22 of the significantly regulated SLCs have previously been linked to erythrocytes, erythropoiesis or anemia. For instance, Mitoferrin-1 (SLC25A37), with a continuous upregulation during erythroid maturation, is a mitochondrial iron importer essential for heme biosynthesis in erythroblasts [54]. For some SLCs, roles in transporting nutrients including glucose and amino acids, and ions such as zinc, and necessary functions as redox regulators in erythropoiesis have already been described (Table S2). In addition, our dataset also contains many transporters – including for

vitamins, lipids, and whose substrates have not yet been identified – vastly extending the repertoire of SLCs and transported molecules associated with erythropoiesis.

Among the prominent observations emerging from our data were the several differentially expressed SLCs attributed to a common ligand. We first focused on hexose/glucose transporters. It has been known that cellular metabolism in mature red blood cells is strictly limited to glycolysis, which makes glucose uptake crucial for erythrocyte development. Glucose uptake during maturation appeared to roughly track with EPOR expression, reaching a maximal value when EPO response was highest, perhaps because of regulation by EPO stimulation in erythroid progenitor cells, as reported previously (Rogers et al., 2010). In line with the growing need for glucose during maturation, two out of four identified SLCs transporting glucose (SLC2A1 and SLC2A4) gradually increased from progenitors to Ortho (Figure 3B-C, Table S2) and their concordant profiles have been described recently (Justus et al., 2019). The other two glucose transporters (SLC2A3 and SLC45A3) are highly expressed specifically in progenitors, and to our knowledge have not been associated with erythropoiesis; their regulation would be interesting to investigate in the future.

A second remarkable example concerns SLCs for transporting metal ions (14 identified in total) with a full eight of them dedicated to zinc import and export. Maintenance of intracellular zinc levels controlled by GATA/heme circuit has recently been discovered as a vital determinant of erythroid maturation [55]. This indicates the adaptation of differentiating cells to stage-specific metabolic requirements and their interaction with the environment. Apart from the zinc importer SLC30A1 and zinc exporter SLC39A8, previously described in a “zinc switch” model reflecting their reciprocal expression during terminal erythropoiesis [55], we here uncovered additional three upregulated exporters (SLC39 family) and three downregulated importers (SLC30 family) during maturation (Figure 3B-C, Table S2). Only one of these had prior implications in erythrocyte homeostasis [56], suggesting even more intricate and possibly redundant regulation of zinc homeostasis.

#### Computational extraction and characterization of stage-specific protein markers

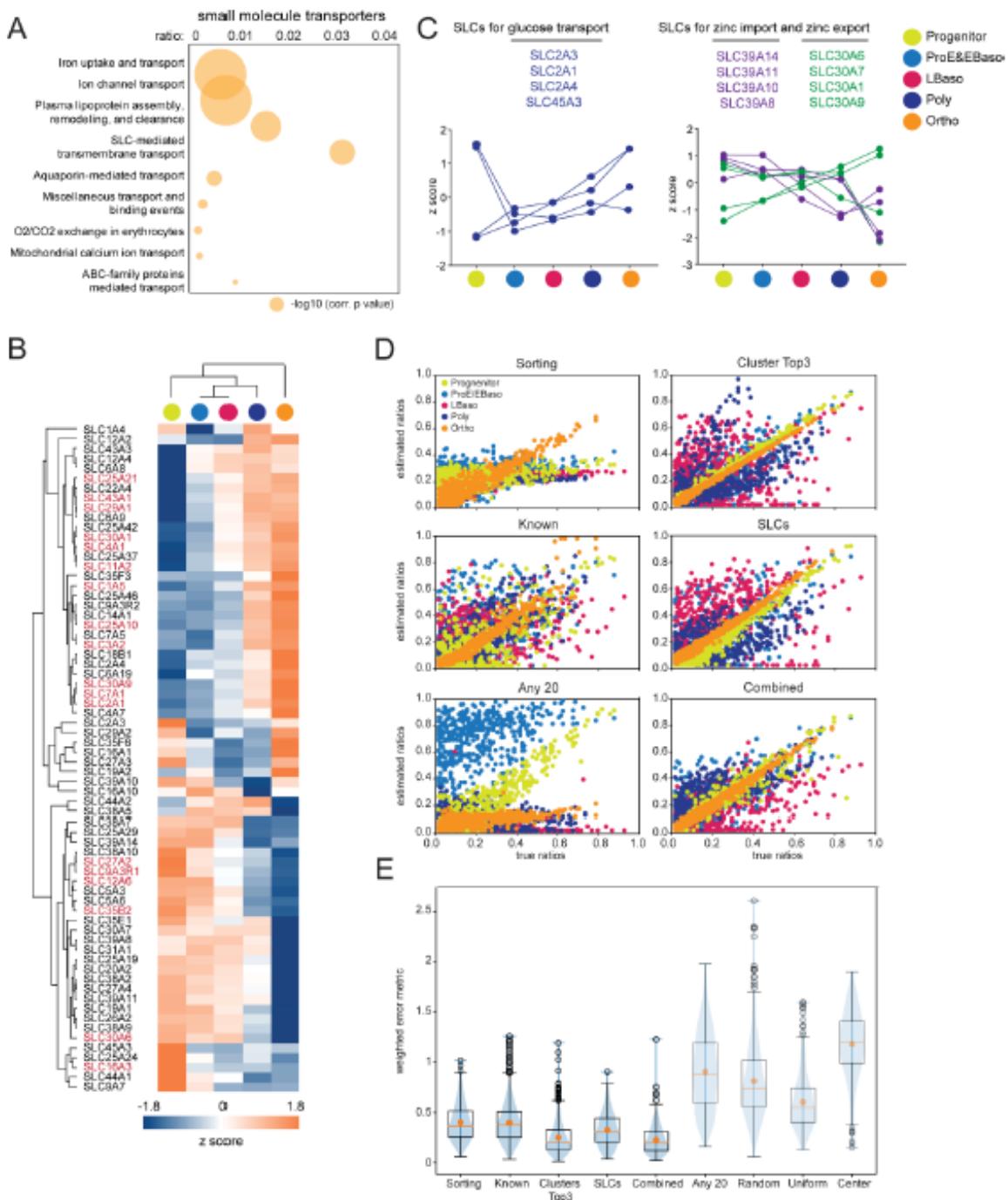
Given the distinct stage-specific expression patterns of the SLCs, we wondered if they could even serve as marker and selection proteins. The standard approaches for distinguishing erythroid developmental stages rely on canonical cell surface markers, including the ones we employed for FACS enrichments [51]. Our proteomics analysis revealed that drastic proteome-wide changes of numerous proteins occurred at transitions, in particular from progenitors to ProE/EBaso and from Poly to Ortho, implying the expression of numerous new stage-specific protein markers that might be exploited for refining the isolation and quantification of each differentiation stage. Panels of proteins with characteristic profiles could also be useful for *in silico* deconvolution of mixed developmental populations.

As a starting point, we investigated a *known marker set* of 22 proteins as well as *FACS sorting markers* (Table S2). These proteins correlated well with their expected expression profiles along the differentiation process. We next constituted a marker set of 18 SLCs on the basis of the most consistent quantification profiles (Figure 3B). As a final set, we selected 18 stage-specific proteins from our proteomics data comprising the top three most significant ones for each of the six clusters in Figure 2A (*cluster Top3 set*, smallest ANOVA q-values) (Table S2). Among those, KLF13, which activates the promoters of several erythroid genes *in vitro*, was gradually upregulated until very late stages, consistent with its reported role in mouse erythroblast maturation [57, 58].

With these four protein panels in hand (*sorting and known markers, cluster Top3, SLCs*), we developed an *in silico* deconvolution approach to distinguish

different developmental stages. Briefly, we used these marker panels to define signature matrices which we applied in 500 random *in silico* mixtures of aggregated abundances from a linear combination of the five differentiation stages at predefined ratios. We evaluated the results by comparing the estimated ratios to the predefined ratios of the *in silico* mixtures (Figure 3D-E).

The *sorting and known markers* reasonably estimated the fraction of the Ortho stage in the mixtures, but performed worse for all other stages (diagonal orange markers in Figure 3D). Remarkably, the *cluster Top3 and SLC markers* better characterized the differentiation process than previously known proteins and produced more accurate estimations for both Ortho and progenitor fractions in the computational mixture populations (diagonal orange and yellow markers in Figure 3D). However, they were still less effective at distinguishing stages from ProE to Poly, in line with their smaller proteome differences in our data (Figure 3D). A combined set of 62 proteins outperformed all others, even in estimating intermediate, adjacent differentiation stages as judged by a quantitative error analysis and compared to random controls (Methods, Figure 3E). In addition to recent advances in single cell transcriptomics [29], our deconvolution approach could further aid the identification of specific populations amongst bulk pools obtained during erythropoiesis, for example in the study of differentiation dynamics from *in vivo* samples. The proteins selected in this analysis, especially the SLCs add to our resource as they are interesting candidates for investigating stage-specific mechanisms in follow up studies.



**Figure 3. Solute carriers in erythroid maturation and computational extraction of stage-specific protein markers**

(A) Plot shows the overrepresented Reactome pathways [59, 60] with their corrected p-values (-log10) and the ratios of given entities from a particular pathway vs all entities from that pathway (n=1,882).

(B) Heat map of z-scored SLC protein abundances (log2 DIA intensities) across differentiation. The proteins in red were used to generate the input matrix for the SLCs marker set used in D and E.

(C) Expression of four different glucose transporters during erythrocyte development (log2 DIA intensities, left panel). Counterveiling expression regulation of zinc importers and exporters during erythrocyte development (log2 DIA intensities, right panel).

(D) Computation sorting quality comparing pre-defined versus estimated ratios of cells in the five differentiation stages.

(E) Accuracy of cell type prediction based on different protein marker sets as measured by a weighted error metric (y-axis, also see Methods). The blue violin plots illustrate the underlying distribution reaching from minimum to maximum. The black box plots depict the quartiles of the distribution with whiskers extending to the quartiles  $\pm$  1.5 x interquartile range. The orange horizontal lines indicate the median and the orange dot highlights the mean of the distribution.

#### An orchestrated network of erythropoietic kinases and their downstream target

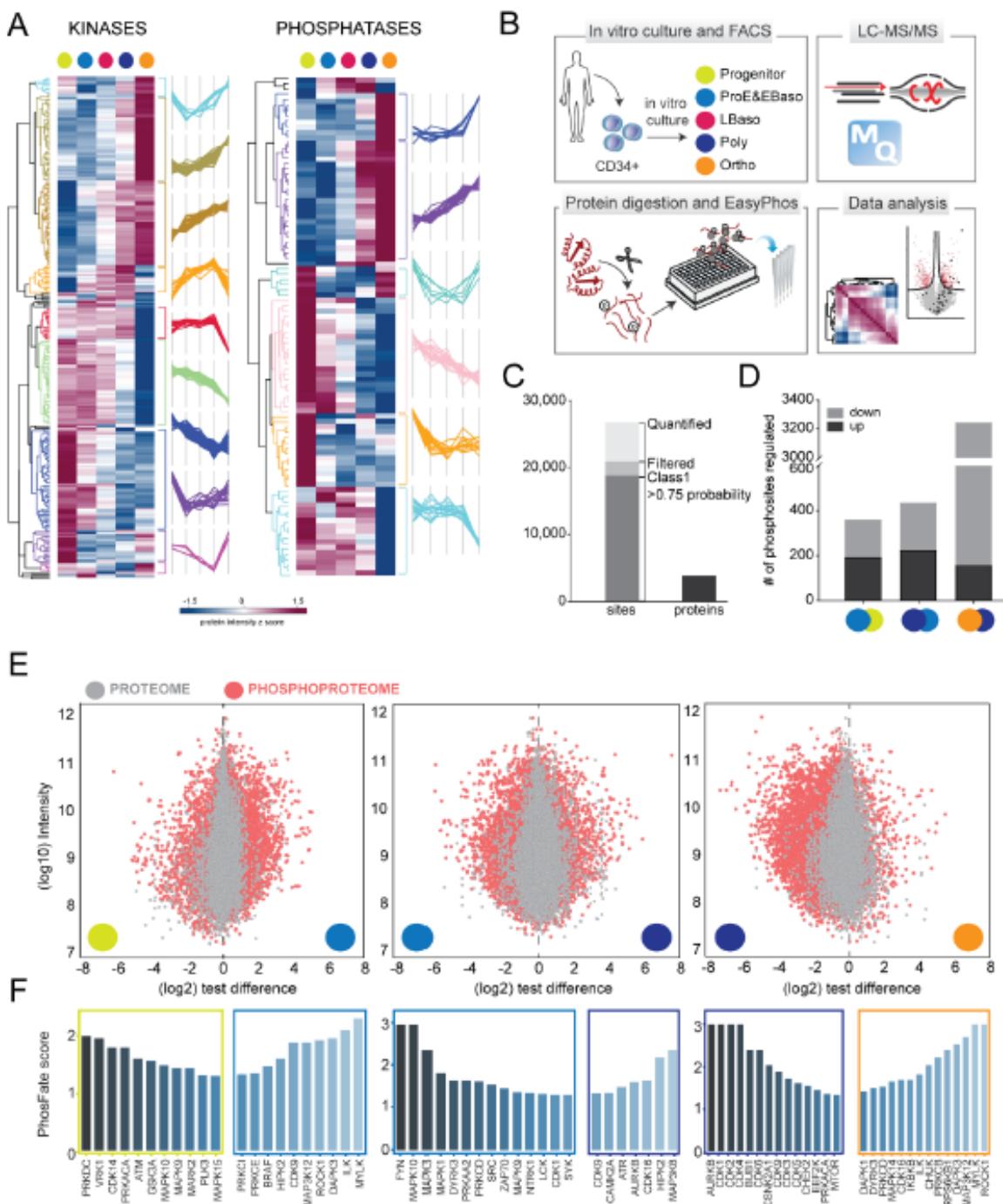
Several kinases act in or have already been implicated in a complex regulatory network in erythropoiesis. To advance our understanding of the dynamic phospho-regulatory network during erythropoiesis, we assessed temporal kinase activities at a global scale across terminal maturation. Mining of our proteome data revealed an astounding 270 kinases and 90 phosphatases that were differentially expressed with clear stage-specific profiles during differentiation (Figure 4A). To investigate their activities, we turned to phosphoproteomics which globally captures their substrates (Figure 4B). We enriched phosphopeptides from the same differentiation stages in biological quadruplicates using the EasyPhos platform (Figure 4B) [43, 44]. This streamlined protocol enabled deep profiling of phosphoproteomes at specific developmental stages in single-run DDA measurements from only 80  $\mu$ g of protein lysates, capturing 27,166 distinct phosphosites on more than 4,200 proteins (Figure 4B). Almost 20,000 sites were identified in more than two replicates of at least one maturation stage and 3,604 were novel sites according to the PhosphoSitePlus database [61] (Figure 4C and Table S3). Given the prominent changes in the plasma membrane proteome, it was interesting to see that 401 of them had phosphosites, often multiple ones within proximity in linear sequence stretches. This encompassed 23 of the aforementioned SLCs, suggesting stage-specific signaling roles [62] in addition to their dynamic expression across stages. Specifically, our phosphoproteomics also identified Ser/Thr phosphorylation sites on Band 3, whose tyrosine phosphorylation is known to enable docking of cytoplasmic signaling molecules [63, 64].

For further statistical analysis, we used a stringently filtered dataset of 12,216 phosphopeptides quantified in all four replicates of at least one differentiation stage. Strikingly, about half of these phosphosites significantly changed in at least one developmental transition (ANOVA, FDR<0.05) and

a quarter of all phosphosites (3,089) were dephosphorylated from Poly to Ortho stage (Figure 4D and Table S3).

To compare the dynamics of the phosphoproteomes to the proteomes, we visualized fold change distributions of quantified proteins (grey) and phosphopeptides (pink) for three pairwise comparisons: (i) progenitor vs ProE/EBaso, (ii) ProE/EBaso vs Poly, and (iii) Poly vs Ortho (Figure 4E). The fold change distributions of phosphopeptides were considerably more scattered than those of proteins in all three comparisons, reflecting dynamic, large scale phosphoregulation. The largest fold change of regulated phosphopeptides occurred between early stages of progenitor to ProE/EBaso and the later stages, Poly to Ortho (Figure 4E). The highly dynamic changes in global phosphorylation landscape likely reflects critical roles for distinct kinases at specific maturation stages.

Next, we inferred kinase activities from the phosphoproteome by stage-dependent enrichment analysis using PhosFate profiler (Figure 4F) [65]. This method predicts changes in kinase activity by testing the enrichment of differentially regulated, annotated kinase-substrate motifs. Substrates peaking during the early stages of differentiation (ProE/EBaso) were enriched with motifs for kinases of the MAPK signaling network (BRAF, MAPK1, MAPK3, FYN, SRC), which are known to promote cell cycle and proliferation [15, 66, 67]. Interestingly, the observed substrate phosphorylations suggest that CDK1 and many other cell cycle associated kinases (AURKB, BUB1, CDK14, CDK16, CDK2, CDK3, CDK4, CDK5, CDK6, and DYRK3) remain active until very late stages (Poly and Ortho). DNA damage checkpoint kinases (ATM, ATR, and CHEK2) were also enriched, presumably to maintain genome stability during erythroid differentiation. Together, our data reveal a rich network of temporally activated kinases during differentiation of human erythrocytes.



(E) Distributions of phosphopeptides and their matching proteins based on their log10 intensities (Y-axis) vs log2 test differences (X-axis) are illustrated for ProE/EBaso vs Progenitor (left), Poly vs ProE/EBaso (middle), and Ortho vs Poly (right). Pink represents phosphopeptides whereas grey represents proteins.

(F) Stage-specific predicted active kinases based on targeted sites identified by PhosFate profiler (<http://phosfate.com>). Left boxes represent kinases whose substrates are preferentially detected at the earlier stage of differentiation and right boxes represent those whose substrates are preferentially detected at the later stage.

#### CRISPR/Cas9 screen reveals critical functions of the erythropoietic kinome

The proteomics analysis established a “kinome atlas” revealing dynamic changes in kinase abundance and activity at distinct stages of erythropoiesis, with a dramatic decrease in the global phosphoproteome during late maturation. To investigate potential functional implications of these changes, we performed a CRISPR/Cas9 screen in HUDEP-2 cells, an immortalized human erythroblast line that proliferates in an immature state and can be induced to undergo terminal maturation by manipulation of culture conditions [68]. HUDEP-2 cells stably expressing Cas9 (HUDEP-2<sup>Cas9</sup>) were transduced at low multiplicity of infection with a lentiviral vector library encoding 3,051 single guide (sg) RNAs targeting the coding regions of most known kinases (n=482) and a green fluorescence protein (GFP) reporter gene [69, 70] (Figure 5A). Two days later, GFP<sup>+</sup> cells were flow cytometry-purified, and split into pools for further expansion or induced maturation, followed by next generation sequencing (NGS) to assess sgRNA abundance (Figure 5A and S5A-D). Compared to cells at two days post-transduction (“day0”), 30 sgRNAs were underrepresented after 10 days of expansion (FDR<0.05), reflecting candidate kinase genes that promote survival and/or proliferation of immature erythroblasts (Figure 5B and Table S4). These genes encoded cyclin-dependent kinases (CDK1, CDK7, CDK9), cell signaling components (KIT, JAK2), DNA damage checkpoint response proteins (ATR, CHEK1) and a regulator of ion flux (OXSR1), several of which exhibited maturation stage-specific expression in our proteome analysis (e.g. CDK1, CDK9, ATR, KIT, and JAK2) (Figure 4F). The KIT and JAK2 genes are essential signaling molecules for erythropoiesis [71-73]. Previous proteomic studies identified OXSR1 (OSR, oxidative stress-responsive kinase 1) as one of the most abundant Ser/Thr kinases in reticulocytes and mature erythrocytes [34]. The OXSR1 protein phosphorylates Na<sup>+</sup>-K<sup>+</sup> and K<sup>+</sup>-Cl<sup>-</sup> membrane co-transporters to activate and inhibit their activities, respectively [74]. Our data suggest a role for OSR1 in the maintenance of erythroid precursors.

Transduced HUDEP-2 cells induced to undergo terminal maturation were cultured for 3 days, fractionated according to their expression of the late-stage erythroid marker Band 3, and analyzed by NGS for sgRNA abundance. Single guide RNAs for five genes were significantly overrepresented in immature (Band3<sup>-</sup>) cells, indicating that these genes are positive effectors of maturation, while sgRNAs for nine genes were overrepresented in mature (Band3<sup>+</sup>) cells, representing candidates that inhibit maturation (Figure 5C and Table S5). There was minimal overlap between genes that affect expansion or maturation (Figure S5E). Notably, eight kinases identified as regulating differentiation in the CRISPR/Cas9 screen are also identified amongst the stage-specific active kinases predicted by phosphorylation of their cognate motifs (Figure 4F and Figure S5F).

We noted that disruption of numerous genes stimulating the Ras/MAPK signaling pathway caused accelerated erythroid maturation (Figure 5C). Three of the corresponding proteins, RAF1, BRAF1 and MAPK1, are members of the canonical Ras/MAPK family, while LYN is known to engage and potentiate RAF1 [75]. In non-erythroid cells, PIM1 kinase has been shown to phosphorylate ERK and activate Ras/MAPK signaling [76]. To validate these candidates, we transduced Cas9-expressing HUDEP-2 cells with individual sgRNAs for each gene and then induced erythroid maturation. Consistent with results of the screen, knockout of RAF1, BRAF1, MAPK3, LYN or PIM1 resulted in significantly accelerated terminal maturation (Figure 5D-E and S5G-I). Together, these findings indicate that downregulation of the Ras/MAPK pathway promotes terminal erythroid maturation. Consistent with this hypothesis, the TGFRB2 gene, identified as positive regulator of maturation (Figure 5C), is known to inhibit MAPK signaling [77].

Results of the screen identified PIM1 as a candidate gene that both drives erythroid precursor expansion and inhibits maturation (Figures 5B-C and S5H).

Consistent with the latter, PIM1 protein levels were undetectable after the LBaso stage (Figure 5F). The *PIM1* gene encodes a serine-threonine kinase oncoprotein that stimulates cell survival and cell cycle progression by phosphorylating numerous substrates that have been identified in non-erythroid cell types. To identify potential effectors of PIM1 signaling in erythroid cells, we searched our phosphoproteomics data for previously-described PIM1 substrates [61] and identified 25 significant phosphosites (out of 100) that coincided with PIM1 expression (Figure 5G). We then generated a PIM1 phosphorylation site consensus motif using the PhosphoSitePlus database [61] (Figure 5H), and investigated whether the motif is enriched in the erythroid maturation stage-dependent phosphoproteome. This identified 79 phosphorylation targets of which 54% have maturation stage-dependent profiles correlating with PIM1 protein abundance (Figure 5I). GO-term analysis revealed significant enrichment of terms associated with “chromatin remodeling”, “transcriptional regulation”, “kinase/phosphatase activity”, and “ubiquitylation” (Figure 5J). Of particular interest were Ras/MAPK family members such as MAP3K1, MAP3K5 and MAP3K2, consistent with a regulatory role of PIM1 in Ras/MAPK signaling.

#### System-wide dissection of c-Kit and EPOR phosphosignaling in erythropoiesis

Phosphoprotein analysis and a CRISPR-Cas9-sgRNA screen defined a dynamic, developmental stage-specific kinome during erythropoiesis and indicated that Ras/MAPK downregulation might be critical for erythroid maturation. We explored this further by examining our proteomics dataset for Ras/MAPK signaling components in relation to the expression and activities of c-Kit and EPOR. The kinetics of Ras/MAPK protein expression varied across erythroid maturation with MAPK1, MAPK3 and RAF1 persisting into late maturation stages (Figure 6A), suggesting that their activity may be suppressed post-translationally. In agreement, activating T185/Y187 phosphorylations on ERK indicated maximal activity during ProE/EBaso and termination by the LBaso stage (Figure 6B) [78, 79]. The activating S63 phosphorylation on ATF1, a distal target of Ras/MAPK signaling, peaked later (at the LBaso stage) and persisted throughout erythroid maturation (Figure 6B). The RSK kinase, which phosphorylates ATF1, is activated by both MAPK and PI3K/Akt-mTOR signaling [80].

The erythroid cytokine receptors c-Kit and EPOR have distinct roles in erythropoiesis, although their signaling pathways overlap considerably (Figure 6C). Our erythroid culture system contained both SCF (c-Kit Ligand) and EPO in the first and second phase (day0-7) and EPO only in the third phase (day12-14) (Figure 1A). The rationale for this culture system is based on findings that persistently elevated SCF-c-Kit signaling inhibits terminal erythroid maturation [81-83]. The levels of c-Kit and EPOR/JAK2 proteins decreased during differentiation but with differing kinetics (Figures 6D). c-Kit levels declined after the ProE stage, similar to ERK activity. Tyrosine phosphorylation in the c-Kit cytoplasmic domain, which reflects receptor activity [84], was maximal in erythroid progenitors and ProE/EBaso and decreased by the LBaso stage, even with SCF present in the culture media (Figure 6D). ERK2 phosphorylation declined at the same stage, even with SCF present in the culture medium (LBaso day7) and was undetectable in LBaso day14 when SCF was not present in the medium. Thus, c-Kit protein levels and its phosphorylation, along with downstream Ras/MAPK signaling are downregulated relatively early in erythroid maturation, consistent with the literature (Figure 6E-F) [40, 85].

Compared to c-Kit, EPOR/JAK2 levels were stable until the Poly stage. We were not able to detect phosphorylation of EPOR or JAK2, perhaps because the levels of these proteins are relatively low. Compared to ERK, STAT5 phosphorylation declined more slowly and persisted until the later stages of erythropoiesis, similar to the kinetics of EPOR expression (Figure 6G). Thus, ERK phosphorylation levels parallel the expression and activation of c-Kit, while phospho-STAT5 levels correlate with expression of EPOR, likely reflecting preferential signaling activities of the two cytokine receptor pathways. Together, our findings suggest that Ras/MAPK activity occurs during early stages of erythropoiesis, delays terminal maturation and is c-Kit driven. This is consistent with the established role for c-Kit in supporting proliferation and survival of early erythroid progenitor cells and a requirement for c-Kit downregulation during normal erythropoiesis [8, 83, 86].

Phosphorylation is also regulated by phosphatases. Of 51 phosphatases implicated in inhibiting the Ras/MAPK pathway [87, 88], 16 were detected in our data, 6 of which were induced during terminal

maturation (Figure 6H). The majority of these phosphatases are novel candidate genes whose roles in erythropoiesis need further exploration.

## DISCUSSION

Here we show that in-depth quantitative proteomic and phosphoproteomic analysis of purified erythroid precursors at distinct maturation stages is now made possible by state-of-the art MS-based proteomics and our EasyPhos technology, allowing us to assess erythroid maturation at the level of the proteins, the main functional cellular entities. The breadth and depth of coverage achieved by these technologies offers unbiased system wide insights into the regulation of erythropoiesis, which we complemented further by performing an unbiased CRISPR/Cas9 screen to interrogate the erythroid kinome.

Our analyses of proteins mediating solute transport and phospho-based signaling highlight two examples by which the data can be mined for hypothesis generating discovery and focused problems related to erythroid biology. The key emerging concept is that proteome-wide changes accompanying differentiation from early erythroid progenitors into nearly mature erythrocytes involves remarkable regulation of and by signaling pathways – at the protein level.

Tracking the levels across particular families of proteins defined numerous distinctive stage-specific profiles, exemplified by coordinated expression of specific cohorts of SLCs, kinases, and phosphatases. For some of these SLCs, previous reports established roles in transporting crucial molecules for erythropoiesis (Table S3). However, the much larger repertoire of SLCs of unknown biological function likely reflects erythroid stage-specific metabolic requirements to be elucidated in future studies. Beyond this, both the substrate diversity of SLCs and their widespread phosphorylation (average of three per SLC) point toward exquisite fine-tuning and coordination of transporters with stage-specific signaling pathways. So far, very few examples of this mode of regulation have been described, most prominently the tyrosine phosphorylation of the SLC Band 3, which mediates docking of cytoplasmic signaling molecules [63]. The prevalence of stage-specific SLC phosphorylation observed here may indicate system-wide coordination of small molecule transport with

cytoplasmic signaling throughout erythroid maturation and across many transporters. SLCs are relevant to therapeutics of several human diseases and to drug discovery, either as drug targets themselves or as mediators of drug uptake [52]. It now seems likely that the varying expression of SLCs with overlapping selectivities, as well as their regulation by post-translational modification, will also contribute to pathology and provide opportunities for therapeutic development [89]. Importantly, our data provide a framework for systemwide studies of SLC small molecule flux and signaling throughout the differentiation process. The dynamics of SLCs, together with more than 700 other quantified membrane proteins may furthermore contribute to our understanding of changing cell membrane properties required for erythropoiesis.

The distinct cohorts of kinases and phosphatases expressed coordinately and with varying kinetics across the erythrocyte maturation pathway likewise reflect extensive protein-level regulation, in this case through post-translational modification. With this notion in mind, we complemented the quantitative stage-specific proteome measurements with a kinase-targeting CRISPR/Cas9 screen and phosphoproteomics, which provides a profile of system-wide signaling across erythroid maturation. Pursuing PIM1, the highest scoring hit for erythroid maturation of HUDEP-2 cells, we defined a composite profile based on (1) stage-specific expression, (2) phosphorylation kinetics of known substrates, and (3) a PIM1 consensus sequence by correlation. Further analysis then provided a list of candidate PIM1 substrates that kinetically parallel PIM1 activity and may coordinate PIM1 activity with that of other diverse signaling effectors including epigenetic regulators, regulators of ubiquitin signaling, and more, whose functional roles in erythropoiesis can now be studied.

We also took advantage of our data to mine the stage-resolved phosphoproteomics of the SCF- and EPO-triggered signaling network, based on its crucial role in erythropoiesis, and opportunities offered by our culturing and FACS-based protocol. Combining phosphoproteomics and a functional screen, our data highlights a general decrease in kinase activity across the erythroid proteome during terminal maturation and in particular a critical role for downregulation of Ras/MAP kinases activity is suggested. Previous studies have demonstrated

defective terminal maturation in systems expressing either constitutively active forms of c-Kit or Ras proteins [81, 85]. Our systems-wide analysis of erythropoiesis derived from primary healthy donor human CD34<sup>+</sup> cells suggests c-Kit drives expansion of early erythroid precursors and inhibits terminal maturation via Ras/MAPK signaling whereas EPOR drives signaling predominantly via the JAK-STAT5 pathway to foster later stages of terminal maturation. These examples demonstrate the ability to gain insight on regulation of complex signaling systems during erythropoiesis using our phosphoproteomics dataset and provide a framework for future studies to interrogate stage-specific erythroid cytokine signaling and regulatory pathways that dampen this signaling.

Given the unexpectedly large role of phosphosignaling during erythropoiesis defined by our unbiased global, high resolution proteomic study, it will now be interesting to investigate other post translational protein modifications, which could employ different enrichment steps but similar strategies for bioinformatic and functional follow up. In this regard, we already observed distinct regulation of more than a hundred members of the ubiquitin machinery, making this post-translational modification particularly exciting for further explorations.

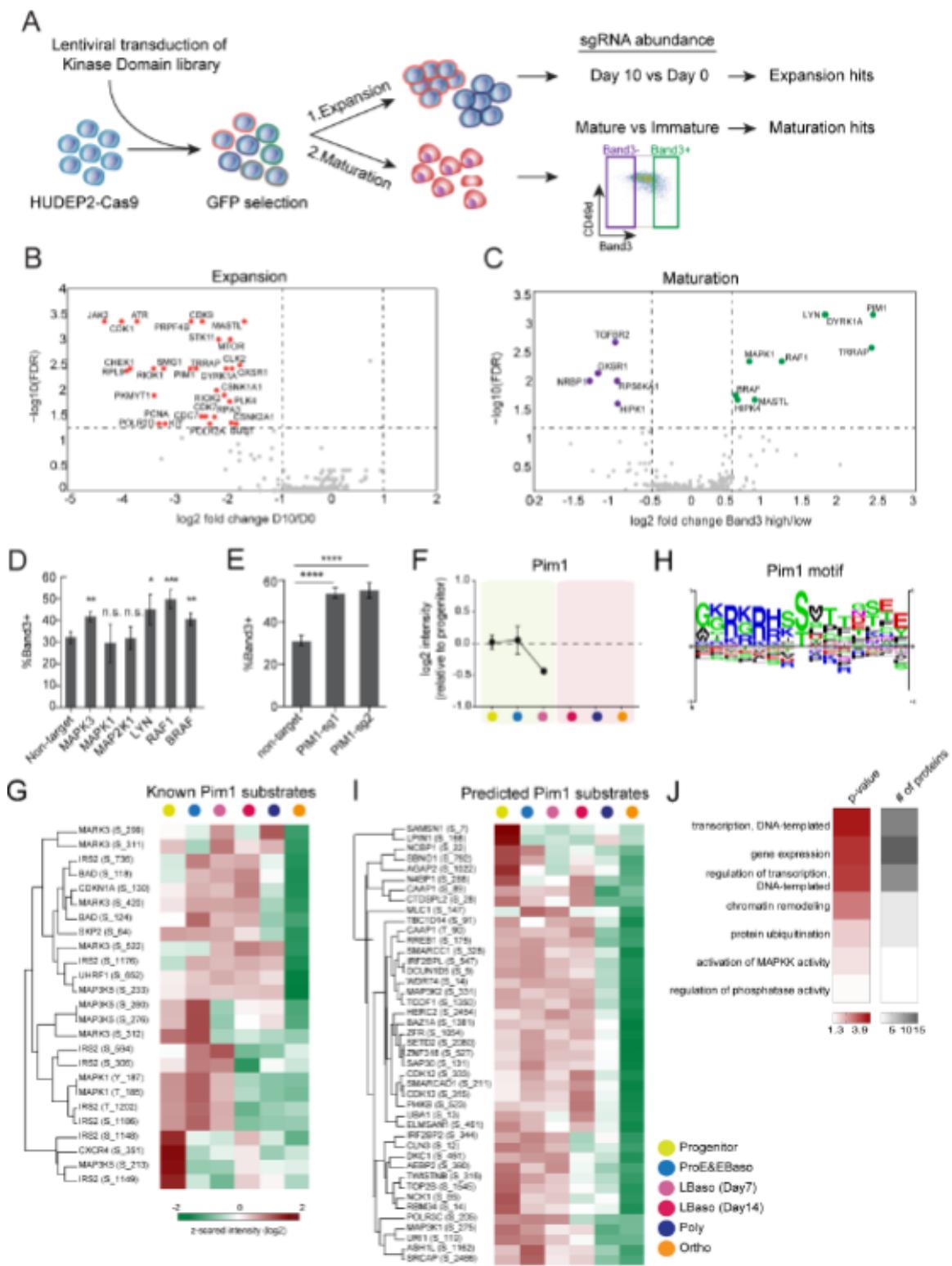


Figure 5. Kinome-targeting CRISPR/Cas9 screen in HUDEP-2 cells

(A) Workflow of a CRISPR/Cas9 screen with an sgRNA library targeting 482 human kinase genes to identify those that alter erythroid precursor expansion or terminal maturation.

(B) Volcano plots showing FDR vs log<sub>2</sub> fold-change in sgRNA abundance between Day 0 and Day 10 of expansion. Results were analyzed using Mageck (Methods). Each dot represents a single kinase gene based on the enrichment of four sgRNAs. Significantly different genes (FDR<0.05; log<sub>2</sub>fold-change<-1) are shown in red.

(C) Volcano plots showing FDR vs log2 fold-change in sgRNA abundance between Band3 high and low fractions after three days of maturation. Results were analyzed using Mageck (Methods). Each dot represents a single kinase gene based on the enrichment of four sgRNAs. The significant positive or negative regulators for maturation (FDR <0.05; log<sub>2</sub> fold-change<-0.5 and log<sub>2</sub> fold-change>0.5) are shown in purple and green, respectively.

(D) HUDEP-2 cells expressing Cas9 were transduced with lentiviral vectors encoding single sgRNAs targeting the indicated genes, induced to undergo terminal maturation and analyzed after 3 days. Graph shows fraction of Band3<sup>+</sup> cells. Error bars represent mean  $\pm$  SEM of 3 biological replicates. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.005$ ; n.s., not significant; unpaired t-test.

(E) Effects of two different PIM1-targetign sgRNAs on erythroid maturation of HUDEP-2 cells, performed as described for panel D. Error bars represent mean  $\pm$  SEM of 3 technical replicates. \*\*\* $P < 0.005$ , \*\*\*\* $P < 0.0001$ ; unpaired t-test.

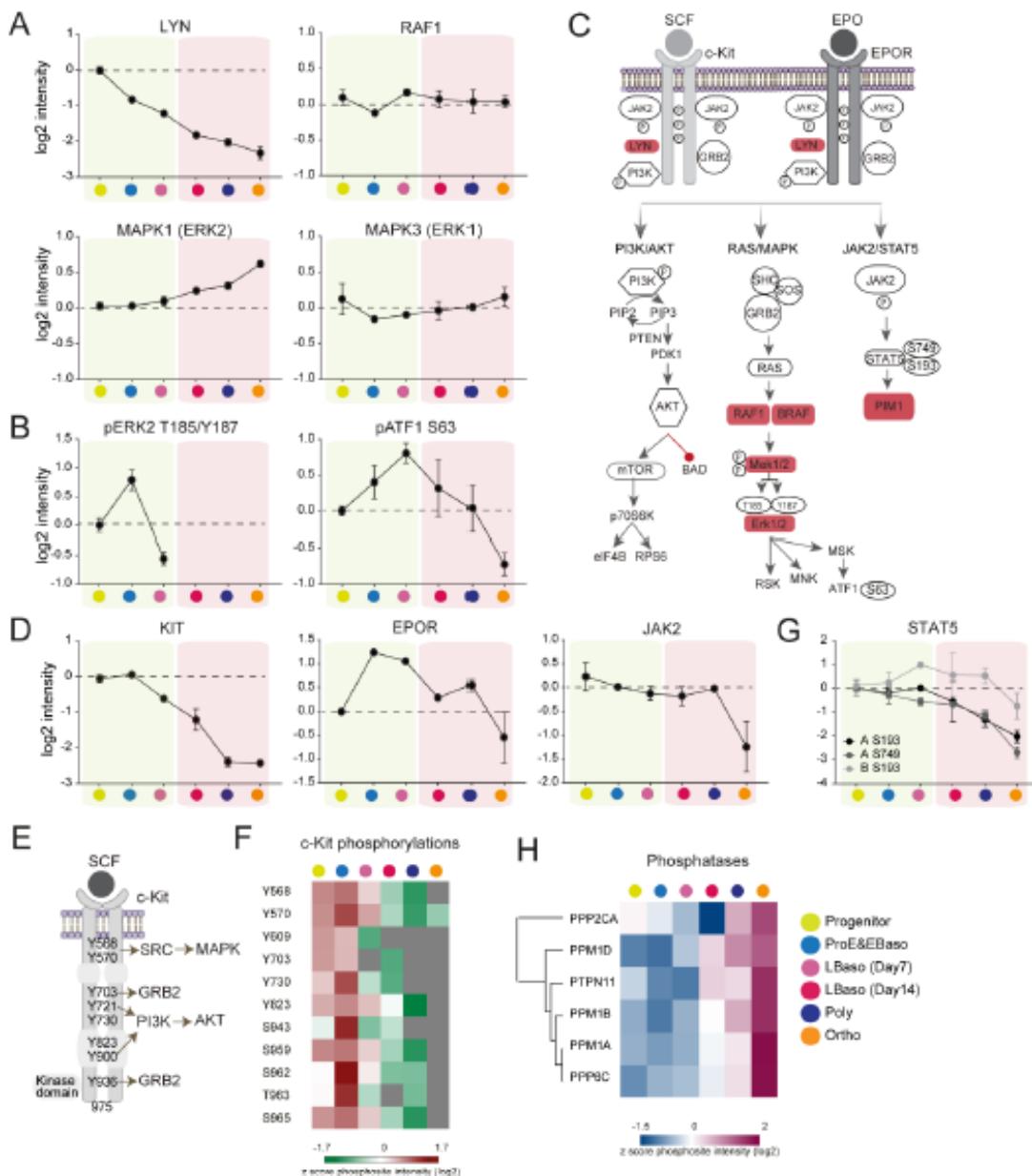
(F) Protein abundances (log2) of PIM1 during terminal differentiation of primary erythroblasts. Yellow highlighting indicates that SCF and EPO were present in the culture medium, while pink indicates EPO only.

(G) Heat map showing z-scored (log2) phosphopeptide intensities detected in known PIM1 substrates.

(H) Consensus PIM1 kinase motif from PhosphoSitePlus database [61].

(I) Heat map of z-scored (log2) phosphopeptide intensities of potential PIM1 kinase targets identified by motif analysis.

(J) Gene Ontology (GO) enrichment analysis of potential PIM1 kinase targets, performed using Fischer's exact test. 5% threshold was applied to Benjamini-Hochberg FDR to determine the significance.



**Figure 6. System-wide dissection of c-Kit and EPOR phosphosignaling**

(A) Protein abundances (log2 DIA intensities) normalized to progenitor stage. In all panels, stages shaded yellow were cultured with SCF and EPO, while those shaded pink were cultured with EPO only. Data is plotted if quantified in at least 50% of biological replicates. Error bars represent mean  $\pm$  SEM of at least two biological replicates.

(B) Profiles of phosphorylations (log2 DDA intensities) normalized to progenitor stage.

(C) Major signaling pathways downstream of c-Kit and EPOR activation by their corresponding ligands stem cell factor (SCF) and erythropoietin (EPO). Shaded genes indicate those identified to inhibit erythroid maturation in the CRISPR/Cas9 screen described in Figure 5.

(D) Protein abundances (log2 DIA intensities) of c-Kit, EPOR, and JAK2, normalized to progenitor stage.

(E) Following activation by SCF, phosphorylated tyrosine residues on c-Kit receptor serve as binding sites to key signal transduction molecules (SRC, GRB2, and PI3K) resulting in activation of downstream signaling pathways.

(F) Heat map of z-scored (log2) phosphopeptide intensities of c-Kit receptor.

(G) Profiles of STAT5A/B phosphorylations (log2 DDA intensities) normalized to progenitor stage.

(H) Heat map of z-scored protein abundances (log2 DIA intensities) of phosphatases.

#### ACKNOWLEDGEMENT

This work was supported by the Max-Planck Society for the Advancement of Science and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SCHU 3196/1-1. We thank Florian Meier, Igor Paron, Christian Deiml, Philipp Geyer, Johannes B Mueller, Fynn M Hansen, Sebastian Virreira Winter and all the members of the departments of Proteomics and Signal Transduction and Molecular Machines and Signaling at Max-Planck-Institute of Biochemistry for their assistances and helpful discussions. We also thank the NCI Cancer Center grant to St. Jude (NIHP30CA021765) and St Jude Core facilities including Flow Core for cell sorting, Hartwell Center for NGS, and Image Core for Cytospin Scanning.

#### AUTHOR CONTRIBUTIONS

OK performed proteomics experiments and analyzed the data. PX and YY performed FACS sorting, tissue culture experiments and CRISPR/Cas9 screen and biological validation assay for PIM1. IB developed the bioinformatics deconvolution approach to validate markers. ARFC and ASD helped with the bioinformatics analysis of phosphoproteome data. SVB helped with the analysis and the interpretation of the data. OK, PX, AFA, SVB, BAS, MW and MM designed the study and wrote the paper. BAS, AFA, MW and MM coordinated and supervised.

#### DECLARATION OF INTERESTS

The authors declare no competing interests.

#### METHODS

##### CD34<sup>+</sup> cell culture and manipulation

Human CD34<sup>+</sup> cells were obtained under human subject research protocols that were approved by local ethical committees: St. Jude Children's Research Hospital protocol "Bone marrow for hemoglobinopathy research" (NCT00669305). CD34<sup>+</sup> hematopoietic stem and progenitor cells (HSPCs) were mobilized from normal subjects by granulocyte colony-stimulating factor, collected by apheresis, and enriched by immunomagnetic bead selection using an autoMACS Pro Separator (Miltenyi Biotec), according to the manufacturer's protocol. At least 95% purity was achieved, as assessed by flow cytometry using a PE-conjugated anti-human CD34 antibody (Miltenyi Biotec, clone AC136, #130-081-002). A 3-phase culture protocol was used to promote erythroid differentiation and maturation. In phase 1 (days 0–7), cells were cultured

at a density of 10<sup>5</sup>–10<sup>6</sup> cells/mL in IMDM with 2% human AB plasma, 3% human AB serum, 1% penicillin/streptomycin, 3 IU/mL heparin, 10 µg/mL insulin, 200 µg/mL holo-transferrin, 1 IU EPO, 10 ng/mL SCF, and 1 ng/mL IL-3. In phase 2 (days 8–12), IL-3 was omitted from the medium. In phase 3 (days 12–18), cells were cultured at a density of 10<sup>6</sup>/mL, with both IL-3 and SCF being omitted from the medium and the holo-transferrin concentration increased to 1 mg/ml. Erythroid differentiation and maturation were monitored by flow cytometry, using FITC-conjugated anti-CD235a (BD Biosciences, clone GA-R2, #561017), APC-conjugated anti-Band3 (gift from Xiuli An Lab in New York Blood Center), and VioBlue-conjugated anti-CD49d (Miltenyi, clone MZ18-24A9, #130-099-680).

##### CRISPR/Cas9 screen with kinase-domain library

The kinase domain-focused sgRNA library was designed based on the human kinase gene list from a previous study [90]. The kinase enzymatic domain information was retrieved from NCBI database conserved domain annotation. Six independent sgRNAs were designed for targeting each individual domain regions. All the sgRNAs were designed using the same design principle reported previously and the sgRNAs with the prediction of high off-target effect were excluded [91]. Domain targeting and positive/negative control sgRNAs were synthesized in duplicate or triplicate in a pooled format on an array platform (Twist Bioscience) and then PCR cloned into the BsmB1-digested LRG2.1 vector (Addgene: #108098) using Gibson Assembly kit (NEB). Approximately 12 × 10<sup>6</sup> HUDEP-2 cells stably expressing Cas9 were transduced at a multiplicity of infection (MOI) of ~0.3 to minimize the transduction of any cell with more than 1 vector particle and achieve an approximately 1000-fold library coverage (100ul Virus per 2 M cells) such that 40% cell were GFP positive. Two days after infection, GFP<sup>+</sup> cells were sorted by FACS and then maintained in the expansion medium for 6 days (total 8 days post-infection). Then, half of total cells were kept in expansion culture for additional 10 days and the other half were transitioned to differentiation media and induced maturation for 3 days. Erythroid maturation was monitored by flow cytometry, using FITC-conjugated anti-CD235a (BD Biosciences, clone GA-R2), APC-conjugated anti-Band3 (gift from Xiuli An Lab in New York Blood Center), and Violet Blue-conjugated anti-CD49d (Miltenyi, clone MZ18-24A9). Band3<sup>+</sup> and Band3<sup>-</sup> cell populations from the CD235a<sup>+</sup> cell fraction was

purified by fluorescence-activated cell sorting (FACS). Library preparation and deep sequencing were performed as previously described 46,47. Briefly, genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen). Reactions were done with 24 cycles of amplification with 200 ng of gDNA in 25  $\mu$ L CloneAmp enzyme system and 8 parallel reactions were performed to maintain sgRNA library representation. PCR reactions were then pooled for each sample and column purified with QIAGEN PCR purification kit. PCR products were analyzed on an agarose gel, and the DNA band of expected size was excised and purified. MiSeq 250-bp paired-end sequencing (Illumina) was performed. For data analysis, FastQ files obtained after MiSeq sequencing were demultiplexed using the MiSeq Reporter software (Illumina). Paired reads were trimmed and filtered using the CLC Genomics Workbench (Qiagen) and matched against sgRNA sequences within the library. Read counts for each sgRNA were normalized against total read counts across all samples. Mageck method were used for differential analysis for sgRNA and Gene ranking [92, 93]. A  $P < 0.05$  was considered to be statistically significant.

**HUDEP-2 cell culture and induced maturation**  
Mycoplasma-free HUDEP-2 cells were cultured as described [68]. Immature cells were expanded in the StemSpan serum-free medium (SFEM; Stem Cell Technologies) supplemented with 1  $\mu$ M dexamethasone, 1  $\mu$ g/mL doxycycline, 50 ng/mL human stem cell factor (SCF), 3 units/mL erythropoietin (EPO), and 1% penicillin-streptomycin. To induce erythroid maturation, HUDEP-2 cells were cultured in a differentiation medium composed of IMDM base medium (Invitrogen) supplemented with 2% FBS, 3% human serum albumin, 3 units/mL EPO, 10  $\mu$ g/mL insulin, 1000  $\mu$ g/mL holo-transferrin, and 3 units/mL heparin. Erythroid differentiation and maturation were monitored by flow cytometry, using FITC-conjugated anti-CD235a (BD Biosciences, clone GA-R2, #561017), APC-conjugated anti-Band3 (gift from Xiuli An Lab in New York Blood Center), and VioBlue-conjugated anti-CD49d (Miltenyi, clone MZ18-24A9, #130-099-680).

#### CRISPR/Cas9-mediated genome editing of HUDEP-2 cells

The sgRNA sequences were selected from the CRISPR library, generated as oligonucleotides. After annealing, construct was cloned into the *Bbs*I

or *Bsm*BI site of the pXPR\_003 vector. Lentivirus supernatant were prepared from 293T cells. For cell pool genome editing, HUDEP-2 cells stably expressing Cas9 were transduced with lentiviral vector (pXPR\_003) encoding individual sgRNAs. Cells were incubated for 7–10 days with 10  $\mu$ g/mL blasticidin and 1  $\mu$ g/mL puromycin to select for transduction with sgRNA and Cas9 vectors, respectively. On-target insertion/deletion mutations were characterized by PCR, followed by next-generation sequencing or TIDE-seq analysis from Sanger sequencing datasets.

#### Cell lysates and immunoblot analysis

Cells were suspended in Thermo Scientific Pierce IP Lysis Buffer (ThermoFisher #87787) supplemented with 1 mM phenylmethylsulfonyl fluoride, and 1:500 protease inhibitor cocktail (Sigma-Aldrich). Proteins were resolved on polyacrylamide gels (BioRad), transferred to a PVDF membrane, and incubated in blocking buffer (5% milk in TBST). Antibody staining was visualized using the Odyssey CLx Imaging System.

#### (Phospho)proteome sample preparation for MS analysis

All MS experiments were performed in biological quadruplicates. Cell pellets were lysed in SDC buffer (4% Sodium deoxycholate in 100 mM Tris pH 8.5) and heated for 5 min at 95°C. Lysates were cooled on ice and sonicated. Protein concentration was determined by Tryptophan assay [94]. We reduced disulphide bonds and carbamidomethylate cysteine residues by adding TCEP and 2-Chloroacetamide to the final volumes of 10 mM and 40 mM, respectively, for 5 min at 45°C. Proteins were subsequently digested by the addition of 1:100 LysC and Trypsin overnight at 37°C with agitation (1,500 rpm). Next day, 20  $\mu$ g of protein material was aliquoted and processed using an in-StageTip (iST) protocol [94]. Approximately 500 ng peptide was used for single shot DIA analysis, while the rest (~10 $\mu$ g) of clean peptides were fractionated using the high-pH reversed-phase 'Spider fractionator' into 8 fractions to generate deep proteomes to build spectral library [95]. 80  $\mu$ g of peptides were used for phosphopeptide enrichment using the EasyPhos workflow as described previously [43, 44]. After mixing peptides with Isopropanol and EP enrichment buffer (48% TFA, 8 mM KH<sub>2</sub>PO<sub>4</sub>), they were enriched with 5mg of TiO<sub>2</sub> beads which were prepared at a concentration of 1 mg/ $\mu$ l in loading buffer (6% TFA/80% ACN (vol/vol)) and incubated

at 40°C with shaking (2,000 rpm) for 5 min. Afterwards, the phosphopeptide containing TiO<sub>2</sub> beads were further washed with 4ml wash buffer (5% TFA/60% ISO (vol/vol)), and treated with elution buffer (40% ACN, 15% NH<sub>4</sub>OH). Eluted phosphopeptides were concentrated in a SpeedVac for 20min at 45 °C, and using StageTips loaded with SDB-RPS discs. 6  $\mu$ l MS loading buffer (0.2% TFA/2% ACN (vol/vol)) was added to the dried samples prior to LC-MS/MS analysis.

#### Liquid chromatography-MS analysis

Nanoflow LC-MS/MS measurements were carried out on an EASY-nLC 1200 system (ThermoFisher Scientific) combined with the latest generation linear quadrupole Orbitrap instrument (Q Exactive HF-X) coupled to a nano-electrospray ion source (Thermo Fisher Scientific). We always used a 50 cm HPLC column (75  $\mu$ m inner diameter, in-house packed into the tip with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin (Dr. Maisch GmbH)). Column temperature was kept at 60°C by a Peltier element containing in-house developed oven.

500 ng peptides were analyzed with a 100 min gradient. Peptides were loaded in buffer A (0.1% formic acid (FA) (v/v)) and eluted with a linear 80 min gradient of 5-30% of buffer B (80% acetonitrile (ACN) plus 0.1% FA (v/v)), followed by a 4 min increase to 60% of buffer B and a 4 min increase to 95% of buffer B, and a 4 min wash of 95% buffer B at a flow rate of 300 nl/min. Buffer B concentration was decreased to 4% in 4 min and stayed at 4% for 4 min.

For the analysis of the fractions, the instrument was operated in the DDA mode (Top12). The resolution of the Orbitrap analyzer was set to 60,000 and 15,000 for MS1 and MS2, with a maximum injection time of 20 ms and 60 ms, respectively. The mass range monitored in MS1 was set to 300–1,650 m/z. The automatic gain control (AGC) target was set to 3e6 and 1e5 in MS1 and MS2, respectively. The fragmentation was accomplished by higher energy collision dissociation at a normalized collision energy setting of 27%. Dynamic exclusion was 20 sec.

For single shot samples, the instrument was operated in the DIA mode. Every MS1 scan (350 to 1650 m/z, 120,000 resolution at m/z 200, AGC target of 3e6 and 60 ms injection time) was followed by 33 MS2 windows ranged from 300.5 m/z (lower boundary of

first window) to 1649.5 m/z (upper boundary of 33rd window). This resulted in a cycle time of 3.4 s. MS2 settings were an ion target value of 3 x 106 charges for the precursor window with an Xcalibur-automated maximum injection time and a resolution of 30,000 at m/z 200. The fragmentation was accomplished by higher energy collision dissociation with stepped collision energies of 25.5, 27 and 30%. The spectra were recorded in profile mode. The default charge state for the MS2 was set to 3. Data were acquired with Xcalibur 4.0.27.10 and Tune Plus version 2.1 (Thermo Fisher).

Phosphopeptides were analyzed with a 100 min gradient. Peptides were loaded in buffer A (0.1% formic acid (FA) (v/v)) and eluted with a linear 60 min gradient of 3-19 of buffer B (80% acetonitrile (ACN) plus 0.1% FA (v/v)), followed by a 30 min increase to 41% of buffer B and a 5 min increase to 90% of buffer B, and a 5 min wash of 90% buffer B at a flow rate of 350 nl/min. The instrument was operated in the DDA mode (Top10). The resolution of the Orbitrap analyzer was set to 60,000 and 15,000 for MS1 and MS2, with a maximum injection time of 120 ms and 50 ms, respectively. The mass range monitored in MS1 was set to 300–1,600 m/z. The automatic gain control (AGC) target was set to 3e6 and 1e5 in MS1 and MS2, respectively. The fragmentation was accomplished by higher energy collision dissociation at a normalized collision energy setting of 27%. Dynamic exclusion was 30 sec.

#### MS data analysis

The fractions (DDA) and the single shot samples (DIA) were used to generate a DDA-library and direct-DIA-library, respectively, which were combined into a hybrid library in Spectromine version 1.0.21621.8.15296 (Biognosys AG). The hybrid spectral library was subsequently used to search the MS data of the single shot samples in Spectronaut version 12.0.20491.9.26669 (Biognosys AG) for final protein identification and quantification. All searches were performed against the Human UniProt FASTA database (2017). Carbamidomethylation was set as fixed modification and acetylation of the protein N-terminus and oxidation of methionine as variable modifications. Trypsin/P proteolytic cleavage rule was used with a maximum of two miscleavages permitted and a peptide length of 7-52 amino acids. When generating the spectral library generation, minimum and maximum of number of fragments per peptide were

set to 3 and 6, respectively. A protein and precursor FDR of 1% were used for filtering and subsequent reporting in samples (q-value mode).

For the phosphoproteome, raw MS data were processed using MaxQuant version 1.6.2.10 [96, 97] with an FDR < 0.01 at the peptide and protein level against the Human UniProt FASTA database (2017). Enzyme specificity was set to trypsin, and the search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein and oxidation of methionine and phosphorylation (SYT) as variable modifications. Up to two missed cleavages were allowed for protease digestion, and peptides had to be fully tryptic.

#### Bioinformatics data analysis

We mainly performed data analysis in the Perseus (version 1.6.0.9) [98], Microsoft Excel and data visualized using GraphPad Prism (GraphPad Software) or RStudio (<https://www.rstudio.com/>). Apart from coefficient of variation, log<sub>2</sub>-transformed protein intensities were used for further analysis. Coefficients of variations were calculated for raw protein intensities between replicates individually. Phosphopeptides that were identified in the decoy reverse database were not considered for data analysis. Both data sets were filtered to make sure that identified proteins and phosphopeptides showed expression in all biological triplicates of at least one differentiation stage and the missing values were subsequently replaced by random numbers that were drawn from a normal distribution (width=0.3 and down shift=1.8). PCA analysis of differentiation stages and biological replicates was performed as previously described in [99]. Multi-sample test (ANOVA) for determining if any of the means of differentiation stages were significantly different from each other was applied to protein data set. For truncation, we used permutation-based FDR which was set to 0.05 in conjunction with an S0-parameter of 0.1. For hierarchical clustering of significant proteins, median protein abundances of biological replicates were z-scored and clustered using Euclidean as a distance measure for row clustering. Gene ontology (GO) enrichments in the clusters were calculated by Fischer's exact test using Benjamini-Hochberg false discovery rate for truncation, setting a value of 0.02 as threshold. Mean log<sub>2</sub> ratios of biological triplicates and the corresponding p-values were visualized with volcano plots. We chose a significance cut-off based on a FDR<0.05 or 0.01 in volcano plots.

#### Copy number calculation

Intensities were converted to copy number estimations using the proteomic ruler [48]. The proteomic ruler plug-in v.0.1.6 was downloaded from the Perseus plugin store, for use with Perseus version 1.5.5.0. Protein intensities were filtered for 100% data completeness in at least one stage. Protein groups were annotated with amino acid sequence and tryptic peptide information for the leading protein ID, using the .FASTA file used for processing data. Copy numbers were estimated using the following settings; averaging mode – ‘All columns separately’, molecular masses - ‘average molecular mass’, scaling mode – ‘Histone proteomic ruler’, ploidy ‘2’, total cellular protein concentration – ‘200 g/l’.

#### *In silico* deconvolution approach

All marker sets (except of ‘any20’) were filtered for an ANOVA q-value < 0.01. For the cluster markers, the top three most significant (smallest ANOVA q-value) proteins for each cluster were selected. The combined markers contain all proteins from the sorting markers, known markers, cluster markers and SLC markers. The ‘any20’ markers were selected by randomly picking 20 proteins from the unfiltered protein list, excluding proteins included in any of the other marker lists.

A signature matrix was generated for each of the six marker sets. Only two out of the four cell type replicates, replicate 2 and 4, were used for generating the signature matrices containing averaged, non-logged intensity values of each marker protein. The other two replicates (replicates 1 and 3) were subsequently used for creating *in silico* mixture populations.

The ratios for mixing the different cell types were determined by randomly picking 500 combinations of 5 values (corresponding to the 5 cell types) that add up to 1:

$$(1) \sum_{i=1}^5 a_i = 1$$

The mixture intensity  $I_{\text{mix}}^k$  of each protein k was then determined by summing up the intensity of protein k in cell type i,  $I_i^k$ , multiplied by the fraction of cells from cell type  $a_i$ :

$$(2) I_{\text{mix}}^k = \sum_{i=1}^5 I_i^k a_i$$

Next, using the mixture intensities  $\underline{i}_{\text{mix}}$  of each marker protein  $k$  (Eq. 2) as well as the signature matrix, we set out to estimate the fractions of cells contributed by each of the five cell types  $\underline{a}_i$ . The closer the estimations  $\underline{a}_i$  are to the true mixing ratios  $\underline{a}_i$ , the better the marker set is for deconvoluting and differentiating the five cell types.

Writing equation 2 in matrix form when several proteins are evaluated at the same time, the model expands to the following:

$$(3) \quad \underline{i}_{\text{mix}} = \underline{I}\underline{a}$$

Here,  $\underline{i}_{\text{mix}}$  is a vector of mixture intensities for each evaluated marker protein  $k$ ,  $\underline{I}$  is the signature matrix containing the intensities of each evaluated marker protein  $k$  (rows) in each cell type  $i$  (columns), and  $\underline{a}$  is the vector with the fractions of cells in each cell type  $i$  that we aim to estimate.

To estimate  $\underline{a}$ , we can solve the linear equation system (Eq. 3) by using a minimum least squares optimization (python `scipy.optimize.minimize`). Boundary conditions set the minimum possible values of  $\underline{a}$  to zero in order to avoid negative values.

After an initial estimation of  $\underline{a}$ , only the top 90<sup>th</sup>

percentile of marker proteins for which the estimates fit best are kept for solving the linear equation system (Eq. 3) in a second iteration. The estimated vector of fractions,  $\underline{a}$ , is finally normalized by the Manhattan (L1) norm.

To evaluate the results of the deconvolution step, we implemented a weighted error metric to estimate how close the estimated ratios  $\underline{a}$  are to the true ratios  $\underline{a}$ . Here, mistakes in assigning neighboring cell types (e.g. between P3 and P4) contribute less to the overall error than mistakes between far distant cell types (e.g. Progenitors and P5). In addition to evaluating the six sets of protein markers, three controls were generated: 'random' uses a random ratio estimation; 'uniform' assumes a uniform ratio estimation ( $\underline{a}_i = 0.2$ ); and 'center' assumes that all cells are of type P3 ( $\underline{a}_3 = 1$ ).

#### Data availability

All MS proteomics data have been deposited on ProteomeXchange via the PRIDE database with the dataset identifier PXD017276. All other data supporting findings of this study are available within this article and in supplementary tables.

## REFERENCES

1. Palis, J., *Primitive and definitive erythropoiesis in mammals*. Front Physiol, 2014. 5: p. 3.
2. Moras, M., S.D. Lefevre, and M.A. Ostumi, *From Erythroblasts to Mature Red Blood Cells: Organelle Clearance in Mammals*. Front Physiol, 2017. 8: p. 1076.
3. Nguyen, A.T., et al., *UBE2O remodels the proteome during terminal erythroid differentiation*. Science, 2017. 357(6350).
4. Zhao, B., J. Yang, and P. Ji, *Chromatin condensation during terminal erythropoiesis*. Nucleus, 2016. 7(5): p. 425-429.
5. Seo, Y., et al., *Current Advances in Red Blood Cell Generation Using Stem Cells from Diverse Sources*. Stem Cells Int, 2019. 2019: p. 9281329.
6. Broudy, V.C., *Stem cell factor and hematopoiesis*. Blood, 1997. 90(4): p. 1345-64.
7. Ingle, E., *Integrating novel signaling pathways involved in erythropoiesis*. IUBMB Life, 2012. 64(5): p. 402-10.
8. Nocka, K., et al., *Expression of c-kit gene products in known cellular targets of W mutations in normal and W mutant mice--evidence for an impaired c-kit kinase in mutant mice*. Genes Dev, 1989. 3(6): p. 816-26.
9. Wu, H., et al., *Functional interaction of erythropoietin and stem cell factor receptors is essential for erythroid colony formation*. Proc Natl Acad Sci U S A, 1997. 94(5): p. 1806-10.
10. Wu, H., et al., *Interaction of the erythropoietin and stem-cell-factor receptors*. Nature, 1995. 377(6546): p. 242-6.
11. Zhang, C.C. and H.F. Lodish, *Cytokines regulating hematopoietic stem cell function*. Curr Opin Hematol, 2008. 15(4): p. 307-11.
12. Klingmuller, U., *The role of tyrosine phosphorylation in proliferation and maturation of erythroid progenitor cells--signals emanating from the erythropoietin receptor*. Eur J Biochem, 1997. 249(3): p. 637-47.
13. Wojchowski, D.M., et al., *Signal transduction in the erythropoietin receptor system*. Exp Cell Res, 1999. 253(1): p. 143-56.

14. Bouscary, D., et al., *Critical role for PI 3-kinase in the control of erythropoietin-induced erythroid progenitor proliferation*. Blood, 2003. 101(9): p. 3436-43.
15. Carroll, M.P., et al., *Erythropoietin induces Raf-1 activation and Raf-1 is required for erythropoietin-mediated proliferation*. J Biol Chem, 1991. 266(23): p. 14964-9.
16. Ghaffari, S., et al., *AKT induces erythroid-cell maturation of JAK2-deficient fetal liver progenitor cells and is required for Epo regulation of erythroid-cell differentiation*. Blood, 2006. 107(5): p. 1888-91.
17. Linnekin, D., et al., *Stem cell factor, the JAK-STAT pathway and signal transduction*. Leuk Lymphoma, 1997. 27(5-6): p. 439-44.
18. Miura, Y., et al., *Activation of the mitogen-activated protein kinase pathway by the erythropoietin receptor*. J Biol Chem, 1994. 269(47): p. 29962-9.
19. Socolovsky, M., et al., *Fetal anemia and apoptosis of red cell progenitors in Stat5a-/5b-/- mice: a direct role for Stat5 in Bcl-X(L) induction*. Cell, 1999. 98(2): p. 181-91.
20. Wandzioch, E., et al., *Activation of the MAP kinase pathway by c-Kit is PI-3 kinase dependent in hematopoietic progenitor/stem cell lines*. Blood, 2004. 104(1): p. 51-7.
21. Akashi, K., et al., *Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis*. Blood, 2003. 101(2): p. 383-9.
22. Cantor, A.B. and S.H. Orkin, *Transcriptional regulation of erythropoiesis: an affair involving multiple partners*. Oncogene, 2002. 21(21): p. 3368-76.
23. Cross, M.A. and T. Enver, *The lineage commitment of haemopoietic progenitor cells*. Curr Opin Genet Dev, 1997. 7(5): p. 609-13.
24. Hattangadi, S.M., et al., *From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications*. Blood, 2011. 118(24): p. 6258-68.
25. Perkins, A.C., A.H. Sharpe, and S.H. Orkin, *Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF*. Nature, 1995. 375(6529): p. 318-22.
26. Pevny, L., et al., *Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1*. Nature, 1991. 349(6306): p. 257-60.
27. Shvidasani, R.A., E.L. Mayer, and S.H. Orkin, *Absence of blood formation in mice lacking the T-cell leukaemia oncogene tal-1/SCL*. Nature, 1995. 373(6513): p. 432-4.
28. Tsang, A.P., et al., *FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation*. Cell, 1997. 90(1): p. 109-19.
29. Tusi, B.K., et al., *Population snapshots predict early haematopoietic and erythroid hierarchies*. Nature, 2018. 555(7694): p. 54-60.
30. An, X., et al., *Human and murine erythropoiesis*. Curr Opin Hematol, 2015. 22(3): p. 206-11.
31. Amon, S., et al., *Sensitive Quantitative Proteomics of Human Hematopoietic Stem and Progenitor Cells by Data-independent Acquisition Mass Spectrometry*. Mol Cell Proteomics, 2019. 18(7): p. 1454-1467.
32. Bell, A.J., et al., *Protein distribution during human erythroblast enucleation in vitro*. PLoS One, 2013. 8(4): p. e60300.
33. Dhekne, H.S., et al., *A pathway for Parkinson's Disease LRRK2 kinase to block primary cilia and Sonic hedgehog signaling in the brain*. 2018. 7: p. e40202.
34. Gautier, E.F., et al., *Absolute proteome quantification of highly purified populations of circulating reticulocytes and mature erythrocytes*. Blood Adv, 2018. 2(20): p. 2646-2657.
35. Gillespie, M.A., et al., *Absolute Quantification of Transcription Factors Reveals Principles of Gene Regulation in Erythropoiesis*. Mol Cell, 2020.
36. Liu, X., et al., *Regulation of mitochondrial biogenesis in erythropoiesis by mTORC1-mediated protein translation*. Nat Cell Biol, 2017. 19(6): p. 626-638.
37. Pasini, E.M., et al., *In-depth analysis of the membrane and cytosolic proteome of red blood cells*. Blood, 2006. 108(3): p. 791-801.
38. Roux-Dalvai, F., et al., *Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry*. Mol Cell Proteomics, 2008. 7(11): p. 2254-69.
39. Wilson, M.C., et al., *Comparison of the Proteome of Adult and Cord Erythroid Cells, and Changes in the Proteome Following Reticulocyte Maturation*. Mol Cell Proteomics, 2016. 15(6): p. 1938-46.
40. Gautier, E.F., et al., *Comprehensive Proteomic Analysis of Human Erythropoiesis*. Cell Rep, 2016. 16(5): p. 1470-1484.
41. Aebersold, R. and M. Mann, *Mass-spectrometric exploration of proteome structure and function*. Nature, 2016. 537(7620): p. 347-55.
42. Bekker-Jensen, D.B., et al., *An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes*. Cell Syst, 2017. 4(6): p. 587-599 e4.
43. Humphrey, S.J., S.B. Azimifar, and M. Mann, *High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics*. Nat Biotechnol, 2015. 33(9): p. 990-5.

44. Humphrey, S.J., et al., *High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform*. Nat Protoc, 2018. 13(9): p. 1897-1916.
45. Hu, J., et al., *Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo*. Blood, 2013. 121(16): p. 3246-53.
46. Li, J., et al., *Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E*. Blood, 2014. 124(24): p. 3636-45.
47. Yan, H., et al., *Distinct roles for TET family proteins in regulating human erythropoiesis*. Blood, 2017. 129(14): p. 2002-2012.
48. Wisniewski, J.R., et al., *A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards*. Mol Cell Proteomics, 2014. 13(12): p. 3497-506.
49. Zhao, B., et al., *Nuclear Condensation during Mouse Erythropoiesis Requires Caspase-3-Mediated Nuclear Opening*. Dev Cell, 2016. 36(5): p. 498-510.
50. Liljeholm, M., et al., *Congenital dyserythropoietic anemia type III (CDA III) is caused by a mutation in kinesin family member, KIF23*. Blood, 2013. 121(23): p. 4791-9.
51. Chen, K., et al., *Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis*. Proc Natl Acad Sci U S A, 2009. 106(41): p. 17413-8.
52. Cesar-Razquin, A., et al., *A Call for Systematic Research on Solute Carriers*. Cell, 2015. 162(3): p. 478-87.
53. Kopito, R.R. and H.F. Lodish, *Primary structure and transmembrane orientation of the murine anion exchange protein*. Nature, 1985. 316(6025): p. 234-8.
54. Shaw, G.C., et al., *Mitoferrin is essential for erythroid iron assimilation*. Nature, 2006. 440(7080): p. 96-100.
55. Tanimura, N., et al., *GATA/Heme Multi-omics Reveals a Trace Metal-Dependent Cellular Differentiation Mechanism*. Dev Cell, 2018. 46(5): p. 581-594 e4.
56. Ryu, M.S., et al., *Zinc transporters ZnT1 (Slc30a1), Zip8 (Slc39a8), and Zip10 (Slc39a10) in mouse red blood cells are differentially regulated during erythroid development and by dietary zinc deficiency*. J Nutr, 2008. 138(11): p. 2076-83.
57. Asano, H., X.S. Li, and G. Stamatoyannopoulos, *FKLF-2: a novel Kruppel-like transcriptional factor that activates globin and other erythroid lineage genes*. Blood, 2000. 95(11): p. 3578-84.
58. Gordon, A.R., et al., *Splenomegaly and modified erythropoiesis in KLF13-/- mice*. J Biol Chem, 2008. 283(18): p. 11897-904.
59. Jassal, B., *Pathway annotation and analysis with Reactome: the solute carrier class of membrane transporters*. Hum Genomics, 2011. 5(4): p. 310-5.
60. Jassal, B., et al., *The reactome pathway knowledgebase*. Nucleic Acids Res, 2020. 48(D1): p. D498-D503.
61. Horbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse*. Nucleic Acids Res, 2012. 40(Database issue): p. D261-70.
62. Taylor, P.M., *Amino acid transporters: eminences grises of nutrient signalling mechanisms?* Biochem Soc Trans, 2009. 37(Pt 1): p. 237-41.
63. Brunati, A.M., et al., *Sequential phosphorylation of protein band 3 by Syk and Lyn tyrosine kinases in intact human erythrocytes: identification of primary and secondary phosphorylation sites*. Blood, 2000. 96(4): p. 1550-7.
64. Yannoukakos, D., et al., *Phosphorylation sites in human erythrocyte band 3 protein*. Biochim Biophys Acta, 1991. 1061(2): p. 253-66.
65. Ochoa, D., et al., *An atlas of human kinase regulation*. Mol Syst Biol, 2016. 12(12): p. 888.
66. Sakamoto, H., T. Kitamura, and A. Yoshimura, *Mitogen-activated protein kinase plays an essential role in the erythropoietin-dependent proliferation of CTLL-2 cells*. J Biol Chem, 2000. 275(46): p. 35857-62.
67. Geest, C.R. and P.J. Coffer, *MAPK signaling pathways in the regulation of hematopoiesis*. J Leukoc Biol, 2009. 86(2): p. 237-50.
68. Kurita, R., et al., *Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells*. PLoS One, 2013. 8(3): p. e59890.
69. Grevet, J.D., et al., *Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells*. Science, 2018. 361(6399): p. 285-290.
70. Tarumoto, Y., et al., *LKB1, Salt-Inducible Kinases, and MEF2C Are Linked Dependencies in Acute Myeloid Leukemia*. Mol Cell, 2018. 69(6): p. 1017-1027 e6.
71. Munugalavadla, V. and R. Kapur, *Role of c-Kit and erythropoietin receptor in erythropoiesis*. Crit Rev Oncol Hematol, 2005. 54(1): p. 63-75.
72. Neubauer, H., et al., *Jak2 deficiency defines an essential developmental checkpoint in definitive hematopoiesis*. Cell, 1998. 93(3): p. 397-409.

73. Parganas, E., et al., *Jak2 is essential for signaling through a variety of cytokine receptors*. Cell, 1998. 93(3): p. 385-95.

74. de Los Heros, P., et al., *The WNK-regulated SPAK/OSR1 kinases directly phosphorylate and inhibit the K<sup>+</sup>-Cl<sup>-</sup> co-transporters*. Biochem J, 2014. 458(3): p. 559-73.

75. Tilbrook, P.A., et al., *Erythropoietin-stimulated Raf-1 tyrosine phosphorylation is associated with the tyrosine kinase Lyn in J2E erythroleukemic cells*. Arch Biochem Biophys, 2001. 396(1): p. 128-32.

76. Wang, J., et al., *Pim1 kinase is required to maintain tumorigenicity in MYC-expressing prostate cancer cells*. Oncogene, 2012. 31(14): p. 1794-803.

77. Li, W., et al., *Tgfb2 disruption in postnatal smooth muscle impairs aortic wall homeostasis*. J Clin Invest, 2014. 124(2): p. 755-67.

78. Michaud, N.R., et al., *14-3-3 is not essential for Raf-1 function: identification of Raf-1 proteins that are biologically activated in a 14-3-3- and Ras-independent manner*. Mol Cell Biol, 1995. 15(6): p. 3390-7.

79. Gupta, P. and R. Prywes, *ATF1 phosphorylation by the ERK MAPK pathway is required for epidermal growth factor-induced c-jun expression*. J Biol Chem, 2002. 277(52): p. 50550-6.

80. Koh, H., et al., *Cloning and characterization of a nuclear S6 kinase, S6 kinase-related kinase (SRK); a novel nuclear target of Akt*. Oncogene, 1999. 18(36): p. 5115-9.

81. Haas, N., et al., *Kit transduced signals counteract erythroid maturation by MAPK-dependent modulation of erythropoietin signaling and apoptosis induction in mouse fetal liver*. Cell Death Differ, 2015. 22(5): p. 790-800.

82. Munugalavadla, V., et al., *Repression of c-kit and its downstream substrates by GATA-1 inhibits cell proliferation during erythroid maturation*. Mol Cell Biol, 2005. 25(15): p. 6747-59.

83. Muta, K., et al., *Stem cell factor retards differentiation of normal human erythroid progenitor cells while stimulating proliferation*. Blood, 1995. 86(2): p. 572-80.

84. Lennartsson, J. and L. Ronnstrand, *Stem cell factor receptor/c-Kit: from basic science to clinical implications*. Physiol Rev, 2012. 92(4): p. 1619-49.

85. Matsuzaki, T., et al., *Induction of erythroid differentiation by inhibition of Ras/ERK pathway in a friend murine leukemia cell line*. Oncogene, 2000. 19(12): p. 1500-8.

86. Bernstein, A., et al., *The murine W/c-kit and Steel loci and the control of hematopoiesis*. Semin Hematol, 1991. 28(2): p. 138-42.

87. Kondoh, K. and E. Nishida, *Regulation of MAP kinases by MAP kinase phosphatases*. Biochim Biophys Acta, 2007. 1773(8): p. 1227-37.

88. Li, X., et al., *Elucidating human phosphatase-substrate networks*. Sci Signal, 2013. 6(275): p. rs10.

89. Noomuna, P., et al., *Inhibition of Band 3 tyrosine phosphorylation: a new mechanism for treatment of sickle cell disease*. Br J Haematol, 2020.

90. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. 298(5600): p. 1912-34.

91. Hsu, P.D., et al., *DNA targeting specificity of RNA-guided Cas9 nucleases*. Nat Biotechnol, 2013. 31(9): p. 827-32.

92. Li, W., et al., *MAGECK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens*. Genome Biol, 2014. 15(12): p. 554.

93. Wang, B., et al., *Integrative analysis of pooled CRISPR genetic screens using MAGECKFlute*. Nat Protoc, 2019. 14(3): p. 756-780.

94. Kulak, N.A., et al., *Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells*. Nat Methods, 2014. 11(3): p. 319-24.

95. Kulak, N.A., P.E. Geyer, and M. Mann, *Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics*. Mol Cell Proteomics, 2017. 16(4): p. 694-705.

96. Cox, J. and M.J.N.b. Mann, *MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification*. 2008. 26(12): p. 1367.

97. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment*. 2011. 10(4): p. 1794-1805.

98. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data*. Nat Methods, 2016. 13(9): p. 731-40.

99. Deeb, S.J., et al., *Machine Learning-based Classification of Diffuse Large B-cell Lymphoma Patients by Their Protein Expression Profiles*. Mol Cell Proteomics, 2015. 14(11): p. 2947-60.

### 3.1.3. Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis

Martin Steger<sup>1†</sup>, Federico Diez<sup>2†</sup>, Herschel S Dhekne<sup>3</sup>, Paweł Lis<sup>2</sup>, Raja S Nirujogi<sup>2</sup>, **Ozge Karayel<sup>1</sup>**, Francesca Tonelli<sup>2</sup>, Terina N Martinez<sup>4</sup>, Esben Lorentzen<sup>5</sup>, Suzanne R Pfeffer<sup>3</sup>, Dario R Alessi<sup>2\*</sup>, Matthias Mann<sup>1\*</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Martinsried, Germany <sup>2</sup> Medical Research Council Protein Phosphorylation and Ubiquitylation Unit, School of Life Sciences, University of Dundee, Dundee, United Kingdom <sup>3</sup> Department of Biochemistry, Stanford University School of Medicine, Stanford, United States <sup>4</sup> The Michael J. Fox Foundation for Parkinson's Research, New York, United States <sup>5</sup> Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark \*Corresponding authors

Published in *Elife* (2017)

Using MS-based proteomics, our group, together with Dario Alessi's group from the University of Dundee, discovered three Rab GTPases as the first true *in vivo* substrate of the Parkinson's disease (PD) kinase LRRK2. This study was published in *Elife* in 2016 [105]. To follow up on our findings, we systematically analyzed which of the Rab protein family members are targeted by LRRK2 by proteomics and other techniques. In a HEK293 cell overexpression system, we showed the phosphorylation of the equivalent residues to Thr72 and Thr73 on Rab8A and Rab10, respectively, in isoforms of Rab3 (3A, 3B, 3C & 3D), Rab5 (5A, 5B, 5C), Rab8B, Rab7L1 (both on Thr72 and Ser72) as well as Rab35 and Rab43 by LRRK2. Our results establish that LRRK2 phosphorylates at least 14 Rab proteins.

Next, we aimed to investigate which of these 14 Rab proteins are phosphorylated by LRRK2 endogenously. To this end, we developed a high sensitivity MS method which uses isotope-labeled (SIL) peptides equivalent to the endogenous Rab phosphopeptides of interest to guide targeted quantification. We recorded both light and SIL counterpart phosphopeptides in a multiplexed selected ion monitoring (mxSIM) scan mode on a quadrupole Orbitrap analyzer. This method allowed very sensitive quantification of low abundant endogenous Rab phosphorylation and showed that six out of 14 Rab proteins clearly were kinase substrates in an endogenous context. These findings were well accepted in the community and have now been independently verified by several groups.



# Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis

Martin Steger<sup>1†</sup>, Federico Diez<sup>2†</sup>, Herschel S Dhekne<sup>3</sup>, Paweł Lis<sup>2</sup>, Raja S Nirujogi<sup>2</sup>, Ozge Karayel<sup>1</sup>, Francesca Tonelli<sup>2</sup>, Terina N Martinez<sup>4</sup>, Esben Lorentzen<sup>5</sup>, Suzanne R Pfeffer<sup>3</sup>, Dario R Alessi<sup>2\*</sup>, Matthias Mann<sup>1\*</sup>

<sup>1</sup>Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Martinsried, Germany; <sup>2</sup>Medical Research Council Protein Phosphorylation and Ubiquitylation Unit, School of Life Sciences, University of Dundee, Dundee, United Kingdom; <sup>3</sup>Department of Biochemistry, Stanford University School of Medicine, Stanford, United States; <sup>4</sup>The Michael J. Fox Foundation for Parkinson's Research, New York, United States; <sup>5</sup>Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

**Abstract** We previously reported that Parkinson's disease (PD) kinase LRRK2 phosphorylates a subset of Rab GTPases on a conserved residue in their switch-II domains (Steger et al., 2016) (PMID: 26824392). Here, we systematically analyzed the Rab protein family and found 14 of them (Rab3A/B/C/D, Rab5A/B/C, Rab8A/B, Rab10, Rab12, Rab29, Rab35 and Rab43) to be specifically phosphorylated by LRRK2, with evidence for endogenous phosphorylation for ten of them (Rab3A/B/C/D, Rab8A/B, Rab10, Rab12, Rab35 and Rab43). Affinity enrichment mass spectrometry revealed that the primary ciliogenesis regulator, RILPL1 specifically interacts with the LRRK2-phosphorylated forms of Rab8A and Rab10, whereas RILPL2 binds to phosphorylated Rab8A, Rab10, and Rab12. Induction of primary cilia formation by serum starvation led to a two-fold reduction in ciliogenesis in fibroblasts derived from pathogenic LRRK2-R1441G knock-in mice. These results implicate LRRK2 in primary ciliogenesis and suggest that Rab-mediated protein transport and/or signaling defects at cilia may contribute to LRRK2-dependent pathologies.

DOI: <https://doi.org/10.7554/eLife.31012.001>

\*For correspondence:  
d.r.alessi@dundee.ac.uk (DRA);  
mmann@biochem.mpg.de (MM)

<sup>†</sup>These authors contributed  
equally to this work

**Competing interest:** See  
page 18

**Funding:** See page 18

**Received:** 04 August 2017

**Accepted:** 09 November 2017

**Published:** 10 November 2017

**Reviewing editor:** Ivan Dikic,  
Goethe University Frankfurt,  
Germany

© Copyright Steger et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

## Introduction

LRRK2 encodes a large protein kinase and a number of LRRK2 mutations cause autosomal dominant Parkinson's disease (PD) (Bardien et al., 2011; Funayama et al., 2005; Paisán-Ruiz et al., 2004; Zimprich et al., 2004). Furthermore, genome-wide association studies (GWAS) have pinpointed LRRK2 as a risk factor for idiopathic PD, indicating that it is a master regulator of the molecular pathways controlling both hereditary and sporadic forms of PD (Nalls et al., 2014). LRRK2 is expressed at low levels in brain neurons and at higher levels in lung, kidney, pancreas and immune cells, therefore multiple cell types in different tissues might act in concert during PD pathogenesis (Giesert et al., 2013; Thévenet et al., 2011). The most frequent LRRK2 mutations that segregate with familial PD (R1441C, R1441G, Y1699C, G2019S and I2020T) all map to its catalytic domains, namely the ROC-COR (GTPase activity) and kinase domains (Cookson, 2010). The kinase domain mutations (G2019S and I2020T) increase kinase activity both in vitro and in vivo, and interestingly, R1441C, R1441G and Y1699C enhance this activity only in vivo (Sheng et al., 2012; Steger et al., 2016). This indicates that (i) the kinase and the GTPase domains communicate with each other and

(ii) LRRK2-dependent pathogenicity is triggered by an increase in kinase activity, regardless of the mutation position.

Multiple lines of evidence suggest that LRRK2 regulates intracellular vesicular traffic and that mutant LRRK2-associated trafficking defects contribute to PD pathogenesis (Cookson, 2016). First, LRRK2 translocates to membranes following Toll-like receptor stimulation of immune cells (Schapansky et al., 2014). Second, LRRK2 is reported to regulate the autophagosome/lysosome system, which is particularly interesting as defects in proteins coordinating this pathway are found in multiple forms of PD (Beilina and Cookson, 2016; Giaime et al., 2017; Tang, 2017). Third, LRRK2 phosphorylates at least three Rab GTPases (Rab8A, Rab10 and Rab12) in cultured human and murine cells as well as in mice, and this reduces their binding affinity for different Rab regulatory proteins (Ito et al., 2016; Steger et al., 2016; Thirstrup et al., 2017).

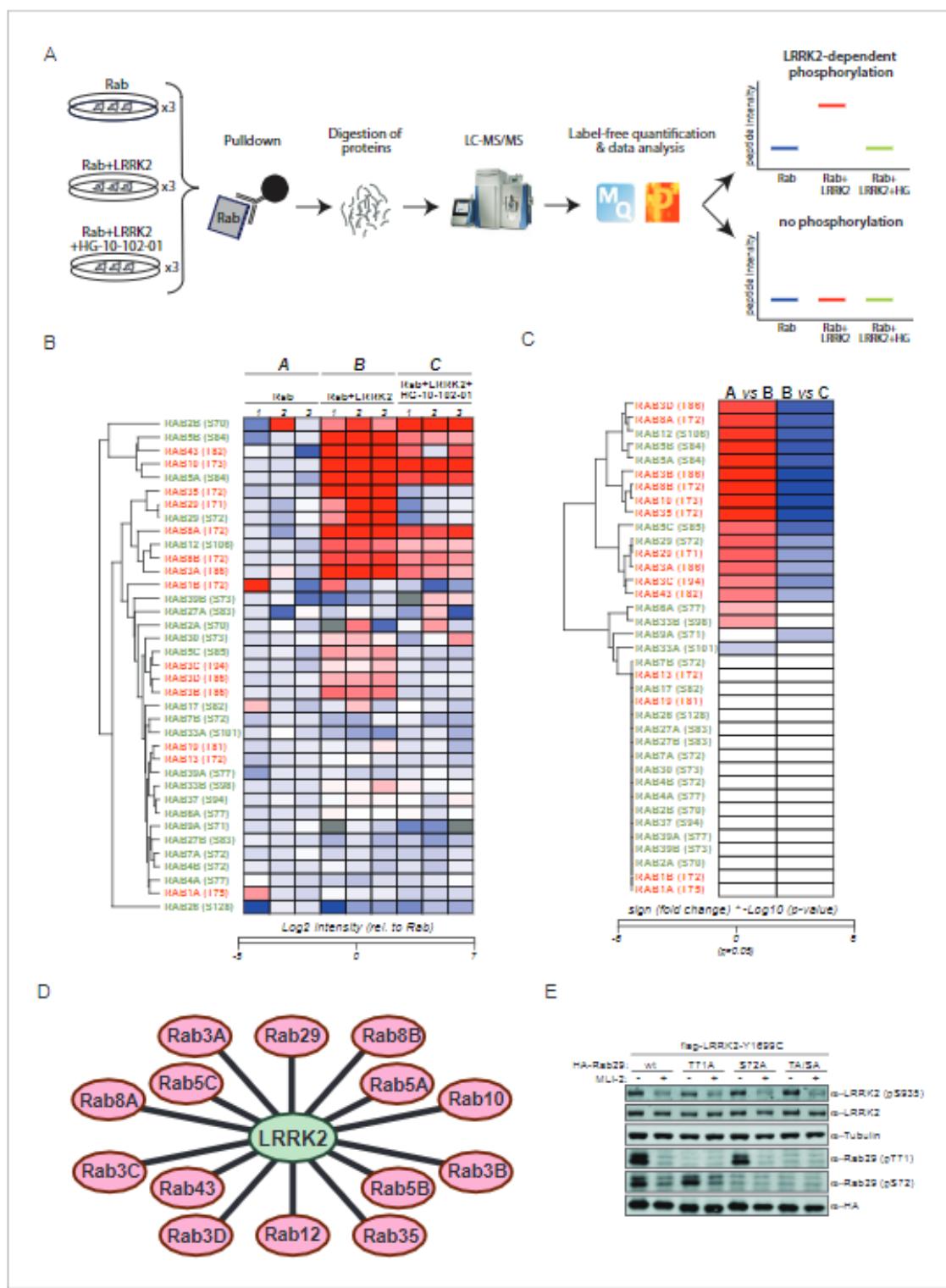
Rab GTPases are master regulators of intracellular vesicle trafficking and about 70 family members are expressed in humans (Wandinger-Ness and Zerial, 2014). In addition to the above-mentioned three Rabs that can be phosphorylated in a LRRK2-dependent manner, Rab29 (also known as Rab7L1) and Rab39B have been implicated in PD. Missense mutations in Rab39B lead to intellectual disability and PD-like clinical features (Wilson et al., 2014). Rab29 localizes to the trans-Golgi network (TGN) and Rab29 knockout mice mimic a phenotype of LRRK2 knockouts – an age-associated lysosomal defect in the kidney (Baptista et al., 2013; International Parkinson's Disease Genomics Consortium et al., 2014; Kuwahara et al., 2016). Our previous work showed that LRRK2 phosphorylates Rab8A, Rab10 and Rab12 at a conserved site in the switch II domain (Rab8A-T72, Rab10-T73 and Rab12-S106), which is conserved in 50 Rabs (Steger et al., 2016). We also found that LRRK2 specifically phosphorylates overexpressed Rab3A, but not Rab7A, despite their high-sequence homology adjacent to the site of phosphorylation. A prediction of whether or not a given Rab protein can be phosphorylated by LRRK2 in cells is thus not possible, based solely on the primary amino acid sequence. Although this work suggests that Rabs are involved in PD pathogenesis, the molecular mechanisms remain to be determined. Here, we systematically investigate whether Rabs other than 8A, 10 and 12 are LRRK2 targets. Subsequent modification-specific interaction experiments led us to establish an unsuspected effect of mutant LRRK2 on ciliogenesis via phosphorylated Rab GTPases.

## Results and discussion

### 14 overexpressed rab GTPases are phosphorylated by LRRK2

We set out to determine which of the 50 Rab proteins that contain a conserved, predicted LRRK2 target site are indeed phosphorylated by LRRK2. For this purpose, we overexpressed each of the 50 Rabs in triplicate in HEK293 cells, either alone or in combination with LRRK2 harboring the activating G2019S mutation (Figure 1A and Figure 1—figure supplement 1). To determine phosphorylation specificity, we included a third condition in which both kinase and substrate were expressed in the presence of the selective LRRK2 inhibitor, HG-10-102-01 (Choi et al., 2012). Although Rab32 and Rab38 do not contain the conserved phosphorylation site (Figure 1—figure supplement 2A), we included them because they have previously been reported to interact with LRRK2 (Waschbüsch et al., 2014).

Expression of all 52 constructs and inhibitor efficacy was demonstrated by immunoblotting (Figure 1—figure supplement 2B). To identify LRRK2 target candidates, we enriched the individual epitope-tagged Rab family members by immunoprecipitation before digestion, peptide purification and high-resolution, label-free LC-MS/MS analysis (Figure 1A). To generate peptides containing the phosphorylation site with an optimal length for MS analysis, we selected the most appropriate protease among trypsin, chymotrypsin, Asp-N and Glu-C (Supplementary file 1). This generated 468 LC-MS/MS files (52 Rabs under three conditions with triplicate analysis), which we stringently processed using the MaxQuant software for peptide identification and quantification (Cox and Mann, 2008). For each Rab, several phosphorylated peptides were generally detected, but only those containing the predicted site showed significant, quantitative differences as a result of LRRK2 overexpression and inhibition. In total, we quantified this site in 37 different Rab proteins. For 12 Rabs, the



**Figure 1.** 14 Rab GTPases are phosphorylated by LRRK2 when overexpressed HEK293s. (A) Workflow for identifying Rabs that are phosphorylated by LRRK2 in an overexpression system. The individual Rab proteins were overexpressed alone or expressed in combination with active (G2019S)- or chemically inhibited LRRK2 (HG-10-102-01, 3  $\mu$ M, 3 hr,  $n = 3$ ). The Rabs were then affinity-enriched, digested and analyzed by label-free LC/MS-MS. Mass spectra database search and data analysis were performed using MaxQuant and Perseus, respectively. Substrates were scored when the Figure 1 continued on next page

Figure 1 continued

predicted Rab phosphopeptide was detected, upregulated upon LRRK2 expression and reduced after inhibitor treatment. (B) Heat map of phosphopeptide intensities (Log2) and (C) Tukey adjusted p-values (-Log10, multiplied by the sign of the fold change of the respective group comparison [A – Rab, B – Rab + LRRK2, C – Rab + LRRK2 + HG-10-102-10]). Values missing in all replicates of one group were imputed and non-imputed, missing values are in grey. (D) Scheme of the 14 Rab proteins that are phosphorylated by LRRK2. (E) Western blot confirming that both Rab29-pT71 and Rab29-pS72 are phosphorylated by LRRK2-Y1699C when overexpressed in HEK293 cells.

DOI: <https://doi.org/10.7554/eLife.31012.002>

The following figure supplements are available for figure 1:

**Figure supplement 1.** Sequence alignment of 50 Rabs in which the predicted LRRK2 phosphorylation site is conserved.

DOI: <https://doi.org/10.7554/eLife.31012.003>

**Figure supplement 2.** Sequence alignment of Rab10, Rab38 and Rab32 and western blot analysis of 52 overexpressed Rab GTPases.

DOI: <https://doi.org/10.7554/eLife.31012.004>

**Figure supplement 3.** Both Rab29-T71 and Rab29-S72 are phosphorylated by overexpressed LRRK2 in HEK293 cells.

DOI: <https://doi.org/10.7554/eLife.31012.005>

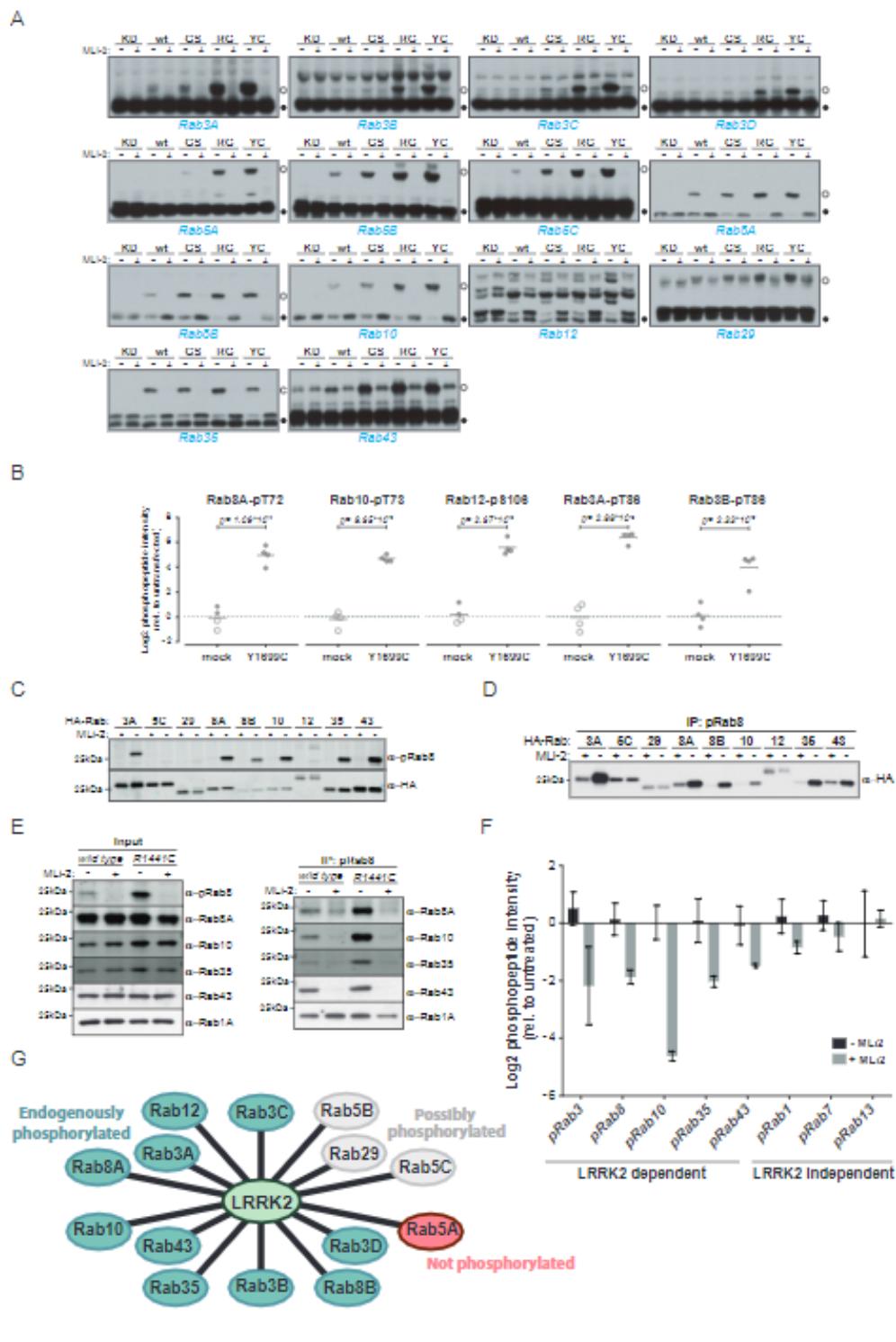
predicted phosphopeptide was not found. Nonetheless, we were able to readily detect and quantify the non-phosphorylated counterpart, suggesting that the modified form should have been detected if it had been present. Thus, we conclude that in these cases, the site is either not phosphorylated at all, or the phosphorylation stoichiometry is below the detection limit. The LC-MS/MS analysis of the remaining Rabs - Rab6C, Rab9B and Rab41 – was inconclusive as we were not able to either identify the phosphorylated- or the non-phosphorylated predicted LRRK2 target peptides (*Supplementary file 1*).

The proteolytic cleavage of some Rabs with very high-sequence homology (e.g. Rab3A/B/C/D) creates peptides with identical primary amino acid sequence, making them indistinguishable to MS and MS/MS. However, as we individually overexpressed and immunoprecipitated Rabs with monoclonal antibodies directed against the epitope tag, it is very unlikely that several such proteins are enriched at equal efficiencies. Our analysis revealed 14 phosphopeptides that were regulated in a LRRK2-dependent manner, in that their abundance was significantly increased upon kinase expression and decreased when cells were treated with HG-10-102-01 (Tukey's multiple comparisons test,  $p<0.05$ ) (Figure 1B–D). This includes all the previously identified LRRK2 substrates: Rab8A, Rab10 and Rab12 (Steger et al., 2016), and adds Rab3A/B/C/D, Rab5A/B/C, Rab8B, Rab29, Rab35 and Rab43.

Rab29 was recently shown to interact with LRRK2, both physically and genetically, and Rab29 knockout in mice phenocopies LRRK2 knockouts, indicating that both proteins might act in the same pathway (International Parkinson's Disease Genomics Consortium et al., 2014; Kuwahara et al., 2016; Simón-Sánchez et al., 2009). Interestingly, we identified a doubly-phosphorylated peptide (pT71 and pS72) that was regulated by LRRK2. To confirm this finding, we individually mutated both residues to non-phosphorylatable alanine and expressed the constructs in HEK293 cells, along with active or chemically inhibited LRRK2. We quantified the levels of the peptides with either alanine substitution by MS and found that the regulation persisted in each case (Figure 1—figure supplement 3). Finally, we raised two phospho-specific antibodies recognizing either Rab29-pT71 or Rab29-pS72, which further substantiated the evidence that both residues are modified by LRRK2 in this system (Figure 1E). Interestingly, the 14 Rabs that are phosphorylated by LRRK2 are widely dispersed over the Rab phylogenetic tree and further analysis is required to understand the determinants of phosphorylation. It is likely that they co-localize with LRRK2 and this may account for the phosphorylation specificity.

## LRRK2 phosphorylates at least ten endogenous Rab GTPases in cells

Because pathogenic LRRK2 mutations occurring in different functional domains increase the phosphorylation of Rab8A, Rab10 and Rab12 in cells (Ito et al., 2016; Steger et al., 2016), we next investigated if the 14 Rabs regulated in our screen were equally affected by mutations in LRRK2. For this, we used the phos-tag assay, a method complementary to MS, in which phosphorylated Rabs are resolved from their non-phosphorylated counterparts by SDS-PAGE (Ito et al., 2016). As judged by the increased Rab-specific band-shift intensity, all LRRK2 pathogenic mutations analyzed (G2019S, R1441G, Y1699C) led to increased phosphorylation of 13 Rab substrates (Figure 2A). For



**Figure 2.** Ten endogenous Rabs are phosphorylated by LRRK2 in cells. (A) Phos-tag SDS-PAGE of 14 Rabs co-expressed with different LRRK2 variants (KD – kinase dead, wt – wild type, GS – G2019S, RG – R1441G, YC – Y1699C) in HEK293 cells. Filled circles indicate unmodified Rabs and open circles, phosphorylated forms (MLI-2 = 100 nM, 1 hr). (B) MS-quantified Rab3A-pT86, Rab3B-pT86, Rab8A-pT72, Rab10-pT73 and Rab12-pS106 peptide intensities from mock and LRRK2-Y1699C transfected cells. Open circles indicate imputed values. (C) Western blot analysis of HA-tagged Rabs co-expressed with LRRK2 variants. (D) Western blot analysis of HA-tagged Rabs co-expressed with LRRK2 variants. IP: pRab8 antibody. (E) Western blot analysis of HA-tagged Rabs co-expressed with LRRK2 variants. Input and IP: pRab8 lanes are shown. (F) Bar chart showing Log2 phosphopeptide intensity (n.f. to untreated) for various phosphopeptides. Black bars represent MLI-2, grey bars represent ML2. Asterisks indicate significant differences. (G) Schematic diagram of LRRK2-mediated Rab phosphorylation. LRRK2 (green) phosphorylates various Rabs (light blue circles). Rab5A (red circle) is not phosphorylated. Other Rabs are either endogenously phosphorylated (Rab8A, Rab3A, Rab12, Rab10, Rab43, Rab35, Rab3B, Rab3C, Rab29, Rab5B, Rab5C) or possibly phosphorylated (Rab3D, Rab8B). Figure 2 continued on next page

Figure 2 continued

expressed with LRRK2-Y1699C in HEK293 cells (-/+100 nM MLI-2, 2 hr) using polyclonal anti-phospho-Rab8 and HA antibodies. (D) Same as (C) but Rabs were immunoprecipitated using the anti-pRab8 antibody and HA antibody was used for detection. (E) Western blot of endogenous Rabs in wild type and LRRK2-R1441C MEFs (-/+100 nM MLI-2, 1 hr) using the indicated antibodies (left) and immunoprecipitation of Rabs using anti-pRab8 antibody (right). (F) MS-quantified Rab phosphosite intensities of pRab8 immunoprecipitation in R1441C MEFs (-/+100 nM MLI-2, 2 hr). Error bars are mean  $\pm$  SEM (n = 3). (G) 10 Rab proteins at endogenous expression levels are phosphorylated by LRRK2 in cells (green). Rab5A is not (red) and Rab5B, Rab5C and Rab29 are possibly phosphorylated (grey).

DOI: <https://doi.org/10.7554/eLife.31012.006>

The following figure supplements are available for figure 2:

**Figure supplement 1.** Rab5A was immunoprecipitated from mock or LRRK2-Y1699C-transfected HEK293 cells.

DOI: <https://doi.org/10.7554/eLife.31012.007>

**Figure supplement 2.** Nine Rab GTPases containing a Thr site in their switch-II domain are phosphorylated by LRRK2.

DOI: <https://doi.org/10.7554/eLife.31012.008>

Rab12, the assay resulted in multiple bands and no meaningful interpretation of the migration pattern was possible; however, our previous work had already established this protein as a LRRK2 substrate. In this context, we also investigated Rab6C, Rab9B and Rab41 by the phos-tag method, for which the MS analysis was inconclusive, and found no evidence of LRRK2-induced mobility shifts.

We next investigated which of these 14 Rab proteins are phosphorylated by LRRK2 endogenously. For this, we overexpressed the pathogenic LRRK2-Y1699C form in HEK293 cells, immunoprecipitated endogenous Rab proteins using commercially available- (Rab3A, Rab3B, Rab5A, Rab8A, Rab10) or in-house raised antibodies (Rab12) and quantified LRRK2-mediated phosphorylation by LC-MS/MS. As expected from our previous results, Rab8A-pT72, Rab10-pT73 and Rab12-pS106 phosphopeptide levels were strongly upregulated (~16 fold) in cells expressing LRRK2-Y1699C (Figure 2B). Rab3A-pT86 was also clearly regulated by LRRK2 in this system, but we were not able to detect any phosphopeptides mapping to either Rab3B-pT86 or Rab5A-pS84, despite the fact that the corresponding unmodified peptides were sequenced multiple times and our initial overexpression screen had established detectability of these phosphopeptides. Therefore, to further increase mass spectrometric sensitivity, we used stable isotope labeled (SIL) peptides equivalent to the endogenous Rab peptides of interest to guide targeted quantification. In contrast to the absolute quantification (AQUA) strategy (Gerber et al., 2003) that relies on fragment ions for quantification, we recorded both light and heavy counterparts in a multiplexed selected ion monitoring (SIM) scan mode on a quadrupole Orbitrap analyzer and used the heavy form to indicate the retention time (heavy) and the light form for quantification. This combination of the spike-in SIL peptide together with a very sensitive MS method should allow quantification of very low abundant endogenous Rab phosphorylation (Materials and methods). With this increased sensitivity, we found LRRK2-specific modification of Rab3B, but still failed to observe clear LRRK2-mediated phosphorylation of Rab5A-S84 (Figure 2B and Figure 2—figure supplement 1). Thus, Rab5A phosphorylation by LRRK2 in HEK293 cells occurs only when both the Rab and kinase are overexpressed exogenously. It is possible that accessory factors in other cell types are required for efficient LRRK2 phosphorylation of this protein.

For the remaining Rab proteins for which no suitable antibodies for total-Rab protein enrichment were available, we raised a rabbit polyclonal Rab antibody against the LRRK2-phosphorylated Rab8 switch-II motif. This efficiently recognized and immunoprecipitated LRRK2 phosphorylated Rabs containing a Thr site: namely, Rab3, Rab8, Rab10, Rab35 and Rab43 (Figure 2C,D). As an exception, the antibody failed to recognize pRab29 or to immunoprecipitate LRRK2-phosphorylated Rab29 protein. This could be either due to the particular linear sequence surrounding T71 (~1 position compared to the equivalent site on other Rabs) or due to double phosphorylation (T71 and S72) by LRRK2. Next, we treated wild type and LRRK2-R1441C knock-in mouse embryonic fibroblasts (MEFs) with the selective LRRK2 inhibitor MLI-2 (Fell et al., 2015), subjected the cell lysates to immunoprecipitation with the phospho-Rab8 antibody and analyzed the immunoprecipitates by immunoblotting or LC-MS/MS. Immunoblotting of the input cell extracts confirmed that compared to wild type, the signal detected by the newly developed anti-phospho-Rab8 antibody strongly increased in LRRK2-R1441C MEFs, and this was blocked by MLI-2 treatment (Figure 2E). For detecting immunoprecipitates by western blotting, we used CRISPR/CAS9 knock-out-validated antibodies (Figure 2—figure

supplement 2A). This revealed specific enrichment of phospho-Rab8A, -Rab10, -Rab35 and -Rab43, which increased in LRRK2-R1441C MEFs and was completely prevented by MLi-2 treatment (Figure 2E). In contrast, Rab1A phosphorylation was not sensitive to MLi-2 in this system, confirming the data from our initial overexpression screen.

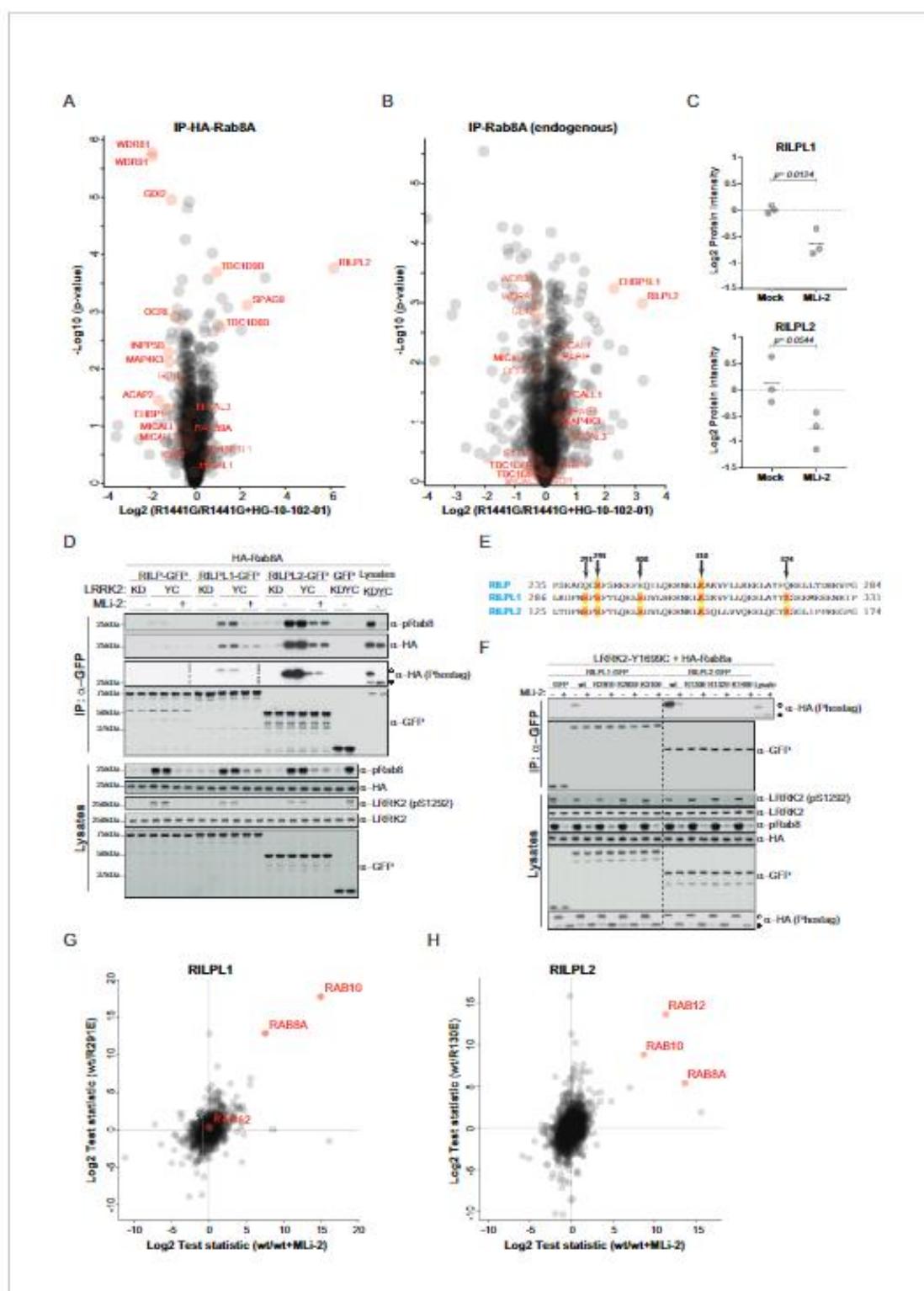
Using the anti-phospho-Rab8 antibody in LRRK2-R1441C fibroblasts, we analyzed immunoprecipitates by LC-MS/MS. Quantification of 32 Rab protein intensities showed that Rab3A, Rab8A, Rab8B, Rab10, Rab35 and Rab43 were enriched more than twofold in a LRRK2-dependent manner. Rab3B and Rab3D were also detected, however, compared to MLi-2 treatment, their enrichment in untreated cells was less pronounced (about 1.5-fold), probably due to lower protein expression levels (Figure 2—figure supplement 2B,C). In concordance with this, quantification of the switch-II phosphorylated peptides of Rab3, Rab8, Rab10, Rab35 and Rab43 in the same experiment revealed a more than two-fold decrease upon MLi-2 treatment (Figure 2F). For Rab1A, Rab7 and Rab13, however, there was no pronounced difference in the corresponding protein and phospho-peptide levels when comparing untreated with MLi-2-treated samples, suggesting that these Rabs are phosphorylated by a kinase other than LRRK2. Finally, we immunoprecipitated phosphorylated Rabs from HEK293 cells expressing LRRK2-Y1699C using the same antibody. Again, we confirmed LRRK2-specific enrichment of nine threonine phosphorylated Rab proteins. These were Rab3A/B/C/D, Rab8A/B, Rab10, Rab35 and Rab43 and the protein sequence coverage ranged from 33% (Rab3B) to 90% (Rab10). Compared to the other family members, this indicates that they are recognized by the antibody in a direct manner (Figure 2—figure supplement 2D).

In summary, our study demonstrates that 14 Rab proteins are phosphorylated by LRRK2 at the switch-II site in an overexpression system (Rab3A/B/C/D, Rab5A/B/C, Rab8A/B, Rab10, Rab12, Rab29, Rab35 and Rab43). For ten of these (Rab3A/B/C/D, Rab8A/B, Rab10, Rab12, Rab35 and Rab43), endogenous LRRK2-dependent phosphorylation is clearly established; for one (Rab5A), we observed phosphorylation in an overexpression but not an endogenous setting. In vivo analysis of the remaining three Rabs (Rab5B, Rab5C and Rab29) was hampered by the unavailability of suitable antibodies for protein enrichment or detection (Figure 2G).

### Phosphorylation-dependent binding of regulatory proteins to Rabs

Different residues in the switch-II domain of Rabs can be modified post-translationally, and this interferes with binding to partner proteins such as Rab GDP dissociation inhibitor alpha/beta (GDI1/2). For example, Rab1B is AMPylated on Y77 (Y80 in Rab1A) by the Legionnaires' disease protein, DrA, resulting in a constitutively active Rab protein that fails to interact with GTPase-activating proteins (GAPs) and GDIs (Müller et al., 2010; Oesterlin et al., 2012). Similarly, the switch-II domain of Rab1A can be phosphocholinated on S79 during *Legionella* infection, which results in strongly decreased GDI binding (Goody et al., 2012; Mukherjee et al., 2011). We showed previously that phosphorylation of Rab8A-T72 and the equivalent sites, Rab10-T73 and Rab12-S106 by LRRK2 decreases GDI binding (Steger et al., 2016). Substitution of Rab7A-T72 with a phosphomimetic glutamic acid has been shown to abrogate GDI interaction (Satpathy et al., 2015). Here, we set out to test systematically, the effect of replacing the predicted LRRK2 target site with a negatively charged, phosphomimetic glutamic acid residue on partner protein binding. For this, we mutated these sites to either non-phosphorylatable alanine or glutamic acid in all 14 Rabs (both T71 and S72 for Rab29) and expressed them in HEK293 cells ( $n = 3$ ). Following affinity-enrichment, we digested bound proteins and analyzed the resulting peptides by label-free LC-MS/MS. As expected, the S/T→E mutations strongly reduced partner binding for all tested Rabs. The S/T→A Rab mutants instead stably bound to GDI1/2 and the Rab escort proteins, CHM and CHML, except for both tested Rab29 constructs (Figure 3—figure supplement 1A,B). In that case, neither the T71A/E nor the S72A/E constructs interacted with GDI, indicating either low binding affinities, an alteration in nucleotide-binding properties, or protein misfolding.

To investigate more specifically the effect of LRRK2-induced Rab phosphorylation on protein interactions, we expressed LRRK2 in HEK293 cells and determined the interactomes of both HA-tagged and endogenous Rab8A by AE-MS, before and after chemical LRRK2 inhibition. To increase phosphosite occupancy at T72, we used the pathogenic ROC-COR domain LRRK2-R1441G mutant, which confers strong intracellular kinase activity (Sheng et al., 2012; Steger et al., 2016). Upon LRRK2 inhibition, Rab8A-pT72 levels decreased about eight fold as shown by quantitative MS, establishing inhibitor efficacy (Figure 3—figure supplement 2). As we had reported previously, GDIs



**Figure 3.** Phosphorylation-specific protein binding to Rabs. (A) AE-MS of HA-Rab8A and (B) endogenous Rab8A using extracts of GFP-LRRK2-R1441G expressing Flp-In T-Rex HEK293 cells. Expression of the kinase was induced for 48 hr by doxycycline (1  $\mu$ g/ml) and LRRK2 inhibited using HG-10-102-01 (3  $\mu$ M, 3 hr). LRRK2-regulated, known and unknown Rab8A-binding proteins in both AE-MS experiments are highlighted in red. (C) Label-free (LFO) [Cox et al., 2014], MS-quantified RILPL1 and RILPL2 levels after immunoprecipitation of Rab8A from mock- or MLi-2 (200 nM, 2 hr) treated LRRK2. Figure 3 continued on next page

Figure 3 continued

R1441C MEFs. (D) Pulldown of GFP-tagged RILP, RILPL1 or RILPL2, transiently expressed with HA-Rab8A and the indicated LRRK2 variants (KD-Y1699C/D2017A) in HEK293 cells. Western blot after Phos-tag SDS-PAGE was used to detect interacting proteins using the indicated antibodies. (MLi-2 = 150 nM, 2 hr). (E) Sequence alignment of the RILP homology (RH) domains of RILP, RILPL1 and RILPL2 showing five conserved basic residues, which are highlighted. (F) Same as (D) but using different RILPL1 and RILPL2 mutants. For phos-tag blots, filled circles indicate non phosphorylated proteins and open circles phosphorylated proteins. (G) AE-MS of GFP-RILPL1 (wt or R291E) and (H) GFP-RILPL2 (wt or R130E), expressed with LRRK2-Y1699C in HEK293 cells, and treated or not with MLi-2 (for wt, 200 nM, 2 hr). The student's two sample test statistic (Log2) of the indicated comparisons was used for plotting.

DOI: <https://doi.org/10.7554/eLife.31012.009>

The following figure supplements are available for figure 3:

**Figure supplement 1.** Phosphomimetic S/T->E mutation of the LRRK2 phosphorylation site abrogates GDI1/2 and CHM/CHML binding in 13 Rab GTPases.

DOI: <https://doi.org/10.7554/eLife.31012.010>

**Figure supplement 2.** MS-quantified pT72-Rab8A peptide levels in HEK293 cells expressing LRRK2-R1441G.

DOI: <https://doi.org/10.7554/eLife.31012.011>

**Figure supplement 3.** Phospho-dependent interaction of Rab8A, Rab10 and Rab12 with RILPL1/2.

DOI: <https://doi.org/10.7554/eLife.31012.012>

associated preferentially with the non-phospho forms of both HA-Rab8A and endogenous Rab8A (Figure 3A,B). However, this effect was more pronounced in the case of overexpressed Rab8A, probably due to the higher fraction of phosphorylated protein.

Several, to-date undescribed interaction partners that showed preferential binding to the non-phospho-forms in both the endogenous and HA-tagged systems (Figure 3A,B). These include the poorly characterized, WD repeat domain (WDR)-containing proteins, WDR81 and WDR91, which have recently been linked to regulation of endosomal phosphatidylinositol 3-phosphate levels in *C. elegans* and mammalian cells (Liu et al., 2016). The inositol-5-phosphatase, OCRL is a known Rab8A effector, and our experiments revealed that phosphorylation abolishes its binding to Rab8A. Mutations in the human OCRL gene cause the oculocerebrorenal syndrome of Lowe and it would be interesting to investigate whether the phosphorylation of Rabs is relevant in this context (Hou et al., 2011).

Strikingly, we found that Rab interacting lysosomal protein like 2 (RILPL2) binds preferentially to the phosphorylated forms of both HA-Rab8A and Rab8A. RILPL2 contains a C-terminal, coiled-coil Rab-binding domain, which is known as RILP homology (RH) domain, and this is conserved in four other members of this protein family (RILP, RILPL1 and JIP3/4) (Matsui et al., 2012; Wu et al., 2005). RILPL1 and RILPL2 are poorly characterized proteins; however, a recent report links both to the regulation of protein localization in primary cilia (Schaub and Stearns, 2013). RILP is much better studied, and interacts with Rab7 to control lysosomal motility in cells (Jordens et al., 2001).

To investigate Rab8A-RILPL2 interaction in an endogenous setting, we immunoprecipitated Rab8A from mock- or MLi-2 treated, knock-in MEFs expressing the hyperactive LRRK2 variant, R1441C. This confirmed our initial LRRK2-specific interaction of Rab8A and RILPL2 and also identified RILPL1 as phospho-dependent Rab8A interactor (Figure 3C).

To further verify the pRab8A interaction with RILPL1 and RILPL2 and to test for RILP interaction, we co-expressed GFP-tagged, full length RILP, RILPL1 and RILPL2 with HA-Rab8A and LRRK2-Y1699C or kinase-dead LRRK2-Y1699C/D2017A. Both RILPL1 and RILPL2, but not RILP, specifically bound to LRRK2-generated phospho-Rab8A, and this interaction was strongly reduced by MLi-2 (Figure 3D). In this side by side comparison, RILPL2-bound phosphorylated Rab8A much stronger than RILPL1, indicating that RILPL2 might be a key effector specific for pRab8A. Phos-tag analysis independently confirmed that only the LRRK2-phosphorylated HA-Rab8A protein was co-immunoprecipitated with either RILPL2 or RILPL1 (Figure 3D).

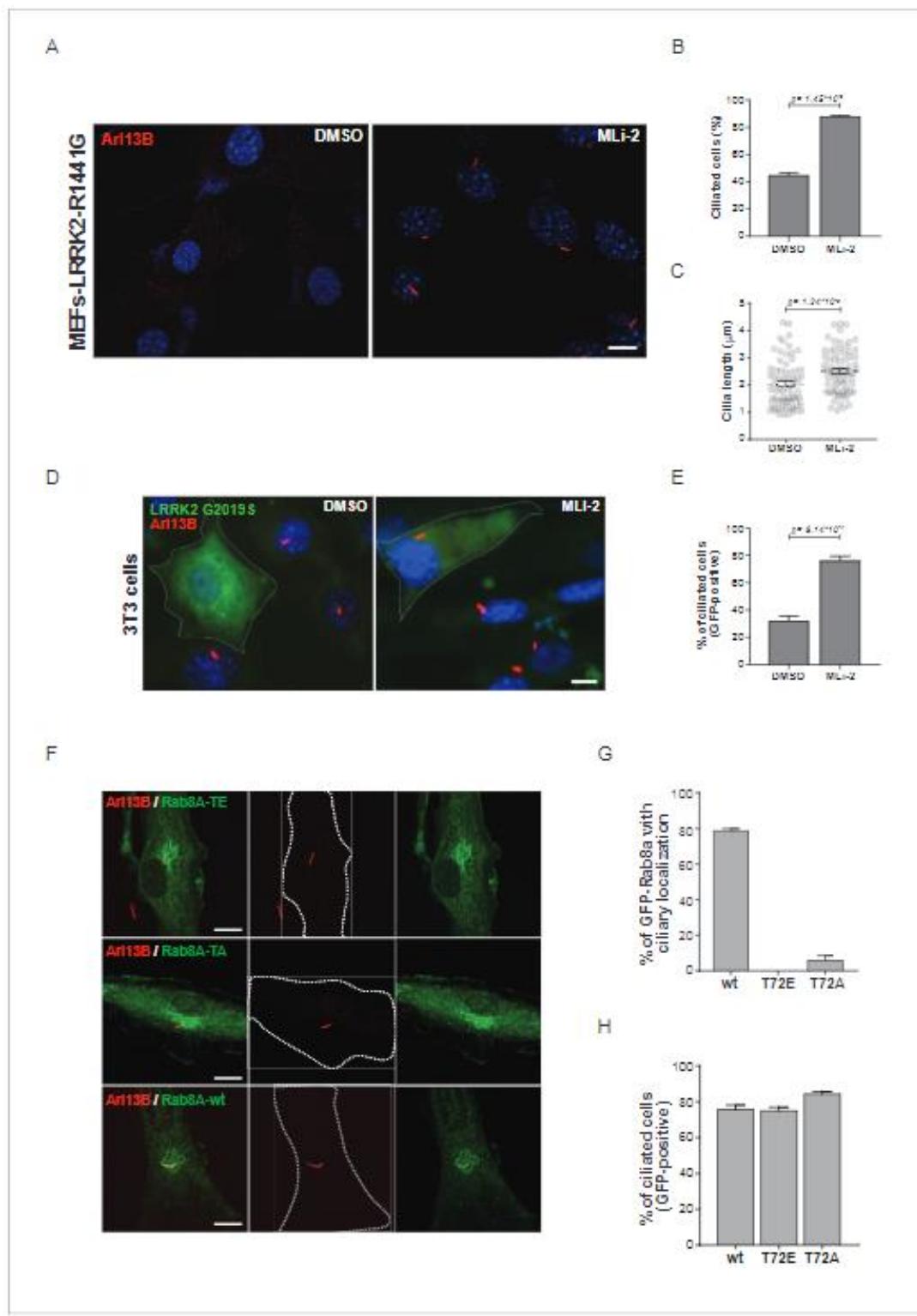
There are five highly conserved basic residues in the RH domain on RILPL1 and RILPL2 that could potentially mediate the interaction with LRRK2-phosphorylated Rab8A (R291, R293, R300, K310 and K324 in RILPL1) (Figure 3E). Individual mutation of R291, R293 or K310, but not R300 or K324, abolished the interaction of N-terminally truncated RILPL1 with LRRK2-phosphorylated Rab8A (Figure 3—figure supplement 3A). Accordingly, the identified mutations abrogated pRab8 interaction with both full-length RILPL1 and RILPL2 (Figure 3F).

To extend our RILPL1/2-Rab interaction analysis reciprocally, we expressed GFP-RILPL1 (wt or R291E) and GFP-RILPL2 (wt or R130E) with LRRK2-Y1699C in HEK293 cells, in the presence or absence of MLi-2, and processed the samples for LC-MS/MS analysis after affinity enrichment. Strikingly, in this setup, RILPL1-wt interacted not only with Rab8A, but also with Rab10, and this was strongly reduced by MLi-2 and the R291E Rab interaction mutant (Figure 3G). RILPL2 instead bound all three of our previously identified bona-fide LRRK2 substrates, Rab8A, Rab10 and Rab12 in a LRRK2-dependent manner (Steger et al., 2016) (Figure 3H). We further confirmed the phospho-specific Rab8A, Rab10 and Rab12 interactions with RILPL1/2 by immunoprecipitation-phos-tag and phospho-Rab directed antibodies (Figure 3—figure supplement 3B–D). Together these results unambiguously establish that both RILPL1/2 are phospho-specific Rab interactors and that this interaction is mediated by the RH domain of RILPL1 and RILPL2. Based on these findings, it will be interesting to investigate possible genetic association of RILPL1/2 mutations with PD.

### Ciliogenesis is attenuated in cells expressing hyperactive LRRK2

Primary cilia are hair-like structures extending from the cell body and are found on nearly all human cells (Malicki and Johnson, 2017). Ciliary defects lead to a variety of human diseases known as ciliopathies, some more severe than others, and more than 200 genes have been associated with these disorders (Reiter and Leroux, 2017). Rab or Rab-regulatory protein dysfunctions play important roles in a number of ciliopathies. For instance, Rab28 is localized at the periciliary membrane of ciliated neurons and mutations in its gene are associated with cone-rod dystrophy, an inherited ocular disorder (Jensen et al., 2016; Roosing et al., 2013). Similarly, Rab23 is involved in ciliary protein transport and Rab23 mutations are associated with a rare congenital disorder known as Carpenter syndrome (Boehlke et al., 2010; Jenkins et al., 2007). Rab8A, Rab10, Rab11, Rab17 and Rab29 are involved in ciliogenesis; however, no specific disorders are associated with mutations in these genes (Knödler et al., 2010; Nachury et al., 2007; Onnis et al., 2015; Sato et al., 2014; Yoshimura et al., 2007). Since LRRK2 phosphorylation of Rab8A, Rab10 and Rab12 results in increased RILPL1/2 interaction (Figure 3) and Rab8A, Rab10 and RILPL1/2 are reported to regulate primary ciliogenesis (Nachury et al., 2007; Sato et al., 2014; Schaub and Stearns, 2013), we reasoned that LRRK2 itself might be a regulator of primary cilia formation. To test this, we induced cilia formation by serum starvation in LRRK2-R1441G knock-in MEFs that harbor increased kinase activity, and cultured them with or without MLi-2. Strikingly, as judged by anti-Arl13B staining, DMSO-treated LRRK2-R1441G expressing cells displayed a twofold decrease in the number of ciliated cells and a reduction in cilia length compared with overnight-MLi-2-treated cells (Figure 4A–C). To confirm this finding using another cell system, we analyzed cilia formation in 3T3 fibroblasts transfected with GFP-LRRK2-G2019S. In this case, about 30% of serum-starved, LRRK2-expressing cells (GFP positive) were ciliated (Arl13B positive). Overnight LRRK2 inhibitor treatment (MLi-2) increased this by twofold, demonstrating kinase dependency of the ciliation defect (Figure 4D,E). Thus, two different cellular assays support our finding that LRRK2 regulation is important for ciliogenesis. To determine whether the observed defect in ciliogenesis seen in hyperactive LRRK2 mutant-expressing cells is due to Rab8A phosphorylation, we expressed GFP-Rab8A-T72A or GFP-Rab8A-T72E in serum-starved RPE-1 cells and assessed both primary cilia formation and exogenous Rab8A protein localization by immunofluorescence microscopy. While GFP-Rab8A-wt was localized at cilia in about 80% of Arl13B-positive cells, this was not true for the T72A nor the T72E mutants, and total numbers of ciliated cells did not change in any of the conditions analyzed (Figure 4F–H). Thus, exogenous expression of the T72A or T72E mutants alone was not sufficient to block cilia formation. This suggests that other LRRK2 targets might play a role in this process or that endogenous Rab8A is sufficient to sustain cilia formation in this experimental setup.

Together, our data demonstrate that LRRK2 activity influences ciliogenesis and kinase hyperactivity interferes with this process. The precise details of how LRRK2-mediated Rab8A phosphorylation controls this process and influences RILPL1 and RILPL2 function will be important areas for future work. Whether phosphorylation of Rabs other than Rab8A by LRRK2 influences cilia formation will also require further analysis.



**Figure 4.** Pathogenic LRRK2 mutations inhibit primary cilia formation. (A) LRRK2-R1441G knock-in MEFs were serum starved overnight and treated with 200 nM MLI-2 (right) or DMSO (left). Primary cilia were stained using mouse anti-Arl13B (red) and nuclei using DAPI. (B) Quantification of primary cilia (Arl13B staining,  $n = 2, >100$  cells per condition) and (C) cilia length (70 per condition). Scale bar = 10  $\mu$ m. (D) NIH3T3 cells transfected with eGFP-LRRK2-G2019S were serum starved for 24 hr in the presence or absence of MLI-2 (200 nM). Scale bar = 10  $\mu$ m. (E) Quantification of primary cilia (Arl13B staining,  $n = 2, >100$  cells per condition). Scale bar = 10  $\mu$ m. (F) NIH3T3 cells expressing GFP-Rab8a with a TE mutation (T72E) or a TA mutation (T72A) were serum starved for 24 hr in the presence or absence of MLI-2 (200 nM). Primary cilia were stained using mouse anti-Arl13B (red) and GFP (green). White dashed lines indicate the cell boundaries. Scale bar = 10  $\mu$ m. (G) Quantification of the percentage of GFP-Rab8a with ciliary localization. (H) Quantification of the percentage of ciliated cells (GFP-positive). Error bars represent standard deviation.  $p$  values are indicated in (B), (C), (E), and (G).

Figure 4 continued

positive) in GFP-positive cells; >50 cells per replicate were counted ( $n = 3$ ). Error bars represent SEM and p-values were determined using unpaired, two-tailed Student's t-tests. (F) RPE-1 cells were infected with a lentivirus encoding GFP-Rab8A (T72E, T72A or wt) and 48 hr later, serum starved overnight to induce ciliation; primary cilia were detected with mouse anti-Arl13B (red). Dotted lines indicate cell outlines. Scale bars = 10  $\mu$ m. (G) Quantification of cells with ciliary (Arl13B positive) GFP-Rab8A localization; (H) Quantitation of primary cilia (Arl13B staining) in GFP-positive cells. Error bars represent SEM from three experiments with >75 cells counted per condition.

DOI: <https://doi.org/10.7554/eLife.31012.013>

## Conclusion

Our systematic analysis of LRRK2-dependent Rab phosphorylation revealed that a distinct subset of the Rab family is subject to regulation by this kinase. When overexpressed a total of 14 Rabs are phosphorylated by LRRK2, and at least ten are phosphorylated when they are present in cells at endogenous levels. New tools such as specific antibodies or sensitive targeted MS methods, as well as more in depth analyses of tissues expressing LRRK2, are needed to clarify whether the remaining three Rabs (Rab5B/C and Rab29) are true endogenous LRRK2 substrates. Phosphomimetic substitution of the identified LRRK2 phosphosites abrogated the binding of Rabs to GDI1/2 and CHM/CHML, demonstrating that the phosphorylation of the switch-II domain is a general regulatory mechanism of the Rab cycle. Our data suggest that kinases other than LRRK2 phosphorylate the switch-II domain of some Rabs in different contexts. In fact, we found 37 overexpressed Rabs to be phosphorylated within this domain, but only 14 of these are catalyzed by LRRK2. Recent reports that Rab7A is phosphorylated on T72 during B cell signaling and that Rab1A-T75 is a TAK1 target (Levin et al., 2016; Satpathy et al., 2015) support this proposal.

Interactome analysis of all LRRK2-regulated Rabs confirmed GDI-binding as a general feature of the non-phospho forms. Unexpectedly, our experiments revealed different binding partners when the phosphomimetic S/T $\rightarrow$ E substitution was compared with LRRK2-induced phosphorylation. This suggests that the S/T $\rightarrow$ E mutation does not accurately substitute for phosphorylation in this context. It will therefore be necessary to analyze systematically, phosphorylated, endogenous Rabs to identify proteins that bind specifically to their phosphorylated forms. The fact that Rabs are only phosphorylated to a low extent in cells may pose challenges for the detection of modification-specific interactors.

Our previous work suggested that LRRK2 phosphorylation of Rab proteins leads to their functional inactivation due to interference with binding to specific effectors and their regulating proteins (Steger et al., 2016). Here, we have discovered a dominant interaction whereby phosphorylated Rabs bind preferentially to RILPL1 and RILPL2, proteins that are key for ciliogenesis (Schaub and Stearns, 2013). We verified this phosphorylation-dependent interaction in a variety of assays and showed that RILPL1 and RILPL2 are important binding partners of LRRK2 phosphorylated-Rab8A, Rab10 and Rab12.

For the first time, our findings establish a connection between LRRK2 and cilia formation. Indeed, expression of activated LRRK2 mutant proteins interferes with ciliogenesis, a process that is influenced by both Rab8A and Rab10 proteins. Future studies are needed to shed light on the mechanistic details of how LRRK2 regulates the Rab-mediated transport of vesicles during ciliogenesis. In particular, it will be interesting to investigate whether the defective dopamine signaling in neurons of PD patients is at least partially due to LRRK2-mediated protein trafficking alterations in primary cilia. Although PD is not classified as a ciliopathy, subtle changes in ciliation may have profound signaling consequences in particular cell types that are critical for PD pathogenesis.

## Materials and methods

### Reagents

MLi-2 was synthesized by Natalia Shpiro (University of Dundee) as described previously (Fell et al., 2015). HG-10-102-01 was from Calbiochem. Doxycycline,  $\gamma$ -S-GTP, HA-agarose and trypsin from Sigma and LysC from Wako. GluC, AspN and Chymotrypsin from Promega. GFP-agarose beads were from Chromotek. Complete protease and phosphatase inhibitor tablets were from Roche.

## Antibodies

Anti-GFP (#2555), anti-GAPDH (#5174), anti-HA (#3724), anti-Rab1A (#13075), anti-Rab8A (#6975), and anti-Rab10 (#8127) were from Cell Signaling Technologies. Rab3A (PA1-770) and Rab5A (PA5-29022) from Thermo. Rab3B from Abnova (H00005865-M02). Rabbit monoclonal antibodies for total LRRK2 and pS935-LRRK2 were purified at the University of Dundee (Dzamko *et al.*, 2012). Anti-Arl13B from Neuromab, goat anti-mouse Alexa 555 from Thermo Scientific. Rabbit monoclonal anti-Rab8A was from Abcam (ab188574). Sheep polyclonal antibodies against Rab29 phospho-Thr71 (S877D), Rab29 phospho-Ser72 (SA136) and Rab12 phospho-Ser106 (S876D) were generated by injection of KLH (keyhole limpet hemocyanin) and bovine serum albumin-conjugated phospho-peptides CIAGQERFT\*SMTRLYYR (Rab29 pT71), CDIAGQERFTS\*MTRLYYRSS (Rab29 pS72) or CAGQERFNS\*ITSAYYR (Rab12 pS106) into sheep and affinity-purified from serum using the same phosphopeptides. Antibodies were used for immunoblotting at final concentrations of 1 µg/ml, in the presence of 10 µg/ml of non-phosphorylated peptide. Sheep polyclonal antibodies against total Rab12 (SA227), Rab35 (SA314) and Rab43 (SA135) were generated by injection of recombinant full-length Rab12 (no epitope tag), GST-Rab35 or GST-Rab43 proteins into sheep and affinity-purified using the proteins with a different tag (6HIS-SUMO-Rab12, MBP-Rab35, HIS-SUMO-Rab43). Antibodies were used for immunoblotting at the final concentration of 1 µg/ml.

The polyclonal phospho-Rab8 antibody that efficiently immunoprecipitated multiple LRRK2 phosphorylated Rab proteins with a Thr site was from Abcam (Burlingame, California). The antibody was raised against two phospho-T72-Rab8A/Rab8B peptides (C-Ahx-AGQERFRT\*ITTAYYR-amide and Ac-AGQERFRT\*ITTAYYR-Ahx-C-amide, corresponding to residues 65–79 of human Rab8). The antibody was affinity purified employing the phospho peptide as a positive selection and the non-phosphorylated C-Ahx-AGQERFRT\*ITTAYYR as a negative selection step. Rabbit monoclonal antibodies for phospho-T72 Rab8A/Rab8B and phospho-T73 Rab10 were custom-made by Abcam in collaboration with the Michael J Fox Foundation (Burlingame, California) (Lis *et al.*, 2017). The phospho-Rab10 antibody was raised against two peptides (C-Ahx-AGQERFHT\*ITTSYYR-amide and Ac-AGQERFHT\*ITTSYYR-Ahx-C-amide, corresponding to residues 66–80 of human Rab10). The phospho-Rab8 antibody was raised against two peptides (C-Ahx-AGQERFRT\*ITTAYYR-amide and Ac-AGQERFRT\*ITTAYYR-Ahx-C-amide, corresponding to residues 65–79 of human Rab8). For immunization, the peptides were coupled to KLH via the Cys residue. \*demark the phosphorylated residues.

## Plasmids

GFP-Rab7A (Addgene plasmid #12605) and GFP-Rab9A (Addgene plasmid #12663) were from Richard Pagano. GFP-Rab6A (Addgene plasmid #49469), GFP-Rab27A (Addgene plasmid #49605), GFP-Rab2A (Addgene plasmid #49541), GFP-Rab4A (Addgene plasmid #49434), GFP-Rab18 (Addgene plasmid #49550), GFP-Rab30 (Addgene plasmid #49607), GFP-Rab3A (Addgene plasmid #49542), GFP-Rab35 (Addgene plasmid #49552), GFP-Rab1A (Addgene plasmid #49467), GFP-Rab13 (Addgene plasmid #49548), GFP-Rab5A (Addgene plasmid #49888) and GFP-Rab10 (Addgene plasmid #49472) from Marci Scidmore. The following Rab constructs were from the University of Dundee: HA-Rab1B (DU25771), HA-Rab1C (DU55042), HA-Rab2B (DU51392), HA-Rab3A (DU51539), HA-Rab3B (DU55007), HA-Rab3C (DU55048), HA-Rab3D (DU26388), HA-Rab4B (DU51607), HA-Rab5A (DU26389), HA-Rab5B (DU26106), HA-Rab5C (DU47956), HA-Rab6B (DU47954), HA-Rab7B (DU54134), HA-Rab8A (DU35414), HA-Rab8B (DU39856), HA-Rab9B (DU54135), HA-Rab10 (DU44250), HA-Rab12 (DU48963), HA-RAB12-S106A (DU48966), GFP-Rab15 (DU26357), HA-Rab17 (DU54205), GFP-Rab26 (DU47921, DU47922), HA-Rab27B (DU53133), HA-Rab29 (DU50222), GFP-Rab31 (DU28238), HA-Rab32 (DU52622), HA-Rab33A (DU28313), HA-Rab33B (DU28314), HA-Rab35 (DU26478), GFP-Rab37 (DU53072), HA-Rab38 (DU52517), GFP-Rab39A (DU51554), GFP-Rab39B (DU54141), HA-Rab40A (DU53132), HA-Rab40B (DU28315), HA-Rab40C (DU53203), HA-Rab40AL (DU53222), GFP-Rab41 (DU28239), GFP-Rab43 (DU53044), HA-Rab43 (DU26392), GFP-Rab44 (DU53224), GFP-Rab45 (DU47924), HA-Rab8A-T72A/E (DU47360/DU47355), HA-Rab10-T73A/E (DU51006/DU51007), HA-Rab12-S106A/E (DU48966/DU48967), Flag-LRRK2-WT (DU6841), Flag-LRRK2-D2017A (DU52725), Flag-LRRK2-G2019S (DU10129), Flag-LRRK2-R1441G (DU13077), Flag-LRRK2-Y1699C (DU13165), Flag-LRRK2-Y1699C/D2017A (DU52703), RILP-GFP (DU27496), RILPL1-GFP (DU27305), RILPL2-GFP (DU27481), RILPL1[E280-END]-GFP (DU27317), RILPL1[E280-END],

R291L-GFP (DU27478), RILPL1[E280-END, R291E]-GFP (DU27479), RILPL1[E280-END, R293L]-GFP (DU27471), RILPL1[E280-END, R293E]-GFP (DU27472), RILPL1[E280-END, R300L]-GFP (DU27473), RILPL1[E280-END, R300E]-GFP (DU27480), RILPL1[E280-END, K310L]-GFP (DU27474), RILPL1[E280-END, K310E]-GFP (DU27475), RILPL1[E280-END, K324L]-GFP (DU27476), RILPL1[E280-END, K324E]-GFP (DU27477), RILPL1[R291E] (DU27525), RILPL1[R293E] (DU27526), RILPL1[K310E] (DU27527), RILPL2[R130E] (DU27520), RILPL2[R132E] (DU27522), RILPL2[K149E] (DU27523) Full datasheets and reagents are available on <https://mrcppureagents.dundee.ac.uk/>. Mutagenesis on Rab3A/B/C/D, Rab5A/B/C, Rab8B, Rab29, Rab35 and Rab43 was carried out using Q5 site-directed mutagenesis kit (NEB, #E0554S).

### Culture and transfection of cells

HEK293, MEFs and NIH3T3 cells were cultured in Dulbecco's modified Eagle medium (Glutamax, Gibco) supplemented with 10% fetal calf serum (FCS), 100 U/ml penicillin and 100 µg/ml streptomycin. A549 cells were cultured in DMEM supplemented with 10% FBS, 2 mM L-glutamine, 100 units/ml penicillin and 100 µg/ml streptomycin. The HEK293-t-rex-flpIn stable cell lines with doxycycline-inducible mutant forms of LRRK2 have been described previously (Nichols et al., 2010). LRRK2 expression in HEK293-t-rex-flpIn was induced by doxycycline (1 µg/ml). NIH3T3 cells were transfected with eGFP-LRRK2-G2019S at 40% confluence and cilia formation induced by serum starvation after 24 hr. At the same time, 200 nM MLI-2 (or DMSO as control) was added and cells incubated for an additional 24 hr before fixation. LRRK2-R1441G MEFs were plated at 40% confluence and serum starved for 24 hr after 1 day. MLI-2 was added at 100 nM or 200 nM over night before fixation of the cells. Transient transfections were performed 24–48 hr prior to cell lysis using either FuGene 6, Fugene HD (Promega) or polyethylenimine PEI (Polysciences). hTERT-RPE cells were cultured in DMEM + F12 medium with 10% FBS. GFP-Rab8 (wt, T72A and T72E) were expressed using lentivirus vector pSLO1371. After 48 hr of infection, cells were plated onto collagen coated coverslips and 48 hr later treated with serum-free medium to initiate ciliation. Cells were fixed with PFA and stained using mouse anti-Arl13B (Neuromab, 1:1000). All cells were tested for mycoplasma contamination and overexpressing lines were verified by western blot analysis.

### Generation of CRISPR-CAS9 knock-outs

The A549 Rab10 knockout cell line has been described previously (Ito et al., 2016). Rab8A, Rab35 and Rab43 knockout cell lines were generated using the following constructs (available at <https://mrcppureagents.dundee.ac.uk/>): DU48928/DU48942 (Rab8A), DU54332/DU54340 (Rab35), DU54344/DU54350 (Rab43). Cells at about 80% confluence were co-transfected in a six-well plate with the pairs of corresponding vectors using Lipofectamine LTX, with a final amount of 9 µl Lipofectamine LTX and 2.5 µg DNA per well. 24 hr after transfection, the medium was replaced and fresh medium supplemented with puromycin (2 µg/ml). After 24 hr selection, the medium was replaced with medium without puromycin and the cells were left to recover for 48 hr before performing single-cell sorting using an Influx cell sorter (Becton Dickinson). Single cells were placed in individual wells of a 96-well plate containing DMEM supplemented with 10% FBS, 2 mM L-glutamine, 100 units/ml penicillin and 100 µg/ml streptomycin and 100 µg/ml normocin (InvivoGen). At about 80% confluence individual clones were transferred into six-well plates and screened for Rab8A, Rab35 or Rab43 knock-out by western blotting.

### Generation of mouse embryonic fibroblasts (MEFs)

LRRK2-R1441C knock-in mice were obtained from The Jackson Laboratory and maintained on a C57BL/6J background. Littermate matched wild-type and homozygous LRRK2-R1441C mouse embryonic fibroblasts (MEFs) were isolated from mouse embryos at day E12.5 resulting from crosses between heterozygous LRRK2-R1441C/WT mice as described (Wiggin et al., 2002). Cells were spontaneously immortalized by prolonged passaging. Genotyping of mice and MEFs was performed by PCR using genomic DNA. Primer 1 (5' -CTGCAGGCTACTAGATGGTCAAGGT -3') and Primer 2 (5' -CTAGATAGGACCGAGTGTGAGAG- 3') were used to detect the wild-type and knock-in alleles. The PCR program consisted of 5 min at 95°C, then 35 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C, and 5 min at 72°C.

### Immunofluorescence microscopy

Cells were fixed (3% paraformaldehyde, 15 min room temperature), permeabilized using Triton X-100 (0.1%), blocked with 2% BSA (in PBS) and stained with mouse anti-Arl13B and goat anti-mouse Alexa 555 antibodies (Both 1:1000). Nuclei were stained with DAPI (Sigma). Coverslips were mounted with Mowiol and NIH3T3 cilia visualized using an Olympus IX70 microscope with 60  $\times$  1.4 NA Plan-Apochromat oil immersion objective (Olympus, Center Valley, PA) and a charge-coupled device camera (CoolSNAP HQ, Photometrics, Tucson, AZ). MEF-R1441G cells were visualized using an SP8 laser scanning confocal microscope (Leica Microsystems, Germany) using a 63  $\times$  1.4 NA oil immersion objective. When needed, maximum intensity projections were made and images analyzed using ImageJ.

### Phos-tag SDS-PAGE and western blotting

Cells were lysed in ice-cold lysis buffer containing 50 mM Tris/HCl, pH 7.4, 1% (v/v) Triton X-100, 10% (v/v) Glycerol, 1 mM sodium orthovanadate, 50 mM NaF, 10 mM 2-glycerophosphate, 5 mM sodium pyrophosphate, 0.1  $\mu$ g/ml mycrocystin-LR (Enzo Life Sciences, Switzerland) and complete EDTA-free protease inhibitor cocktail (Roche). Lysates were clarified by centrifugation at 20,800 g and then mixed with 4  $\times$  SDS/PAGE sample buffer (250 mM Tris/HCl, pH 6.8, 8% (w/v) SDS, 40% (v/v) glycerol, 0.02% (w/v) Bromophenol Blue and 4% (v/v) 2-mercaptoethanol) and heated at 95°C for 5 min. For SDS/PAGE, 4–20  $\mu$ g sample was loaded on to NuPAGE Bis-Tris 4–12% gel (Life Technologies) and electrophoresed at 150 V. For Phos-tag SDS-PAGE, samples were supplemented with 10 mM MnCl<sub>2</sub> before loading gels. Phos-tag SDS-PAGE was performed following the protocol described previously with slight modifications (Ito et al., 2016). Gels for Phos-tag SDS/PAGE consisted of a stacking gel [4% (w/v) acrylamide, 125 mM Tris/HCl, pH 6.8, 0.1% (w/v) SDS, 0.2% (v/v) N,N,N',N'-tetramethylethylenediamine (TEMED) and 0.08% (w/v) ammonium persulfate (APS)] and a separating gel [12% (w/v) acrylamide, 375 mM Tris/HCl, pH 8.8, 0.1% (w/v) SDS, 100  $\mu$ M Phos-tag acrylamide, 200  $\mu$ M MnCl<sub>2</sub>, 0.1% (v/v) TEMED and 0.05% (w/v) APS]. 4–12  $\mu$ g of lysate or pull-down samples were electrophoresed at 70 V for the stacking gel and at 135 V for the separating gel with the running buffer [25 mM Tris/HCl, 192 mM glycine and 0.1% (w/v) SDS]. After SDS-PAGE, gels were washed three times with transfer buffer [48 mM Tris/HCl, 39 mM glycine and 20% (v/v) methanol] containing 10 mM EDTA and 0.05% (w/v) SDS, followed by a 10-min wash with transfer buffer containing 0.05% SDS. Blotting to nitrocellulose membranes was carried out at 90 V for 180 min on ice in the transfer buffer without SDS/EDTA. Transferred membranes were blocked with 5% (w/v) non-fat dry milk (NFDM) dissolved in TBS-T [20 mM Tris/HCl, pH 7.5, 150 mM NaCl and 0.1% (v/v) Tween 20] at room temperature for 45 min. Membranes were incubated with primary antibodies in 5% NFDM/TBS-T overnight at 4°C. After washing in TBS-T, membranes were incubated at room temperature for 1 hr with either horseradish peroxidase-labelled secondary antibody or with near-infrared fluorescent antibodies labeled with different fluorophores (680 and 800 nm, both diluted in 5% NFDM/TBS-T). After washing in TBS-T, protein bands were detected by exposing films [Medical Film (Konica Minolta)] to the membranes using an ECL solution [Amersham ECL Western Blotting Detection Reagents (GE Healthcare)] or by Li-COR scanning.

### Pull-downs

Cells were lysed either in ice cold NP-40 buffer (50 mM Tris-HCl, pH 7.5, 120 mM NaCl, 1 mM EDTA, 6 mM EGTA, 15 mM sodium pyrophosphate and 1% NP-40 supplemented with protease and phosphatase inhibitors [Roche]), in buffer containing  $\gamma$ -S-GTP (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5 mM MgCl<sub>2</sub>, 1 mM EDTA, 1 mM DTT, 10 mM sodium-glycerophosphate, 10 mM sodium pyrophosphate, 100  $\mu$ M  $\gamma$ -S-GTP and 1% NP-40 supplemented with protease and phosphatase inhibitors) in case of endogenous Rab IPs or in 50 mM Tris/HCl, pH 7.4, 1% (v/v) Triton X-100, 10% (v/v) Glycerol, 1 mM sodium orthovanadate, 50 mM NaF, 10 mM 2-glycerophosphate, 5 mM sodium pyrophosphate, 0.1  $\mu$ g/ml mycrocystin-LR (Enzo Life Sciences) and complete EDTA-free protease inhibitor cocktail.

Lysates were clarified by centrifugation at 14,000 rpm after a liquid nitrogen freeze-thaw cycle. For  $\alpha$ -GFP and  $\alpha$ -HA pulldowns, lysates were incubated with GFP/HA-agarose resin for 2 hr (20–30  $\mu$ l beads/pulldown). Bound complexes were recovered by washing the beads twice with NP-40 buffer and twice with 50 mM Tris-HCl (pH 7.5) before overnight on-bead digestion at 37°C using

trypsin (~250 ng/pulldown in 2 M urea [dissolved in 50 mM Tris-HCl pH 7.5]). Digestion with enzymes other than trypsin (chymotrypsin, GluC, AspN) were performed according to manufacturer's instructions. The resulting peptides were processed as described in 'LC-MS-MS sample preparation'.

GFP-tagged RILP, RILPL1 and RILPL2 were affinity-purified from lysates (1–2 mg of protein) by incubation with 5 µl GFP-agarose resin (GFP-Trap\_A, Chromotek) for 1 hr at 4°C. Immunoprecipitates were washed twice with lysis buffer and samples eluted from the beads by addition of 2 × SDS/PAGE sample buffer. The mixture was then incubated at 100°C for 10 min and the eluent collected by centrifugation through a 0.22 µm pore-size Spinex column before adding 2-Mercaptoethanol (1%). Samples were incubated for 5 min at 70°C prior to normal SDS-PAGE or Phos-tag analysis.

For immunoprecipitation of phospho-Rab from MEFs, the antibody was cross-linked to Protein-G-agarose beads at a ratio of 1 µg of antibody per 1 µl of beads using dimethyl pimelimidate as the cross-linking reagent. Lysates were incubated with the antibody-bead complexes for 2 hr at 4°C (20 µl of beads/2 mg of cell extract). The immunocomplexes were washed three times with PBS and samples eluted from the beads using 2 × SDS/PAGE sample buffer. The mixture was incubated at 100°C for 10 min and the eluent was collected by centrifugation through a 0.22 µm pore-size Spinex column before addition of 2-Mercaptoethanol (1%). Samples were incubated for 5 min at 70°C prior to western blotting.

### LC-MS/MS sample preparation

Peptides were purified using in-house prepared SDB-RPS (Empore) stage tips (Rappaport et al., 2003) before LC-MS/MS analysis as described previously (Kulak et al., 2014). Briefly, stage tips were prepared by inserting two layers of SDB-RPS matrix into a 200 µl pipette tip using an in-house prepared syringe device. Stage-tips were first activated with 100 µl of 30% Methanol/1% Trifluoroacetic acid (TFA) and then washed with 100 µl 2% Acetonitrile (ACN)/0.2% TFA before loading of the acidified peptides (1% TFA v/v). After centrifugation, the stage-tips were washed trice (200 µl each) with 2% ACN/0.2% TFA. Elution was performed using 60 µl of 60% ACN/1.25% NH<sub>4</sub>OH. Eluates were collected in 200 µl PCR tubes and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus) at 60°C. Peptides were resuspended in buffer A\* (2% ACN/0.1% TFA) and briefly sonicated (Branson Ultrasonics) before LC/MS-MS analysis.

### LC-MS/MS measurements

Peptides were loaded on a 20 or 50 cm reversed phase column (75 µm inner diameter, packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin [Dr. Maisch GmbH]). Column temperature was maintained at 60°C using a homemade column oven. An EASY-nLC 1200 system (Thermo Fisher Scientific) was directly coupled online with the mass spectrometer (Q Exactive HF, Q Exactive HF-X, LTQ Orbitrap, Thermo Fisher Scientific) via a nano-electrospray source, and peptides were separated with a binary buffer system of buffer A (0.1% formic acid [FA]) and buffer B (80% acetonitrile plus 0.1% FA), at a flow rate of 250 or 350 nL/min. Peptides were eluted with a nonlinear 45–180 min gradient of 5–60% buffer B (0.1% (v/v) formic acid, 80% (v/v) ACN). After each gradient, the column was washed with 95% buffer B for 5 min. The mass spectrometer was programmed to acquire in a data-dependent mode (Top10–Top15) using a fixed ion injection time strategy. Full scans were acquired in the Orbitrap mass analyzer with resolution 60,000 at 200 m/z (3E6 ions were accumulated with a maximum injection time of 25 ms). The top intense ions (N for TopN) with charge states ≥ 2 were sequentially isolated to a target value of 1E5 (maximum injection time of 120 ms, 20% underfill), fragmented by HCD (NCE 27%, Q Exactive) or CID (NCE 35%, LTQ Orbitrap) and detected in the Orbitrap (Q Exactive, R = 15,000 at m/z 200) or the ion trap detector (LTQ Orbitrap).

Figure 2 (Rab3B) and Figure 2—figure supplement 1 (Rab5A): Xcalibur (v4.0.27.19) was used to build a method in which every full scan with resolution 120,000 at 200 m/z (3E6 ions accumulated with a maximum injection time of 20 ms) was followed by five data-dependent MS/MS scans with resolution 15,000 at 200 m/z (the target value of 1E5 with a maximum injection time of 32 ms, 1.4 m/z isolation window and NCE 27%) and a multiplexed SIM scan employed using the multiple ion injection method in which both light (endogenous) and heavy counterpart phosphopeptides of Rab5 and Rab3B were recorded simultaneously (amount of spike-in = 300 fmol per injection). Each SIM scan covered a range of m/z 150–2,000 with resolution 120,000 (1E5 ions accumulated with a

maximum injection time of 240 ms for both light and heavy counterparts, 1.4 m/z isolation window and 0.4 m/z isolation offset). For Rab5, m/z values of doubly ( $z = 2$ )- and triply ( $z = 3$ )-charged oxidized or non-oxidized YH<sub>p</sub>SLAPMYYR ions were defined as follows: 699.891, 691.894, 467.297, 461.965. For Rab3b, only the doubly ( $z = 2$ )-charged non-oxidized YR<sub>p</sub>TITAYYR was targeted with an m/z value of 695.424.

LC-MS/MS analysis of pRab8 IPs (Figure 2—figure supplement 2B): Samples were analyzed using an Orbitrap Fusion Tribrid mass spectrometer coupled to Dionex Ultimate 3000 RSLC nano system (Thermo Scientific). Peptides were loaded on a pre-column (C18, 100A°, 2 cm PN: 164564) and resolved on 50 cm analytical column (C18, 100A°, 75μX 50 cm PN: ES803, Easy -Spray column, Thermo Scientific). Peptides were separated with a binary buffer system of buffer A (0.1% formic acid [FA]) and buffer B (0.1% (v/v) FA, 90% Acetonitrile (v/v)) by applying a non-linear gradient of 5% to 35% for 65 min and a total run of 85 min. Mass spectrometry data were acquired in a data-dependent mode targeting top 10 precursor peptides from the survey scan for MS2 fragmentation using HCD (32% NCE). Full scans were acquired in the Orbitrap mass analyzer with 120,000 resolution at 200 m/z and the MS/MS was acquired with Orbitrap with 30,000 resolution at 200 m/z. AGC was set at 2E5 and 5E4 for MS and MS/MS, respectively.

## Data processing and analysis

Raw mass spectrometry data were processed using MaxQuant version 1.5.3.15 or 1.5.8.3 (pRab8 IP of Figure 2—figure supplement 2B) (Cox and Mann, 2008; Cox et al., 2011) with an FDR < 0.01 at the level of proteins, peptides and modifications. Searches were performed against the Mouse or Human UniProt FASTA database (September 2015). Enzyme specificity was set to trypsin, GluC, AspN, LysC or chymotrypsin depending on the employed enzyme. The search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein, oxidation of methionine, and/or phosphorylation of Ser, Thr, Tyr residue (PhosphoSTY) as variable modifications. Up to two missed cleavages were allowed for protease digestion. Protein quantification was performed by MaxQuant (Label free quantification MaxLFQ [Cox et al., 2014]), 'match between runs' was enabled, with a matching time window of 0.7 min. Ion chromatograms of the SIM-MS files were extracted using Skyline (version 3.7.0.10940) (MacLean et al., 2010). Bioinformatic analyses were performed with Perseus ([www.perseus-framework.org](http://www.perseus-framework.org)) and Microsoft Excel and data visualized using Graph Prism (GraphPad Software Inc) or Perseus (Tyanova et al., 2016). Significance was assessed using one sample t-test, two-sample student's t-test and ANOVA analysis, for which replicates were grouped, and statistical tests performed with permutation-based FDR correction for multiple hypothesis testing. Where indicated, missing data points were replaced by data imputation after filtering for valid values (all valid values in at least one experimental group). Error bars are mean  $\pm$  SEM or mean  $\pm$  SD.

## Acknowledgements

This work was supported by The Michael J. Fox Foundation for Parkinson's Research (grant ID 6986.05), the Max-Planck Society for the Advancement of Science and the Medical Research Council (grant numbers MC\_UU\_12016/2 (to DRA)). We thank S Uebel, M Spitaler and S Pettera from the MPIB Biochemistry Core Facility and S Kroiss, G Sowa, K Mayr and I Paron from the department of Proteomics and Signal Transduction for technical assistance. Members of the Dundee MRC-PPU Cloning, DNA Sequencing, antibody, protein production, tissue culture and mass spectrometry teams for technical assistance. Sandy Chou at Abcam for generation of the phospho-Rab8 polyclonal antibody. LRRK2 R1441G MEFs were kindly provided by Dr Shu-Leong Ho (Division of Neurology, Department of Medicine, University of Hong Kong) and have been described previously (Ito et al., 2016).

## Additional information

### Competing interests

Suzanne R Pfeffer: Reviewing editor, eLife. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Max-Planck-Gesellschaft		Martin Steger Ozge Karayel Matthias Mann
Michael J. Fox Foundation for Parkinson's Research	6986.05	Martin Steger Federico Diez Herschel S Dhekne Pawel Lis Raja S Nirujogi Ozge Karayel Francesca Tonelli Terina N Martinez Suzanne R Pfeffer Dario R Alessi Matthias Mann
Medical Research Council	357811350 R60	Federico Diez Pawel Lis Raja S Nirujogi Francesca Tonelli Dario R Alessi

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Martin Steger, Conceptualization, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Federico Diez, Herschel S Dhekne, Investigation, Visualization, Methodology; Pawel Lis, Raja S Nirujogi, Ozge Karayel, Investigation, Visualization; Francesca Tonelli, Validation, Investigation; Terina N Martinez, Resources; Esben Lorentzen, Conceptualization, Writing—original draft; Suzanne R Pfeffer, Conceptualization, Supervision, Funding acquisition, Investigation, Visualization, Writing—review and editing; Dario R Alessi, Conceptualization, Supervision, Funding acquisition, Writing—original draft, Project administration, Writing—review and editing; Matthias Mann, Conceptualization, Supervision, Funding acquisition, Investigation, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Martin Steger  <http://orcid.org/0000-0003-1637-8190>  
Esben Lorentzen  <http://orcid.org/0000-0001-6493-7220>  
Suzanne R Pfeffer  <http://orcid.org/0000-0002-6462-984X>  
Dario R Alessi  <http://orcid.org/0000-0002-2140-9185>  
Matthias Mann  <http://orcid.org/0000-0003-1292-4799>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.31012.017>  
Author response <https://doi.org/10.7554/eLife.31012.018>

---

## Additional files

Supplementary files

- Supplementary file 1 Summary of used proteases and identified peptides for the proteomic analysis of 52 Rab GTPases. Proteomics raw data have been deposited to the ProteomeXchange Consortium (Vizcaíno *et al.*, 2014) via the PRIDE partner repository with the data set identifier PXD007214.  
DOI: <https://doi.org/10.7554/eLife.31012.014>
- Transparent reporting form  
DOI: <https://doi.org/10.7554/eLife.31012.015>

## References

Baptista MA, Dave KD, Frasier MA, Sherer TB, Greeley M, Beck MJ, Varsho JS, Parker GA, Moore C, Churchill MJ, Meshul CK, Fiske BK. 2013. Loss of leucine-rich repeat kinase 2 (LRRK2) in rats leads to progressive abnormal phenotypes in peripheral organs. *PLoS ONE* **8**:e80705. DOI: <https://doi.org/10.1371/journal.pone.0080705>, PMID: 24244710

Bardien S, Lesage S, Brice A, Carr J. 2011. Genetic characteristics of leucine-rich repeat kinase 2 (LRRK2)-associated Parkinson's disease. *Parkinsonism & Related Disorders* **17**:501–508. DOI: <https://doi.org/10.1016/j.parkreldis.2010.11.008>, PMID: 21641266

Beilina A, Cookson MR. 2016. Genes associated with Parkinson's disease: regulation of autophagy and beyond. *Journal of Neurochemistry* **139**:91–107. DOI: <https://doi.org/10.1111/jnc.13266>, PMID: 26223426

Beilina A, Rudenko IN, Kaganovich A, Civiero L, Chau H, Kalia SK, Kalia LV, Lobbestael E, Chia R, Ndukwe K, Ding J, Nalls MA, International Parkinson's Disease Genomics Consortium, North American Brain Expression Consortium, Olszewski M, Hauser DN, Kumaran R, Lozano AM, Baekelandt V, Greene LE, Taymans JM, Greggio E, et al. 2014. Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease. *PNAS* **111**:2626–2631. DOI: <https://doi.org/10.1073/pnas.1318306111>, PMID: 24510904

Boehlke C, Bashkurov M, Buescher A, Krick T, John AK, Nitschke R, Walz G, Kuehn EW. 2010. Differential role of Rab proteins in ciliary trafficking: Rab23 regulates smoothened levels. *Journal of Cell Science* **123**:1460–1467. DOI: <https://doi.org/10.1242/jcs.058883>, PMID: 20375059

Choi HG, Zhang J, Deng X, Hatcher JM, Patricelli MP, Zhao Z, Alessi DR, Gray NS. 2012. Brain penetrant LRRK2 inhibitor. *ACS Medicinal Chemistry Letters* **3**:658–662. DOI: <https://doi.org/10.1021/ml300123a>, PMID: 23066449

Cookson MR. 2010. The role of leucine-rich repeat kinase 2 (LRRK2) in Parkinson's disease. *Nature Reviews Neuroscience* **11**:791–797. DOI: <https://doi.org/10.1038/nrn2935>, PMID: 21088684

Cookson MR. 2016. Cellular functions of LRRK2 implicate vesicular trafficking pathways in Parkinson's disease. *Biochemical Society Transactions* **44**:1603–1610. DOI: <https://doi.org/10.1042/BST20160228>, PMID: 27913668

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics* **13**:2513–2526. DOI: <https://doi.org/10.1074/mcp.M113.031591>, PMID: 24942700

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**:1367–1372. DOI: <https://doi.org/10.1038/nbt.1511>, PMID: 19029910

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10**:1794–1805. DOI: <https://doi.org/10.1021/pr101065j>, PMID: 21254760

Dzamko N, Inesta-Vaquera F, Zhang J, Xie C, Cai H, Arthur S, Tan L, Choi H, Gray N, Cohen P, Pedrioli P, Clark K, Alessi DR. 2012. The IkappaB kinase family phosphorylates the Parkinson's disease kinase LRRK2 at Ser935 and Ser910 during toll-like receptor signaling. *PLoS One* **7**:e39132. DOI: <https://doi.org/10.1371/journal.pone.0039132>, PMID: 22723946

Fell MJ, Mirescu C, Basu K, Cheewatrakoolpong B, DeMong DE, Ellis JM, Hyde LA, Lin Y, Markgraf CG, Mei H, Miller M, Poulet FM, Scott JD, Smith MD, Yin Z, Zhou X, Parker EM, Kennedy ME, Morrow JA. 2015. MLi-2, a potent, selective, and centrally active compound for exploring the therapeutic potential and safety of LRRK2 kinase inhibition. *Journal of Pharmacology and Experimental Therapeutics* **355**:397–409. DOI: <https://doi.org/10.1124/jpet.115.227587>, PMID: 26407721

Funayama M, Hasegawa K, Ohta E, Kawashima N, Komiya M, Kowa H, Tsuji S, Obata F. 2005. An LRRK2 mutation as a cause for the parkinsonism in the original PARK8 family. *Annals of Neurology* **57**:918–921. DOI: <https://doi.org/10.1002/ana.20484>, PMID: 15880653

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *PNAS* **100**:6940–6945. DOI: <https://doi.org/10.1073/pnas.0832254100>, PMID: 12771378

Glaime E, Tong Y, Wagner LK, Yuan Y, Huang G, Shen J. 2017. Age-dependent dopaminergic neurodegeneration and impairment of the autophagy-lysosomal pathway in Lrrk-deficient mice. *Neuron*. DOI: <https://doi.org/10.1016/j.neuron.2017.09.036>, PMID: 29056298

Giesert F, Hofmann A, Bürger A, Zerle J, Kloos K, Hafen U, Ernst L, Zhang J, Vogt-Weisenhorn DM, Wurst W. 2013. Expression analysis of Lrrk1, Lrrk2 and Lrrk2 splice variants in mice. *PLoS ONE* **8**:e63778. DOI: <https://doi.org/10.1371/journal.pone.0063778>, PMID: 23675505

Goody PR, Heller K, Oesterlin LK, Müller MP, Itzen A, Goody RS. 2012. Reversible phosphocholinatation of Rab proteins by Legionella pneumophila effector proteins. *The EMBO Journal* **31**:1774–1784. DOI: <https://doi.org/10.1038/emboj.2012.16>, PMID: 22307087

Hou X, Hagemann N, Schoebel S, Blankenfeldt W, Goody RS, Erdmann KS, Itzen A. 2011. A structural basis for Lowe syndrome caused by mutations in the Rab-binding domain of OCRL1. *The EMBO Journal* **30**:1659–1670. DOI: <https://doi.org/10.1038/emboj.2011.60>, PMID: 21378754

Ito G, Katsemonova K, Tonelli F, Lis P, Baptista MA, Shpiro N, Duddy G, Wilson S, Ho PW, Ho SL, Reith AD, Alessi DR. 2016. Phos-tag analysis of Rab10 phosphorylation by LRRK2: a powerful assay for assessing kinase

function and inhibitors. *Biochemical Journal* **473**:2671–2685. DOI: <https://doi.org/10.1042/BCJ20160557>, PMID: 27474410

Jenkins D, Seelow D, Jehee FS, Perlyn CA, Alonso LG, Bueno DF, Donnai D, Josifova D, Josifova D, Mathijssen IM, Morton JE, Orstavik KH, Sweeney E, Wall SA, Marsh JL, Nurnberg P, Passos-Bueno MR, Wilkie AO. 2007. RAB23 mutations in Carpenter syndrome imply an unexpected role for hedgehog signaling in cranial-suture development and obesity. *The American Journal of Human Genetics* **80**:1162–1170. DOI: <https://doi.org/10.1086/518047>, PMID: 17503333

Jensen VL, Carter S, Sanders AA, Li C, Kennedy J, Timbers TA, Cai J, Scheidel N, Kennedy BN, Morin RD, Leroux MR, Blacque OE. 2016. Whole-organism developmental expression profiling identifies rab-28 as a novel ciliary gtpase associated with the bbsome and intraflagellar transport. *PLOS Genetics* **12**:e1006469. DOI: <https://doi.org/10.1371/journal.pgen.1006469>, PMID: 27930654

Jordens I, Fernandez-Borja M, Marsman M, Dusseljee S, Janssen L, Calafat J, Janssen H, Wubbolts R, Neefjes J. 2001. The Rab7 effector protein RILP controls lysosomal transport by inducing the recruitment of dynein-dynactin motors. *Current Biology* **11**:1680–1685. DOI: [https://doi.org/10.1016/S0960-9822\(01\)00531-0](https://doi.org/10.1016/S0960-9822(01)00531-0), PMID: 11696325

Knödler A, Feng S, Zhang J, Zhang X, Das A, Peränen J, Guo W. 2010. Coordination of Rab8 and Rab11 in primary ciliogenesis. *PNAS* **107**:6346–6351. DOI: <https://doi.org/10.1073/pnas.1002401107>, PMID: 20308558

Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M. 2014. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods* **11**:319–324. DOI: <https://doi.org/10.1038/nmeth.2834>, PMID: 24487582

Kuwayahara T, Inoue K, D'Agati VD, Fujimoto T, Eguchi T, Saha S, Wolozin B, Iwatsubo T, Abeliovich A. 2016. LRRK2 and RAB7L1 coordinately regulate axonal morphology and lysosome integrity in diverse cellular contexts. *Scientific Reports* **6**:29945. DOI: <https://doi.org/10.1038/srep29945>, PMID: 27424887

Levin RS, Hertz NT, Burlingame AL, Shokat KM, Mukherjee S. 2016. Innate immunity kinase TAK1 phosphorylates Rab1 on a hotspot for posttranslational modifications by host and pathogen. *PNAS* **113**:E4776–E4783. DOI: <https://doi.org/10.1073/pnas.1608355113>, PMID: 27482120

Lis P, Burel S, Steger M, Mann M, Brown F, Diez F, Tonelli F, Holton JL, Ho PW, Ho S-L, Chou M-Y, Polinski NK, Martinez TN, Davies P, Alessi DR. 2017. Development of phospho-specific Rab protein antibodies to monitor in vivo activity of the LRRK2 Parkinson's disease kinase. *Biochemical Journal* **BCJ20170802**. DOI: <https://doi.org/10.1042/BCJ20170802>

Liu K, Jian Y, Sun X, Yang C, Gao Z, Zhang Z, Liu X, Li Y, Xu J, Jing Y, Mitani S, He S, Yang C. 2016. Negative regulation of phosphatidylinositol 3-phosphate levels in early-to-late endosome conversion. *The Journal of Cell Biology* **212**:181–198. DOI: <https://doi.org/10.1083/jcb.201506081>, PMID: 26783301

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**:966–968. DOI: <https://doi.org/10.1093/bioinformatics/btq054>, PMID: 20147306

Malicki JJ, Johnson CA. 2017. The Cilium: Cellular Antenna and Central Processing Unit. *Trends in Cell Biology* **27**:126–140. DOI: <https://doi.org/10.1016/j.tcb.2016.08.002>, PMID: 27634431

Matsui T, Ohbayashi N, Fukuda M. 2012. The Rab interacting lysosomal protein (RILP) homology domain functions as a novel effector domain for small GTPase Rab36: Rab36 regulates retrograde melanosome transport in melanocytes. *The Journal of biological chemistry* **287**:28619–28631. DOI: <https://doi.org/10.1074/jbc.M112.370544>, PMID: 22740695

Mukherjee S, Liu X, Arasaki K, McDonough J, Galán JE, Roy CR. 2011. Modulation of Rab GTPase function by a protein phosphocholine transferase. *Nature* **477**:103–106. DOI: <https://doi.org/10.1038/nature10335>, PMID: 2182290

Müller MP, Peters H, Blümmer J, Blankenfeldt W, Goody RS, Itzen A. 2010. The *Legionella* effector protein DrrA AMPylates the membrane traffic regulator Rab1b. *Science* **329**:946–949. DOI: <https://doi.org/10.1126/science.1192276>, PMID: 20651120

Nachury MV, Loktev AV, Zhang O, Westlake CJ, Peränen J, Merdes A, Slusarski DC, Scheller RH, Bazan JF, Sheffield VC, Jackson PK. 2007. A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. *Cell* **129**:1201–1213. DOI: <https://doi.org/10.1016/j.cell.2007.03.053>, PMID: 17574030

Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, DeStefano AL, Kara E, Bras J, Sharma M, Schulte C, Keller MF, Arepalli S, Letson C, Edsall C, Stefansson H, Liu X, Pliner H, Lee JH, Cheng R, et al. 2014. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics* **46**:989–993. DOI: <https://doi.org/10.1038/ng.3043>, PMID: 25064009

Nichols RJ, Dzamko N, Morrice NA, Campbell DG, Deak M, Ordureau A, Macartney T, Tong Y, Shen J, Prescott AR, Alessi DR. 2010. 14-3-3 binding to LRRK2 is disrupted by multiple Parkinson's disease-associated mutations and regulates cytoplasmic localization. *Biochemical Journal* **430**:393–404. DOI: <https://doi.org/10.1042/BJ20100483>, PMID: 20642453

Oesterlin LK, Goody RS, Itzen A. 2012. Posttranslational modifications of Rab proteins cause effective displacement of GDP dissociation inhibitor. *PNAS* **109**:5621–5626. DOI: <https://doi.org/10.1073/pnas.1121161109>, PMID: 22411835

Onnis A, Finetti F, Patrucci L, Gottardo M, Cassioli C, Spanò S, Baldari CT. 2015. The small GTPase Rab29 is a common regulator of immune synapse assembly and ciliogenesis. *Cell Death and Differentiation* **22**:1687–1699. DOI: <https://doi.org/10.1038/cdd.2015.17>, PMID: 26021297

Paisán-Ruiz C, Jain S, Evans EW, Gilks WP, Simón J, van der Brug M, López de Munain A, Aparicio S, Gil AM, Khan N, Johnson J, Martinez JR, Nicholl D, Carrera IM, Pena AS, de Silva R, Lees A, Martí-Massó JF, Pérez-Tur J, Wood NW, et al. 2004. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron* **44**:595–600. DOI: <https://doi.org/10.1016/j.neuron.2004.10.023>, PMID: 15541308

Rappaport J, Ishihama Y, Mann M. 2003. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical Chemistry* **75**:663–670. DOI: <https://doi.org/10.1021/ac026117i>, PMID: 12585499

Reiter JF, Leroux MR. 2017. Genes and molecular pathways underpinning ciliopathies. *Nature Reviews Molecular Cell Biology* **18**:533–547. DOI: <https://doi.org/10.1038/nrm.2017.60>, PMID: 28698599

Roosing S, Rohrschneider K, Beryozkin A, Sharon D, Weisschuh N, Staller J, Kohl S, Zelinger L, Peters TA, Neveling K, Strom TM, van den Born LI, Hoyng CB, Klaver CC, Roepman R, Wissinger B, Banin E, Cremers FP, den Hollander AJ, European Retinal Disease Consortium. 2013. Mutations in RAB28, encoding a farnesylated small GTPase, are associated with autosomal-recessive cone-rod dystrophy. *The American Journal of Human Genetics* **93**:110–117. DOI: <https://doi.org/10.1016/j.ajhg.2013.05.005>, PMID: 23746546

Sato T, Iwano T, Kunii M, Matsuda S, Mizuguchi R, Jung Y, Hagiwara H, Yoshihara Y, Yuzaki M, Harada R, Harada A. 2014. Rab8a and Rab8b are essential for several apical transport pathways but insufficient for ciliogenesis. *Journal of Cell Science* **127**:422–431. DOI: <https://doi.org/10.1242/jcs.136903>, PMID: 24213529

Satpathy S, Wagner SA, Belli P, Gupta R, Kristiansen TA, Malinova D, Francavilla C, Tolar P, Bishop GA, Hostager BS, Choudhary C. 2015. Systems-wide analysis of BCR signalosomes and downstream phosphorylation and ubiquitylation. *Molecular Systems Biology* **11**:810. DOI: <https://doi.org/10.1525/msb.20145880>, PMID: 26038114

Schapansky J, Nardozzi JD, Felizia F, LaVoie MJ. 2014. Membrane recruitment of endogenous LRRK2 precedes its potent regulation of autophagy. *Human Molecular Genetics* **23**:4201–4214. DOI: <https://doi.org/10.1093/hmg/ddu138>, PMID: 24682598

Schaub JR, Stearns T. 2013. The Rilp-like proteins Rilpl1 and Rilpl2 regulate ciliary membrane content. *Molecular Biology of the Cell* **24**:453–464. DOI: <https://doi.org/10.1091/mbc.E12-08-0598>, PMID: 23264467

Sheng Z, Zhang S, Bustos D, Kleinheinz T, Le Pichon CE, Dominguez SL, Solanoy HO, Drummond J, Zhang X, Ding X, Cai F, Song Q, Li X, Yue Z, van der Brug MP, Burdick DJ, Gunzner-Toste J, Chen H, Liu X, Estrada AA, et al. 2012. Ser1292 autophosphorylation is an indicator of Irrk2 kinase activity and contributes to the cellular effects of pd mutations. *Science Translational Medicine* **4**:3004485. DOI: <https://doi.org/10.1126/scitranslmed.3004485>

Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisán-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Geiger C, Houlden H, et al. 2009. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics* **41**:1308–1312. DOI: <https://doi.org/10.1038/ng.487>, PMID: 19915575

Steger M, Tonelli F, Ito G, Davies P, Trost M, Vetter M, Wachter S, Lorentzen E, Duddy G, Wilson S, Baptista MA, Fiske BK, Fell MJ, Morrow JA, Reith AD, Alessi DR, Mann M. 2016. Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *eLife* **5**:e12813. DOI: <https://doi.org/10.7554/eLife.12813>, PMID: 26824392

Tang BL. 2017. Rabs, membrane dynamics, and parkinson's disease. *Journal of Cellular Physiology* **232**:1626–1633. DOI: <https://doi.org/10.1002/jcp.25713>, PMID: 27925204

Thirstrup K, Dächsel JC, Oppermann FS, Williamson DS, Smith GP, Fog K, Christensen KV. 2017. Selective LRRK2 kinase inhibition reduces phosphorylation of endogenous Rab10 and Rab12 in human peripheral mononuclear blood cells. *Scientific Reports* **7**:10300. DOI: <https://doi.org/10.1038/s41598-017-10501-z>, PMID: 28860483

Thévenet J, Pescini Gobert R, Hooft van Huijsduijnen R, Wiessner C, Sagot YJ. 2011. Regulation of LRRK2 expression points to a functional role in human monocyte maturation. *PLoS One* **6**:e21519. DOI: <https://doi.org/10.1371/journal.pone.0021519>, PMID: 21738687

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**:731–740. DOI: <https://doi.org/10.1038/nmeth.3901>, PMID: 27348712

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martínez-Bartolomé S, et al. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32**:223–226. DOI: <https://doi.org/10.1038/nbt.2839>, PMID: 24727771

Wandinger-Ness A, Zerial M. 2014. Rab proteins and the compartmentalization of the endosomal system. *Cold Spring Harbor Perspectives in Biology* **6**:a022616. DOI: <https://doi.org/10.1101/cshperspect.a022616>, PMID: 25341920

Waschbüsch D, Michels H, Strassheim S, Ossendorf E, Kessler D, Gloeckner CJ, Barmekow A. 2014. LRRK2 transport is regulated by its novel interacting partner Rab32. *PLoS One* **9**:e111632. DOI: <https://doi.org/10.1371/journal.pone.0111632>, PMID: 25360523

Wiggin GR, Soloaga A, Foster JM, Murray-Tait V, Cohen P, Arthur JS. 2002. MSK1 and MSK2 are required for the mitogen- and stress-induced phosphorylation of CREB and ATF1 in fibroblasts. *Molecular and Cellular Biology* **22**:2871–2881. DOI: <https://doi.org/10.1128/MCB.22.8.2871-2881.2002>, PMID: 11909979

Wilson GR, Sim JC, McLean C, Giannandrea M, Galea CA, Riseley JR, Stephenson SE, Fitzpatrick E, Haas SA, Pope K, Hogan KJ, Gregg RG, Bromhead CJ, Wargowski DS, Lawrence CH, James PA, Churchyard A, Gao Y, Phelan DG, Gillies G, et al. 2014. Mutations in RAB39B cause X-linked intellectual disability and early-onset

### 3.1.4. Accurate MS-based Rab10 phosphorylation stoichiometry determination for LRRK2 activity in Parkinson's disease

Ozge Karayel<sup>1</sup>, Francesca Tonelli<sup>2</sup>, Sebastian Virreira Winter<sup>1</sup>, Phillip E. Geyer<sup>1</sup>, Ying Fan<sup>2</sup>, Esther M. Sammler<sup>2,3</sup>, Dario R. Alessi<sup>2</sup>, Martin Steger<sup>1,\*</sup> and Matthias Mann<sup>1,\*</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany <sup>2</sup> Medical Research Council Protein Phosphorylation and Ubiquitylation Unit, University of Dundee, Dundee, United Kingdom <sup>3</sup> Department of Neurology, School of Medicine, Ninewells Hospital, Ninewells Drive, Dundee, United Kingdom

\* Corresponding authors

Accepted (in press) in *Molecular & Cellular Proteomics* (2020)

The pathogenic gain of function mutations in the kinase LRRK2 is the predominant genetic cause of Parkinson's disease (PD). They increase its activity, resulting in augmented Rab10-Thr73 phosphorylation, and conversely, LRRK2 inhibition decreases pRab10 levels. Monitoring of the levels of Rab10-Thr73 phosphorylation could thus be used to track PD pathogenesis or onset and assess LRRK2 inhibitor efficacy in cells and various tissues. This is a crucial or even desperate need as clinical trials cannot proceed without such knowledge. In this study, we developed a high accuracy and sensitivity targeted mass spectrometry (MS)-based assay for determining Rab10-Thr73 phosphorylation stoichiometry in human samples. Our assay uses synthetic stable isotope-labeled (SIL) analogues for both phosphorylated and non-phosphorylated counterpart tryptic peptides to derive the percentage of Rab10 phosphorylation. It employs multiplexed selected ion monitoring (mxSIM) scans in which the endogenous peptides are separately admitted into an Orbitrap analyzer with the appropriate injection times that result in a rough equality between them for analysis. This measurement mode precisely and accurately quantified attomole amounts of the endogenous peptides. Using our assay, I measured Rab10-Thr73 phosphorylation stoichiometry in neutrophils of LRRK2 mutation carriers before and after LRRK2 inhibition. Compared to healthy controls, the pathogenic mutation carriers displayed about two fold increased pRab10 levels, from  $1.8 \pm 0.3\%$  to  $3.36 \pm 0.4\%$ . Our generic MS-based assay established the significance of pRab10 as a treatment efficacy and stratification marker. To our knowledge, this is the first time that phosphorylation stoichiometry is measured accurately in patient-derived cells and used as a surrogate parameter to monitor the activity of a pathogenic mutant enzyme. The use of our assay in clinical trials is already on the way.

## Accurate MS-based Rab10 phosphorylation stoichiometry determination as readout for LRRK2 activity in Parkinson's disease

Özge Karayel<sup>1</sup>, Francesca Tonelli<sup>2</sup>, Sebastian Virreira Winter<sup>1</sup>, Phillip E. Geyer<sup>1</sup>, Ying Fan<sup>2</sup>, Esther M. Sammler<sup>2,3</sup>, Dario R. Alessi<sup>2</sup>, Martin Steger<sup>1,†,\*</sup> and Matthias Mann<sup>1,\*</sup>

<sup>1</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

<sup>2</sup>Medical Research Council Protein Phosphorylation and Ubiquitylation Unit, University of Dundee, Dundee, United Kingdom

<sup>3</sup>Department of Neurology, School of Medicine, Ninewells Hospital, Ninewells Drive, Dundee, United Kingdom

<sup>†</sup>Current address: Evotec (München) GmbH, Am Klopferspitz 19a, Martinsried, Germany

\*To whom correspondence should be addressed:

[martin.steger@evotec.com](mailto:martin.steger@evotec.com) and [mmann@biochem.mpg.de](mailto:mmann@biochem.mpg.de)

**Running title:** MS-based clinical assay for PTM occupancy calculation

## Summary

Pathogenic mutations in the Leucine-rich repeat kinase 2 (LRRK2) are the predominant genetic cause of Parkinson's disease (PD). They increase its activity, resulting in augmented Rab10-Thr73 phosphorylation and conversely, LRRK2 inhibition decreases pRab10 levels. Currently, there is no assay to quantify pRab10 levels for drug target engagement or patient stratification. To meet this challenge, we developed an high accuracy and sensitivity targeted mass spectrometry (MS)-based assay for determining Rab10-Thr73 phosphorylation stoichiometry in human samples. It uses synthetic stable isotope-labeled (SIL) analogues for both phosphorylated and non-phosphorylated tryptic peptides surrounding Rab10-Thr73 to directly derive the percentage of Rab10 phosphorylation from attomole amounts of the endogenous phosphopeptide. The SIL and the endogenous phosphopeptides are separately admitted into an Orbitrap analyzer with the appropriate injection times. We test the reproducibility of our assay by determining Rab10-Thr73 phosphorylation stoichiometry in neutrophils of LRRK2 mutation carriers before and after LRRK2 inhibition. Compared to healthy controls, the PD predisposing mutation carriers LRRK2 G2019S and VPS35 D620N display 1.9-fold and 3.7-fold increased pRab10 levels, respectively. Our generic MS-based assay further establishes the relevance of pRab10 as a prognostic PD marker and is a powerful tool for determining LRRK2 inhibitor efficacy and for stratifying PD patients for LRRK2 inhibitor treatment.

Parkinson's disease (PD) is the second most common neurodegenerative conditions, and no disease-modifying therapies exist to date<sup>1,2</sup>. Although most PD cases are idiopathic, mutations in several genes have been linked to familial forms of the disease<sup>3</sup>. Among those, mutations in the Leucine-rich repeat kinase 2 (LRRK2) comprise the predominant genetic cause of PD and account for 1% of sporadic and 4% of familial cases worldwide, and much higher in some populations<sup>4</sup>. At least six pathogenic missense mutations in LRRK2, including the most frequent G2019S substitution, have been identified<sup>5</sup> and several studies confirmed that these mutations increase its kinase activity<sup>5-8</sup>. LRRK2-associated PD is clinically largely indistinguishable from idiopathic PD, suggesting that LRRK2 inhibition may be a useful a therapy for a larger group of patients<sup>4</sup>. Clinical trials with selective LRRK2 kinase inhibitors are ongoing and have already passed phase 1.

Downstream targets of the LRRK2 kinase have long been enigmatic and controversial, but using genetic mouse models, specific inhibitors and a novel phosphoproteomics workflow, we have recently identified and verified a subset of Rab GTPases (Rab3A/B/C/D, Rab5A/B/C, Rab8A/B, Rab10, Rab12, Rab29, Rab35, and Rab43) as bona fide substrates<sup>8,9</sup>. Among these, Rab10 appears to be a key physiological kinase substrate<sup>8-14</sup>. We found that LRRK2 directly phosphorylates Rab10 at Thr73 and all known pathogenic forms of LRRK2 enhance this phosphorylation<sup>8,9</sup>. Intriguingly, the PD-associated D620N mutation of the retromer complex protein VPS35 also activates LRRK2 kinase activity, which in turn results in augmented Rab10 phosphorylation<sup>15</sup>. Thus, multiple PD-associated factors are interconnected and dysregulation of a common LRRK2-Rab signaling pathway appears to be an underlying cause of PD.

The LRRK2 autophosphorylation site Ser1292 has been widely used for assessing LRRK2 kinase activity<sup>16-19</sup>. However, its stoichiometry (percentage of molecules phosphorylated) appear to be extremely low and there is no sensitive phospho-specific antibody available to reliably detect and quantify phosphorylation at this site<sup>6,20,21</sup>. We recently developed several high-affinity antibodies for detecting pRab proteins in cells and in tissues<sup>14</sup>. Among those, a highly specific clone detects pThr73 levels in human peripheral blood cells of (mutant) LRRK2 G2019S and VPS35-D620N carriers with Parkinson's disease<sup>11,15</sup>. While pRab10 levels were markedly increased in VPS35 D620N carriers, no statistically significant differences in pRab10 levels were detected by immunoblotting analysis when comparing controls to LRRK2-G2019S carriers in peripheral blood mononuclear cells (PBMCs) and neutrophils<sup>11,15</sup>. The reproducible quantification of immunoblots, particularly the detection of small (<2 fold-) changes is challenging. In contrast, mass spectrometry-based quantification has become a gold standard and has several advantages over traditional biochemical methods, as it is more specific and potentially more accurate. Importantly, MS allows simultaneous detection of both the phosphorylated peptide and the total protein pool and hence enables the

direct calculation of the absolute fraction of the phosphorylated protein, also known as phosphorylation stoichiometry or occupancy<sup>22</sup>.

There are several strategies differing in their accuracy, throughput, and applicability for measuring phosphorylation stoichiometry by MS<sup>22-27</sup>. One way is to compare the intensities of modified and unmodified peptides by label-free proteomics<sup>27</sup>. However, due to potential differences in ionization efficiencies, the MS signals of individual peptides cannot be directly compared to each other. Stoichiometry determination based on heavy-to-light ratios of stable isotope-labeled (SIL) analogues of phosphorylated and non-phosphorylated proteolytic peptides and their endogenous counterparts can overcome this problem and allow a much more sensitive and precise readout for monitoring kinase activity<sup>22,23</sup>. Alternatively, a SIL recombinant protein that is chemically or enzymatically phosphorylated can also be used as the spike-in standard<sup>28,29</sup>. MS strategies determining stoichiometry that are combined with tailored targeted methods are especially suited for accurate quantification of low levels of a given phosphorylated analyte<sup>30</sup>. In particular, instruments with an Orbitrap mass analyzer can be operated in targeted scan modes such as high resolution selected ion monitoring (SIM) or parallel reaction monitoring (PRM). In both methods, precursor ions are isolated with a narrow m/z range by a quadrupole mass filter before introduction into the Orbitrap analyzer, thereby providing increased signal to noise ratio for the ion of interest and allowing attomole-level limits of detection<sup>31,32</sup>. As no fragmentation is involved in SIM, quantification of selected ions relies on the high resolution in the Orbitrap mass analyzer alone. In contrast, in PRM, fragment ions are used for quantification of the peptide. While that approach is more specific, the overall sensitivity of PRM can be lower, as the signal of a given precursor ion is distributed across multiple fragments<sup>32</sup>.

Here we describe an accurate and highly-sensitive MS-assay for determining the Rab10-Thr73 phosphorylation stoichiometry and how it changes in Parkinson's disease. We evaluate it by comparing Rab10-Thr73 phosphorylation levels in healthy controls, idiopathic PD patients, and PD patients with defined genetic cause. Our assay enables the detection of subtle changes in Rab10 phosphorylation, which are beyond what is detectable by typical immunoassays. We show that pRab10 stoichiometry when measured precisely using our assay can serve as a robust target engagement and patient stratification marker in clinical studies.

## EXPERIMENTAL PROCEDURE

### Experimental Design and Statistical Rationale

The experimental design and statistical rationale for each of the experiments conducted in this study are described in detail in each subsection and figure legend. Briefly, all neutrophil samples isolated from individuals were measured with technical replicates (n=3). The LOQ experiment and IP experiments were also measured in technical replicates (n=3). For CV experiments, we repeated MS measurements (analytical), and the workflow in the same gel (intra-assay) or in different gels (inter-assay) using the same phosphoprotein standard (n=6).

### Study participants and blood sample collection

Our assay was applied in the neutrophils collected from 14 healthy controls, three idiopathic PD patients, four LRRK2 G2019S and three VPS35 D620N mutation carriers with PD.

For setting up and validating our assay, we recruited seven volunteers from within the Department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry who kindly donated blood for our study. The data shown in Figures 1 and 3 and Figures S1-3 are derived blood samples from healthy donors, which provided a written informed consent, with prior approval of the ethics committee of the Max Planck Society.

For the experiments shown in Figure 4, neutrophil lysates derived from either idiopathic PD patients, PD patients carrying a heterozygous LRRK2 G2019S or VPS35 D620N mutation or non-PD controls were used that had been previously used for publication (<sup>11</sup> and <sup>15</sup>, respectively, Table S4). All procedures were performed in compliance with the local ethics review boards and all participants provided informed consent. These lysates were subjected to MS analysis in a blinded experimental set-up, with the identity of the lysates only being revealed after completion of the MS analysis.

### Reagents

MLi-2 was purchased from Tocris Bioscience (cat# 5756). DIFP, HA-agarose and trypsin were from Sigma and LysC was from Wako. Microcystin-LR was from Enzo Life Sciences. Complete protease and phosphatase inhibitor tablets were from Roche. SIL peptides in absolute quantities (AQUA) were purchased from Sigma-Aldrich.

Antibodies

Anti-GAPDH (#5174), anti-HA (#3724), and anti-Rab10 (#8127) were from Cell Signaling Technologies. Rabbit monoclonal antibody for total LRRK2 (UDD3, ab133518), Rab10 (T73) antibody [MJF-R21] and polyclonal phospho-Rab (initially raised against Rab8) antibody [MJF-R20] that efficiently immunoprecipitated multiple LRRK2 phosphorylated Rab proteins were purchased from Abcam (Burlingame, California). They were custom-made by Abcam in collaboration with the Michael J. Fox Foundation (Burlingame, California)<sup>14,33</sup>.

Plasmids

HA-Rab10 (DU44250) and Flag-LRRK2-Y1699C (DU13165). Full datasheets and reagents are available on <https://mrcppureagents.dundee.ac.uk/>.

Stable isotope labeling and purification of the Rab10 protein standard

Buffer A: 50 mM HEPES pH 8.0, 500 mM LiCl, 1 mM MgCl<sub>2</sub>, 100 μM GDP, 1 mM TCEP and Elution buffer: Buffer A + 500 mM imidazole

To obtain unlabeled Rab10 standard, human His-tagged Rab10 (residues 1-175) construct from Rai et al.<sup>34</sup> was expressed in *E. coli* BL21 (DE3) harboring the GroEL/S plasmid and protein expression was induced with isopropyl β-D-1-thiogalactopyranoside overnight. Cells were lysed by sonication in a buffer containing 50 mM HEPES pH 8.0, 500 mM LiCl, 1 mM MgCl<sub>2</sub>, 100 μM GDP, 1 mM TCEP (Buffer A) supplemented with several protease inhibitors including 30 μM Antipain and 50 μM Chymostain. Proteins were purified by Ni-NTA affinity chromatography. Briefly, bound proteins eluted with an imidazole gradient (25mM-500mM) and imidazole was removed by washing with 500 μM ATP containing Buffer A. Further purification was done by ion-exchange chromatography (Q-Sepharose) followed by size exclusion chromatography using a Superdex 200 column. Peak fractions containing recombinant protein were pooled. Identity and purity of the standard protein were assessed by Maldi-TOF MS and SDS-PAGE. To obtain labeled Rab10 standard, we used an auxotrophic expression strain for arginine and lysine<sup>29,35</sup>. Cultures were grown in PA5052 minimal autoinduction media containing heavy Arg10 and Lys8 and cells were harvested for purification of SIL Rab10 protein standard as described earlier.

In vitro LRRK2 kinase assays

Recombinant LRRK2 G2019S (Invitrogen, PV4881) and Rab proteins were incubated at 30°C for 30 min in kinase assay buffer (100 mM Tris-HCl, pH 7.5, 50 mM MgCl<sub>2</sub>, 5 mM EGTA, 1 mM GDP, 10 mM DTT, 25 mM β-glycerol phosphate, 5mM Sodium orthovanadate, and 50 μM ATP). The reaction was terminated by addition of 2 μM HG-10-102-0.

Cell culture and transfection

HEK293 and MEFs (WT and LRRK2-R1441G) cells were cultured in Dulbecco's modified Eagle medium (Glutamax, Gibco) supplemented with 10% fetal calf serum, 100 U/ml penicillin and 100 μg/ml streptomycin. Transient transfections were performed 48 hours prior to cell lysis using polyethylenimine PEI (Polysciences). Transfected cells were subjected to DMSO or MLi-2 (dissolved in DMSO) treatments at concentrations and periods of time as indicated in each figure legend. All cells were tested for mycoplasma contamination and overexpressing lines were verified by western blot analysis.

Neutrophil isolation, characterization, treatments and lysis

The procedure for neutrophil collection was done as described in detail in our video article<sup>36</sup>. Pipetting of human blood were undertaken in a biological safety cabinet. Briefly, 10 mL of blood was collected into a blood collection tube and mixed gently by inverting tubes. Next, it was transferred into a 50 mL conical tube and mixed gently with 100 μL of EDTA Stock Solution (100 mM EDTA in phosphate-buffered saline (PBS)). Neutrophils were isolated by immune-magnetic negative isolation using the MACSxpress® Neutrophil Isolation Kit (Miltenyi Biotec, Cat# 130-104-434). For 8 ml of blood, one vial of 'Isolation Cocktail' (magnetic beads) from the neutrophil isolation kit, delivered as a lyophilized pellet, was reconstituted by adding 0.25 mL of Buffer A and 0.25 mL of Buffer B in that order. The cocktail was mixed by gently pipetting and added to 8 ml of whole blood in 15 ml falcon tube. The blood sample containing the cocktail was gently mixed by inversion and incubated at room temperature for 5 min. The falcon tube was next placed into the magnetic field of the MACSxpress Separator (# 130-098-308) for 15 min. The magnetically labeled cells (non-neutrophils) adhere to the wall of the tube while the aggregated erythrocytes sediment to the bottom. The supernatant (~ 7 ml) containing the enriched neutrophils was carefully pipetted into a new 15 ml falcon tube, avoiding touching the magnetic beads attached to the sides of the falcon tube as well as the red blood cells at the bottom of the tube. The supernatant containing the isolated neutrophils was centrifuged at 300×g for 10 min at room temperature (acceleration and deceleration is both 5 using a Beckman Coulter Allegra X-15R Centrifuge). To ensure the removal of erythrocytes, the pelleted neutrophil cells were resuspended in 10 ml of 1× Red Blood Cell Lysis Solution (# 130-094-183), incubated for 10

min at room temperature and centrifuged at 300×g for 10 min at room temperature. To assess the purity and recovery of the enriched neutrophils an aliquot was taken at this stage and we later fluorescently stained cells with CD14-PerCP, CD15-PE, CD16-APC, and CD193-FITC and analyzed by flow cytometry. Cell debris, dead cells were excluded from the analysis based on scatter signals and DAPI. Finally, the cell pellet was resuspended with 10 ml room temperature RPMI 1640 media by gentle pipetting. At this stage, purified cells were subjected to MLi-2 (dissolved in DMSO) treatment at concentrations and periods of time as indicated in each figure legend. An equivalent volume of DMSO was added to negative control samples. Following treatment, cells were pelleted through centrifugation at 500 g for 5 min. Cells were then resuspended in 5 ml of ice-cold PBS and centrifuged again at 500 g for 5 min. The supernatant was carefully removed by pipetting. Cells were lysed in 300  $\mu$ l of ice-cold NP-40 buffer (50 mM Tris-HCl, pH 7.5, 1% (v/v) Triton X-100, 1 mM EGTA, 1 mM sodium orthovanadate, 50 mM NaF, 0.1% (v/v) 2-mercaptoethanol, 10 mM 2-glycerophosphate, 5 mM sodium pyrophosphate, 0.1  $\mu$ g/ml microcystin-LR, 270 mM sucrose, 0.5 mM DIFP (Sigma, Cat# D0879) supplemented with protease and phosphatase inhibitors (Roche)). Lysate was clarified by centrifugation at 16,000 rpm for 15 min at 4°C after a liquid nitrogen freeze-thaw cycle. Protein concentration was measured using Bradford assay (Thermo Scientific), snap-frozen and stored at -80°C. The whole procedure from the collection of blood to the freezing of neutrophil lysates was up to 4 hours.

#### Immunoblot analysis

Cell lysates were mixed with 4× SDS-PAGE loading buffer [250 mM Tris-HCl, pH 6.8, 8% (w/v) SDS, 40% (v/v) glycerol, 0.02% (w/v) Bromophenol Blue and 4% (v/v) 2-mercaptoethanol] to final total protein concentration of 1  $\mu$ g/ $\mu$ l and heated at 85°C for 10 min. Samples were loaded onto NuPAGE Bis-Tris 4–12% gel (Life Technologies) and electrophoresed at 180 V for 1 h with MOPS SDS running buffer followed by transfer onto the nitrocellulose membrane (GE Healthcare, Amersham Protran Supported 0.45  $\mu$ m NC) at 100 V for 90 min on ice in the transfer buffer (48 mM Tris-HCl and 39 mM glycine). Membrane was then cropped into pieces: from top of the membrane to 75 kDa to incubate with rabbit anti-LRRK2 UDD3 antibody, from 75 to 30 kDa to incubate with rabbit anti-GAPDH antibody and from 30 kDa to the bottom of the membrane to incubate with rabbit monoclonal antibodies for anti-Rab10-pThr73 or anti-Rab10. Antibodies diluted in 5% (w/v) BSA in TBS-T to a final concentration of 1  $\mu$ g/ml. All blots were incubated in primary antibody overnight at 4°C. Prior to secondary antibody incubation, membranes were washed three times with TBS-T (20 mM Tris/HCl, pH 7.5, 150 mM NaCl and 0.2% (v/v) Tween 20) for 10 min each. Membranes were incubated with secondary antibody multiplexed with goat anti-rabbit diluted in 5% (w/v) non-fat dry milk (NFDM) in TBS-T (1:5000 dilution) for 1 h at room temperature. They were next

washed with TBS-T three times for 10 min. Protein bands were detected using an ECL solution (Amersham ECL Western Blotting Detection Reagents (GE Healthcare)) and the ImageQuant LAS 4000 imaging system. For the immunoblots shown in figure 4A and 4D, the membranes were developed using the LI-COR Odyssey CLx Western Blot imaging system.

#### Cells lysis and pull-downs

Cells were lysed in either NP-40 buffer (50 mM Tris-HCl, pH 7.5, 120 mM NaCl, 1 mM EDTA, 6 mM EGTA, 20mM NaF, 15 mM sodium pyrophosphate and 1 % NP-40 supplemented with protease and phosphatase inhibitors (Roche)) or Triton X-100 buffer (50 mM Tris-HCl, pH 7.5, 1% (v/v) Triton X-100, 1 mM EGTA, 1 mM sodium orthovanadate, 50 mM NaF, 0.1% (v/v) 2-mercaptoethanol, 10 mM 2-glycerophosphate, 5 mM sodium pyrophosphate, 0.1 µg/ml microcystin-LR, 270 mM sucrose, 0.5 mM DIFP (Sigma, Cat# D0879) supplemented with protease and phosphatase inhibitors (Roche)). Lysates were clarified by centrifugation at 16,000 rpm for 15 min at 4°C after a liquid nitrogen freeze-thaw cycle. Protein concentrations were measured using Bradford assay (Thermo Scientific), snap-frozen and stored at -80°C. For HA pulldowns, lysates were incubated with HA-agarose resin for 2 h (25 µl resin per 100 µg of lysates). For immunoprecipitation using total Rab10 or phospho-Rab, lysates were incubated with antibodies in manufacturer's recommend dilutions or concentrations overnight at 4°C and subsequently incubated with 25 µl of Protein-A/G-agarose beads for 2 h at 4°C. To remove unspecific binders, beads were washed twice with matching lysis buffer and twice with 50 mM Tris-HCl (pH 7.5). SIL phosphorylated Rab10 protein was mixed with the beads at this step if used as standard. Washes were followed by on-bead digestion overnight at 37°C with trypsin (~500 ng/pulldown in urea buffer [2M urea dissolved in 50 mM ammonium bicarbonate] or SDC buffer [1% (w/v) SDC in 100 mM Tris-HCl pH 8.5]). The resulting peptides were processed as described in 'LC-MS/MS sample preparation'.

#### In-Gel digestion protocol

Cell lysates were mixed with 4 × SDS/PAGE sample buffer (250 mM Tris/HCl, pH 6.8, 8% (w/v) SDS, 40% (v/v) glycerol, 0.02% (w/v) Bromophenol Blue and 4% (v/v) 2-mercaptoethanol) and heated at 85°C for 5 min. SIL Rab10 phosphoprotein was mixed with samples at this step if used as standard. 20-30 µg sample (5 µg per MS analysis) was loaded onto NuPAGE Bis-Tris 4-12% gel (Life Technologies) and electrophoresed at 180 V. After SDS-PAGE, gel was washed once with deionized water and stained with 0.1% Coomassie Blue R250 in 10% acetic acid, 40% methanol and 60% deionized water for 20 min and subsequently destained by soaking for at least 2 h in 10% acetic acid, 40% methanol, and 60% deionized

water with at least two changes of the solvent (until the background is nearly clear). Gel band corresponding to 20-30 kDa is excised and chopped into smaller pieces (~ 1 x 1 mm) and placed in clean 1.5 ml tubes. Gel pieces are washed two or three times with 50% 50 mM ABC / 50% EtOH for 20 min at RT and then completely dehydrated by incubating for 10 min in absolute EtOH. Samples were dried in a speed-vac for 10 min (45°C) until the gel pieces were bouncing in the tube. Gel pieces were rehydrated in 300 µl of 1% (w/v) SDC buffer (10 mM TCEP, 40 mM CAA, trypsin in 100 mM Tris-HCl pH 8.5) per sample and placed at 37°C overnight. The next day, 300 µl of isopropanol buffer (1% TFA in isopropanol) was added and shaken for 10 min and spin down. The liquid was transferred into a fresh tube. Further 200 µl of isopropanol buffer was added to the gel pieces and shaken vigorously for 20 min at RT. It was combined with that from the previous step. Samples were directly loaded onto SDB-RPS stage tips and processed as described in 'LC-MS/MS sample preparation'.

#### Human Neutrophil Proteome Digestion, In-StageTip Purification and Fractionation

The neutrophil cell pellet was prepared with the iST Kit for proteomic sample preparation (P.O. 00001, PreOmics GmbH). In brief, this involved denaturation, alkylation, digestion and peptide purification. Reduction of disulfide bridges, cysteine alkylation and protein denaturation was performed at 95°C for 10 min. After a 5 min cooling step at room temperature, trypsin and LysC were added to the mixture at a ratio of 1:100 micrograms of enzyme to micrograms of protein. Digestion was performed at 37°C for 1 h. 20 µg of peptides was loaded on two 14-gauge SDB-RPS StageTip plugs. Samples were directly loaded onto SDB-RPS StageTips processed as described in 'LC-MS/MS sample preparation'. Clean peptides were separated using the high-pH reversed-phase 'Spider fractionator' into 24 fractions as described previously to generate deep proteomes<sup>37</sup>.

#### LC-MS/MS sample preparation

StageTips<sup>38</sup> were prepared by inserting two 16-gauge layers of a SDB-RPS matrix (Empore) into a 200 µl pipette tip using an in-house prepared syringe device as described previously<sup>39</sup>. The StageTips were centrifuged using an in-house 3D-printed StageTip centrifugal device at 1500 g. The acidified peptides (1% TFA v/v) were loaded onto the StageTips that were later washed with 1% TFA in isopropanol and subsequently 2% ACN/0.2% TFA. Elution was performed using 60 µl of 50% ACN/1.25% NH<sub>4</sub>OH or 80% ACN/1.25 % NH<sub>4</sub>OH. Eluates were collected in PCR tubes and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus) at 60°C. Peptides were resuspended in buffer A\* (2% ACN/0.1% TFA) and briefly sonicated (Branson Ultrasonics) before LC/MS-MS analysis. To calculate absolute Rab10-pThr73

occupancy, we spiked SIL phosphorylated and non-phosphorylated counterpart peptides into samples at this step. Light contamination in heavy standards was not detected up to 100 fmol of spike-in amount (the highest amount tested). Given that the detection limits of these peptides with our mxSIM assay were below 50 amol, light contamination is determined to be <0.05%.

LC-MS/MS measurements

Peptides were loaded on a 20 or 50 cm reversed phase column (75  $\mu$ m inner diameter, packed in house with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin (Dr. Maisch GmbH)). Column temperature was maintained at 60°C using a homemade column oven. An EASY-nLC 1200 system (Thermo Fisher Scientific) was directly coupled online with the mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) via a nano-electrospray source, and peptides were separated with a binary buffer system of buffer A (0.1% formic acid (FA)) and buffer B (80% acetonitrile plus 0.1% FA), at a flow rate of 300 nL/min. Peptides were eluted with a 45 min gradient of 5-60% buffer B (0.1% (v/v) FA, 80% (v/v) ACN). After each gradient, the column was washed with 95% buffer B for 5 min.

The mass spectrometer was programmed to acquire in targeted scan mode in which every full scan with resolution 60,000 at 200 m/z ( $3 \times 10^6$  ions accumulated with a maximum injection time of 20 ms) was followed by two multiplexed selected ion monitoring (mxSIM) scans employing multiplexing degree of two to record both light (endogenous) and heavy counterpart simultaneously for either phosphorylated or non-phosphorylated Rab10-pT73 tryptic peptides. Each SIM scan covered a range of m/z 150–2000 with resolution 120,000 ( $10^5$  ions accumulated with a maximum injection time of 230 ms for both light and heavy counterparts, 1.4 m/z isolation window and 0.4 m/z isolation offset). m/z values of doubly-charged Rab10-pT73 tryptic (FH<sup>3</sup>TITTSYYR) light and heavy peptides (arginine labeled (13C, 15N)) were defined as follows: 684.8028 and 689.8070 whereas doubly-charged Rab10 non-phosphorylated tryptic (FHTITTSYYR) light and heavy peptides (arginine labeled (13C, 15N)) were targeted at m/z of 644.8197 and 649.8238.

For the LOD experiment (Figure 1), the mass spectrometer was programmed to acquire in either full scan mode alone or SIM or PRM combined with a full scan. Full scan acquisition was performed with a resolution of 120,000 at 200 m/z ( $3 \times 10^6$  ions accumulated with a maximum injection time of 230 ms) to cover the scan range of 350-1,650 m/z. SIM acquisition was performed using a resolution of 120,000 at 200 m/z, isolation windows of 1.4 m/z with 0.4 m/z offset, target AGC values of  $2 \times 10^5$ , and a maximum injection time of 230 ms. PRM acquisition was performed using a resolution of 60,000 at 200 m/z, isolation

windows of 1.4 m/z with 0.4 m/z offset, target AGC values of  $2 \times 10^5$ , and a maximum injection time of 130 ms. Fragmentation was performed with a normalized collision energy of 27.

The neutrophil proteomes were analyzed using an LC-MS instrumentation consisting of an EASY-nLC 1200 system (Thermo Fisher Scientific) combined with a Q Exactive HF Orbitrap (Thermo Fisher Scientific) and a nano-electrospray ion source (Thermo Fisher Scientific). The purified peptides were separated on a 50 cm HPLC column (75  $\mu$ m inner diameter, in-house packed into the tip with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin (Dr. Maisch GmbH)). Of each of the 24 fractions around 0.5  $\mu$ g peptides were analyzed with a 45 min gradient. Peptides were loaded in buffer A (0.1% FA, 5% DMSO (v/v)) and eluted with a linear 35 min gradient of 3-30% of buffer B (0.1% FA, 5% DMSO, 80% (v/v) ACN), followed by a 7 min increase to 75% of buffer B and a 1 min increase to 98% of buffer B, and a 2 min wash of 98% buffer B at a flow rate of 450 nL/min. Column temperature was kept at 60°C by a Peltier element containing in-house developed oven. MS data were acquired with a Top15 data-dependent MS/MS scan method (topN method). Target values for the full scan MS spectra was  $3 \times 10^6$  charges in the 300-1,650 m/z range with a maximum injection time of 55 ms and a resolution of 120,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27. MS/MS scans were performed at a resolution of 15,000 at m/z 200 with an ion target value of  $5 \times 10^4$  and a maximum injection time of 25 ms.

For intact mass analysis, proteins were loaded on a reversed-phase column (Phenomenex AerisTM 3.6  $\mu$ m WidePore C4 100 mm x 2.1 mm inner diameter, 200 Å pore size). An Agilent 1100 HPLC system was coupled online with the mass spectrometer (microTOF, Bruker Daltonik) and masses were recorded from 800-3,000 m/z. Proteins were separated with a binary buffer system of buffer A (0.05% TFA in H<sub>2</sub>O, pH 2.0) and buffer B (0.05% TFA in can, pH 2.0) at a flow rate of 250 nL/min. Proteins were eluted with a gradient of 20-80% buffer B in 20 min. After each gradient, the column was washed with 95% buffer B for one minute and 20% buffer B for 3 min. Data were processed using the CompassTM 'DataAnalysis' software from Bruker Daltonik, deconvoluted with 'MaximumEntropy' and an instrument resolving power of 10,000.

#### Data analysis and development of mxSIM assay to calculate phosphorylation stoichiometry

The type of our targeted occupancy assay is a Tier 2, according to previously published guidelines <sup>40</sup>. For the calculation of absolute Rab10-pThr73 occupancies, raw MS data were processed using Skyline 4.2 (0.19072) <sup>41</sup> which is an open source software project and can be freely installed. Raw files were directly imported into Skyline in their native file format. After data import, graphical displays of chromatographic

traces were manually inspected for proper peak picking of MS1 filtered endogenous peptides based on co-eluting stable isotope-labeled peptides. All quantitation performed for phosphorylation occupancy calculations in this study were done on the precursor ion level. Only the most abundant first two peaks of the isotope cluster were used for quantitation. Peptide areas (AUCs) for the non-phosphorylated tryptic Rab10 peptide (FHTITTSYYR, m/z 644.8197++) and the phosphorylated tryptic Rab10 peptide (FH<sup>p</sup>TITTSYYR, m/z 684.8028++) with their R=13C615N4 heavy analogues were extracted to derive light-to-heavy ratios. The absolute quantification was determined by comparing the abundance of the known SIL internal standard peptides with the native peptides. The phosphorylation stoichiometry/occupancy is calculated by taking the ratio of the total amount of phosphorylated fraction to the total amount of both phosphorylated and non-phosphorylated forms, which is always represented as percentage (%). When using mxPRM method, the median ratio of all unique y ions that belong to the phosphorylated and nonphosphorylated tryptic Rab10 peptides was used to calculate occupancy. All details for occupancy calculations in this study are provided in the Table S3.

For the deep proteome of human neutrophils, raw MS data were processed using MaxQuant version 1.5.6.8<sup>41,42</sup> against the Human UniProt FASTA database containing 21,052 entries (UniProt, release 2017) and a list of 245 potential contaminants. Enzyme specificity was set to trypsin, and the search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein and oxidation of methionine as variable modifications. Up to two missed cleavages were allowed for protease digestion, and peptides had to be fully tryptic. MaxQuant uses individual mass tolerances for each peptide, whereas the initial maximum precursor mass tolerances were set to 20 ppm in the first search and 4.5 ppm in the main search, and the fragment mass tolerance was set to 20 ppm. The false discovery rate was controlled with a target-decoy approach at less than 1% for peptide spectrum matches and less than 1% for protein group identifications.

Bioinformatic analyses in this study were performed with Perseus ([www.perseus-framework.org](http://www.perseus-framework.org))<sup>43</sup>, Microsoft Excel and data visualized using GraphPad Prism (GraphPad Software) or RStudio (<https://www.rstudio.com/>).

#### Data Availability

Proteomics raw data have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the data set identifier PXD015219 (Username: [reviewer48252@ebi.ac.uk](mailto:reviewer48252@ebi.ac.uk) and Password: GCE1hjuR) and PXD019814 (Username: [reviewer00895@ebi.ac.uk](mailto:reviewer00895@ebi.ac.uk) and Password: gVVv406k) and at Panorama Web

(<https://panoramaweb.org/pRab10.url>, username: `panorama+mpi3@proteinms.net`, password: `1kHhN#Pk`). Annotated spectra for the results can be viewed through MS-Viewer using the search key `zbcwt6pszv`<sup>44</sup>.

## RESULTS

### Rab10-pThr73 serves as a readout for LRRK2 activity in human peripheral blood neutrophils

We decided to monitor Rab phosphorylation in human peripheral blood neutrophils, as these can be sampled in a minimally invasive way. To explore which Rab GTPases were expressed in these cells and which of them could serve as a readout for LRRK2 activity in a quantitative MS-based assay, we first isolated neutrophils from whole blood using a negative selection approach. We assessed the recovery and the purity of the enriched cells by flow cytometry and found it to be more than 98%, with > 96% of them viable (Figure S1). Proteomic analysis resulted in 5,488 quantified proteins, for which we estimated the copy numbers per cell using the proteomic ruler approach<sup>45</sup> (Figure 1A and Table S1). Encouragingly, LRRK2, the vesicular protein VPS35 and many of the Rab proteins that are LRRK2 substrates were in the highest quartile of the abundance-ranked proteome (Figure 1B). The bona-fide LRRK2 substrate Rab10 had the highest copy number among all detected Rab proteins with an estimated 1,820,000 copies per cell, suggesting its suitability as a marker for LRRK2 activity in PD (Figure S2).

To determine whether Rab10 is the only Rab family member that is phosphorylated by LRRK2 in neutrophils, we first treated freshly isolated cells with the selective LRRK2 inhibitor MLi-2 and confirmed the downregulation of Rab10-pThr73 by immunoblotting<sup>11</sup> (Figure 1C). In the same lysates, we enriched phosphorylated Rab proteins with the previously described pThr-specific Rab antibody and subjected the eluates to LC-MS/MS analysis<sup>9</sup>. We identified 19 Rab GTPases in total, of which 7 were previously shown to be LRRK2 targets<sup>9</sup> (Figure 1D, Figure S3, Table S2). In two independent pThr-specific Rab antibody pulldown experiments, Rab10, Rab43 and Rab8a phospho-protein levels decreased significantly upon MLi-2 treatment, indicating that these Rabs are targeted by LRRK2 in this system. To directly confirm LRRK2-mediated phosphorylation of these proteins, we immunoprecipitated them and quantified Rab10-pThr73, Rab43-pThr82 and Rab8a-pThr72 phosphopeptides by targeted mass spectrometry (Selected Ion Monitoring, SIM, see below). Only Rab10 and Rab43 phosphorylation sites were downregulated more than 4-fold after LRRK2 inhibitor treatment, demonstrating that the phosphorylation status of these Rabs can be used as readout for LRRK2 kinase activity (Figure 1E).

The high abundance of Rab10 in neutrophils and the promise of pThr73 as a biomarker for Parkinson's disease encouraged us to develop a highly accurate and sensitive targeted MS-based assay for quantifying phosphorylated Rab10 in human cells at the peptide level. To maximize sensitivity, we explored a multiplexed SIM (mxSIM) setup on a linear quadrupole Orbitrap instrument (Q Exactive HF-X)<sup>46</sup>. In our strategy, the SIL analogue of the phosphorylated tryptic Rab10-Thr73 peptide acts as a sentinel peptide, as it can be spiked-in in high amounts and elutes simultaneously with its endogenous light counterpart. Light (endogenous) as well as SIL counterpart phosphopeptides are consecutively isolated by narrow quadrupole isolation windows but simultaneously injected into the Orbitrap mass analyzer. We set the maximum total ion accumulation time to 230 ms, but allocate 90% of this to the endogenous phosphopeptide, thus boosting its signal and increasing the sensitivity of our assay many-fold.

To compare the sensitivity and the accuracy of (mx)SIM with regular full-MS scanning and with PRM, we mixed variable amounts of the SIL Rab10 phosphopeptide (10 amol to 50 fmol) with 50 ng of a tryptic HeLa digest and measured the SIL pRab10 peptide intensity using our different scan protocols. Relative to full-MS scanning, in which the entire mass range (300-1650 m/z) is analyzed, SIM, either multiplexed or not, provided a 20-fold increase in sensitivity with a limit of detection (LOD) of 50 amol (Figure 1F). PRM performed equally well in terms of sensitivity, however, SIM had a somewhat higher quantification accuracy ( $R^2$  of 0.992 vs. 0.983). For this reason, and because it was sufficiently specific in our system, we decided to develop a Rab10-pThr73 quantification assay based on mxSIM.

To determine the maximum heavy-to-light ratio and the limit of quantification (LOQ) of our method, we mixed 25 fmol of the light pRab10 peptide with variable amounts of its heavy counterpart (10 amol to 50 fmol) in 50 ng of HeLa digest. The results for mxSIM indicate excellent reproducibility of quantification also in the multiplexed case ( $R^2 = 0.997$ ) (Figure 1G). Due to the differential filling strategy, we accurately quantified heavy-to-light ratios of Rab10 phosphopeptides of up to 1:500 (25 fmol light and 50 amol heavy Rab10 peptide).

#### **mxSIM precisely accurately determines Rab10-Thr73 phosphorylation stoichiometry**

To evaluate protein and peptide-centric approaches for determining Rab10-Thr73 phosphorylation stoichiometry, we expressed and purified Rab10 (1-175 aa) from an auxotroph *E.coli* strain, which allowed for incorporation of SIL lysine and arginine into the newly synthesized protein<sup>29,35</sup>. We were able to phosphorylate 50% of the recombinant protein by LRRK2, as shown by intact mass analysis, and bottom-

up proteomics confirmed Thr73 as the phosphorylation site (Figure S4A and S4B). This is a very suitable proportion as both phosphorylated and non-phosphorylated peptides are needed as standards and we therefore decided to use this SIL recombinant phosphoprotein for quantifying the percentage of pRab10 in cells. For this purpose, we immunoprecipitated HA-Rab10 from LRRK2-Y1699C-expressing HEK293 cells, either treated with MLi-2 or not, and mixed the enriched protein with our SIL standard before joint tryptic digestion (Figure S5A-B). We then derived the Rab10-Thr73 phosphorylation stoichiometry, as described above, to be  $66.6 \pm 1.5\%$ . Subsequently, we used the same measurement but calculated the occupancy with the peptide-centric approach, yielding  $67.2 \pm 1.1\%$  occupancy. As both approaches gave nearly identical results, we decided to use the simpler peptide centric approach for all further calculations (Figure 2B).

Detection of sub-stoichiometric, post-translationally modified peptides in complex mixtures by proteomics is challenging, even with very sensitive targeted methods, and requires one or more upfront enrichment steps. To select a suitable antibody for enriching Rab10, we compared one recognizing the total protein with one recognizing the HA-epitope tag in HA-Rab10 expressing cells. Unexpectedly, the Rab10 directed antibody yielded significantly lower apparent pThr73 occupancy as compared to the antibody directed against the epitope tag ( $50.9 \pm 0.5\%$  vs  $67.2 \pm 1.1\%$ ). An likely explanation for this discrepancy could be that the anti-Rab10 antibody preferentially recognizes the non-phosphorylated fraction of the total protein pool (Figure 2B and C). To address our challenge with a different approach that does not rely on antibodies, we separated the cell lysate mixed with the SIL phosphoprotein standard on SDS-PAGE, excised the region of  $\sim 15\text{-}30$  kDa and digested the proteins using trypsin, followed by mxSIM analysis. This resulted in a measured Rab10-pThr73 occupancy of  $70.8 \pm 1.0\%$ , in almost perfect agreement with the occupancy obtained by the anti-HA immunoprecipitation approach (Figure 2B-C). Finally, we used SIL phosphorylated and non-phosphorylated peptides for deriving the Rab10-Thr73 phosphosite occupancy and again found that  $69.7 \pm 1.3\%$  of the protein was phosphorylated in the same cell lysate; again an almost identical occupancy value as the one obtained from spike-in of the SIL phosphoprotein (Figure 2C). We therefore decided to combine gel-separation, SIL peptide spike-in and the peptide-centric calculation method. With these tools in hand, we extended our in-gel digestion workflow to endogenous Rab10-pThr73 occupancy determination. We found that  $58.0 \pm 0.9\%$  of the protein was phosphorylated in LRRK2-Y1699C-transfected cells, whereas the treatment with MLi-2 almost completely abolished the phospho-occupancy ( $0.89 \pm 0.03\%$ ) (Figure 2D).

To further benchmark our assay, we incubated recombinant Rab10 (1-175 aa) with LRRK2 and stopped the phosphorylation reaction by adding the LRRK2 inhibitor HG-10-102-01 at defined time intervals. Next, we

determined the percentages of the phosphorylated Rab10 proteins by intact mass spectrometry and compared these values to the occupancies obtained using our mxSIM method (Figure S6A-D). This experiment revealed an excellent correlation between these methods in two independent experiments with  $R^2$  of 0.984 and 0.994 (Figure 2E). To specifically test the SIL peptide spike-in part of the approach, we mixed SIL phosphorylated and non-phosphorylated Rab10 peptides in 1:9, 1:2.3, 1:1, 2.3:1 and 9:1 ratios to mimic phosphosite occupancies of 10, 30, 50, 70 and 90%, respectively, and mixed them with our recombinant Rab10 phosphoprotein. We then derived the Rab10-pThr73 occupancy based on the heavy-to-light ratios of both phosphorylated and non-phosphorylated peptides (Figure 2F). The calculated mean occupancy for our recombinant Rab10 phosphoprotein was  $75.3 \pm 1.5\%$ , for mxSIM and 70% for intact mass analysis. We also measured the Rab10-pThr73 occupancy of the same recombinant protein and obtained  $77.4 \pm 0.4\%$  for PRM and  $77.6 \pm 0.6\%$  for mxSIM (Figure S7A). Finally, we determined the analytical, the intra- and the inter-assay variabilities, which yielded excellent coefficients of variations (CVs) of 1.9%, 8.7% and 13.0%, respectively (Figure 2G). This demonstrates that our mxSIM assay is highly accurate and reproducible.

#### In-gel digestion combined with mxSIM can detect subtle changes in Rab10-Thr73 phosphorylation within cells

Most pathogenic LRRK2 mutations, including R1441C/G/H, increase LRRK2 kinase activity and significantly stimulate Rab10 protein phosphorylation in mouse and human cells and tissues<sup>6-9</sup>. To determine whether our assay was sufficiently accurate and robust to detect small differences of LRRK2 activity in cells, we treated WT and LRRK2 R1441G knock-in mouse embryonic fibroblasts (MEFs) with increasing concentrations of MLi-2 and determined Rab10-Thr73 phosphorylation occupancies (Figure 3A). In parallel, we controlled LRRK2 inhibitor efficacy by immunoblotting and probing for Rab10-pThr73 (Figure 3B). Compared to WT, in which the Rab10-pThr73 occupancy was  $12.1 \pm 0.6\%$ , we found a 2.45-fold increase in R1441G ( $29.7 \pm 0.9\%$ ) (Figure 3C). Rab10-pThr73 occupancy already decreased by 1.5-fold ( $18.8 \pm 1.5\%$ ) upon treatment with 1 nM of MLi-2 and by almost 2-fold ( $15.4 \pm 0.7\%$ ) with 3 nM of the inhibitor (Figure 3B-C). The corresponding Rab10-pThr73 immunoblot signals also showed some decrease, however, reliable and precise quantification of these bands was difficult.

We reasoned that our data should also allow determination Rab10-pThr73 IC<sub>50</sub> of MLi-2 directly from the occupancies. We extracted a value of 3 nM, well within the 3-10 nM range estimated by our previous phosphotag Rab10 analysis (Figure 3D)<sup>12</sup>. Together, our results establish that our assay can reliably detect very small differences in phosphorylation occupancy in cultured cells.

We next addressed our main objective: to determine whether our assay is suitable and sufficiently sensitive for quantifying the percentage of Rab10-pThr73 in human peripheral blood. To this end, we isolated neutrophils from five healthy volunteers. Upon treatment with either 30 nM or 100 nM of MLi-2, we found a decrease in Rab10-pThr73 by immunoblotting (Figure 3E). Judged by this, the level of phosphorylated Rab10 protein was very low in all individuals. Applying our workflow in the manner described above, showed that the median occupancy was  $2.3 \pm 0.2\%$  in DMSO-treated cells and that it decreased to  $1.3 \pm 0.1\%$  after 30 nM and to  $0.6 \pm 0.1\%$  after 100 nM MLi-2 treatment (Figure 3F). These occupancies correspond to  $225,500 \pm 15,500$  (DMSO),  $115,270 \pm 6,190$  (30 nM MLi-2) and  $60,160 \pm 4,600$  (100 nM MLi-2) phosphorylated Rab10 molecules per cells (Table S3). We independently confirmed these results by PRM (Figure S7B).

Intriguingly, the residual Rab10-pThr73 signal after MLi-2 treatment suggested that a kinase other than LRRK2 acts on this site. This finding raises interesting questions about the disease mechanism but in our current context, it presents a source of variability. To account for this, we corrected the occupancies of each untreated sample by subtracting the corresponding baseline occupancy (after MLi-2-treatment). This better separated disease phenotypes in clinical cohorts (Figure 3F). Finally, we measured a MLi-2 dose-response curve (1-100 nM) using neutrophils from two donors and determined their occupancy based- $IC_{50}$  values. These were very similar in both donors (2.4 nM and 3 nM), further demonstrating the high reproducibility of our method even *in vivo* (Fig 3G and H).

#### PD patients have increased Rab10-Thr73 phosphorylation stoichiometry

The frequency of the LRRK2 G2019S mutation is 1% in patients with sporadic PD and 4% in hereditary PD patients with up to 30-40% in certain populations such as Ashkenazi Jews and North African Berbers<sup>4</sup>. To investigate the central question of whether our assay can stratify PD patients with elevated LRRK2 activity in clinical trials, we analyzed Rab10-pThr73 levels in neutrophils of four LRRK2-G2019S carriers with PD, together with an equal number of healthy controls in a blinded experimental setup (see Experimental Procedure, Table S4). Each neutrophil population was treated with DMSO or 100 nM MLi-2 and all samples were subjected to quantitative immunoblot analysis for Rab10-pThr73 levels, revealing a clear reduction in Rab10 phosphorylation upon LRRK2 inhibition (Figure 4A). LRRK2-pSer935 levels were also monitored by immunoblot analysis in the lysates to confirm the efficacy of MLi-2 treatment. Total Rab10 levels did not differ between the various genotypes (Figure 4A). We quantified Rab10 phosphorylation by normalizing it with either total Rab10 or GAPDH (Figure 4B). Although the data show

that Rab10-pThr73 levels are slightly increased in G2019S carriers when normalized with GAPDH, the observed differences did not reach statistical significance (Figure 4B). The median occupancy of normalized Rab10-pThr73 was  $1.8 \pm 0.3\%$  in healthy controls and increased by almost 2-fold ( $3.36 \pm 0.4\%$ ) in G2019S carriers (Figure 4C-D). In MLi-2 treated samples, the occupancy decreased 5-fold on average for both groups. Despite a good correlation between the two measurements ( $R = 0.59$ ), MS-based quantification was better at separating controls from G2019S carriers ( $p = 0.019$ ) (Figure 4E).

Our previous work demonstrated that VPS35 controlled LRRK2 activity and that the D620N substitution resulted in increased LRRK2 activity as assessed by monitoring phospho-Rabs levels<sup>15</sup>. To test if our assay can determine LRRK2 activity in PD patients with the VPS35 D620N mutation, we measured Rab10-Thr73 phosphorylation occupancies in neutrophils isolated from three heterozygous patients, as well as three age-matched idiopathic PD patients and three non-PD controls (see Experimental Procedure, Table S4). We treated the neutrophils from each subject with either DMSO or 200 nM MLi-2 before cell lysis. In accordance with our previous work, immunoblot analysis revealed significantly elevated Rab10-pThr73 levels in VPS35 D620N patients (Figure 4F-G). The median occupancies were  $0.6 \pm 0.1\%$ ,  $0.8 \pm 0.1\%$  and  $1.9 \pm 0.5\%$  in healthy controls, idiopathic and heterozygous VPS35 D620N patients, respectively, and decreased 6.9-fold on average in MLi-2 treated cells (Figure 4H). Our MS-based assay also confirmed significantly higher levels of Rab10-pThr73 levels in VPS35 D620N patients compared to controls and idiopathic cases and we observed a very good correlation between the two measurements ( $R = 0.79$ ) (Figure 4I-J). Compared to controls, there was 1.5-fold increase in Rab10-pThr73 levels in idiopathic PD with both assays, however, the coefficient of variation in the control group was 30% with the mxSIM occupancy assay and 51% by immunoblotting (Figure 4G- and 4I).

In conclusion, we here determined pRab10 occupancies in neutrophils of 7 healthy controls and 7 PD patients with defined mutations (four LRRK2 G2019S and three VPS35 D620N). Compared to healthy controls, neutrophils of mutation carriers with PD robustly displayed a higher fraction of phosphorylated Rab10, namely a 2.2-fold increase in pRab10 levels ( $p$ -value  $<0.05$ ) (Figure 4K). These results demonstrate that our assay can accurately ascertain the LRRK2 kinase activity in human derived samples, including its increase in mutation carriers, and suggest that it can be applied to stratify patients.

## DISCUSSION

Here, we have established an accurate and highly-sensitive MS-based assay for determining the percentage of Rab10-Thr73 phosphorylation in samples collected from PD patients. Using stable-isotope labeled spike-

in peptides and a differential filling strategy for an Orbitrap analyzer, our range of quantification extended down to 50 attomoles and we were able to capture phosphorylation differences as small as those characterize controls and mutation carriers (1.26% to 2.73 in our human neutrophil samples). We confirmed Rab10-Thr73 phosphorylation as a direct readout for LRRK2 activity and specifically applied our assay to monitor LRRK2 activity in human neutrophils. We found that these cells contain relatively high levels of both LRRK2 and Rab10 and that their isolation is straightforward, although rapid isolation with minimal protease activation is important<sup>11</sup>. As a next step, it will be interesting to investigate whether pRab10 can be detected in post-mortem brain tissues or bodily fluids such as cerebrospinal fluid (CSF) and urine. Large and readily accessible collections of these bio-fluids could be used to further establish pRab10 as a bona-fide biomarker for PD disease progression.

After careful evaluation of spike-in standard type (SIL phosphoprotein vs peptides) and occupancy calculation approach (protein- vs peptide-centric approach), we opted for the peptide-centric approach in conjunction with the use of SIL peptides standards. Although spike-in of a labeled protein has the advantage of automatically accounting for the digestion efficiency, addition of labeled peptides makes the occupancy determination more generally applicable, as labeled peptides can readily be obtained at very high purity.

Furthermore, our assay utilizes separation of Rab10 by SDS-PAGE, which is followed by in-gel digestion. This strategy substantially enriches Rab10 and allows quantifying Rab phosphorylation with high sensitivity. When applied to very large cohorts, a limitation of our assay is its limited throughput. To solve this, the in-gel step in our workflow should be eliminated and as alternative antibodies that recognize both GDP and GTP bound Rab10 could be employed. There are considerable efforts currently being invested into producing these tools<sup>14</sup>. Alternatively, non-hydrolysable GTP analogues could be used to enrich Rab proteins before MS analysis.

We determined pRab10 occupancy in neutrophils of 14 healthy controls, three idiopathic PD patients, four LRRK2 G2019S and three VPS35 D620N mutation carriers with PD. Despite the small number of patients and the large intragroup variability, the difference between control and PD cases with LRRK2 G2019S and VPS35 D620N mutations reached statistical significance. Analysis of larger cohorts and inclusion of patient samples with higher LRRK2 activity, such as the R1441G/C mutations, should add statistical power and further establish pRab10 occupancy as a bona-fide marker of LRRK2 activity in PD. It will be exciting to investigate whether PD patients with increased pRab10 levels would also benefit from LRRK2 inhibitor treatment.

We find that the levels of Rab10-Thr73 phosphorylation are stable but very low in human neutrophils. This is also true for PD-associated LRRK2-G2019S mutation carriers where activation of the LRRK2 kinase is modest while individuals with a VPS35 D620N mutation demonstrate much higher Rab10-Thr73 phosphorylation, reflecting its much greater catalytic effect on the LRRK2 kinase. Our results show that changes in pRab10 can be used as direct read-out for LRRK2 activity and the very low percentage of pRab10 that we observe *ex vivo* may explain the age-dependent, incomplete penetrance of the LRRK2 G2019S mutation. It will be interesting to compare pRab10 percentages of matched disease-manifesting and non-manifesting LRRK2 G2019S mutation carriers at younger and older ages.

Our results could explain how minimal changes in the total pool of a phosphorylated protein could trigger pathogenesis. Stratification for Rab10-pThr73 levels in idiopathic and mutation carrying PD patients may identify individuals with increased LRRK2 kinase activity who would most likely benefit from LRRK2 kinase inhibitor treatment. Furthermore, our mass spectrometric assay is completely generic and not restricted to a particular disease. It could be applied to study any other kinase-substrate relation in clinical research. For example, phosphosite occupancies of prominent oncogenic factors could give important information on how phosphorylation stoichiometry influences tumorigenesis.

<sup>14</sup>

In summary, our generic and very sensitive MS-based assay accurately measures phosphorylation stoichiometry of substrates of interest. Presently, it can be used for ongoing clinical LRRK2 inhibitor studies, in which the target engagement, dosing and efficacy of the compounds needs to be evaluated, and we are currently pursuing this. We envision that our assay can also be applied to further investigate the role of LRRK2 in the PD, potentially revealing new upstream players of LRRK2 pathway, and determining whether they have a potential in treating PD.

#### ACKNOWLEDGEMENT

We thank Sabine Suppmann, Leopold Urich, Stephan Uebel, Stefan Pettera, Martin Spitaler, Nagarjuna Nagaraj, Victoria Sanchez and Antonio Piras from the MPIB Biochemistry Core Facility. We also thank Florian Meier, Susanne Kroiss, Gabriele Sowa, Igor Paron, Christian Deiml, Johannes B. Mueller and all the members of the department of Proteomics and Signal Transduction for their assistances and helpful discussions. We value the contributions of Shalini Padmanabhan, Marco Baptista (all from the Michael J. Fox Foundation for Parkinson's research), Kalpana Merchant (TransThera Consulting) and Suzanne Pfeffer (University of Stanford). We are grateful that the patients and healthy volunteers kindly donated blood for the present study. We would like to thank Thomas Gasser (University of Tubingen, Tubingen, Germany) and Alexander Zimprich (Medical University of Vienna, Wien, Austria) for providing access to LRRK2

[G2019S] and VPS35 [D620N] mutation carriers respectively, as well as controls and idiopathic PD patients, that donated blood for neutrophil isolation. GroEL/S and C terminal truncated construct of Rab10 (1-175) were kindly provided by Dr. Matthias Müller from the Max Planck Institute of Molecular Physiology in Dortmund. LRRK2 R1441G MEFs were kindly provided by Dr Shu-Leong Ho (Division of Neurology, Department of Medicine, University of Hong Kong).

## FUNDING

This work was supported by The Michael J. Fox Foundation for Parkinson's Research (grant ID. 6986), the Max-Planck Society for the Advancement of Science and the Medical Research Council (MC\_UU\_12016/2).

## COMPETING INTEREST

The authors declare no competing interests.

## REFERENCES

- 1 Alessi, D. R. & Sammler, E. J. S. LRRK2 kinase in Parkinson's disease. *360*, 36-37 (2018).
- 2 Eyers, P. A. J. B. J. Back to the future: new target-validated Rab antibodies for evaluating LRRK2 signalling in cell biology and Parkinson's disease. *475*, 185-189 (2018).
- 3 Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *49*, 1511 (2017).
- 4 Healy, D. G. *et al.* Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *7*, 583-590 (2008).
- 5 Paisán-Ruiz, C. J. H. m. LRRK2 gene variation and its contribution to Parkinson disease. *30*, 1153-1160 (2009).
- 6 Sheng, Z. *et al.* Ser1292 autophosphorylation is an indicator of LRRK2 kinase activity and contributes to the cellular effects of PD mutations. *4*, 164ra161-164ra161 (2012).
- 7 West, A. B. *et al.* Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity. *102*, 16842-16847 (2005).
- 8 Steger, M. *et al.* Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *Elife* **5** (2016).
- 9 Steger, M. *et al.* Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis. *Elife* **6** (2017).
- 10 Atashrazm, F. *et al.* LRRK2-mediated Rab10 phosphorylation in immune cells from Parkinson's disease patients. (2018).
- 11 Fan, Y. *et al.* Interrogating Parkinson's disease LRRK2 kinase pathway activity by assessing Rab10 phosphorylation in human neutrophils. *Biochemical Journal* **475**, 23-44 (2018).

12 Ito, G. *et al.* Phos-tag analysis of Rab10 phosphorylation by LRRK2: a powerful assay for assessing kinase function and inhibitors. *BCJ20160557* (2016).

13 Thirstrup, K. *et al.* Selective LRRK2 kinase inhibition reduces phosphorylation of endogenous Rab10 and Rab12 in human peripheral mononuclear blood cells. *7*, 10300 (2017).

14 Lis, P. *et al.* Development of phospho-specific Rab protein antibodies to monitor in vivo activity of the LRRK2 Parkinson's disease kinase. *Biochemical Journal* **475**, 1-22 (2018).

15 Mir, R. *et al.* The Parkinson's disease VPS35 [D620N] mutation enhances LRRK2-mediated Rab protein phosphorylation in mouse and human. *475*, 1861-1883 (2018).

16 Deng, X. *et al.* Characterization of a selective inhibitor of the Parkinson's disease kinase LRRK2. *7*, 203 (2011).

17 Dzamko, N., Chua, G., Ranola, M., Rowe, D. B. & Halliday, G. M. J. J. o. P. s. d. Measurement of LRRK2 and Ser910/935 phosphorylated LRRK2 in peripheral blood mononuclear cells from idiopathic Parkinson's disease patients. *3*, 145-152 (2013).

18 Dzamko, N. *et al.* Inhibition of LRRK2 kinase activity leads to dephosphorylation of Ser910/Ser935, disruption of 14-3-3 binding and altered cytoplasmic localization. *430*, 405-413 (2010).

19 Nichols, R. J. *et al.* 14-3-3 binding to LRRK2 is disrupted by multiple Parkinson's disease-associated mutations and regulates cytoplasmic localization. *430*, 393-404 (2010).

20 Kluss, J. H. *et al.* Detection of endogenous S1292 LRRK2 autophosphorylation in mouse tissue as a readout for kinase activity. *4*, 13 (2018).

21 Reynolds, A., Doggett, E. A., Riddle, S. M., Lebakken, C. S. & Nichols, R. J. J. F. i. m. n. LRRK2 kinase activity and biology are not uniformly predicted by its autophosphorylation and cellular phosphorylation site status. *7*, 54 (2014).

22 Olsen, J. V. *et al.* Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **3**, ra3-ra3 (2010).

23 Wu, R. *et al.* A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nature methods* **8**, 677 (2011).

24 Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences* **100**, 6940-6945 (2003).

25 Tsai, C.-F. *et al.* Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. *6*, 6622 (2015).

26 Domanski, D., Murphy, L. C. & Borchers, C. H. J. A. c. Assay development for the determination of phosphorylation stoichiometry using multiple reaction monitoring methods with and without phosphatase treatment: application to breast cancer signaling pathways. *82*, 5610-5620 (2010).

27 Steen, H., Jebanathirajah, J. A., Springer, M. & Kirschner, M. W. J. P. o. t. N. A. o. S. Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS. *102*, 3948-3953 (2005).

28 Hanke, S., Besir, H., Oesterhelt, D. & Mann, M. J. J. o. p. r. Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *7*, 1118-1130 (2008).

29 Zeiler, M. *et al.* A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *11*, O111. 009613 (2012).

30 Domon, B. & Gallien, S. J. P. C. A. Recent advances in targeted proteomics for clinical applications. *9*, 423-431 (2015).

31 Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *mcp. M111. 011015* (2011).

32 Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. **11**, 1709-1723 (2012).

33 Dzamko, N. *et al.* The IkappaB kinase family phosphorylates the Parkinson's disease kinase LRRK2 at Ser935 and Ser910 during Toll-like receptor signaling. **7**, e39132 (2012).

34 Rai, A. *et al.* bMERB domains are bivalent Rab8 family effectors evolved by gene duplication. **5**, e18675 (2016).

35 Matic, I. *et al.* Absolute SILAC-compatible expression strain allows Sumo-2 copy number determination in clinical samples. **10**, 4869-4875 (2011).

36 Fan, Y. *et al.* Human Peripheral Blood Neutrophil Isolation for Interrogating the Parkinson's Associated LRRK2 Kinase Pathway by Assessing Rab10 Phosphorylation. (2020).

37 Kulak, N. A., Geyer, P. E., Mann, M. J. M. & Proteomics, C. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. **16**, 694-705 (2017).

38 Rappaport, J., Ishihama, Y. & Mann, M. J. A. c. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. **75**, 663-670 (2003).

39 Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. J. N. m. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. **11**, 319 (2014).

40 Carr, S. A. *et al.* Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. **13**, 907-917 (2014).

41 Cox, J. & Mann, M. J. N. b. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. **26**, 1367 (2008).

42 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. **10**, 1794-1805 (2011).

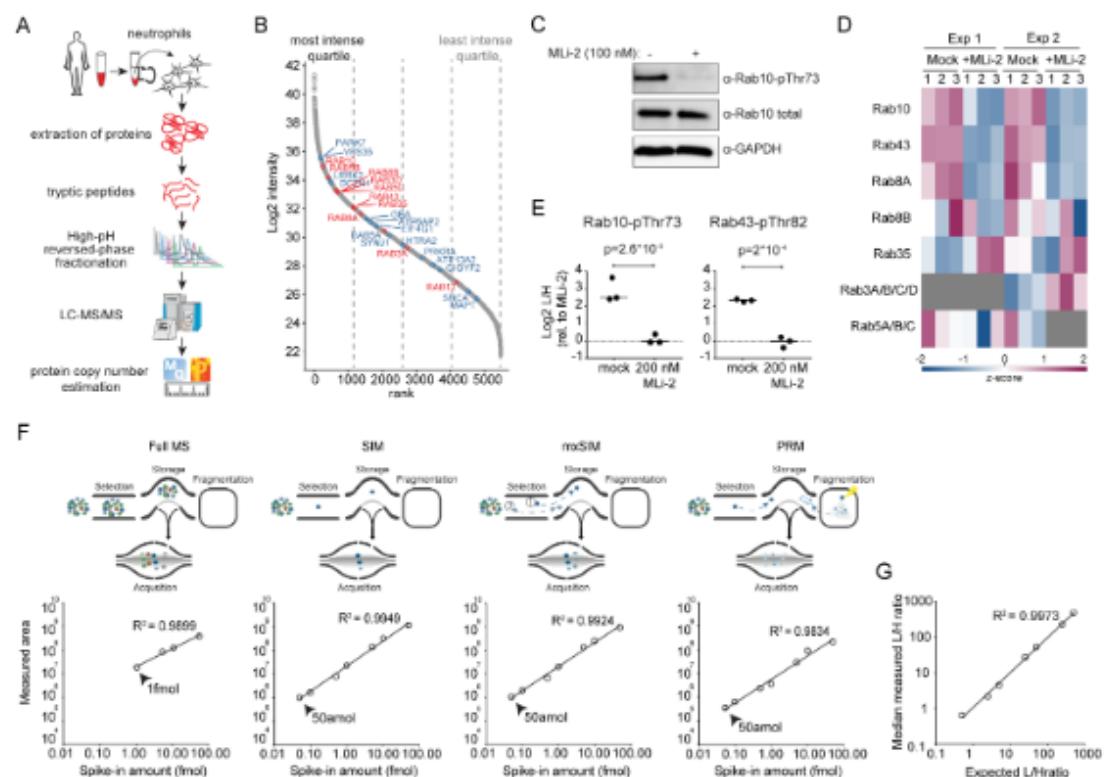
43 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote) omics data. **13**, 731 (2016).

44 Baker, P. R., Chalkley, R. J. J. M. & Proteomics, C. MS-viewer: a web-based spectral viewer for proteomics results. **13**, 1392-1396 (2014).

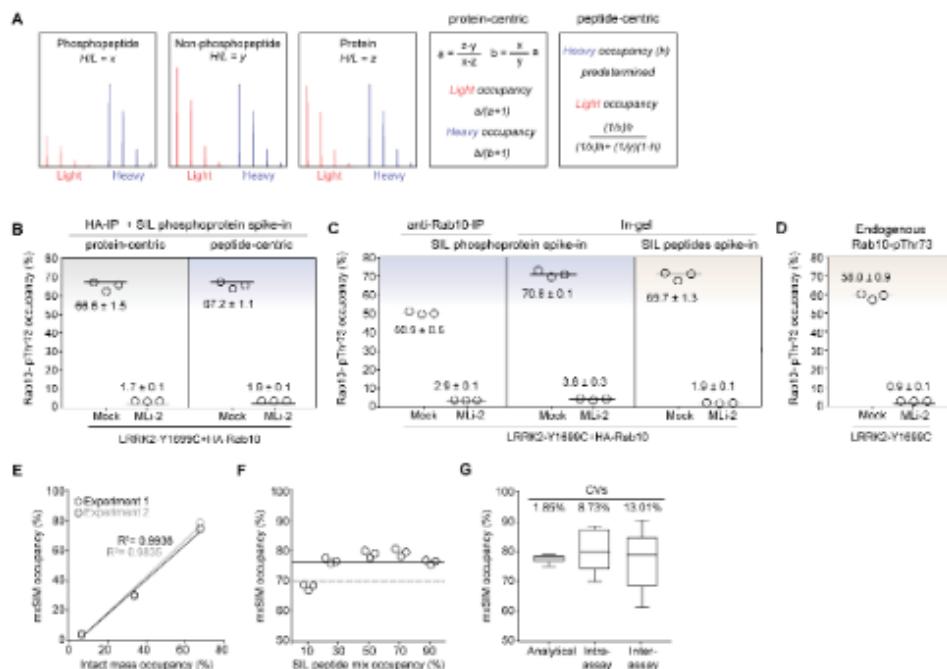
45 Winiewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Molecular & cellular proteomics, mcp*. M113. 037309 (2014).

46 Kelstrup, C. D. *et al.* Performance evaluation of the Q Exactive HF-X for shotgun proteomics. **17**, 727-738 (2018).

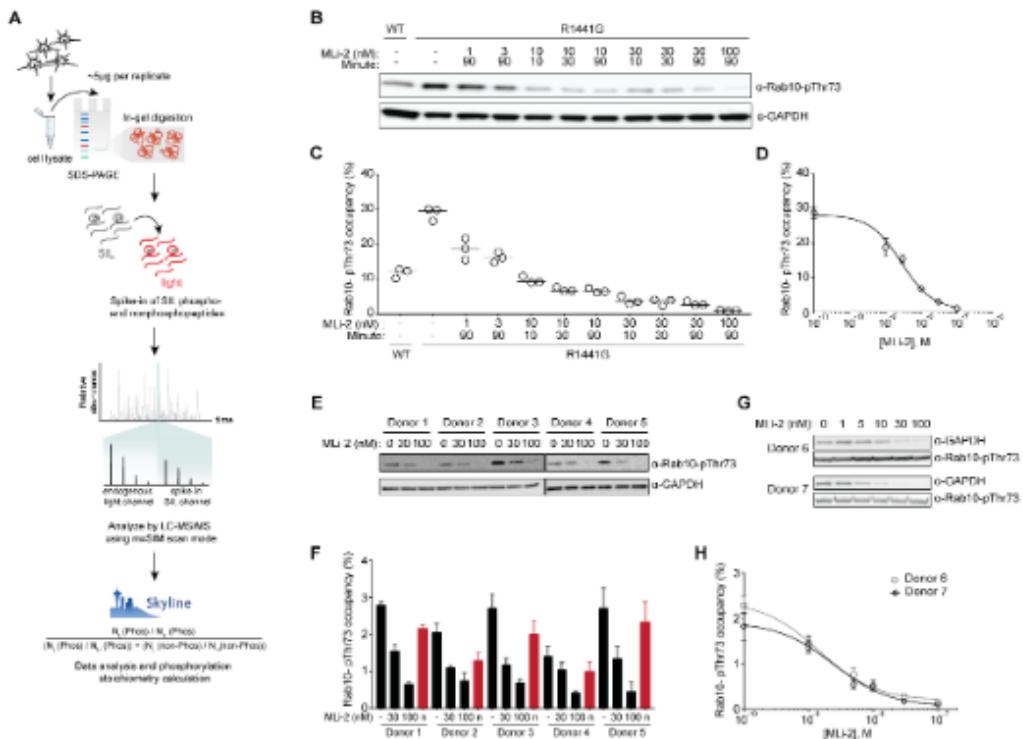
FIGURES AND FIGURE LEGENDS



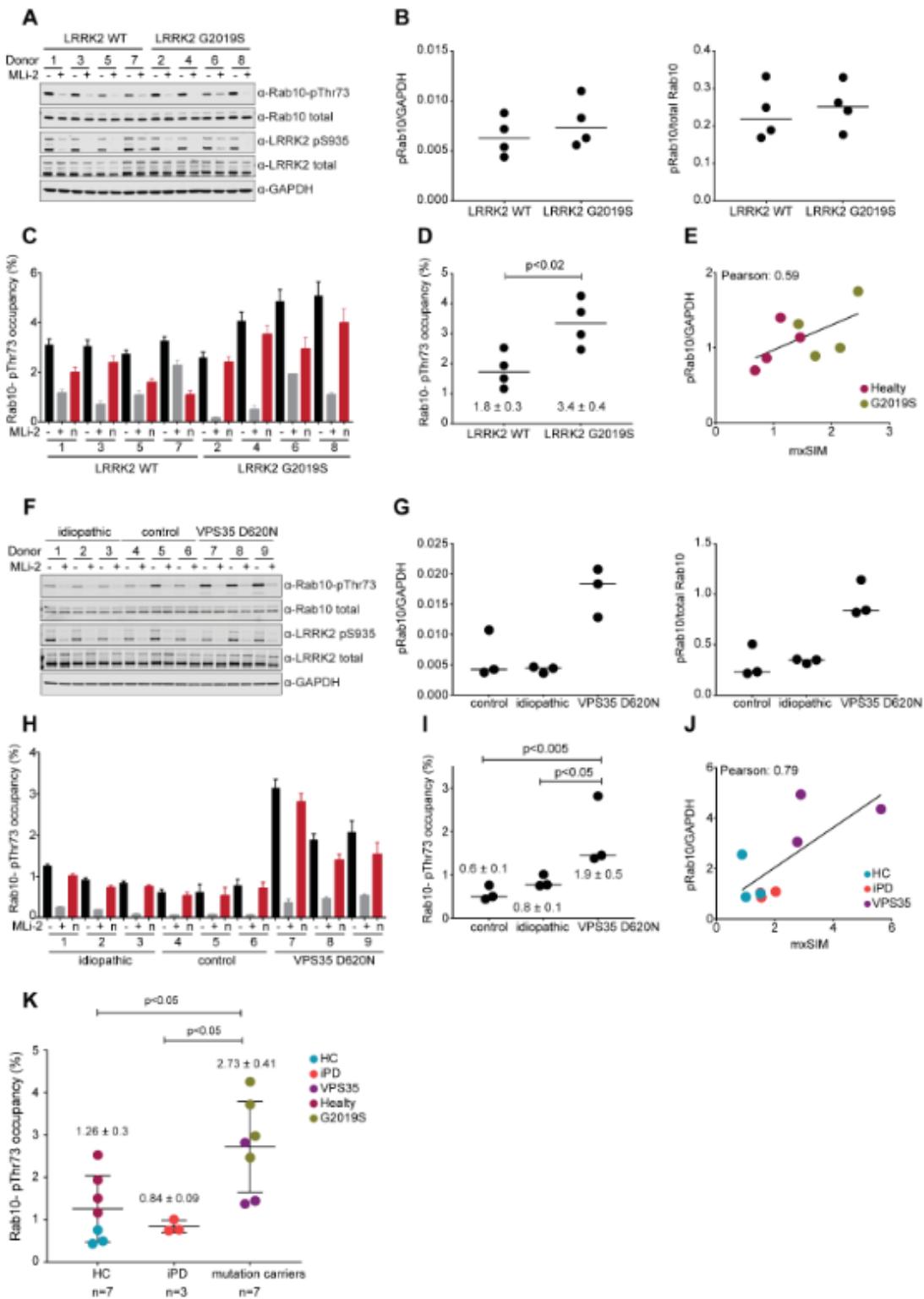
**Figure 1. Rab10-pThr73 as a readout for LRRK2 activity in human peripheral blood neutrophils.** (A) Workflow for global proteome analysis of human peripheral blood neutrophils. (B) Proteins ranked according to their abundances across the global neutrophil proteome. Quartiles are indicated with dashed lines. LRRK2 phosphorylated Rab proteins and other PD-associated proteins are highlighted in red and blue, respectively. (C) Western blot analysis of neutrophils (-/+200 nM MLi-2, 30 min) using anti-MJFF-pRAB10 (pThr73), Rab10 total and GAPDH antibodies. (D) Heat map of z-scored phosphopeptide intensities of LRRK2-phosphorylated Rab proteins from pRab immunoprecipitation of neutrophil lysates (-/+ 200 nM MLi-2, 30 min, 3 technical replicates,  $n=2$ ). Missing values are in grey. (E) Targeted MS-quantified Rab10-pThr73 and Rab43-pThr82 peptide intensities in immunoprecipitations of individual Rab10 and Rab43 proteins from neutrophils (-/+ 200 nM MLi-2, 30 min). P-values were assessed using unpaired Student's t-test analysis. (F) Limit of detection (LOD) of SIL Rab10-pThr73 tryptic peptide (FHpTITTSYYR) with various acquisition methods; full MS, SIM, mxSIM and PRM. Linear ranges of the dilution curves were assessed with singlets analyses of the SIL Rab10-pThr73 phosphopeptide (10 amol to 50 fmol, spiked-in a digest of HeLa). (G) Limit of quantification (LOQ) for the Rab10-pThr73 tryptic peptide measured in mxSIM mode. 25 fmol of light phosphopeptide was mixed with variable amounts of its SIL counterpart (10 amol to 50 fmol) and spiked into a background of HeLa digest (3 technical replicates). Median ratios extracted from Skyline were plotted against the expected ratios.



**Figure 2. mxSIM accurately determines Rab10-Thr73 phosphorylation stoichiometry.** (A) Heavy-to-light ratios and the formulas used for protein- vs peptide-centric approaches. (B) HA-Rab10-pThr73 occupancy in LRRK2-Y1699C-expressing HEK293 cells was determined using the protein- vs peptide-centric approaches after HA-IP. (C) HA-Rab10-pThr73 occupancy determined using the peptide-centric approach with either SIL phosphoprotein or SIL peptide standard spike-in after enrichment by anti-Rab10-IP or SDS-PAGE followed by in-gel digestion in HEK293 cells expressing LRRK2-Y1699C (-/+ 200 nM MLi-2, 60 min). (D) Endogenous Rab10-pThr73 occupancy determined using SIL peptide standards in mock and LRRK2-Y1699C expressing HEK293 cells (-/+ 200 nM MLi-2, 60 min). Samples from (B), (C) and (D) were analyzed in triplicates using the mxSIM method and the phosphorylation occupancies are presented as means  $\pm$  SEM. (E) Benchmarking our method using Thr73-phosphorylated unlabeled recombinant Rab10 proteins (1-175 aa) as standards. Correlation of the median occupancies determined either by intact mass analysis or by mxSIM in triplicates for two independent experiments. (F) SIL phosphorylated and non-phosphorylated Rab10 peptides mixed to mimic 10, 30, 50, 70 and 90% occupancies and spiked into the digested recombinant pRab10 protein were measured using our mxSIM method in triplicate. The solid line represents the median occupancy ( $75.4 \pm 1.5\%$ ) whereas the dashed line shows the estimated phosphorylation occupancy of the standard protein (70%) by intact mass analysis. (G) CVs were calculated by repeating MS measurements (analytical), the workflow in the same gel (intra-assay) or in different gels (inter-assay) using the same phosphoprotein standard (n=6).



**Figure 3. Reliably determination of Rab10-pThr73 occupancy in cell lysates.** (A) Workflow for the Rab10-pThr73 occupancy assay. (B) Immunoblotting of WT and R1441G knock-in MEFs treated with the indicated concentrations of MLi-2 for different intervals using monoclonal MJFF-pRAB10 (pThr73) and GAPDH antibodies. (C) Rab10-pThr73 occupancies were determined by mxSIM in the same lysates (n=3). (D) Dose-response curve of Rab10-pThr73 occupancy in R1441G knock-in MEFs to generate occupancy based- $IC_{50}$  values for MLi-2. Each data point represents the median of triplicate measurements in the samples treated with MLi-2 for 90 min (E) Immunoblotting of neutrophils isolated from healthy individuals (DMSO or 30 and 100 nM MLi-2 treated) using monoclonal MJFF-pRAB10 (pThr73) and GAPDH antibodies. (F) Rab10-pThr73 occupancies were determined by mxSIM in the same lysates. Normalized Rab10-pThr73 occupancies by subtracting the MLi-2 treated values are shown in red. (G) Immunoblotting of the neutrophils isolated from healthy individuals and treated with the indicated concentrations of MLi-2 using monoclonal MJFF-pRAB10 (pThr73) and GAPDH antibodies. (H) Individual-specific dose-response curves to generate occupancy based- $IC_{50}$  for MLi-2. Error bars represent SEM.



**Figure 4. Rab10-pThr73 occupancy in PD patient-derived neutrophils.** (A) Immunoblotting of neutrophils isolated from four PD patients with the G2019S LRRK2 mutation and four healthy controls (-/+100 nM MLi-2, 30 min) using anti total LRRK2, pSer935 LRRK2, total Rab10, MJFF-pRAB10 (pThr73) and GAPDH antibodies. (B) Quantitation of immunoblots by analysing phospho-Thr73 Rab10/total Rab10 ratio (right) and phospho-Thr73 Rab10/GAPDH ratio (left panel). (C) Blind application of our mxSIM assay to the same mock (black) and MLi2-treated (grey) lysates (n=3). Normalized Rab10- pThr73 occupancies by subtracting the MLi-2 treated values are also shown (red). Error bars represent SEM. (D) Quantification of Rab10-pThr73 occupancy in controls and PD patients with the G2019S LRRK2 mutation. Each data point represents the median of triplicate measurements of untreated samples that are normalized to MLi2. One-way ANOVA with Bonferroni's multiple comparisons test was applied and the occupancies are presented as means  $\pm$  SEM. (E) Pearson correlation between immunoblotting and mxSIM assay. Plotted are fold changes of normalized Rab10-pThr73 levels (occupancies in mxSIM) relative to control cases. (F) Immunoblotting of neutrophils isolated from controls, idiopathic PD patients and heterogeneous VPS35 D620N mutation carriers (-/+ 200 nM MLi-2, 30 min) using anti total LRRK2, pSer935 LRRK2, total Rab10, MJFF-pRAB10 (pThr73) and GAPDH antibodies. (G) Quantitation of immunoblots by analysing phospho-Thr73 Rab10/total Rab10 ratio (right) and phospho-Thr73 Rab10/GAPDH ratio (left panel). (H) Blind application of our assay to the same mock (black) and MLi2-treated (grey) lysates (n=3). Normalized Rab10- pThr73 occupancies by subtracting the MLi-2 treated values are shown in red. Error bars represent SEM. (I) Quantification of Rab10-pThr73 occupancy in controls, idiopathic PD and PD patients with heterogeneous VPS35 D620N mutation. The significant analysis was done as in (D). (J) Same as (E) with healthy controls (HC), idiopathic PD (iPD) and VPS35 D620N (VPS35). (K) Quantification of Rab10-pThr73 occupancy between controls, idiopathic cases and PD patients with a defined mutation. The significant analysis was done as in (D).

### 3.2. Ubiquitin or Ubiquitin-like modification

#### 3.2.1. Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology

Fynn M. Hansen<sup>1</sup>, Maria C. Tanzer<sup>1</sup>, Franziska Brüning<sup>1,2</sup>, Isabell Bludau<sup>1</sup>, Brenda A. Schulman<sup>3</sup>, Maria S. Robles<sup>2,\*</sup>, Ozge Karayel<sup>1,\*</sup> and Matthias Mann<sup>1,\*,‡</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany <sup>2</sup> Institute of Medical Psychology, Faculty of Medicine, LMU, Munich, Germany <sup>3</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, Germany \*Corresponding authors

Under peer review in *Nature Communications* (2020)

Protein ubiquitination is a highly versatile post-translational modification regulating a plethora of complex cellular processes. Dysregulation of components of the ubiquitin system contributes to a variety of diseases, including cancer and neurodegeneration. To better understand the biological function of ubiquitination and ubiquitin-related mechanisms in health and disease, it is crucial to identify the targets of the ubiquitin machinery and to quantify the ubiquitination events. MS-based proteomics has become the gold standard for the analysis PTM-specific proteomes. However, the large-scale study of the ubiquitinome remains challenging mostly due to technological reasons. Data independent acquisition (DIA), especially on the Orbitrap mass spectrometers, holds great promise in increasing the reproducibility and depth of proteome analysis. Therefore, in this study, we investigated the power of DIA to improve data completeness and sensitivity in a single run analysis format. Compared to the DDA method, our optimized DIA method substantially increased the number diGly peptides quantified in single run MS measurements (~35,000) while also dramatically improving quantitative accuracy and reproducibility. Application of our workflow to the well-studied TNF signaling pathway retrieved the sites described over the years in a single experiment and even identified novel and biologically meaningful ones. We then analyzed the rhythmic ubiquitinome in a circadian model. Bioinformatics analysis of cycling sites unexpectedly discovered a high proportion of rhythmically modified membrane proteins associated with transporter functions opening up an entirely novel area for chronobiologists to investigate. We believe that our DIA-based diGly workflow will be of interest to the community to elucidate the complexity of protein ubiquitination in cell signaling and disease states.

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology

Fynn M. Hansen<sup>1</sup>, Maria C. Tanzer<sup>1</sup>, Franziska Brüning<sup>1,2</sup>, Isabell Bludau<sup>1</sup>, Brenda A. Schulman<sup>3</sup>, Maria S. Robles<sup>2,\*</sup>, Ozge Karayel<sup>1,\*</sup> and Matthias Mann<sup>1</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

<sup>2</sup> Institute of Medical Psychology, Faculty of Medicine, LMU, Munich, Germany

<sup>3</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, Germany

\*To whom correspondence may be addressed: charo.robles@med.uni-muenchen.de, karayel@biochem.mpg.de or mmann@biochem.mpg.de

### SUMMARY

Protein ubiquitination is involved in virtually all cellular processes. Enrichment strategies employing antibodies targeting ubiquitin-derived diGly remnants combined with mass spectrometry (MS) have enabled investigations of ubiquitin signaling at a large scale. However, so far the power of data independent acquisition (DIA) with regards to sensitivity in single run analysis and data completeness have not yet been explored. We developed a sensitive workflow combining diGly antibody-based enrichment and optimized Orbitrap-based DIA with comprehensive spectral libraries together containing more than 90,000 diGly peptides. This approach identified 35,000 diGly peptides in single measurements of proteasome inhibitor-treated cells – double the number and quantitative accuracy of data dependent acquisition. Applied to TNF-alpha signaling, the workflow comprehensively captured known sites while adding many novel ones. A first systems-wide investigation of ubiquitination of the circadian cycle uncovered hundreds of cycling ubiquitination sites and dozens of cycling ubiquitin clusters within individual membrane protein receptors and transporters, highlighting novel connections between metabolism and circadian regulation.

### KEYWORDS

Ubiquitin, DIA, TNF, circadian clock, ubiquitinome, mass spectrometry, systems biology, post-translational modifications

## INTRODUCTION

Ubiquitination is a reversible and highly versatile post-translational modification (PTM) involved in virtually all cellular processes. A ubiquitin conjugation cascade, involving ubiquitin activating (E1), conjugating (E2) and ligating (E3) enzymes, mediates the covalent attachment of the 76 amino acid long ubiquitin molecule to a ε-amine group of a lysine residue on a substrate protein. Its removal is mediated by an enzyme family called deubiquitinating enzymes (DUB). Ubiquitin itself can be ubiquitinated N-terminally or via one of its seven lysine residues, giving rise to a plethora of chain topologies, which encode a diverse and specific set of biological functions<sup>1,2</sup>. Deregulation of this highly complex process has been linked to numerous diseases including neurodegenerative diseases<sup>3,4</sup>, autoimmunity<sup>5,6</sup> and inflammatory disorders<sup>7-9</sup>.

Protein ubiquitination is one of the most widely studied PTMs in the field of mass spectrometry (MS)-based proteomics. However, due to low stoichiometry of ubiquitination and varying ubiquitin chain topologies, comprehensive profiling of endogenous ubiquitination is challenging and requires one or more enrichment steps prior to MS analysis<sup>10</sup>. Early reports to catalog ubiquitin conjugated proteins from yeast and human described various enrichment methods including the use of epitope-tagged ubiquitin or ubiquitin-associated domains (UBA)<sup>11-13</sup>. After trypsinization, previously ubiquitinated or NEDDylated peptides bear a signature diGly remnant that can be targeted by a specific antibody<sup>14</sup>. Enrichment strategies employing such antibodies have enabled identification of thousands of ubiquitination sites by MS<sup>15-17</sup>. A recently described antibody targets a longer remnant generated by LysC digestion to exclude ubiquitin-like modifications such as NEDD8 or ISG15<sup>18</sup>, however, the contribution of diGly sites derived from ubiquitin-like modifications is very low (< 6%)<sup>15</sup>.

The commercialization of such antibodies has accelerated MS-based ubiquitinome analysis and enabled a variety of quantitative, systems-wide studies<sup>19-23</sup>. However, large-scale analysis of ubiquitination events to study key signaling components remains challenging since in-depth diGly proteome coverage requires relatively large sample amounts and extensive peptide fractionation. These requirements, which largely stem from the low stoichiometry of the modification, come at the expense of throughput, robustness and quantitative accuracy.

Thus far, ubiquitinome studies have employed data-dependent acquisition (DDA) methods combined with label-free or isotope-based quantification<sup>24</sup>. Recently, Data Independent Acquisition (DIA) has become a compelling alternative to DDA for proteomics analysis with greater data completeness across samples<sup>25-28</sup>. In contrast to intensity-based precursor picking of DDA, DIA fragments all co-eluting peptide ions within predefined mass-to-charge (*m/z*) windows and acquires them simultaneously<sup>29</sup>. This leads to more precise and accurate quantification with fewer missing values across samples and higher identification rates over a larger dynamic range. DIA usually requires a comprehensive spectral library, from which the peptides are matched into single run MS analyses. Recently, superior performance of DIA for sensitive and reproducible MS measurements has also been demonstrated for global protein phosphorylation analysis<sup>30</sup>. Given the central importance of ubiquitination, we here set out to investigate the power of DIA for improving data completeness and sensitivity in a single run analysis format.

For sensitive and reproducible analysis of the ubiquitin-modified proteome, we devised a workflow combining diGly antibody-based enrichment with a DIA method tailored to the unique properties of the library peptides and to

the linear quadrupole Orbitrap mass analyzer. We acquired extensive spectral libraries that all together contain more than 90,000 diGly peptides allowing us to reproducibly analyze 35,000 distinct diGly peptides in a single measurement of proteasome inhibitor-treated cells. The DIA-based diGly workflow markedly improved the number and quantitative accuracy compared to DDA. To investigate if our new method would have advantages in the exploration of biological signaling systems, we first applied it to the well-studied TNF-signaling pathway, where it retrieved known ubiquitination events and uncovered novel ones. We then extended it to the analysis of circadian post-translational dynamics, so far poorly studied globally with regards to ubiquitination. This uncovered a remarkable extent and diversity of ubiquitination events. These include closely spaced clusters with the same circadian phase, which are likely pointing to novel mechanisms. Together, our design and results establish a sensitive and accurate DIA-based workflow suitable for investigations of ubiquitin signaling at a systems-wide scale.

## RESULTS

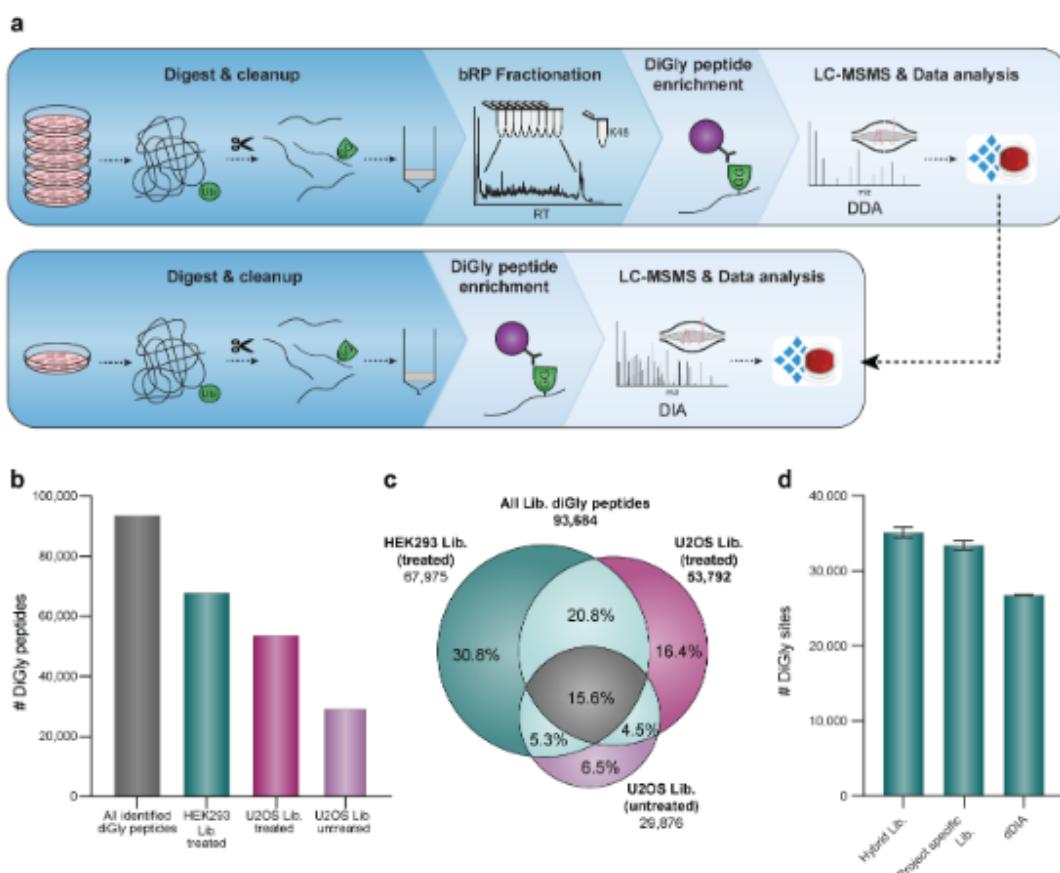
**DIA quantification enables in-depth diGly proteome coverage in single shot experiments**  
To obtain a comprehensive, in-depth spectral library for efficient extraction of diGly peptides in single shot DIA analysis, we treated two human cell lines (HEK293 and U2OS) with a common proteasome inhibitor (10  $\mu$ M MG132, 4 h). After extraction and digestion of proteins, we separated peptides by basic reversed-phase (bRP) chromatography into 96 fractions, which were concatenated into 8 fractions (Methods, Supplementary Fig. 1a). Here, we isolated fractions containing the highly abundant K48-linked ubiquitin-chain derived diGly peptide (K48-peptide) and processed them separately to reduce excess amounts of K48-peptides in individual pools, which compete for antibody

binding sites during enrichment and interfere with the detection of co-eluting peptides (Supplementary Fig. 1b). We found this to be a particular issue for MG132 treatment, as blockage of the proteasome activity further increases K48-peptide abundance in these samples. The resulting 9 pooled fractions were enriched for diGly peptides, which were separately analyzed using a DDA method (PTMScan Ubiquitin Remnant Motif (K-e-GG) Kit, CST) (Fig. 1a and Supplementary Fig. 1a-b). This identified more than 67,000 and 53,000 diGly peptides in MG132 treated HEK293 and U2OS cell lines, respectively (Fig. 1b). Furthermore, to fully cover diGly peptides of an unperturbed system, we also generated a third library using the same workflow but with untreated U2OS cells (used later for biological applications). This added a further 6,000 distinct diGly peptides (Fig. 1b). In total, we obtained 89,650 diGly sites corresponding to 93,684 unique diGly peptides, 43,338 of which were detected in at least two libraries (Fig. 1c, see also source data at PRIDE: PXD019854). To our knowledge, this represents the deepest diGly proteome to date. According to the PhosphositePlus database<sup>31</sup>, 57% of the identified diGly sites were not reported before and 7.3% of them were previously found to be acetylated or methylated, indicating that different PTMs can act on the same sites.

In possession of these large diGly spectral libraries, we evaluated DIA method settings for best performance in single shot diGly experiments (Supplementary Table 1). Impeded cleavage C-terminal to modified lysine residues frequently generates longer peptides with higher charge states, resulting in diGly precursors with unique characteristics. Guided by the empirical precursor distributions, we first optimized DIA window widths – the transmission windows that together cover the desired precursor peptide range. This increased the number of identified

diGly peptides by 6% (Supplementary Fig. 2a-b). Next, we tested different window numbers and fragment scan resolution settings, to strike an optimal balance between data quality and a cycle time that sufficiently samples eluting chromatographic peaks. We found that a method with relatively high MS2 resolution of 30,000 and 46 precursor isolation windows performed best (13% improvement compared to the standard full proteome method that we started with)

(Supplementary Fig. 2c). Furthermore, we determined the optimal antibody and peptide input combination to maximize peptide yield and depth of coverage in single DIA experiments. To mimic endogenous cellular levels, we used peptide input from cells not treated with MG132. From titration experiments, enrichment from 1 mg of peptide material using 1/8<sup>th</sup> of an anti-diGly antibody vial (31.25 µg) turned out to be optimal (Methods and Supplementary



**Fig. 1 In-depth diGly proteomics for DIA identification.**

**a** Experimental workflow for in-depth diGly peptide library construction (upper panel) and our single run DIA-based workflow (lower panel). Protein digestion and peptide extraction are followed by bRP fractionation and diGly peptide enrichment. For library construction, samples were measured by DDA and computationally processed (Spectronaut Pulsar). Individual samples are measured by our DIA workflow, including matching against a library for identification (Spectronaut software). **b** Number of identified diGly peptides in three different spectral libraries. **c** Commonly and exclusively identified diGly peptides for different libraries. **d** Number of identified diGly sites ( $\pm$  SEM) of three workflow replicates of MG132 treated HEK293 cells measured in analytical duplicates using different DIA library search strategies.

**Fig. 2d,e).** With the improved sensitivity by DIA, only 25% of the total enriched material needed to be injected (**Supplementary Fig. 2f**).

Using our optimized DIA-based workflow, we identified a remarkable  $33,350 \pm 1,481$  distinct diGly sites in single measurements of MG132 treated HEK293 samples. This implies that about half of the sites in the deep, cell line specific spectral library were matched into the single runs. Interestingly, even without using any library, a search of six single runs identified  $26,800 \pm 145$  diGly sites (direct DIA, **Methods**). Finally, employing a hybrid spectral library - generated by merging the DDA library with a direct-DIA search - resulted in  $34,733 \pm 1,670$  diGly sites in the same samples (**Fig. 1d, Supplementary Table 2**). Compared to recent reports in the literature<sup>24</sup>, these numbers double diGly peptide identifications in a single run format.

#### DIA improves diGly proteome quantification accuracy

To evaluate the reproducibility of the entire DIA-based diGly workflow, we used MG132 treated HEK293 cells and performed three independent diGly peptide enrichments followed by DIA analysis in duplicates. This identified around 36,000 distinct diGly peptides in all replicates, 45 and 77% of which had coefficients of variations CVs below 20 and 50%, respectively (**Fig. 2a-b, Supplementary Table 3**). In contrast, a DDA method identified substantially fewer distinct diGly peptides and a smaller percentage with good CVs (16,000 sites; 15% with CVs < 20%; **Fig. 2a-b**). Overall, the six DIA experiments yielded almost 48,000 distinct diGly peptides, while the corresponding DDA experiments resulted in 19,000 sites.

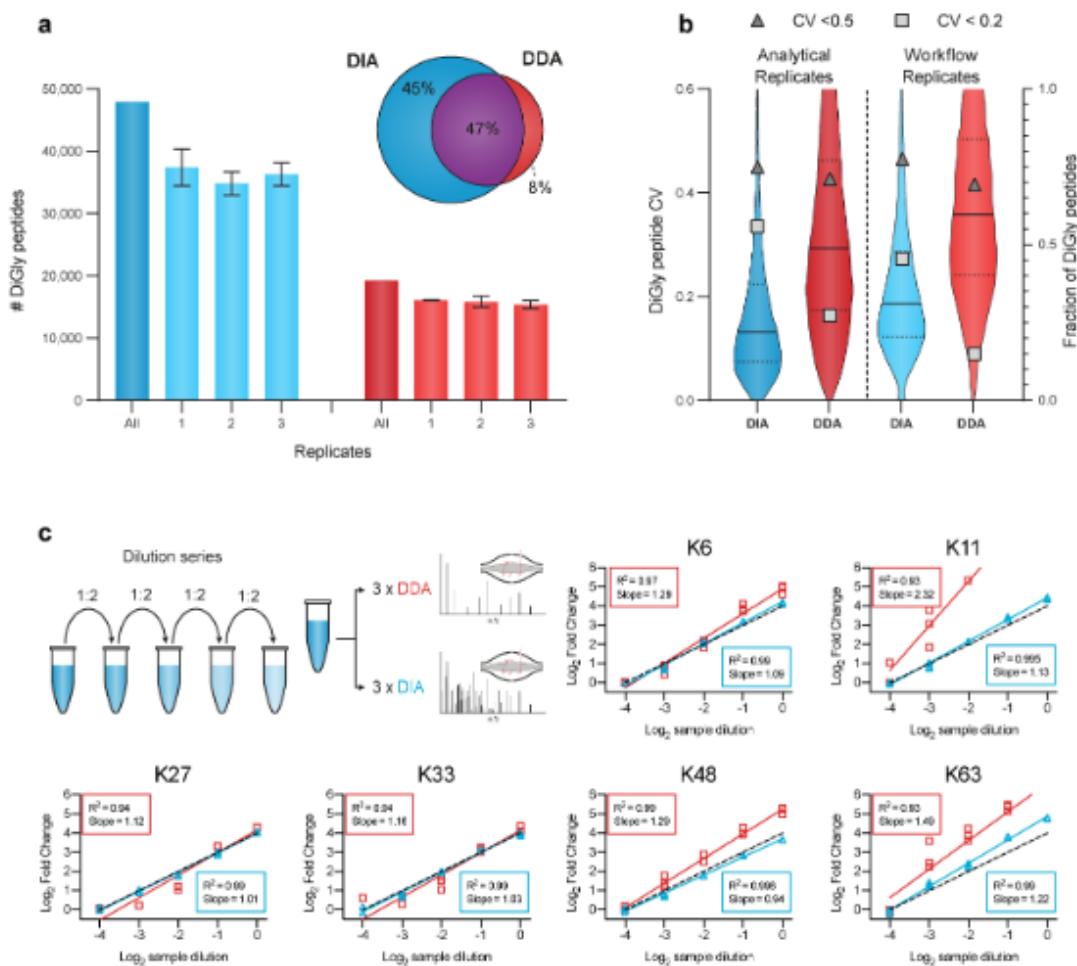
To further investigate the quantitative precision and accuracy of our method, we turned to ubiquitin-chain linkage derived diGly peptides. These are the most abundant diGly peptides, all

ranking in the top 20 by abundance and spanning three orders of magnitude in MS signal (**Supplementary Fig. 2g**). Diverse chain linkages confer various functions to proteins; hence, accurate quantification is important to decode the cellular roles of different ubiquitin linkage types. We performed a dilution series of a diGly sample and analyzed each dilution sample using both DIA and DDA methods in triplicates. Linear regression resulted in excellent correlations with  $R^2$  values higher than 0.99 for all six chain peptides assessed, much higher than the corresponding values for DDA ( $R^2$  0.93-0.99; **Fig. 2c, Supplementary Table 3**). Importantly for quantification purposes, the experimentally observed slope for DIA was much closer to 1 than for DDA.

Together, these analytical results establish that the DIA-based workflow substantially increased the number of diGly peptides identified while markedly improving the precision and accuracy of quantification compared to a DDA-based workflow.

#### In-depth ubiquitinome analysis of the TNF signaling pathway

The pro-inflammatory properties of TNF are heavily regulated by dynamic ubiquitination of its receptor-signaling complex (RSC)<sup>32,33</sup> and global ubiquitinome changes upon TNF stimulation were described previously in a proteomics study<sup>34</sup>. Encouraged by the technical capabilities of our DIA-based diGly workflow, we here aimed to test our DIA-based diGly workflow on this well studied system, to demonstrate benefits of DIA over DDA based on accurate ubiquitination site quantification and, if possible, to extend the current knowledge of the TNF regulated ubiquitinome (**Fig. 3a**). Applying both DIA- and DDA-based diGly workflows together quantified over 10,000 diGly sites in TNF-stimulated U2OS cells (**Fig. 3b, Supplementary Table 4**). Both



**Fig. 2** Accurate and reproducible diGly proteomics for DIA quantification

a Number of identified diGly peptides ( $\pm$  SD) for DIA (blue, HEK293 hybrid library) and DDA (red) strategies (three workflow replicates with each replicate analytical duplicates). Venn diagram depicts the proportion of shared and exclusively identified diGly sites between DIA and DDA approaches. b CV value distribution for DIA and DDA approaches. Solid and dotted lines denote median and 1<sup>st</sup> or 3<sup>rd</sup> quartile, respectively (left axis). Fractions of CV values (right axis) below 50% and 20% are shown with filled triangles and filled squares, respectively. c Dilution series of diGly enriched sample. Plots show simple linear regression fits of individual ubiquitin-chain linkage type peptides measured in triplicates using DIA (blue) or DDA (red). Dotted black lines indicate the expected 1:1 slope.

methods quantified a comparable number of ubiquitination sites (10,300 in DIA and 9,500 in DDA experiment, Fig. 3b). However, the DIA experiment resulted in 248 significantly upregulated ubiquitination sites (5% FDR, median fold change 2.5), of which 37 mapped to 23 proteins known to be involved in TNF $\alpha$ /NF $\kappa$ B signaling (Fig. 3c). In stark contrast, the DDA approach identified only 38 significant

ubiquitination sites (5% FDR and median fold change 4.1), of which 15 mapped to 7 TNF $\alpha$ /NF $\kappa$ B signaling proteins. In line with these numbers, gene ontology (GO) enrichment analysis had lower FDR values and larger group sizes for terms related to the TNF $\alpha$ /NF $\kappa$ B pathway in the DIA experiment compared to DDA (Fig. 3d).

# Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219056>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

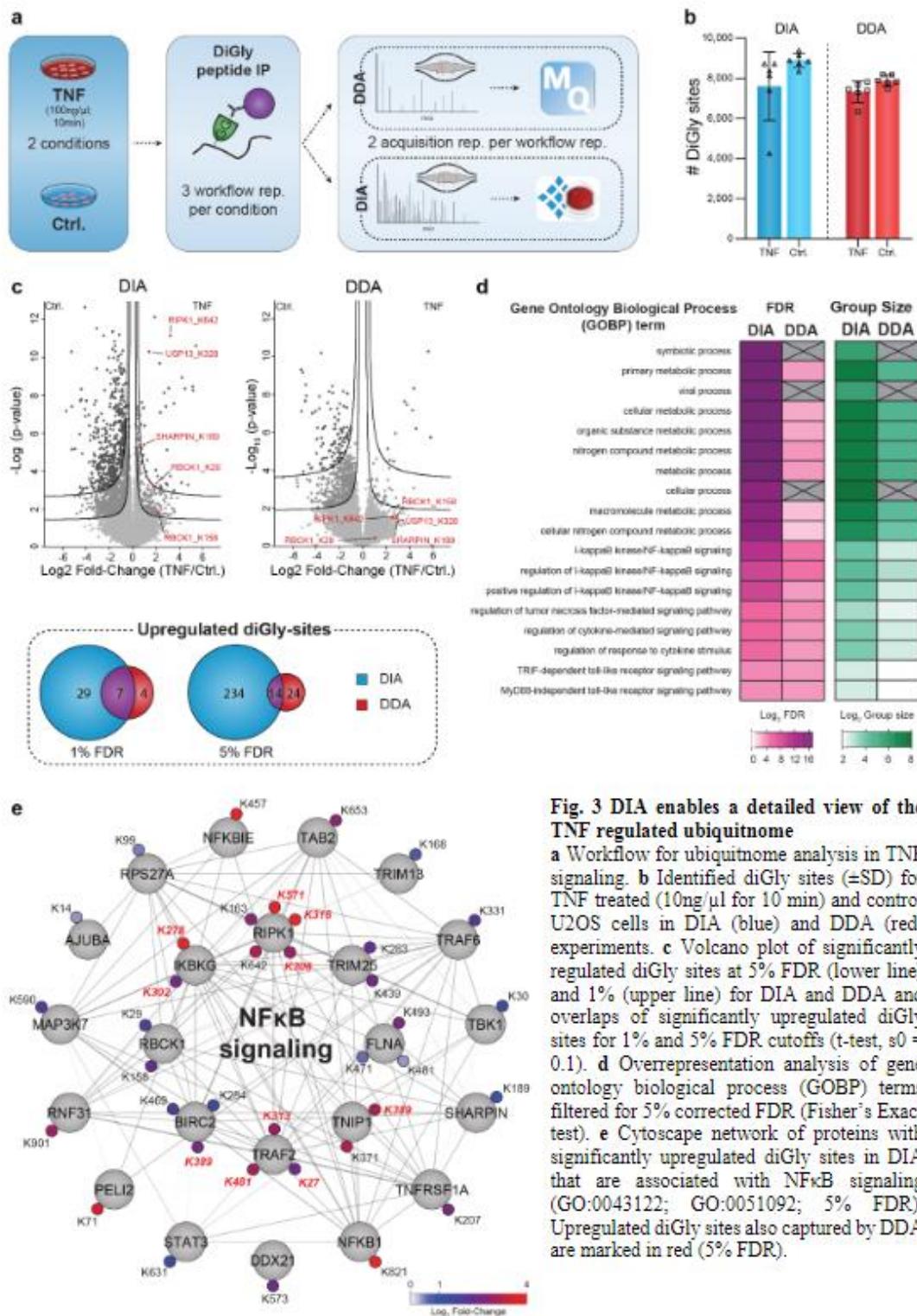


Fig. 3 DIA enables a detailed view of the TNF regulated ubiquitome

**a** Workflow for ubiquitome analysis in TNF signaling. **b** Identified diGly sites ( $\pm$ SD) for TNF treated (10ng/ $\mu$ l for 10' min) and control U2OS cells in DIA (blue) and DDA (red) experiments. **c** Volcano plot of significantly regulated diGly sites at 5% FDR (lower line) and 1% (upper line) for DIA and DDA and overlaps of significantly upregulated diGly sites for 1% and 5% FDR cutoffs (t-test,  $s_0 = 0.1$ ). **d** Overrepresentation analysis of gene ontology biological process (GOBP) terms filtered for 5% corrected FDR (Fisher's Exact test). **e** Cytoscape network of proteins with significantly upregulated diGly sites in DIA that are associated with NF- $\kappa$ B signaling (GO:0043122; GO:0051092; 5% FDR). Upregulated diGly sites also captured by DDA are marked in red (5% FDR).

Several members of the TNF signaling pathway have been implicated in viral infection and TNF-receptor blockage increases susceptibility to viral infection<sup>35,36</sup>. The ‘viral processes’ term was significantly enriched in our DIA analysis, in line with the literature on involvement of TNF-mediated ubiquitination regulation during viral infection. Underscoring the depth of the DIA analysis, the same term failed to reach significance in the DDA analysis (Fig. 3d, Supplementary Fig. 3). In agreement with previous studies, both DIA and DDA analyses revealed increased ubiquitination of prominent members of the TNF-RSC, including TRAF2, RIPK1 and BIRC2<sup>37,38</sup> (Fig. 3e). DIA allowed the detection of further ubiquitination events associated with the TNF $\alpha$ /NF $\kappa$ B signaling (Fig. 3c). For instance, the death domain (DD) of RIPK1 mediates interaction with FADD and TRADD<sup>39</sup> and we found K642 in this domain to be ubiquitinated upon TNF $\alpha$  stimulation. Furthermore, DIA but not DDA established regulated ubiquitination of all members - HOIP/RNF31, HOIL-1/RBCK1 and Sharpin- of the LUBAC complex, a critical E3 ligase complex in TNF signaling<sup>40,41</sup> in agreement with a previous study that showed LUBAC auto-ubiquitination during inflammation<sup>42</sup> (Fig. 3e). p105/NFKB1, is a precursor for p50 and inhibitor of NF $\kappa$ B signaling<sup>43</sup> and we observed a striking 16-fold upregulation of K821 in its DD. Proteasome mediated limited proteolysis of p105 during NF $\kappa$ B signaling yields the active p50 subunit<sup>44-47</sup> and the strong regulation of the K821 site suggests its involvement in this process.

DIA-based diGly analysis also uncovered TNF-regulated ubiquitination of numerous proteins known to be involved in other immune pathways. For instance, Peli2, an E3 ligase important for TLR and IL-1 signaling pathways<sup>48</sup> and its interaction partner TRAF6 were ubiquitinated upon TNF stimulation. We also found that STAT2, which mediates signaling by type I

interferons<sup>49</sup>, and USP13, which is involved in the antiviral response by deubiquitinating STING<sup>50</sup>, were ubiquitinated at K161 and K3218, respectively. Our results thus suggest further molecular mechanisms for the crosstalk or cross priming function by TNF to other immune pathways during viral and bacterial infections. In summary, our DIA-based ubiquitin workflow provides an in-depth view on the dynamic ubiquitination of core and peripheral members of TNF-stimulation. Apart from validating the advantages of DIA over DDA, our results provide novel regulatory ubiquitination sites, conveying a more complete picture of the various aspects of TNF $\alpha$  signaling.

#### Circadian rhythm is globally regulated by ubiquitination

In mammals, circadian clocks are driven by interlocked transcription-translation feedback-loops. At the cellular and tissue level, they regulate oscillations of gene expression, protein abundance and post-translation modifications<sup>51-53</sup>. Ubiquitination plays a pivotal role in the core clock machinery (reviewed in<sup>54</sup>), exemplified by the ubiquitin-dependent spatiotemporal regulation of CRY proteins, the major negative clock regulators<sup>55</sup>. Focused studies have provided insights into several ubiquitin-dependent events modulating core clock proteins and their effects<sup>56-58</sup>. Given the unexpected degree of phosphorylation-mediated signaling temporally regulated *in vivo*<sup>51</sup>, we wondered if ubiquitination shows similar oscillations. With the high accuracy and reproducibility of our DIA-based diGly workflow, we reasoned that it would now be possible to obtain high coverage ubiquitinome quantification across a large time series sample set to answer this question.

To this end, we measured the proteome and ubiquitinome of synchronized U2OS cells - a well-established model to study the cell autonomous circadian clock - collected every

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219056>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4 hours in biological quadruplicates across 32 hours (Fig. 4a). Synchronization was validated by assessing the expression profile of core clock transcripts (*Bmal1* and *Per1*) and further confirmed by PER1 and CLOCK oscillations in our proteome data (Supplementary Fig. 5a-b). After filtering for ubiquitinated peptides present in at least half the samples, we obtained 10,886 ubiquitination sites mapping to 3,238 proteins (Fig. 4b, Supplementary Table 5). Measurements were highly reproducible with median Pearson coefficients greater than 0.95 for biological replicates (Supplementary Fig. 5b-c). A total of 7,590 proteins were quantified in the proteome, of which at most 143 oscillated (*q*-value < 0.33). This small percentage of circadian regulation at the proteome level is in line with our previous proteomics results in tissues<sup>52</sup> and with transcriptomics results in this cellular system<sup>59</sup>. Next, we normalized the intensities of the diGly peptides encompassing each ubiquitination site to their corresponding protein abundance. The resulting quantitative values represent the occupancy of the ubiquitin sites irrespective of changes in protein abundance (Methods, Supplementary Fig. 5d).

Periodicity analysis showed that 8% of the ubiquitination sites on 18% of the proteins oscillated in a circadian manner (856 sites; 590 proteins, Methods, *q*-value < 0.1, Fig. 4c and Supplementary Fig. 5e). A large proportion of rhythmic sites peaked with phases clustered around 16-20 hours after synchronization (Fig. 4c, Supplementary Fig. 5e). Remarkably, 59% of these were annotated to be in membrane proteins, many more than expected by chance ( $p < 10^{-12}$ ; Supplementary Table 5). Overrepresentation analysis revealed that these proteins are predominantly involved in transport of a small molecules, such as ions, amines and organic acids (Fig. 4d). These findings point to a

potential metabolic function of circadian membrane protein ubiquitination.

A full quarter of rhythmic ubiquitinated proteins harbored more than one oscillating site (150 sites; Fig. 4e). To investigate the spatial arrangement of them, we developed a bioinformatic proximity analysis tool (available as part of our website for browsing and analyzing the cellular ubiquitinome <http://cyclingubi.biochem.mpg.de>). In 17% of proteins rhythmic ubiquitination sites were closer together than expected by chance ( $p < 0.05$ ) and 73% were annotated as membrane proteins. Interestingly these adjacent sites were mostly located in regions with potential regulatory function, such as N- and C-termini, cytosolic loops and interaction domains (Table 1). For instance, K4, K30 and K37 of the sodium independent cystine-glutamate transporter (SLC7A11, 501 aa) are rhythmically ubiquitinated with similar phases (13.8; 13.3; 13.1 h, respectively, Fig. 4f). Likewise, the potassium chloride symporter NKCC1 (SLC12A2) has a cluster of eight rhythmically ubiquitinated sites in its C-terminal domain with similar phases (K948, K958, K966, K971, K976, K983, K991, K992; Supplementary Fig. 5f). This widely expressed solute carrier plays a key role in the regulation of ionic balance and cell volume<sup>60</sup>. We also discovered novel oscillating ubiquitin modifications in the MAGE domain of MAGED1, a protein that directly interacts with the core clock protein ROR $\alpha$ , to regulate *Bmal1*, Rev-erba and E4bp4 gene expression (Fig. 4g). Interestingly, despite these rhythmic outputs neither the *Maged1* transcript, protein expression nor its binding to ROR $\alpha$  oscillates<sup>61</sup>. Our results now suggest that MAGED1 activity could instead be rhythmically controlled post-translationally through the multiple ubiquitinations in its MAGE domain.

Together, this first in-depth view of the circadian ubiquitinome, made possible by our DIA-based

diGly workflow, reveals this PTM to be a major regulatory mechanism driving rhythmic processes, which include essential cellular processes such as ion transport and osmotic balance.

## DISCUSSION

We here developed a sensitive and robust DIA-based workflow, capable of identifying 35,000 diGly peptides in single run measurements. Both the depth of coverage and the quantitative accuracy are unprecedented and are doubled compared to otherwise identical DDA experiments. Importantly the workflow requires no extra labeling step or offline fractionation, making it streamlined and easy to implement. Furthermore, it could be used for quantification of other PTMs relying on antibody-based enrichment such as lysine acetylation and tyrosine phosphorylation. A current limitation of the DIA method is that, like for any DIA-based analysis, including phosphoproteome analysis<sup>25,30</sup>, the best coverage and quantification is obtained with custom-made, project-specific spectral libraries. Construction of such spectral-libraries requires some effort, specialized equipment for fractionation and may not always be possible for samples with low amounts such as primary cells. Library-free approaches would greatly aid to simplify DIA workflows in the future and there are considerable efforts currently being invested into producing prediction tools for MS/MS spectra and retention time<sup>62</sup>.

Compared to single run TMT-based workflows, DIA suffers from lower throughput. However, the latest advances in nanoflow liquid chromatography now increasingly allow rapid, robust and deep DIA-based proteome and phosphoproteome profiling, which is likely applicable to DIA-based ubiquitinome analysis as well. Furthermore, the LC MS/MS analysis of our workflow requires only a few hundred µg and it

already enables the analysis of systems such as human primary cell culture models where protein material is limited. However, further sensitivity advances are limited by the initial antibody-based enrichment, which currently requires 0.5-1 mg of sample. If this step could be scaled down and the subsequent peptide purification eliminated altogether, sample size requirements could become much smaller yet. A workflow without a peptide-clean-up step would also aid to further improve throughput and reproducibility, making the entire workflow more streamlined.

By converting from a DDA to a DIA workflow we demonstrate a dramatic increase in the number of ubiquitination sites that can consistently and significantly be quantified. Given the inherent sensitivity of our single run approach allowing system-wide investigations of ubiquitination dynamics of biological processes, we applied it to TNF signaling. This provided an in depth view on the ubiquitination dynamics of TNF signaling, covering core and peripheral signaling members, which a parallel DDA analysis failed to provide. Apart from validating advantages of DIA over DDA, our results show that, like phosphorylation, ubiquitination signaling events are rapidly induced after TNF stimulation. Unexpectedly, we still pinpointed novel TNF-regulated sites on proteins that were not previously described in this well-studied pathway. The rich resource provided here could be further explored to investigate the functions of these ubiquitination events in TNF $\alpha$  signaling in health and disease.

System-wide circadian proteomics studies have so far been limited to the dynamic regulation of protein and phosphorylation levels – largely for technological reasons. Our in-depth quantitative diGly analysis of the cell autonomous circadian clock now extends those studies by providing the first cell intrinsic circadian map of ubiquitination dynamics. Quantifying more than 10,000 unique

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ubiquitination sites in synchronized U2OS cells, a standard cellular model of the circadian rhythm, revealed that 8% of them - located on 18% of the quantified ubiquitinated proteins - oscillated in abundance. Many of the cycling sites match into the DIA library of untreated, rather than the library of proteasome inhibited cells suggesting they could have regulatory, non-degradative functions.

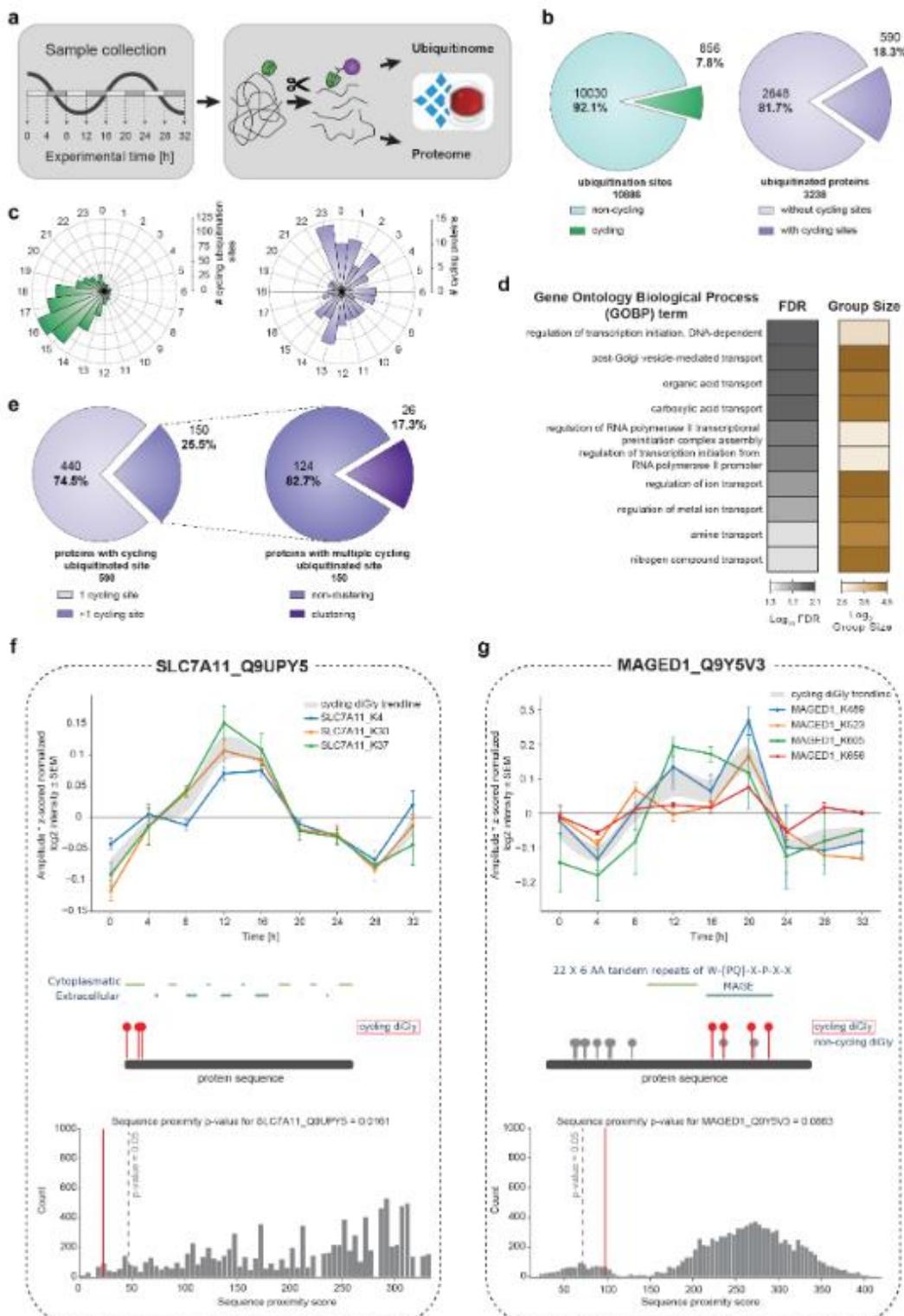
Our data reveal wide-spread ubiquitination of membrane proteins, transporters and receptors, proteins that regulate major cellular processes such as cell volume, ion balance and osmotic homeostasis. Intriguingly, often these cycling ubiquitination sites on membrane proteins are not randomly distributed over the protein sequence but rather cluster in certain regions such as the N- and C-terminus. Circadian rhythms in Mg<sup>2+</sup> and K<sup>+</sup> cellular levels and their transport have been reported in a range of eukaryotic cell types suggesting an evolutionary conservation of this mechanism. Moreover, K<sup>+</sup> transport is a key mechanism driving electrical excitability oscillations in the mammalian master clock and *Drosophila* neurons<sup>63,64</sup> and in turn, plasma membrane potential feeds back to the cellular clock<sup>65,66</sup>. Despite their fundamental cellular role, little is known about the regulatory mechanisms controlling rhythms of ion levels and size in cells<sup>67,68</sup>. Our system-level data suggest that ubiquitination plays a major role in the rhythmic transport of ions and other compounds in the cell by temporally modulating the activity of membrane transporters. Such a mechanism would, for instance, explain the observation that red blood cells lose their daily electrophysiological rhythm after proteasome treatment<sup>67</sup>.

We speculate that ubiquitin-dependent temporal regulation of transporter function for various substrates (e.g. sodium/phosphate/chloride - SLC20A1/SLC20A2, monocarboxylates -

SLC16A1, Sodium/Potassium - ATP1A1, various amino acids - SLC3A2, SLC7A5, SLC7A11, and organic anions - ABCC3) and other receptors (e.g. TGFBR2 or PLXNB2) may serve as temporal cellular switches to sense and respond to daily changes in nutrient availability. Interestingly, in our recent phosphoproteomics study of the synaptic compartment we observed that many of the ubiquitination-related proteins had rhythmic phosphorylation sites<sup>69</sup>. This suggests an interplay between post-translational modifications that together could fine tune daily cycles of membrane-mediated processes essential for proper cellular and tissue metabolism. Given the central role of transporters in chronopharmacology<sup>70-72</sup>, ubiquitin-dependent dynamic regulation of specific membrane transporters are an important functional aspect to consider for drug administration and patient health, both key goals of chronotherapy. The data of our rhythmic ubiquitinome analysis is accessible at <http://cyclingubi.biochem.mpg.de>, opening up new avenues for mechanistic investigations.

# Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



**Fig. 4 Quantification of the rhythmic ubiquitinome**

**a** Experimental workflow for rhythmic ubiquitinome analysis. **b** Proportion of oscillating ubiquitination sites ( $q$ -value  $< 0.1$ ) quantified in  $> 50\%$  of all samples (left panel) and proteins with cycling ubiquitin sites ( $q$ -value  $< 0.1$ ) (right panel). **c** Rose plots indicate phase peaks for cycling ubiquitination sites (left panel) and proteins (right panel). **d** Overrepresentation analysis of gene ontology biological processes (GOBP) filtered for top 10 significant terms. Significance is determined by 5% corrected FDR (Fisher's Exact test). **e** Proportions of proteins with a single and multiple cycling ubiquitination sites (left panel) and those displaying cycling diGly site clusters (right panel). Examples of proximity analysis of cycling ubiquitin clusters (<http://cyclingubi.biochem.mpg.de>). Cycling sites ( $q$ -value  $< 0.1$ ) (top) and their location in the protein sequence along with the domain annotation (middle) and proximity score (average distance,  $p$ -value  $< 0.1$ ) (bottom) for **f** SLC7A11 and **g** MAGED1.

**Table 1 Ubiquitination clusters with potential regulatory circadian functions**

Proteins with multiple cycling ubiquitination sites ( $q$ -value  $< 0.1$ ) in close proximity to each other ( $p$ -value  $< 0.05$ ). Membrane protein annotation by Gene Ontology Cellular Compartment (GOCC) term "membrane" and region/domain classifications are derived from UniProt and manual annotation.

Gene name	UniprotID	Proximity Score	cycling ubiquitination sites	Membrane (GOCC)	Region/Domain
SLC7A5 (LAT1)	Q01650	0.0015	K19, K25, K30	x	N-terminus
SLC16A1 (MCT1)	P53985	0.0036	K216, K223, K224	x	cytosolic loop
LAYN	Q6UX15	0.0064	K272, K273, K297, K311		
EPB41L5	Q9HCM4	0.0073	K508, K510	x	
MYH9	P35579	0.0079	K555, K651, K760, K821	x	Myosin motor, interaction with LIMCH1
RTN4 (NOGO)	F8W914	0.0094	K327, K332, K336	x	Reticulon, C-terminus
ABHD17B	Q5VST6	0.0098	K206, K207	x	
SLC3A2 (MDU1)	F5GZS6	0.0153	K114, K116	x	
SLC7A11 (xCT)	Q9UPY5	0.0161	K4, K30, K37	x	N-terminus
PCNP	Q8WW12	0.0182	K94, K96		
SCRIB (LAP4)	A0A0G2JNZ2	0.0184	K53, K63	x	N-terminus
PLXNB2	O15031	0.0189	K1743, K1757		C-terminus
TOM1	O60784-2	0.0209	K443, K446	x	C-terminus
VLDLR	P98155	0.0220	K828, K839	x	C-terminus
H2AFY	O75367	0.0240	K292, K295		Macro
SLC20A1 (GLVR1, PIT1)	Q8WUM9	0.0262	K286, K320, K394, K399, K456	x	cytosolic loop
KSR1	A0A0A0MQW1	0.0284	K92, K101	x	
HAS2	Q92819	0.0311	K73, K80		
SMARCD2	B9EGA3	0.0318	K200, K207		
TAX1BP1 (T6BP)	Q86VP1-2	0.0321	K561, K571		
SLC20A2 (GLVR2, PIT2)	Q08357	0.0394	K262, K272	x	cytosolic loop
HSP90AB1 (HSP90B)	P08238	0.0411	K568, K577	x	Interaction with NR3C1
SLC12A2 (NKCC1)	G3XAL9	0.0420	K237, K948, K958, K966, K971, K976, K983, K991, K992, K1125	x	SLC12
PCDHB5	Q9Y5E4	0.0426	K767, K784	x	C-terminus
PPAP2B (LPP3)	O14495	0.0455	K8, K15	x	N-terminus
SCAMP1	A0A087WXB0	0.0465	K63, K71	x	N-terminus

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219056>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## ACKNOWLEDGEMENT

FB and MSR were supported by the Volkswagen Foundation (93071). MSR also received funding from the German Research Foundation DFG (Project 329628492- SFB1321, INST 86/1800-1 FUGG and project 428041612). We thank all the members of the Department of Signal Transduction and Proteomics, in particular, Igor Paron and Christian Deiml for MS technical assistance, Bianca Splettstoesser for technical help with experimental work and Johannes Müller, Alex Strasser and Lisa Schweitzer for columns. We also thank Sharah Meszaros and Steve Dewitz for cell culture assistance. We are grateful to Arno F. Alpi, Jesper Olsen, Chuna Choudhary, André Michaelis and Jakob Bader for constructive and insightful discussions. We specially thank Florian Gnad and Cell Signaling Technologies (CST) for gifting of PTMScan® Ubiquitin Remnant Motif (K-e-GG) Kits.

## AUTHOR CONTRIBUTIONS

FMH, OK, MT, FB and MSR designed experiments. TNF experiments were performed by FMH and MT. Circadian experiments were conducted by FMH and FB. Computational proximity analysis was performed by IB. Data was analyzed by FMH. FMH, OK, MT, MSR, BAS and MM wrote, reviewed and edited the manuscript. All authors read and commented on the manuscript.

## DECLARATION OF INTERESTS

All authors declare that they have no conflict of interest.

## METHODS

### Cell culture, treatment, harvest and lysis

HEK293 (human, DMSZ, ACC 635) and U2OS (human, American Type Culture Collection [ATCC], HTB-96) cell were cultivated in DMEM (Gibco, Invitrogen) supplemented with 10% fetal bovine serum (Gibco, Invitrogen), 100 U/ml penicillin (Gibco, Invitrogen) and 100 µg/ml streptomycin (Gibco, Invitrogen) at 37°C in a humidified incubator with a 5% CO<sub>2</sub> atmosphere. For cell harvest, cells were washed twice with ice-cold PBS (Gibco, Invitrogen), centrifuged, snap frozen in liquid nitrogen and stored at -80°C until lysis. Frozen cell pellets were lysed by adding lysis buffer (1% SDC in 100 mM Tris/HCl, pH 8.5) directly onto frozen cell pellets, followed by repeated aspiration and boiling for 5 min at 95°C.

For proteasome inhibition, HEK293 or U2OS cells were treated with 10 µM MG132 (InvivoGen) at approximately 80% confluence for 4 h and successively harvested. For circadian cycle experiments, cells were synchronized, when they reached at least 90% confluence, with dexamethasone (1 µM) for 1 h. Following this, U2OS were washed once with PBS and the medium was replaced. The first time point was collected after 24 hours of synchronization continuing the collection every 4 hours across 32 hours for each of the 4 biological replicates. Collected cells, stored and lysed as described. For TNF stimulation of U2OS cells, confluent cultures were either stimulated with 100 ng/ml TNF for 10 minutes or left unstimulated. Cells were washed 3 x with ice-cold PBS, directly lysed with lysis buffer and boiled for 5 min at 95°C.

### Western blot analysis

U2OS cells were plated in 6 well plates and when confluent stimulated for 5, 10, 15, 30, 60 minutes with 100 ng/ml TNF or left untreated. After stimulation, cells were washed in PBS and lysed

in 4% SDS in 100 mM Tris/HCl, pH 8. Lysates were boiled, sonicated and protein concentrations were estimated using BCA. SDS sample loading buffer (450 mM Tris-HCl, pH 8, 60% (v/v) glycerol, 12% (w/v) SDS, 0.02% (w/v) bromophenol blue, 600 mM DTT) was added to lysates before separation on 12% Novex Tris-glycine gels (Thermo Fisher Scientific, XP00120BOX). Separated proteins were transferred onto PVDF membranes (Merck Millipore, IPVH00010). Membranes were blocked in 5% BSA in PBST and antibodies diluted in 2% BSA in PBST. Antibodies used for immunoblotting were as follows: anti phospho p65 (CST, 3033P), anti p65 (CST, 4764P), anti I $\kappa$ B $\alpha$  (CST, 92424792), anti phospho p38 (CST, 9215), anti p38 (CST, 9212), anti  $\beta$ -actin (CST, 4970).

#### RNA isolation and QPCR

RNA was isolated from 3 biological replicates of each U2OS time point according to manufacture instruction using the RNeasy Plus Mini Kit (QIAGEN, #74134). Isolated RNA was reversely transcribed by using first-strand cDNA synthesis kit (Thermo Fisher Scientific, #K1612). QPCR was performed at the C1000 Thermal Cycler (Bio-Rad) with iQ<sup>TM</sup> SYBR Green Supermix (Bio-Rad, #170-8862) with following primers:

Target gene	Forward primer	Reverse primer
Bmal1	caggaaaaatag gccgaatg	gcgatgacccctct tatcctg
Per1	ggacactctgc gaccag	gggagtggaggtag aagatctaa
Gapdh	agccacatcgct cagacac	gccaatacgacc aaatcc

The in-build analysis tool of the CFX Manager Software (Version 3.1, Bio-Rad) was used to determine the normalized expression with the  $\Delta\Delta Cq$  method of Bmal1 and Per1 compared to Gapdh in technical triplicates for all 3 biological replicates of each time point. The technical

triplicates were further averaged and adjusted so that the highest value was set to 1. Following this, the average of all biological replicates and the SEM (standard error of the mean) was calculated for all the time points.

#### Protein digestion and peptide cleanup

Lysates were sonicated for 1 min (Branson Sonifier) and protein concentrations were estimated by tryptophan assay. After addition of CAA and TCEP to a final concentration of 10 mM and 40 mM, respectively, samples were incubated for 5 min at 45°C for protein reduction and alkylation. Thereafter, Samples were digested overnight at 37°C using trypsin (1:100 w/w, Sigma-Aldrich) and LysC (1/100 w/w, Wako).

For proteome analysis, sample aliquots (~15 $\mu$ g) were desalting in SDB-RPS StageTips (Empore). Briefly, samples were first diluted with 1% TFA in isopropanol to a final volume of 200  $\mu$ l. Thereafter, samples were loaded onto StageTips and sequentially washed with 200  $\mu$ l of 1% TFA in isopropanol and 200  $\mu$ l 0.2% TFA/ 2% ACN. Peptides were eluted with 60  $\mu$ l of 1.25% Ammonium hydroxide (NH<sub>4</sub>OH)/ 80% ACN and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus). Dried peptides were resuspended in buffer A\* (2% ACN/ 0.1% TFA) supplemented with iRT peptides (1/30 v/v) (iRT Standard, Biognosys).

For diGly peptide enrichment, samples were four-fold diluted with 1% TFA in isopropanol and loaded onto SDB-RPS cartridges (Strata<sup>TM</sup>-X-C, 30 mg/ 3 ml or Strata<sup>TM</sup>-X-C, 200 mg/ 6 ml, Phenomenex Inc.). Before peptide loading, cartridges were equilibrated with 8 bed volumes (BV) 30% MeOH/1% TFA and washed with 8 BV of 0.2% TFA. Samples were loaded by gravity flow and sequentially washed twice with 8 BV 1% TFA in isopropanol and once with 8 BV 0.2% TFA/ 2% ACN. Peptides were eluted twice

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with 4 BV 1.25% NH<sub>4</sub>OH/ 80% ACN and diluted with ddH<sub>2</sub>O to a final ACN concentration of 35% ACN. Thereafter, samples were snap frozen in liquid nitrogen, lyophilized and store at 4°C until diGly peptide enrichment.

#### DiGly peptide enrichment

Lyophilized peptides were resuspended in immunoaffinity purification buffer (IAP) (50 mM MOPS, pH 7.2, 10 mM Na<sub>2</sub>HPO<sub>4</sub>, 50 mM NaCl) and sonicated for 2.5 min (Bioruptor plus, Diagenode). Peptide concentration was estimated by tryptophan assay. DiGly remnant containing peptides were enriched using the PTMScan® Ubiquitin Remnant Motif (K-ε-GG) Kit (Cell Signaling Technology (CST)), which was kindly provided by CST. First, antibodies were crosslinking to beads as described by Udeshi et al.<sup>22</sup>. Unless otherwise stated, 1/8 of a vial of crosslinked antibody beads and 1 mg of peptide material were used for diGly peptide enrichments. For this, peptides were added to crosslinked antibody beads and the volume was adjusted to 1 ml with IAP buffer. After 1h of incubation at 4°C and gentle agitation, beads were washed twice with cold IAP and 5 times with cold ddH<sub>2</sub>O. For this, beads were transferred into GF-filter StageTips and for each wash step, the according wash solution was added and passed through by centrifugal force. Thereafter, GF-StageTips were stacked onto SDB-RPS StageTips and peptides were directly eluted into SDB-RPS StageTips. For this, 50 µl 0.15% TFA were added twice onto the beads and passed through by centrifugation for 5 min at 100 g. Thereafter, 100 µl 0.2% TFA were added on top of peptide eluates, followed by sample loading onto the stationary material of SDB-RPS StageTips. Peptides were washed, eluted and dried as described for proteomes samples, with the difference, that 0.2% TFA was used for the first wash step. Dried peptides were resuspended in 9 µl buffer A\*, supplemented

with iRT peptides (1/30 v/v) for LC/MS-MS analysis.

#### Basic reversed phase fractionation

Basic reversed phase (bRP) fractionation for diGly peptide and proteome spectral libraries were performed on an UFC system (Shimadzu) and EASY-nLC 1000 (Thermo Fisher Scientific, Germany), respectively.

For diGly peptide separation, lyophilized samples were resuspended in Buffer A (5 mM NH<sub>4</sub>HCO<sub>3</sub>/ 2% ACN) and 5 mg peptide material (5 mg/ml) was loaded onto a reversed phase column (ZORBAX 300Extend-C18, Agilent). Peptides were separate at a flow rate of 2 ml/min and a constant column temperature of 40 °C using a binary buffer system, consisting of buffer A and buffer B (5 mM NH<sub>4</sub>HCO<sub>3</sub>/ 90% ACN). An elution gradient staring at 0% B stepwise increased to 28 in 53 min and to 78 in 6 min was deployed. Eluting peptides were automatically collected into a 96-deepwell plate while well positions were switched in 40 s intervals.

For peptide fractionation on the EASY-nLC 1000 system, approximately 55 µg peptide material were loaded onto a 30 cm in-house packed, reversed phase columns (250 µm inner diameter, ReproSil-Pur C18-AQ 1.9 µm resin [Dr. Maisch GmbH]). Peptides were separate at a flow rate of 2 µl/min using a binary buffer system of buffer A (PreOmics) and buffer B (PreOmics). An elution gradient staring at 3% B stepwise increased to 30% in 45 min, 60% in 17 min and 95% in 5 min was used. Eluting peptides were concatenated into 24 fractions by switching the rotor valve of an automated concatenation system (Spider fractionator, PreOmics)<sup>73</sup> in 90 s intervals.

#### Library sample preparation

For individual deep diGly libraries 2×5 mg peptide were fractionated by bRP fractionation. For K48-peptide containing fraction

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identification, 100 $\mu$ l aliquots of fractions 46 to 54 were dried in a SpeedVac, resuspended in A\* and measured on an LTQ Orbitrap XL mass spectrometer. K48-peptide containing fractions of both plates, were pooled in sample pool "K48" (Supplementary Fig. 1a). Remaining fractions of both plates were concatenated into P1-P8 (Supplementary Fig. 1a), snap frozen and lyophilized. Lyophilized peptides were resuspended in 1 ml IAP buffer and diGly peptides were enriched as described above. In case of HEK293 library generation, an optional second supernatant IP was conducted. For this, 500 $\mu$ l of previous diGly peptide enrichment supernatants were pooled as indicated (Supplementary Fig. 1a) and used for sequential diGly peptide enrichment.

For the proteome library, aliquots of U2OS samples for proteome cycling analysis were used. Approximately 3  $\mu$ g peptide material of individual time points of two biological replicates, after SDB-RPS cleanup, were pooled and fractionate via bRP fractionation as described above. Fractionated samples were dried using a SpeedVac and resuspended in A\* supplemented with iRT peptides (1/30 v/v) for LC-MS/MS measurement and spectral library generation.

#### Nano-flow LC-MS/MS proteome measurements

Peptides were loaded onto a 50 cm, in-house packed, reversed phase columns (75  $\mu$ m inner diameter, ReproSil-Pur C18-AQ 1.9  $\mu$ m resin [Dr. Maisch GmbH]). The column temperature was controlled at 60°C using a homemade column oven and binary buffer system, consisting of buffer A (0.1% formic acid (FA)) and buffer B (0.1% FA in 80% ACN), was utilized for low pH peptide separation. An EASY-nLC 1200 system (Thermo Fisher Scientific), directly coupled online with the mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) via a nano-electrospray source, was employed for nano-flow

liquid chromatography, at a flow rate of 300 nL/min. For individual measurements, 500 ng of peptide material was loaded and eluted with a gradient starting at 5% buffer B and stepwise increased to 30% in 95 min, 60% in 5 min and 95% in 5 min.

The same general setup was used, for K48-peptide containing fraction identification, while the column and mass spectrometer were changed to a 20 cm column and an LTQ Orbitrap XL, respectively.

For DDA experiments the Thermo Xcalibur (4.0.27.19) and LTQ Tune plus (2.5.5 SP2) software were used for Q Exactive HF-X and LTQ Orbitrap XL instruments, respectively. The Q Exactive HF-X was operated in Top12 mode with a full scan range of 300-1650  $m/z$  at a resolution of 60,000. The automatic gain control (AGC) was set to 3e6 at a maximum injection time of 20 s. Precursor ion selection width was kept at 1.4  $m/z$  and fragmentation was achieved by higher-energy collisional dissociation (HCD) (NCE 27%). Fragment ion scans were recorded at a resolution of 15,000, an AGC of 1e5 and a maximum fill time of 60 ms. Dynamic exclusion was enabled and set to 20 s. The LTQ Orbitrap XL was operated in Top10 mode with a full scan range of 300-1700  $m/z$  at a resolution of 60,000. Precursor ion selection width was kept at 2.0  $m/z$  and fragmentation was achieved by collision-induced dissociation (CID) (NCE 35%).

For DIA analysis, the MaxQuant Live software suite was utilized for data acquisition<sup>74</sup>. The full scan range was set to 300-1650  $m/z$  at a resolution of 120,000. The AGC was set to 3e6 at a maximum injection time of 60 ms. HCD (NCD 27%) was used for precursor fragmentation and fragment ions were analyzed in 33 DIA windows at a resolution of 30,000, while the AGC was kept at 3e6.

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

#### Nano-flow LC-MS/MS diGly measurements

DiGly peptide enriched samples were measured on a Q Exactive HF-X using the same instrumental setup as for proteome analysis. For diGly single run measurements one quarter (2  $\mu$ l) and for diGly library preparation one-half (4  $\mu$ l) of enriched samples were loaded for LC-MS/MS analysis, unless stated otherwise. Loaded peptides were eluted using a gradient starting at 3% buffer B and stepwise increased to 7% in 6 min, 20% in 49 min, 36% in 39 min, 45% in 10 min and 95% in 4 min.

For DDA analysis, the MS was operated in Top12 mode with a full scan range of 300-1350  $m/z$  at a resolution of 60,000. AGC was set to 3e6 at a maximum injection time of 20 s. Precursor ion selection width was kept at 1.4  $m/z$  and fragmentation was achieved by HCD (NCE 28%). Fragment ion scans were recorded at a resolution of 30,000, an AGC of 1e5 and a maximum fill time of 110 ms. Dynamic exclusion was enabled and set to 30 s.

For DIA analysis, the MaxQuant Live software suite was employed for data acquisition<sup>74</sup>. The full scan range was set to 300-1650  $m/z$  at a resolution of 120,000. The AGC was set to 3e6 at a maximum injection time of 60 ms. HCD (NCD 28%) was used for precursor fragmentation and resulting fragment ions were analyzed in 46 DIA windows at a resolution of 30,000 (unless otherwise stated) and an AGC of 3e6. DIA window distribution parameters PdfMu and PdfSigma were set to 6.161865 and 0.348444, respectively, unless stated otherwise.

#### Raw data analysis

DDA raw data used for K48-peptide fraction identification and DIA and DDA comparisons were analyzed with MaxQuant (1.6.2.10) using default settings and enabled match between runs (MBR) functionality. Carbamidomethyl (C) was defined as fixed modification and Oxidation (M),

Acetyl (Protein N-term) and DiGly (K) were set as variable modifications.

DDA raw data, used for spectral library construction, were processed with Spectronauts build in search engine pulsar (13.12.200217.43655)<sup>28</sup>. Default settings were used for proteome spectral libraries. For diGly spectral libraries the “Best N Fragments per peptides” maximum value was adjusted to 25. For hybrid library construction DIA raw files were processed together with DDA library raw files using the same search settings.

DIA raw files were processed using Spectronaut (13.12.200217.43655)<sup>28</sup>. Proteome analysis was performed with default settings. For diGly analysis, diGly (K) was defined as an additional variable modification and PTM localization was enabled and set to 0. For dilution experiments, “XIC RT extraction window” was set to “static” with a window width of 10 min. Direct DIA searches used the same settings as described above.

#### Bioinformatics analysis

Data analysis was primarily performed in the Perseus software suite (1.6.7.0). For diGly site analysis, Spectronaut normal report output tables were aggregated to diGly sites using the peptide collapse plug-in tool for Perseus<sup>30</sup>. DiGly sites were aggregated using the linear model based approach and filtered for a localization probability  $> 0.5$ . Student t-test statistics (FDR cutoff 1% or 5%;  $s_0 = 0.1$ ) for TNF stimulation experiments were performed in Perseus. Fisher’s Exact GOBP Term enrichment of upregulated diGly sites and cycling diGly sites was performed on the pantherdb website (<http://pantherdb.org/>) and in perseus, respectively, with Benjamini Hochberg FDR correction enabled and set to a 5% cutoff. Network representation of upregulated diGly sites was performed with the STRING app (1.5.1) in Cytoscape (3.7.2).

For the cycling analysis of diGly sites, data was first filtered for diGly sites identified in at least 50% across all measurements. Proteins and diGly sites raw intensities were  $\log_2$  transformed and normalized by median subtraction. For diGly site protein normalization the median values of biological quadruplicates were subtracted from normalized diGly sites. Missing values of protein data for subtraction were imputed based on a Gaussian normal distribution with a width of 0.3 and a downshift of 1.8. Cycling analysis of normalized protein and diGly site data was performed as previously described, but in this case with a period time of  $24.8\text{ h}^{51,52}$ . A q-value cutoff of  $< 0.1$  and  $< 0.33$  was used to define cycling DiGly sites and proteins, respectively.

#### Website tool

For profile plots individual z-scores for each protein abundance normalized diGly site and the median z-score and standard error of means (SEM) were subsequently determined for each time point. The resulting median z-scores and SEM values were multiplied with the cycling amplitude of each diGly site (Perseus periodicity analysis output). For sequence visualization and protein domain annotation each diGly site location was mapped to the first UniProt ID of its assigned protein group and was visualized based on its respective protein sequence stored in the fasta file that was used for MS/MS data analysis (human fasta, downloaded 2015). The protein sequences for visualization were obtained using the 'fasta' functions from pteomics<sup>73,76</sup>. Information about protein domains was obtained from UniProt (<https://www.uniprot.org/>, accessed 25.05.2020), including following categories: 'Topological domain', 'Motif', 'Region', 'Repeat', 'Zink finger' and 'Domain [FT]'.

To evaluate whether multiple observed cycling diGly sites are located in a specific region on the

protein, we performed a proximity analysis. Three different metrics were evaluated: (1) the average distance (In amino acids) between all observed cycling diGly sites, (2) the minimum distance between any two observed cycling diGly sites, and (3) the maximum distance between any two observed cycling diGly sites. The observed distance metrics were compared to the distances expected from a random distribution of the diGly sites of a protein across all of its lysines. 10,000 random distributions were considered and an empirical p-value was estimated based on the fraction of random samples with a smaller or equally small distance metric as the observed cycling diGly sites. For the main analysis, diGly sites with a q-value  $\leq 0.1$  were considered as cycling diGly sites.

Data preprocessing and visualization for the dashboard was performed using the python programming language. Following libraries were utilized for data processing: numpy, pandas, re, random and pteomics<sup>73,76</sup>. Several libraries from the HoloViz family of tools were used for data visualization and creation of the dashboard, including panel and holoviews, but also bokeh, plotly and matplotlib.

#### REFERENCES

- 1 Swatek, K. N. & Komander, D. Ubiquitin modifications. *Cell Res* **26**, 399-422, doi:10.1038/cr.2016.39 (2016).
- 2 Yau, R. & Rape, M. The increasing complexity of the ubiquitin code. *Nat Cell Biol* **18**, 579-586, doi:10.1038/ncb3358 (2016).
- 3 Karbowski, M. & Youle, R. J. Regulating mitochondrial outer membrane proteins by ubiquitination and proteasomal degradation. *Curr Opin Cell Biol* **23**, 476-482, doi:10.1016/j.ceb.2011.05.007 (2011).
- 4 Spratt, D. E. *et al.* A molecular explanation for the recessive nature of parkin-linked Parkinson's disease. *Nat*

5 *Commun* 4, 1983, doi:10.1038/ncomms2983 (2013).

6 Seymour, R. E. *et al.* Spontaneous mutations in the mouse Sharpin gene result in multiorgan inflammation, immune system dysregulation and dermatitis. *Genes Immun* 8, 416-421, doi:10.1038/sj.gene.6364403 (2007).

7 Boisson, B. *et al.* Immunodeficiency, autoinflammation and amylopectinosis in humans with inherited HOIL-1 and LUBAC deficiency. *Nat Immunol* 13, 1178-1186, doi:10.1038/ni.2457 (2012).

8 Babu, J. R., Geetha, T. & Wooten, M. W. Sequestosome 1/p62 shuttles polyubiquitinated tau for proteasomal degradation. *J Neurochem* 94, 192-203, doi:10.1111/j.1471-4159.2005.03181.x (2005).

9 Varfolomeev, E. *et al.* c-IAP1 and c-IAP2 are critical mediators of tumor necrosis factor alpha (TNFalpha)-induced NF-kappaB activation. *J Biol Chem* 283, 24295-24299, doi:10.1074/jbc.C800128200 (2008).

10 Zucchelli, S. *et al.* Tumor necrosis factor receptor-associated factor 6 (TRAF6) associates with huntingtin protein and promotes its atypical ubiquitination to enhance aggregate formation. *J Biol Chem* 286, 25108-25117, doi:10.1074/jbc.M110.187591 (2011).

11 Doll, S. & Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem Biol* 10, 63-71, doi:10.1021/cb500904b (2015).

12 Peng, J. *et al.* A proteomics approach to understanding protein ubiquitination. *Nat Biotechnol* 21, 921-926, doi:10.1038/nbt849 (2003).

13 Mayor, T. & Deshaies, R. J. Two-step affinity purification of multiubiquitylated proteins from *Saccharomyces cerevisiae*. *Methods Enzymol* 399, 385-392, doi:10.1016/S0076-6879(05)99026-5 (2005).

14 Hjerpe, R. *et al.* Efficient protection and isolation of ubiquitylated proteins using tandem ubiquitin-binding entities. *EMBO Rep* 10, 1250-1258, doi:10.1038/embor.2009.192 (2009).

15 Xu, G., Paige, J. S. & Jaffrey, S. R. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat Biotechnol* 28, 868-873, doi:10.1038/nbt.1654 (2010).

16 Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* 44, 325-340, doi:10.1016/j.molcel.2011.08.025 (2011).

17 Wagner, S. A. *et al.* A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol Cell Proteomics* 10, M111 013284, doi:10.1074/mcp.M111.013284 (2011).

18 Bustos, D., Bakalarski, C. E., Yang, Y., Peng, J. & Kirkpatrick, D. S. Characterizing ubiquitination sites by peptide-based immunoaffinity enrichment. *Mol Cell Proteomics* 11, 1529-1540, doi:10.1074/mcp.R112.019117 (2012).

19 Akimov, V. *et al.* UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. *Nat Struct Mol Biol* 25, 631-640, doi:10.1038/s41594-018-0084-y (2018).

20 Wagner, S. A. *et al.* Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. *Mol Cell Proteomics* 11, 1578-1585, doi:10.1074/mcp.M112.017905 (2012).

21 Elia, A. E. *et al.* Quantitative Proteomic Atlas of Ubiquitination and Acetylation in the DNA Damage Response. *Mol Cell* 59, 867-881, doi:10.1016/j.molcel.2015.05.006 (2015).

22 Rose, C. M. *et al.* Highly Multiplexed Quantitative Mass Spectrometry Analysis of Ubiquitylomes. *Cell Syst* 3, 395-403 e394, doi:10.1016/j.cels.2016.08.009 (2016).

Udeshi, N. D., Mertins, P., Svinkina, T. & Carr, S. A. Large-scale identification of ubiquitination sites by mass spectrometry. *Nat Protoc* 8, 1950-1960, doi:10.1038/nprot.2013.120 (2013).

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23 van der Wal, L. *et al.* Improvement of ubiquitylation site detection by Orbitrap mass spectrometry. *J Proteomics* **172**, 49-56, doi:10.1016/j.jprot.2017.10.014 (2018).

24 Udeshi, N. D. *et al.* Rapid and deep-scale ubiquitylation profiling for biology and translational research. *Nat Commun* **11**, 359, doi:10.1038/s41467-019-14175-1 (2020).

25 Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol* **14**, e8126, doi:10.15252/msb.20178126 (2018).

26 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).

27 Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J Proteome Res* **17**, 727-738, doi:10.1021/acs.jproteome.7b00602 (2018).

28 Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* **14**, 1400-1410, doi:10.1074/mcp.M114.044305 (2015).

29 Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**, 39-45, doi:10.1038/nmeth705 (2004).

30 Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun* **11**, 787, doi:10.1038/s41467-020-14609-1 (2020).

31 Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520, doi:10.1093/nar/gku1267 (2015).

32 Harhaj, E. W. & Dixit, V. M. Deubiquitinases in the regulation of NF-kappaB signaling. *Cell Res* **21**, 22-39, doi:10.1038/cr.2010.166 (2011).

33 Silke, J. The regulation of TNF signalling: what a tangled web we weave. *Curr Opin Immunol* **23**, 620-626, doi:10.1016/j.co.2011.08.002 (2011).

34 Wagner, S. A., Satpathy, S., Beli, P. & Choudhary, C. SPATA2 links CYLD to the TNF-alpha receptor signaling complex and modulates the receptor signaling outcomes. *EMBO J* **35**, 1868-1884, doi:10.15252/embj.201694300 (2016).

35 Andersen, N. N. & Jess, T. Risk of infections associated with biological treatment in inflammatory bowel disease. *World J Gastroenterol* **20**, 16014-16019, doi:10.3748/wjg.v20.i43.16014 (2014).

36 Kim, S. Y. & Solomon, D. H. Tumor necrosis factor blockade and the risk of viral infection. *Nat Rev Rheumatol* **6**, 165-174, doi:10.1038/nrrheum.2009.279 (2010).

37 Dziedzic, S. A. *et al.* ABIN-1 regulates RIPK1 activation by linking Met1 ubiquitylation with Lys63 deubiquitylation in TNF-RSC. *Nat Cell Biol* **20**, 58-68, doi:10.1038/s41556-017-0003-1 (2018).

38 Weinlich, R. & Green, D. R. The two faces of receptor interacting protein kinase-1. *Mol Cell* **56**, 469-480, doi:10.1016/j.molcel.2014.11.001 (2014).

39 Hsu, H., Huang, J., Shu, H. B., Baichwal, V. & Goeddel, D. V. TNF-dependent recruitment of the protein kinase RIP to the TNF receptor-1 signaling complex. *Immunity* **4**, 387-396, doi:10.1016/s1074-7613(00)80252-6 (1996).

40 Tokunaga, F. & Iwai, K. LUBAC, a novel ubiquitin ligase for linear ubiquitination, is crucial for inflammation and immune responses. *Microbes Infect* **14**, 563-572,

doi:10.1016/j.micinf.2012.01.011 (2012).

41 Ikeda, F. *et al.* SHARPIN forms a linear ubiquitin ligase complex regulating NF- $\kappa$ B activity and apoptosis. *Nature* **471**, 637-641, doi:10.1038/nature09814 (2011).

42 Heger, K. *et al.* OTULIN limits cell death and inflammation by deubiquitinating LUBAC. *Nature* **559**, 120-124, doi:10.1038/s41586-018-0256-2 (2018).

43 Savinova, O. V., Hoffmann, A. & Ghosh, G. The Nfkb1 and Nfkb2 proteins p105 and p100 function as the core of high-molecular-weight heterogeneous complexes. *Mol Cell* **34**, 591-602, doi:10.1016/j.molcel.2009.04.033 (2009).

44 Collins, P. E., Mitxitorena, I. & Carmody, R. J. The Ubiquitination of NF- $\kappa$ B Subunits in the Control of Transcription. *Cells* **5**, doi:10.3390/cells5020023 (2016).

45 Kravtsova-Ivantsiv, Y. *et al.* KPC1-mediated ubiquitination and proteasomal processing of NF- $\kappa$ B1 p105 to p50 restricts tumor growth. *Cell* **161**, 333-347, doi:10.1016/j.cell.2015.03.001 (2015).

46 Lin, L., DeMartino, G. N. & Greene, W. C. Cotranslational biogenesis of NF- $\kappa$ B p50 by the 26S proteasome. *Cell* **92**, 819-828, doi:10.1016/s0092-8674(00)81409-9 (1998).

47 Ciechanover, A. *et al.* Mechanisms of ubiquitin-mediated, limited processing of the NF- $\kappa$ B1 precursor protein p105. *Biochimie* **83**, 341-349, doi:10.1016/s0300-9084(01)01239-1 (2001).

48 Schauvliege, R., Janssens, S. & Beyaert, R. Pellino proteins: novel players in TLR and IL-1R signalling. *J Cell Mol Med* **11**, 453-461, doi:10.1111/j.1582-4934.2007.00040.x (2007).

49 Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat Rev Immunol* **14**, 36-49, doi:10.1038/nri3581 (2014).

50 Sun, H. *et al.* USP13 negatively regulates antiviral responses by deubiquitinating STING. *Nat Commun* **8**, 15534, doi:10.1038/ncomms15534 (2017).

51 Robles, M. S., Humphrey, S. J. & Mann, M. Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metab* **25**, 118-127, doi:10.1016/j.cmet.2016.10.004 (2017).

52 Robles, M. S., Cox, J. & Mann, M. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet* **10**, e1004047, doi:10.1371/journal.pgen.1004047 (2014).

53 Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci USA* **111**, 16219-16224, doi:10.1073/pnas.1408886111 (2014).

54 Stojkovic, K., Wing, S. S. & Cermakian, N. A central role for ubiquitination within a circadian clock protein modification code. *Front Mol Neurosci* **7**, 69, doi:10.3389/fnmol.2014.00069 (2014).

55 Yoo, S. H. *et al.* Competing E3 ubiquitin ligases govern circadian periodicity by degradation of CRY in nucleus and cytoplasm. *Cell* **152**, 1091-1105, doi:10.1016/j.cell.2013.01.055 (2013).

56 Ullah, K. *et al.* The E3 ubiquitin ligase STUB1 attenuates cell senescence by promoting the ubiquitination and degradation of the core circadian regulator BMAL1. *J Biol Chem* **295**, 4696-4708, doi:10.1074/jbc.RA119.011280 (2020).

57 Correia, S. P. *et al.* The circadian E3 ligase complex SCF(FBXL3+CRY) targets TLK2. *Sci Rep* **9**, 198, doi:10.1038/s41598-018-36618-3 (2019).

58 D'Alessandro, M. *et al.* Stability of Wake-Sleep Cycles Requires Robust Degradation of the PERIOD Protein. *Curr Biol* **27**, 3454-3467 e3458, doi:10.1016/j.cub.2017.10.014 (2017).

bioRxiv preprint doi: <https://doi.org/10.1101/2020.07.24.219055>; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

59 Hughes, M. E. *et al.* Harmonics of circadian gene transcription in mammals. *PLoS Genet* **5**, e1000442, doi:10.1371/journal.pgen.1000442 (2009).

60 Demian, W. L. *et al.* The Ion Transporter NKCC1 Links Cell Volume to Cell Mass Regulation by Suppressing mTORC1. *Cell Rep* **27**, 1886-1896 e1886, doi:10.1016/j.celrep.2019.04.034 (2019).

61 Wang, X. *et al.* Interaction of MAGED1 with nuclear receptors affects circadian clock function. *EMBO J* **29**, 1389-1400, doi:10.1038/emboj.2010.34 (2010).

62 Zhang, T. *et al.* High-throughput discovery of genetic determinants of circadian misalignment. *PLoS Genet* **16**, e1008577, doi:10.1371/journal.pgen.1008577 (2020).

63 Kuhlman, S. J. & McMahon, D. G. Rhythmic regulation of membrane potential and potassium current persists in SCN neurons in the absence of environmental input. *Eur J Neurosci* **20**, 1113-1117, doi:10.1111/j.1460-9568.2004.03555.x (2004).

64 Flourakis, M. *et al.* A Conserved Bicycle Model for Circadian Clock Control of Membrane Excitability. *Cell* **162**, 836-848, doi:10.1016/j.cell.2015.07.036 (2015).

65 Nitabach, M. N., Sheeba, V., Vera, D. A., Blau, J. & Holmes, T. C. Membrane electrical excitability is necessary for the free-running larval Drosophila circadian clock. *J Neurobiol* **62**, 1-13, doi:10.1002/neu.20053 (2005).

66 Jones, J. R., Tackenberg, M. C. & McMahon, D. G. Manipulating circadian clock neuron firing rate resets molecular circadian rhythms and behavior. *Nat Neurosci* **18**, 373-375, doi:10.1038/nn.3937 (2015).

67 Henslee, E. A. *et al.* Rhythmic potassium transport regulates the circadian clock in human red blood cells. *Nat Commun* **8**, 1978, doi:10.1038/s41467-017-02161-4 (2017).

68 Martins, B. M. C., Tooke, A. K., Thomas, P. & Locke, J. C. W. Cell size control driven by the circadian clock and environment in cyanobacteria. *Proc Natl Acad Sci U S A* **115**, E11415-E11424, doi:10.1073/pnas.1811309115 (2018).

69 Bruning, F. *et al.* Sleep-wake cycles drive daily dynamics of synaptic phosphorylation. *Science* **366**, doi:10.1126/science.aav3617 (2019).

70 Dallmann, R., Brown, S. A. & Gachon, F. Chronopharmacology: new insights and therapeutic implications. *Annu Rev Pharmacol Toxicol* **54**, 339-361, doi:10.1146/annurev-pharmtox-011613-135923 (2014).

71 Tahara, Y. & Shibata, S. Chronobiology, chrono-pharmacology, and chrono-nutrition. *J Pharmacol Sci* **124**, 320-335, doi:10.1254/jphs.13r06cr (2014).

72 Ozturk, N., Ozturk, D., Kavakli, I. H. & Okyar, A. Molecular Aspects of Circadian Pharmacology and Relevance for Cancer Chronotherapy. *Int J Mol Sci* **18**, doi:10.3390/ijms18102168 (2017).

73 Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol Cell Proteomics* **16**, 694-705, doi:10.1074/mcp.O116.065136 (2017).

74 Wichmann, C. *et al.* MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Mol Cell Proteomics* **18**, 982-994, doi:10.1074/mcp.TIR118.001131 (2019).

75 Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V. & Gorshkov, M. V. Pyteomics--a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J Am Soc Mass Spectrom* **24**, 301-304, doi:10.1007/s13361-012-0516-6 (2013).

76 Levitsky, L. I., Klein, J. A., Ivanov, M. V. & Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J Proteome Res* **18**, 709-714, doi:10.1021/acs.jproteome.8b00717 (2019).

### 3.2.2. UBL3 modification influences protein sorting to small extracellular vesicles

Hiroshi Ageta<sup>1,2</sup>, Natsumi Ageta-Ishihara<sup>3</sup>, Keisuke Hitachi<sup>1</sup>, Ozge Karayel<sup>4</sup>, Takanori Onouchi<sup>5</sup>, Hisateru Yamaguchi<sup>6</sup>, Tomoaki Kahyo<sup>2,7</sup>, Ken Hatanaka<sup>2†</sup>, Koji Ikegami<sup>2,8</sup>, Yusuke Yoshioka<sup>9</sup>, Kenji Nakamura<sup>2†</sup>, Nobuyoshi Kosaka<sup>9†</sup>, Masashi Nakatani<sup>1</sup>, Akiyoshi Uezumi<sup>1</sup>, Tomihiko Ide<sup>10</sup>, Yutaka Tsutsumi<sup>5</sup>, Haruhiko Sugimura<sup>7</sup>, Makoto Kinoshita<sup>3</sup>, Takahiro Ochiya<sup>9</sup>, Matthias Mann<sup>4</sup>, Mitsutoshi Setou<sup>2,8,11,12,13,14,15\*</sup>, Kunihiro Tsuchida<sup>1\*</sup>

<sup>1</sup> Division for Therapies Against Intractable Diseases, Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan <sup>2</sup> Mitsubishi Kagaku Institute of life Sciences (MITILS), 11 Minamiooya, Tokyo 194-8511, Japan <sup>3</sup> Division of Biological Sciences, Department of Molecular Biology, Nagoya University Graduate School of Science, Nagoya 464-8602, Japan <sup>4</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany <sup>5</sup> Department of Pathology, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan <sup>6</sup> Division of Biomedical Polymer Science, Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake, Toyoake, Aichi 470-1192, Japan <sup>7</sup> Department of Tumor Pathology, Hamamatsu University School of Medicine, 1-20-1 Handayama, Hamamatsu, Shizuoka 431-3192, Japan <sup>8</sup> Department of Cellular and Molecular Anatomy and International Mass Imaging Center, Hamamatsu University School of Medicine, 1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan <sup>9</sup> Division of Molecular and Cellular Medicine, National Cancer Center Research Institute, 5-1-1, Tsukiji, Chuo-ku, Tokyo 104-0045, Japan <sup>10</sup> Laboratory of Electron Microscopy, Fujita Health University Joint Research Laboratory, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan <sup>11</sup> Preeminent Medical Photonics, Education & Research Center, 1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan <sup>12</sup> Department of Anatomy, The University of Hong Kong, 6/F, William MW Mong Block 21 Sassoon Road, Pokfulam, Hong Kong SAR, China <sup>13</sup> Division of Neural Systematics, National Institute for Physiological Sciences, 38 Nishigonaka Myodaiji, Okazaki, Aichi, 444-8585, Japan <sup>14</sup> Riken Center for Molecular Imaging Science, 6-7-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan <sup>15</sup> Department of Anatomy II and Cell Biology and Anatomy, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan \* Corresponding authors

Published in *Nature Communications* (2018)

The family of ubiquitin-like proteins (UBLs), including SUMO, NEED8, and ISG15, are structurally similar to ubiquitin and processed by the same enzymatic steps. They also act as post-translational modifiers that influence diverse biological processes. One of the highly evolutionally conserved UBLs, the UBL3/MUB protein, has been identified in *Arabidopsis thaliana*; however, its role as a PTM factor remains poorly understood. Our collaboration partners from Fujita Health University characterized ubiquitin-like 3 (UBL3)/membrane-anchored Ub-fold protein (MUB) and its role in mammalian cells. Unlike other ubiquitin-like molecules, UBL3 modifies target proteins by disulfide bonding through cysteine residues at its C-terminus. To better understand the physiological function of UBL3 modification, I performed a comprehensive MS-based proteomics analysis and identified the proteins that interact with UBL3 in a manner dependent on the two C-terminal cysteine residues. We demonstrated that UBL3 modification primarily influences the sorting of proteins to the small extracellular vesicles.

## ARTICLE

DOI: 10.1038/s41467-018-06197-y

OPEN

# UBL3 modification influences protein sorting to small extracellular vesicles

Hiroshi Ageta<sup>1,2</sup>, Natsumi Ageta-Ishihara<sup>3</sup>, Keisuke Hitachi<sup>1</sup>, Ozge Karayel<sup>4</sup>, Takanori Onouchi<sup>5</sup>, Hisateru Yamaguchi<sup>6</sup>, Tomoaki Kahyo<sup>2,7</sup>, Ken Hatanaka<sup>2,16</sup>, Koji Ikegami<sup>2,8,18,19</sup>, Yusuke Yoshioka<sup>9,10</sup>, Kenji Nakamura<sup>2,17</sup>, Nobuyoshi Kosaka<sup>9,10</sup>, Masashi Nakatani<sup>1</sup>, Akiyoshi Uezumi<sup>1</sup>, Tomihiko Ide<sup>11</sup>, Yutaka Tsutsumi<sup>5</sup>, Haruhiko Sugimura<sup>12</sup>, Makoto Kinoshita<sup>13</sup>, Takahiro Ochiya<sup>9,12</sup>, Matthias Mann<sup>14</sup>, Mitsutoshi Setou<sup>2,8,13,14,15</sup> & Kunihiro Tsuchida<sup>1</sup>

Exosomes, a type of small extracellular vesicles (sEVs), derived from multivesicular bodies (MVBs), mediate cell-to-cell communication by transporting proteins, mRNAs, and miRNAs. However, the molecular mechanism by which proteins are sorted to sEVs is not fully understood. Here, we report that ubiquitin-like 3 (UBL3)/membrane-anchored Ub-fold protein (MUB) acts as a posttranslational modification (PTM) factor that regulates protein sorting to sEVs. We find that UBL3 modification is indispensable for sorting of UBL3 to MVBs and sEVs. We also observe a 60% reduction of total protein levels in sEVs purified from *Ubl3*-knockout mice compared with those from wild-type mice. By performing proteomics analysis, we find 1241 UBL3-interacting proteins, including Ras. We also show that UBL3 directly modifies Ras and oncogenic RasG12V mutant, and that UBL3 expression enhances sorting of RasG12V to sEVs via UBL3 modification. Collectively, these results indicate that PTM by UBL3 influences the sorting of proteins to sEVs.

<sup>1</sup>Division for Therapies Against Intractable Diseases, Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan. <sup>2</sup>Mitsubishi Kagaku Institute of Life Sciences (MITILS), 11 Minamiooya, Tokyo 194-8511, Japan. <sup>3</sup>Division of Biological Sciences, Department of Molecular Biology, Nagoya University Graduate School of Science, Nagoya 464-8602, Japan. <sup>4</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany. <sup>5</sup>Department of Pathology, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan. <sup>6</sup>Division of Biomedical Polymer Science, Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan. <sup>7</sup>Department of Tumor Pathology, Hamamatsu University School of Medicine, 1-20-1 Handayama, Hamamatsu, Shizuoka 431-3192, Japan. <sup>8</sup>Department of Cellular and Molecular Anatomy and International Mass Imaging Center, Hamamatsu University School of Medicine, 1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan. <sup>9</sup>Division of Molecular and Cellular Medicine, National Cancer Center Research Institute, 5-1-1, Tsukiji, Chuo-ku, Tokyo 104-0045, Japan. <sup>10</sup>Department of Translational Research for Extracellular Vesicles, Tokyo Medical University, 6-7-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 160-0023, Japan. <sup>11</sup>Laboratory of Electron Microscopy, Fujita Health University Joint research support promotion facility, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan. <sup>12</sup>Institute of Medical Science, Tokyo Medical University, 6-7-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 160-0023, Japan. <sup>13</sup>Preeminent Medical Photonics, Education & Research Center, 1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan. <sup>14</sup>Department of Anatomy, The University of Hong Kong, 6/F, William MW Mong Block 21 Sassoong Road, Pokfulam, Hong Kong, SAR, China. <sup>15</sup>Department of Anatomy II and Cell Biology and Anatomy, Fujita Health University School of Medicine, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan. <sup>16</sup>Present address: Neurology Tsukuba Research Department, Discovery, Medicine Creation, Neurology Business Group, Eisai Co., Ltd, 1-3 Tokodai 5-Chome, Tsukuba, Ibaraki 300-2635, Japan. <sup>17</sup>Present address: Nitobe bunka College, Department of Clinical Laboratory Sciences, 43-16 Nakano 3-Chome, Nakanoku, Tokyo 164-0001, Japan. <sup>18</sup>Present address: Department of Anatomy and Developmental Biology, Graduate School of Biomedical and Health Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima, Hiroshima 734-8553, Japan. <sup>19</sup>Present address: JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan. Correspondence and requests for materials should be addressed to M.S. (email: [setou@hama-med.ac.jp](mailto:setou@hama-med.ac.jp)) or to K.T. (email: [tsuchida@fujita-hu.ac.jp](mailto:tsuchida@fujita-hu.ac.jp))

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-06197-y

**S**EVs are nanometre-sized vesicles secreted from various cell types<sup>1</sup>. Exosomes, a type of sEVs, derived from multi-vesicular bodies (MVBs)<sup>2,3</sup>, mediate cell-to-cell communication by transporting proteins, mRNAs, and miRNAs<sup>3,4</sup>. The delivery of proteins between cells by sEVs, including exosomes, is related to tumour progression and neurodegenerative diseases<sup>5,6</sup>. Furthermore, neurodegenerative disease-related proteins such as amyloid beta, tau,  $\alpha$ -synuclein, and prions are also packaged inside sEVs, and spread in the brain<sup>6–10</sup>. These results indicate that transportation of the intracellular proteins via sEVs contributes to various types of disease. However, the molecular mechanism by which proteins are sorted to sEVs is not fully understood.

After synthesis, proteins undergo various posttranslational modifications (PTM) that influence a variety of cellular processes. The ubiquitin-dependent modification system, one of the protein degradation systems within the cell, is involved in a variety of cellular processes<sup>11</sup>. We previously found that the level and localisation of the synaptic protein Vesl-1S are controlled by the ubiquitin–proteasome systems<sup>12,13</sup>. Furthermore, using bioinformatics analysis, we identified the SCRAPPER protein as a synaptic E3 ubiquitin ligase<sup>14</sup>. Several proteins are reported to have a ubiquitin-like sequence, which is referred to as a “UBL domain”<sup>15,16</sup>. Some UBLs have been reported to act as post-translational modifiers<sup>15</sup>; these include the small ubiquitin-like modifiers (SUMO)<sup>17</sup> and Nedd8<sup>18</sup>. On the other hand, as one of conserved UBLs, UBL3/MUB protein has been identified in *Arabidopsis thaliana* and is a membrane protein localised by prenylation<sup>19</sup>. However, the role of UBL3/MUB as a PTM factor remains poorly understood.

In this study, we provide evidence that UBL3/MUB acts as a PTM factor that regulates protein sorting to sEVs.

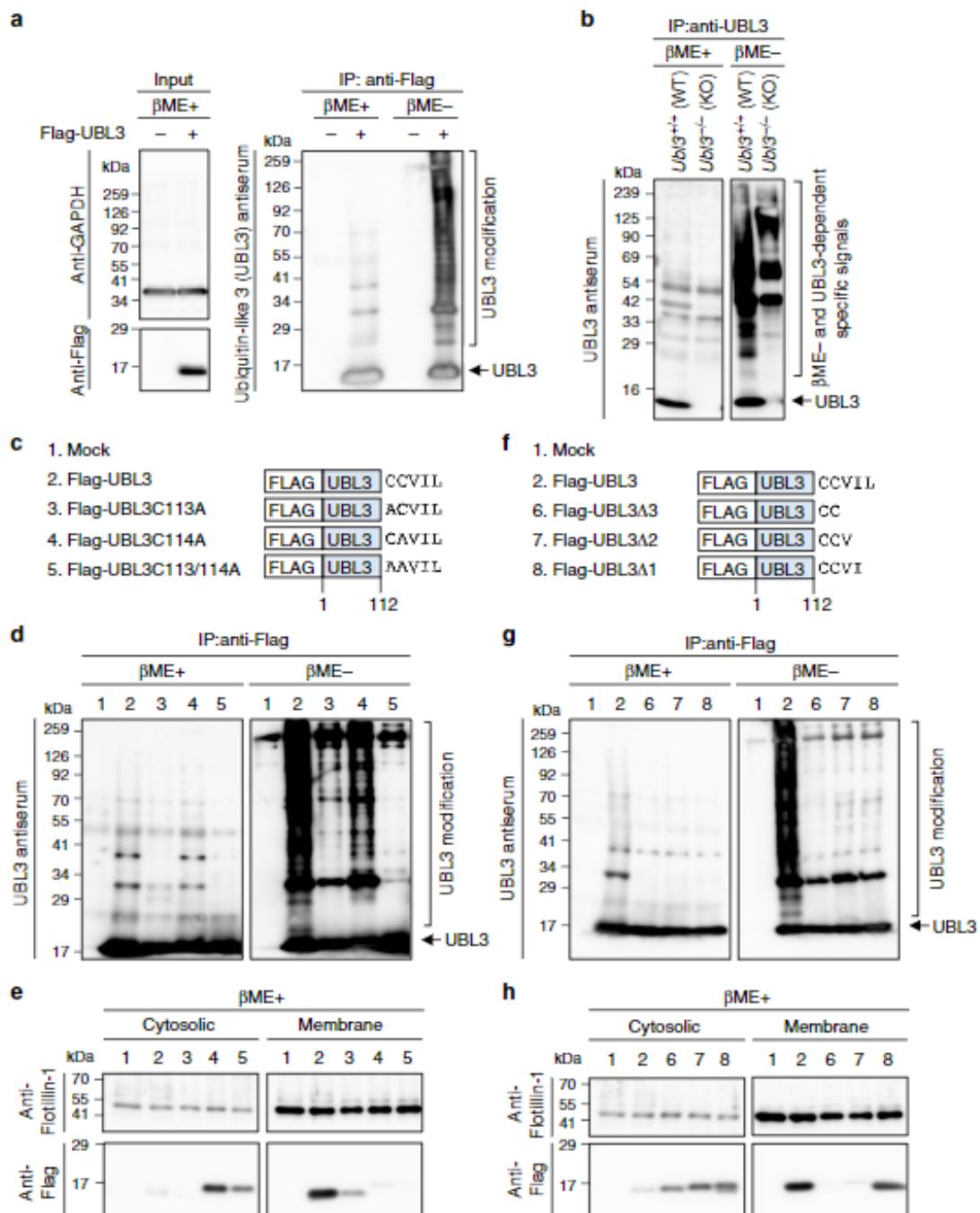
## Results

**Analysis of UBL3 as a post-translational modification factor.** In the present work, to identify PTM factors, we used a bioinformatics method<sup>14,16</sup>. We extracted UBL3/MUB (Supplementary Fig. 1), which is known to contain a ubiquitin-like (UBL) domain and is an evolutionarily conserved membrane protein localised by prenylation in animals, filamentous fungi, and plants<sup>19</sup>. However, the role of UBL3 as a PTM factor remains poorly understood. To clarify the role of UBL3 as a PTM factor, we expressed Flag-UBL3 in MDA-MB-231 breast cancer cells and purified UBL3 proteins by immunoprecipitation, followed by western blotting with UBL3 antiserum. The estimated molecular weight of Flag-UBL3 is 16 kDa. Interestingly, in the Flag-UBL3 expressing cells, the UBL3 signal was observed as a smear band up to a high molecular weight only under non-reducing conditions. The smear signal disappeared after the addition of 2-mercaptoethanol ( $\beta$ ME $+$ ) to the samples, before loading onto an SDS-polyacrylamide gel (Fig. 1a, right panel). To verify whether UBL3 modification occurs in vivo, we established *Ubl3* knockout (KO) mice and analysed PTM in brain lysates, as the expression of UBL3 in the brain is relatively high (Supplementary Fig. 2a–d). We observed endogenous UBL3-related PTM in lysates from the cerebral cortex, and showed that the degree of PTM was reduced in *Ubl3* KO mice (Fig. 1b). We next investigated UBL3 modification using UBL3 mutants; we focused on all cysteine residues as UBL3 modification is dependent on non-reducing conditions. UBL3 has only two cysteine residues at its C-terminus (C113 and C114); therefore we constructed three UBL3 mutants—C113A, C114A, and C113/C114A (Fig. 1c)—and studied these along with the wild-type UBL3 (Fig. 1c–e). In the UBL3C113/C114A mutant, UBL3 modification was completely abolished not only in MDA-MB-231 cells but also in every cell line examined (Fig. 1d, Supplementary

Fig. 3a). However, UBL3 modification was reduced but still retained by the UBL3C113A and UBL3C114A mutants in several cells. A CAAX motif (C, cysteine; A, aliphatic amino acid; X, any amino acid) at the C-terminus is frequently found in membrane-localised proteins<sup>20</sup>. The UBL3/MUB also has a CAAX motif at its C-terminus (Fig. 1c); it has been shown to be prenylated for its anchoring to membranes via C114<sup>19</sup>. The majority of wild-type UBL3 protein, but not the mutants, was selectively found in the membrane fraction (Fig. 1e, Supplementary Fig. 3b).

In order to study the relationship between the presence of UBL3 in the membrane fraction and UBL3 modification, we attempted to identify UBL3 mutants that were still retained in the membrane fraction despite the loss of PTM activity. Intriguingly, UBL3 $\Delta$ 1 which lacked only one C-terminal amino acid, was found to have lost the UBL3 modification (Fig. 1f, g, Supplementary Fig. 3a). However, UBL3 $\Delta$ 1 was retained in the membrane fraction, whereas UBL3 $\Delta$ 2 and UBL3 $\Delta$ 3 were not (Fig. 1h, Supplementary Fig. 3b). These data indicated that UBL3 modification was not necessary for the presence in the membrane fraction.

**UBL3 modification is indispensable for sorting of UBL3 to MVBs and sEVs.** We next examined subcellular localisation of UBL3 by immunocytochemistry, as the degree of UBL3 modification was found to be reduced in the cytoplasmic fraction (Supplementary Fig. 4a). Following the introduction of EGFP-UBL3, colocalisation with various organelles was studied. More than 40% of the total EGFP-UBL3 signal was found to be colocalised with endogenous CD63, a marker for MVBs, supporting that UBL3 is enriched in MVBs (Fig. 2a, b). UBL3 did not show colocalisation with the markers for recycling endosome (Rab11), mitochondria (COXIV), endoplasmic reticulum (Calnexin), Golgi (GM130), peroxisome (PMP70), or nuclear membrane (Lamin B1) (Fig. 2a, Supplementary Fig. 4b). Colocalisation with the early endosome marker (EEA1) or lysosome marker was detected; however, this was weak compared with the degree of colocalisation with the MVB marker (Fig. 2a, b). Then, to elucidate the relationship between UBL3 modification and localisation in MVBs, we examined colocalisation of UBL3C113A, UBL3C114A, UBL3C113/C114A, and UBL3 $\Delta$ 1 with endogenous CD63 (Fig. 2c). Unlike wild-type UBL3, the mutants did not show colocalisation with CD63 (median values: UBL3, UBL3C113A, UBL3C114A, UBL3C113/C114A, UBL3 $\Delta$ 1 = 41.5%, 3.0%, 2.7%, 1.9%, 2.3%, respectively). To clarify the localisation of UBL3 in MVBs and the plasma membrane, we performed immunoelectron microscopic analysis (Fig. 2d, e). When Flag-UBL3 was introduced to MDA-MB-231 cells, the UBL3 signal was clearly detected in the MVBs and the plasma membrane, but not in mitochondria or the nuclear membrane (Supplementary Fig. 4c, d). UBL3 $\Delta$ 1 was localised in the plasma membrane but not in MVBs (Fig. 2d, e). These results indicate that membrane localisation is not sufficient and that UBL3 modification is required for MVB localisation. One fate of MVBs is to fuse with the plasma membrane for release as exosomes. The other fate is to fuse with lysosomes, which degrade the MVB contents<sup>2,3</sup>. In the present study, we focused on the extracellular secretion of UBL3, as this protein colocalises with MVBs to a degree greater than with lysosomes (Fig. 2a, b). Cells release extracellular vesicles (EVs) of various sizes. These vesicles, some of which are derived from the plasma membrane, include large- and medium-sized EVs purified at 2000  $\times$  g (2K) and 10,000  $\times$  g (10K), respectively. sEVs are purified at 100,000  $\times$  g (100K) and are derived in part from the plasma membrane (non-exosomal sEVs) or are derived from MVBs (exosomes)<sup>21</sup>. We have purified EVs from 2K-, 10K-, and 100K-centrifuged pellets from MDA-MB-231 cells and found that

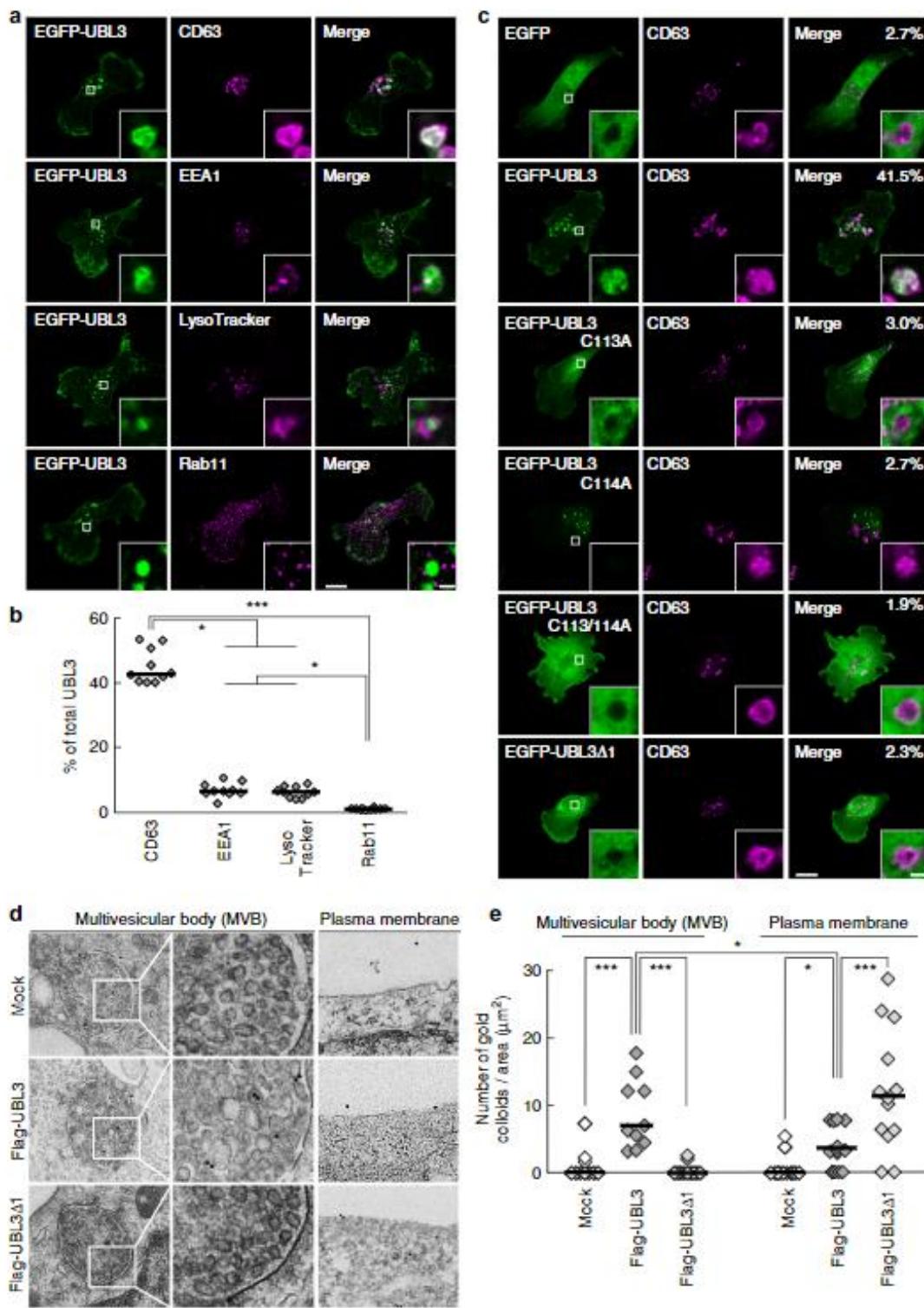


**Fig. 1** Analysis of UBL3 as a post-translational modification factor. **a** UBL3-dependent posttranslational modification was detected by immunoprecipitation (IP) with anti-Flag antibodies from MDA-MB-231 cells transfected with Flag-UBL3, followed by western blotting with UBL3 antiserum (right panel). **b** The tissue extracts from the cerebral cortex of WT and *Ubl3* KO mice were immunoprecipitated with anti-UBL3 antibodies. **c, f** Schematic structures of Flag-tagged wild-type or mutant UBL3. **d, g** Detection of UBL3 modification in these cells. IP products were boiled without 2-mercaptoethanol (βME-). A portion of the samples was treated with 2-mercaptoethanol (βME+). **e, h** Subcellular localisation of UBL3 in MDA-MB-231 cells transfected with Flag-UBL3 (wild-type and mutants). Twenty µg per lane.

UBL3 was more concentrated in the 100K fraction than in the 2K and 10K fractions (Fig. 3a). We verified that sEVs purified from the 100K pellet contained nanometre-sized vesicles ranging from 50 to 100 nm in diameter by negative EM staining (Supplementary Fig. 5a). We found that the majority of UBL3 was

present in sEVs containing CD63 (Supplementary Fig. 5b). Additionally, we found that UBL3 was packaged within the sEVs (Supplementary Fig. 5c). In order to determine whether endogenous UBL3 was released in sEVs, we prepared primary cultured cells from wild-type and *Ubl3* KO mice, purified sEVs from the

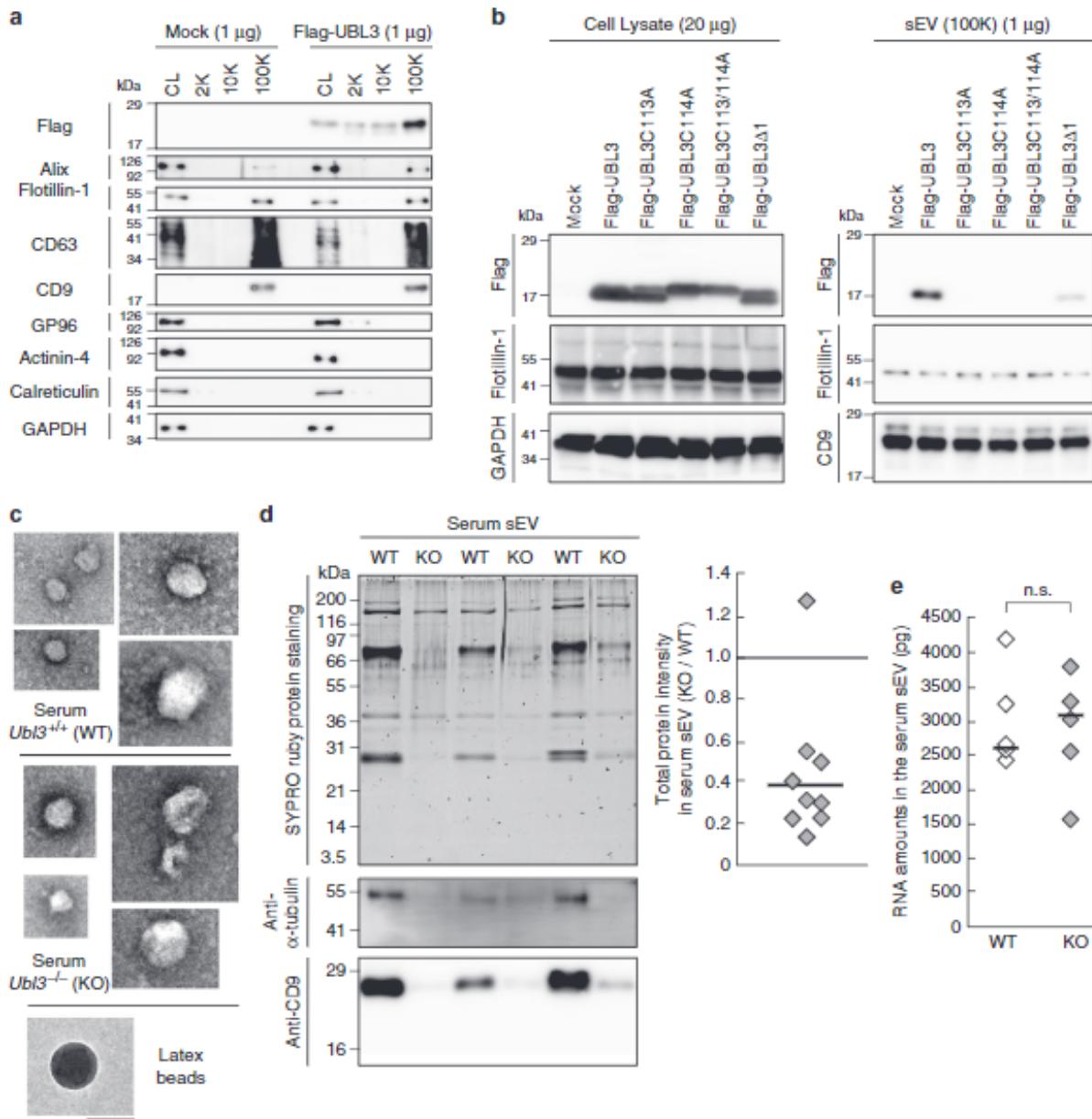
## ARTICLE



culture supernatant, and blotted them using anti-UBL3 antibodies. Endogenous UBL3 was only found in sEVs from wild-type mice (Supplementary Fig. 5d). Finally, UBL3 modification was observed not only in cell lysates but also in sEVs (Supplementary Fig. 5e). In order to clarify that UBL3 modification is

indispensable for sorting to sEVs, the presence of UBL3 and its mutants in 100K sEV was studied. Among these, UBL3 cysteine mutants were not sorted to sEV fractions. On the other hand, UBL3Δ1 was detected at much reduced levels compared with those of the wild-type UBL3 (Fig. 3b). Purified 100K pellets

**Fig. 2** The localisation of UBL3 to MVBs depends on UBL3 modification. **a** Representative projected images of MDA-MB-231 cells transfected with EGFP-UBL3 and co-stained with markers for MVB (CD63,  $n = 10$ ), early endosome (EEA1,  $n = 10$ ), lysosome (LysoTracker,  $n = 10$ ), or recycling endosome (Rab11,  $n = 10$ ). The regions in the dotted box are shown as a single confocal image in the inset. Scale bars, 10 and 1  $\mu$ m. **b** Quantitative analysis of EGFP-UBL3 fluorescence intensity in **a**. \* $p < 0.05$ ; \*\*\* $p < 0.001$  by Kruskal-Wallis/Dunn's multiple-comparisons test. **c** Representative images of MDA-MB-231 cells transfected with EGFP-UBL3 (wild-type,  $n = 5$ ; mutants,  $n = 5$ ) and co-stained with CD63 values shown as % of total UBL3. Scale bars, 10 and 1  $\mu$ m. **d** Immune-EM images of wild-type UBL3 and UBL3 $\Delta$ 1 in MDA-MB-231 cells. Scale bars, 500 nm (left and right panels) and 200 nm (middle panel). **e** Quantification of the numbers of gold colloids per area in **d**. MVB,  $n = 10$ ; Plasma membrane,  $n = 10$ . \* $p < 0.05$ ; \*\*\* $p < 0.001$  by two-tailed Student's and Welch's t-tests



**Fig. 3** The level of total proteins in the sEVs is reduced in *Ubl3* knockout mice. **a** The cell lysate (CL) and pellets from the conditioned medium of MDA-MB-231 cells transfected with 3xFlag-UBL3 vectors were blotted with various antibodies. **b** The presence of UBL3 and its mutants in sEVs. **c** Electron microscopic analyses of purified sEVs by negative staining. Scale bars, 100 nm. **d** Upper left panel, protein staining for the sEVs from the serum in WT and *Ubl3* KO mice. Lower panels, purified serum sEVs were blotted with anti-tubulin and CD9 antibodies with  $\beta$ ME. Right panel, relative intensity of total proteins in serum sEVs.  $n = 9$  pairs. **e** Total RNA levels in the serum sEVs. WT,  $n = 5$ ; KO,  $n = 5$ . n.s.,  $p > 0.05$  by Mann-Whitney test. a,b,  $\beta$ ME + condition: Flag, Flotillin-1, GP96, Actinin-4, Calreticulin, GAPDH, and Alix antibodies.  $\beta$ ME- condition: CD63, and CD9 antibodies

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-06197-y

contained both exosomes and non-exosomal sEVs. The latter include vesicles directly secreted from the plasma membrane<sup>21,22</sup>; this explains why UBL3Δ1, which shows plasma membrane but not MVB localisation, was detected in the 100K pellets in small amounts (Figs. 1h, 2d, e, 3b). This finding indicates that the UBL3 modification is important for sorting of UBL3 to sEVs. Furthermore, when exosome release was inhibited by *Rab27a* shRNA<sup>23</sup>, the levels of UBL3 modification products were increased (Supplementary Fig. 5f). The above results indicate that UBL3 modification is required for localisation of UBL3 in MVBs, and that the majority of UBL3 is secreted in sEVs originating from the fusion of MVBs released as exosomes. As UBL3 was localised in the plasma membrane, UBL3 modification may also influence non-exosomal sEVs.

**The level of total proteins in the sEVs is reduced in *Ubl3* knockout mice.** To elucidate the functional role of UBL3 in the sorting of proteins to sEVs, we quantified the protein contents of purified sEVs derived from *Ubl3*-KO mouse sera. The genotype had no effect on the concentration of sEV particles or the average vesicular diameter (Fig. 3c, Supplementary Fig. 6a, b). Intriguingly, the total level of sEV proteins in the sera from the *Ubl3* KO mice was 60% lower than that from the wild-type mice, although the total protein content in the serum and plasma were not different (Fig. 3d, Supplementary Fig. 6c). We did not find any significant difference in the RNA profiles and RNA levels (Fig. 3e and Supplementary Fig. 6d, e). Specific sEV-enriched miRNAs levels were not statistically different (Supplementary Fig. 6f). These results indicate that the *Ubl3* KO mice produce sEVs with reduced proteins, but with normal levels of miRNAs in the serum. In other words, the sorting of proteins in the serum sEVs is influenced by UBL3.

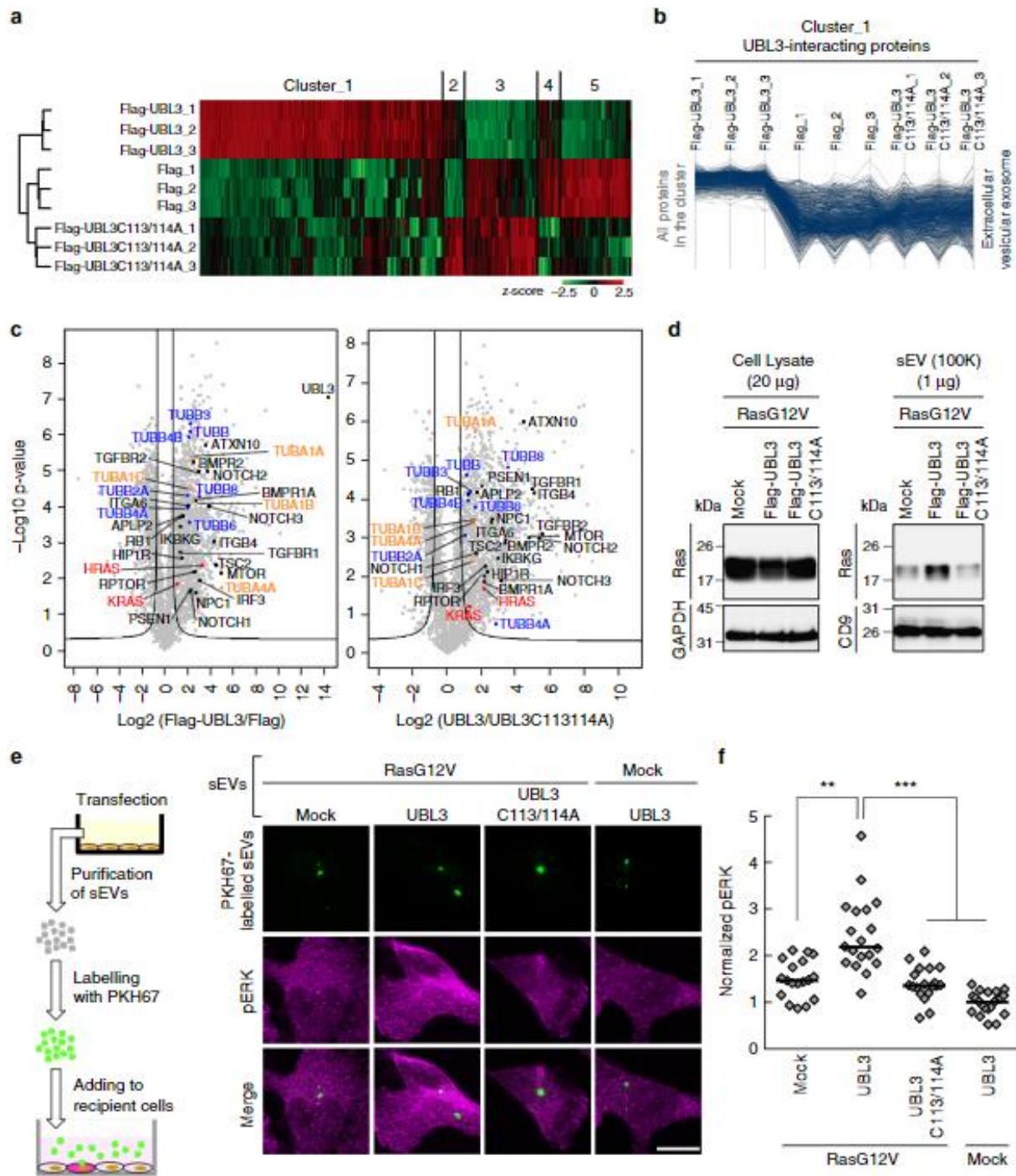
**UBL3 modification influences the sorting of proteins to the sEV.** For a better understanding of the physiological function of UBL3 modification, we performed comprehensive proteomics analysis to identify proteins that interacted with UBL3 in a manner dependent on the two C-terminal cysteine residues. Lysates prepared from 3xFlag-UBL3 expressing in MDA-MB-231 cells were purified with Flag antibody-beads, subjected to on-bead digestion, and analysed by nanoflow liquid chromatography tandem mass spectrometry (LC-MS/MS) gradients on a Q Exactive HF mass spectrometer (Fig. 4a). The dataset covered 3882 proteins in total at a false discovery rate (FDR) of 1%. The average of Pearson correlation coefficients was 0.988 within triplicate pull-downs. Label-free quantification (LFQ)-based unsupervised hierarchical clustering of significantly regulated proteins (ANOVA, FDR = 0.05 & S0 = 1) revealed 1447 possible UBL3-interacting proteins that were upregulated in Flag-UBL3 compared to Flag-UBL3C113/114A or the Flag empty vector (Fig. 4a and Supplementary Table 1). Next, we grouped the UBL3-interacting proteins based on their subcellular localisation as defined by the Gene Ontology Cellular Components (GOCC). We found that 31% of proteins (454 out of 1447 proteins; *p*-value of 0.00405, Fisher exact test) were categorised as the GOCC term 'extracellular vesicular exosome' (Fig. 4b, and Supplementary Data file). In addition, the pairwise comparison of UBL3 with UBL3C113/114A revealed that 1241 proteins significantly interact with UBL3, emphasising the role of cysteine residues in UBL3 modification (two-sample Student's *t*-test, FDR = 0.05 & S0 = 1). Of these proteins, 29% (369) of them were also annotated as extracellular vesicular exosome. Moreover, Fisher exact test for the statistically significant proteins showed an enrichment for the annotation of extracellular vesicular exosome (*p*-value of 0.00828). Among these, tubulins (TUBA1A, 1B, 1C, and 4A; and TUBB, B2A, B3, B4A, B4B, B6, and B8) were

found in 3xFlag-UBL3 samples (Fig. 4c, Supplementary Fig. 7a and Supplementary Data file). Tubulins are reported to be present in sEVs and other EVs<sup>21,24</sup>. Endogenous tubulin alpha can be detected in sEVs with the commonly used specific antibody (DM1A); therefore we chose tubulin alpha as a model case to study sorting of endogenous proteins to sEVs via UBL3 modification. When EGFP-UBL3 introduced to MDA-MB-231 cells was immunopurified from anti-GFP-beads, tubulin was found to be modified by UBL3 and observed as shifted bands only under non-reducing conditions (Supplementary Fig. 7b). We then studied the sorting of tubulin by UBL3 in secreted vesicles prepared by differential centrifugation. Intriguingly, the tubulin signal was specifically augmented by UBL3 expression in the 100K, but not in the 2K or 10K fractions (Supplementary Fig. 7c). Under non-reducing conditions, the majority of tubulin in the 100K sEV pellet was found to have a higher molecular weight than under reducing conditions. We used UBL3 mutants to investigate the relationship between UBL3 modification of tubulin and the sorting to sEVs. The increase in tubulin in the sEVs was specific to wild-type UBL3; the UBL3C113/114A and UBL3Δ1 mutants exhibited no such effects (Supplementary Fig. 7d). We also observed that the protein levels of tubulin in the serum sEVs were reduced in *Ubl3* KO mice (Fig. 3d). We therefore conclude that UBL3 modifies endogenous tubulin posttranslationally and regulates its sorting to sEVs.

We also found that at least 22 disease-related molecules were included as UBL3-interacting proteins (Fig. 4c and Supplementary Fig. 8a). Intriguingly, some of these molecules are related to oncogenesis and tumour progression/metastasis, namely, HRAS, KRAS, TGFBR1, TGFBR2, RB1, ITGA6, ITGB4, mTOR, TSC2 and APLP2. In addition, immune-responsive molecules (IRF3 and IKBKG), mTOR signalling molecules (mTOR, RPTOR and TSC2), Notch signalling molecules (NOTCH1, NOTCH2, NOTCH3), BMP signalling molecules (BMPR1A and BMPR2), and even molecules involved in neurodegenerative/neuronal diseases (PSEN1, ATXN10, HIP1R, APLP2 and NPC1) are identified as UBL3-interacting proteins. Among these, Ras family members are proto-oncogenes<sup>25</sup> and have been reported to be enriched in sEVs<sup>26</sup>. Therefore, we chose H-Ras as a model protein to study the sorting to sEVs via UBL3 modification. We found that exogenous wild-type Ras and constitutively active oncogenic RasG12V mutant<sup>25</sup> interacted with UBL3, and were observed as shifted bands only under non-reducing conditions (Supplementary Fig. 8b). Moreover, exogenous wild-type Ras proteins were more sorted to sEV by UBL3 (Supplementary Fig. 8c). We also found that wild-type UBL3 but not UBL3C113/114A enhanced sorting of RasG12V to sEVs (Fig. 4d).

To examine whether sEVs encapsulated with UBL3 and oncogenic RasG12V caused the activation of Ras signalling in the recipient cells, we added PKH67-labelled sEV purified from MDA-MB-231 cells that were transfected with RasG12V and either mock, 3xFlag-UBL3 or 3xFlag-UBL3C113/114A to recipient MDA-MB-231 cells. As an indicator of Ras activation in the recipient cells, we measured the phosphorylation level of ERK, which is a downstream signalling molecule<sup>25</sup>. Phosphorylated ERK (pERK) activity in PKH67-labelled sEV-incorporated cells was examined by antibody staining. We found that purified sEVs from MDA-MB-231 cells transfected with wild-type UBL3 and RasG12V significantly enhanced the level of pERK in the recipient cells compared to sEVs from UBL3C113/114A and RasG12V transfected cells (Fig. 4e, f). These results indicated that increased sorting of RasG12V to sEVs by UBL3 modification enhanced activation of Ras signalling in the recipient cells.

We also performed a comparative analysis using other UBLs. Unlike UBL3, other UBLs, including ubiquitin, SUMO1, and SUMO2, were not enriched either in MVBs or sEVs (Fig. 5a-c).



Finally, we attempted to determine whether UBL3 serves as a useful tag for protein delivery to sEVs. When EGFP or biotinylated-protein tagged with UBL3 was introduced to MDA-MB-231 cells, EGFP-UBL3 and biotinylated-UBL3 were found in sEVs (Fig. 6). These observations confirmed that UBL3 serves as a tag for the delivery of proteins to sEVs.

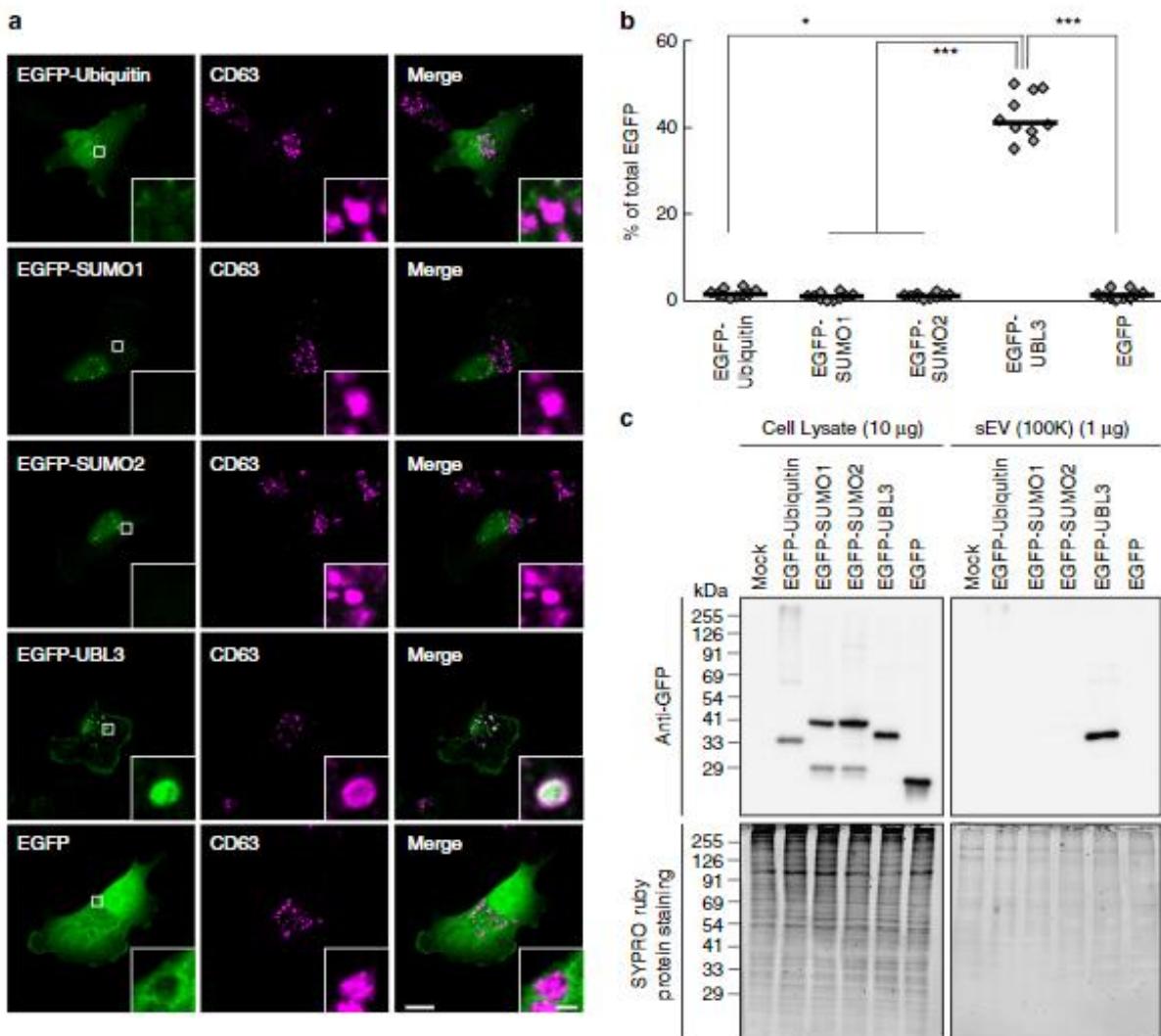
## Discussion

In the current study, we characterised UBL3 as a PTM factor. The glycine residues at the C-termini of ubiquitin, SUMO, and Ned8

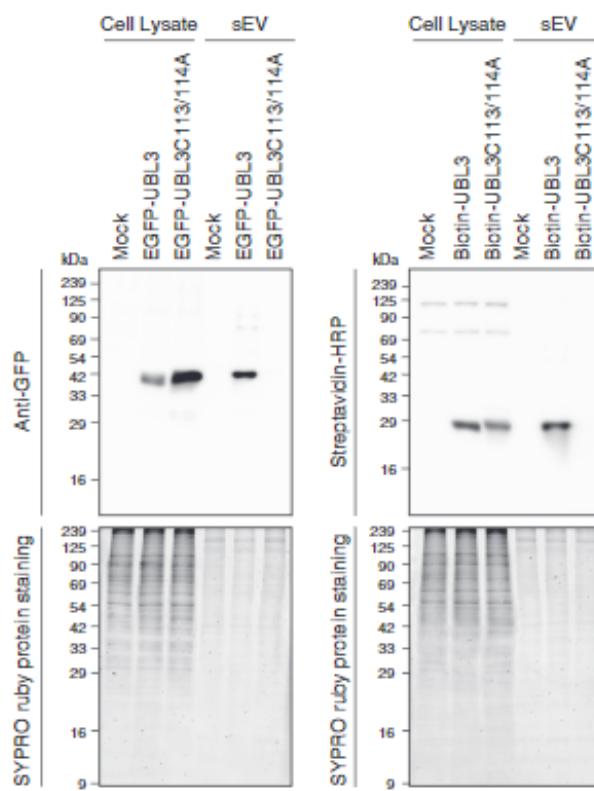
are covalently attached to lysine residues in target proteins in general<sup>15,18,27</sup>. In contrast, our data suggest that UBL3 modifies target proteins by disulfide bonding through cysteine residues at its C-terminus; therefore although UBL3 has a ubiquitin-like domain, UBL3 modification is found to be completely different from conventional ubiquitin and ubiquitin-like modifications. From the result of UBL3Δ1 mutant (Fig. 1g), the membrane localisation of UBL3 alone is not sufficient for the UBL3 modification. Thus, in addition to CAAAX motif for the membrane localisation in the C-terminal of UBL3, it is possible that

## ARTICLE

**Fig. 4** UBL3 modification influences the sorting of proteins to the sEV. **a, b** Heat map of z-scored LFQ intensities of the significantly regulated proteins (ANOVA, FDR = 0.05 & SO = 1) in all three conditions (3xFlag-UBL3, 3xFlag-UBL3C113/114A or 3xFlag empty vector) revealed the UBL3 interacting proteins (Cluster 1). The colour key denotes normalised protein abundances (z-score). Profiles of all proteins (1447) found in Cluster 1 are shown and 454 of those (31%) with 'extracellular vesicular exosome' annotations are highlighted in dark blue. **c** Volcano plots showing the p-values vs. the log2 protein abundance differences in Flag-UBL3 compared with either Flag-UBL3C113/114A or Flag empty vector. The significance cut-off is based on an FDR = 0.05 and SO = 1. Disease-related molecules, black colour. HRAS and KRAS, red colour. TUBA1A, 1B, 1C, and 4A, orange colour. TUBB, B2A, B3, B4A, B4B, B6, and B8, blue colour. **d** sEV pellets were blotted with anti-Ras antibodies. **e** Images of phosphorylated ERK (pERK) in PKH67-labelled sEV-incorporated MDA-MB-231 cells. Purified sEVs from the conditioned medium of MDA-MB-231 cells transfected with RasG12V and either mock, 3xFlag-UBL3, or 3xFlag-UBL3C113/114A or with mock and 3xFlag-UBL3 were labelled with PKH67 dye (green) and added to MDA-MB-231 cells. Scale bars, 10  $\mu$ m. **f** Each plot shows pERK fluorescence in PKH67-labelled sEV-incorporated cells normalised to the average pERK values in the two neighbouring PKH67-labelled sEV-unincorporated cells from each image in e. RasG12V-mock, n = 18; RasG12V-UBL3, n = 19; RasG12V-UBL3C113/114A, n = 19; Mock-UBL3, n = 21. \*\*, p < 0.01; \*\*\*, p < 0.001 by Kruskal-Wallis/Dunn's multiple-comparisons test



**Fig. 5** Subcellular localisation and sorting to sEV of UBLs. **a** Representative images of MDA-MB-231 cells transfected either with EGFP-ubiquitin, -SUMO1, -SUMO2 or -UBL3 and co-stained with markers for MVBs (CD63). Scale bars, 10 and 1  $\mu$ m. **b** Quantitative analysis of EGFP-ubiquitin, -SUMO1, -SUMO2, and -UBL3 fluorescence intensity in a. EGFP-ubiquitin, n = 10; EGFP-SUMO1, n = 10; EGFP-SUMO2, n = 10; EGFP-UBL3, n = 10; EGFP, n = 10. \*, p < 0.05; \*\*\*, p < 0.0001 by Kruskal-Wallis/Dunn's multiple-comparisons test. **c** EGFP-UBL3, but not EGFP, EGFP-ubiquitin, EGFP-SUMO1, or EGFP-SUMO2, was preferentially enriched in the sEVs of the cell culture media. Before sample loading, the samples were boiled with  $\beta$ ME



**Fig. 6** EGFP and biotinylated protein tagged by UBL3 are sorted to the sEVs. EGFP-UBL3 (left panel) and biotinylated protein tagged by UBL3 (right panel), but not EGFP-UBL3C113/114A or biotinylated protein tagged by UBL3C113/114A were preferentially accumulated in the sEVs of the cell culture media. After the western blot analysis, gels were stained with SYPRO Ruby (lower panel). Before loading, the samples were boiled with  $\beta$ ME. The same amounts of protein were loaded on the gels (Cell Lysate, 20  $\mu$ g per lane; sEVs, 1  $\mu$ g per lane)

important amino acid sequences for the UBL3 modification in the C-terminal region exist. In the future UBL3 study, it is important to analyse molecules that recognise the C-terminal sequence for UBL3 modification.

Additionally, we report that UBL3 modification is essential for the sorting of UBL3 to MVBs and sEVs. We identify 1241 UBL3-interacting proteins independently on the C-terminal two cysteine residues, 29% of which are annotated as 'extracellular vesicular exosome'. We also observed a 60% reduction in the total sEV proteins in the sera of *Ubl3* KO mice (Fig. 3d). This result suggested that UBL3 may be involved in the sorting of more than half of all exosomal proteins. It is possible that UBL3 binding to a particular protein is transient, thereby permitting identification of only a subset of UBL3-modified proteins in the comprehensive proteomics analyses of cell lysates. Alternatively, UBL3 modification may be required for only a limited number of target proteins that participate in large protein-protein interaction networks for targeting to exosomes; however, this idea must be confirmed by further investigation.

Specific proteins in sEVs play an important role in tumour progression and organotropic metastasis<sup>5,28,29</sup>. Neurodegenerative disease-related proteins, such as amyloid beta, tau,  $\alpha$ -synuclein, and prions, are also packaged inside sEVs and spread in the brain<sup>6,8,10</sup>. In the current study, we identified 1241 UBL3-interacting proteins that were dependent on the two C-terminal

cysteine residues, including at least 22 disease-related molecules (Fig. 4c), and showed that oncogenic RasG12V proteins were more sorted to sEVs by UBL3 modification (Fig. 4d). These sEVs bearing the oncogenic proteins affected ERK phosphorylation in the recipient cells (Fig. 4e, f). Therefore, the inhibition of UBL3 modification also could be a therapeutic target for sEV-related disorders. Our results also indicate that extraneous proteins are sorted to sEVs by UBL3 tagging (Fig. 6). Thus, for any other type of sEV-related diseases, UBL3 could serve as a useful tool for the generation of sEVs that are modified to act as vehicles for therapeutic cargos.

## Methods

**Cell culture.** MDA MB 231 luc D3H2LN breast cancer cells (Xenogen Co, CA), HEK293T cells (RIKEN Cell Bank, Japan), and HeLa cells (RIKEN Cell Bank, Japan) were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (GIBCO, 11875 093) or Dulbecco's modified Eagle's medium (DMEM, GIBCO, 11965 092) with PS (100 units of penicillin G per mL, 10  $\mu$ g of streptomycin sulphate per mL) and 10% foetal bovine serum (FBS). Cultures were incubated in a 5% CO<sub>2</sub> at 37 °C. All the cell lines were routinely tested for mycoplasma contamination (Biological Industries, EZ PCR Mycoplasma Test kit).

Following previous papers<sup>30,31</sup>, we prepared primary cells from the adipose tissue derived stromal vascular fraction (SVF) and myoblasts from wild type and *Ubl3* KO mice. Briefly, to prepare primary SVF cells, inguinal and visceral adipose tissues were excised from mice and digested with 3.33 mg mL<sup>-1</sup> type II collagenase (Nitta gelatin, type L) with gentle rocking in Hank's balanced salt solution for 60 min at 37 °C. After sequential filtration through 40 and 100  $\mu$ m filters (BD), the isolated SVF cells were suspended in FBS and cultured in DMEM (GIBCO, 11965 092) with 10% FBS and PS. Myoblasts were cultured using the following growth medium in a culture dish coated with collagen type I (IWAKI, 44020 10) for 3–5 days: Ham's F 10 Nutrient Mix (GIBCO, 11550 043) containing 20% FBS, 2.5 ng mL<sup>-1</sup> bFGF (Katayama Chemical Industries, 161 0010 5), L-glutamine, penicillin, and streptomycin (Thermo Fisher Scientific, PS glutamine). For myotube differentiation, the medium was changed to a differentiation medium (5% horse serum in DMEM [GIBCO, 11995 065]) with PS glutamine for 5 days.

**Plasmid construction.** Various expression vectors were subcloned either to pcDNA3 (Invitrogen), pEGFP-C (Clontech), or pcDNA6/BioErase (Invitrogen) using conventional molecular biology techniques and PCR<sup>14,31,32</sup>. The sequence of *Rab27a* shRNA (5' GCTGCCAATGGGACAAACATA 3') was used<sup>23</sup> and inserted into pcDNA 6.2 GW/miR vector (Invitrogen).

In the Ras experiments, we used H1 Ras expression vector set (Clontech, pCMV Ras vector, and pCMV RasV12 vectors). GFP ubiquitin was a gift from Nico Dantuma (Addgene plasmid # 11928)<sup>33</sup>.

**Establishment of *Ubl3*<sup>-/-</sup> (knockout; KO) mice.** The targeting strategy is illustrated in Supplementary Fig. 2a. The targeting construct was linearised by Spel digestion. CCE (129/SvEv) (a gift from Dr. E. Robertson) embryonic stem (ES) cells were electroporated using a Spel linearised targeting vector and selected in G418. Heterozygous ES cells were injected into the blastocysts of C57BL/6J strain mice to generate germ line chimeras. Chimeric mice were backcrossed with C57BL/6J mice for 10 generations. Genotypes were determined by Southern blotting or by PCR using the following three primers: *Ubl3* KO 3', 5' ACCCA GGTCCCTCATGCATCGGTAGA 3'; *Ubl3* KO 5', 5' CCACCCACTGCCITTTCCAGAAC 3'; and *Ubl3* KO Neo, 5' GCTGCAGGGTCGCTCGGTGTT 3'. All procedures related to the care and treatments of animals were in accordance with institutional and National Institutes of Health guidelines and were approved by the Animal Care and Use Committee at Fujita Health University.

**Antibodies.** To obtain the rabbit anti UBL3 antibody, 1.5 mg of glutathione S transferase (GST) fused mouse UBL3 protein expressed in bacteria was injected four times each into nine rabbits. Recombinant, exogenous, and endogenous UBL3 proteins were specifically recognised in only one rabbit's serum. The serum was passed through the GST column three times to remove non-specific antibodies. We called the resulting flow through (FT) the UBL3 antiserum. To obtain a more specific antibody for the UBL3 protein, a polyclonal UBL3 antiserum was passed through the GST column. The resultant FT was further passed through a column conjugated with a brain lysate from *Ubl3* KO mice to exclude the recognition of other UBLs. The FT was further purified using a GST UBL3 column. We called this eluate the anti UBL3 antibody. The antibodies used in the experiments were anti GAPDH (Cell Signalling, 14C10, 1:1000), Flag M2 (Sigma, F3165, 1:1000), Flotillin 1 (BD, Clontech, 18, 1:250), CD63 (Invitrogen, Ts63, 1:1000 for WB, 1:10000 for IF), EEA1 (BD, 610456, 1:5000), Rab11 (BD, 610656, 1:150), COXIV (Cell Signalling, 4850, 1:1000), Calnexin (Enzo, ADI SPA 860, 1:500), GM130 (BD, 610822, 1:100), PMP70 (Thermo, PA1 650, 1:50000), Lamin B1 (Abcam, ab16048, 1:800), GP96 (Enzo, 9G10, 1:500), alpha Actinin 4 (GeneTex, C2C3, 1:250),

## ARTICLE

Calreticulin (Cell Signalling, 2891, 1:500), Alix (Abnova, H100010015 B02, 1:300), CD9 (Affymetrix, 14 0091, 1:500; Invitrogen, Ts9, 1:500), alpha Tubulin (Sigma, DM1A, 1:1000), GFP (MBL, 598, 1:1000; Thermo, A11122, 1:1000), pan Ras (Cell Signalling, 3339, 1:1000; Millipore, MABS195, 1:500), Phospho ERK (Cell Signalling, 4370, 1:200), horse radish peroxidase (HRP) conjugated Streptavidin (Invitrogen, 19534 050, 1:5000), Alexa 488 or 555 conjugated antibodies (Thermo), and HRP conjugated secondary antibody (eBioscience, Cell Signalling, and Abcam). In a differential centrifugation experiment, we used the antibodies against the following markers: sEV markers, CD63 and CD9; multiple EV marker, Flotillin 1 and Alix; and non sEV markers, GP96, Actinin 4, Calreticulin, and GAPDH. Chemiluminescence detection was performed with an image analyser LAS 4000 with Image Reader LAS 4000 (Ver. 2.0) and MultiGauge software (FUJI FILM). Uncropped version of each image is shown in Supplementary Fig. 9.

**Immunoprecipitation.** For the detection of UBL3 modification, MDA MB 231, HEK293T, and HeLa cells were transfected with DNA using Lipofectamine 2000 (Invitrogen). After 24 or 72 h (Supplementary Fig. 5), cells were washed and collected by scraping into ice cold PBS, pelleted by centrifugation at 400  $\times$  g for 1 min at 4 °C, and lysed with 1% Triton buffer (50 mM Tris HCl [pH 7.4], 100 mM NaCl, and 1% (v/v) Triton X 100) for 20 min at 4 °C. Crude nuclei and unbroken cells were excluded by centrifugation at 20,000  $\times$  g for 5 min. In the experiment using hypotonic buffer, after being washed in ice cold PBS, cells were resuspended in hypotonic buffer (10 mM HEPES [pH 7.2], 10 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.1 mM EGTA) and disrupted by a Dounce homogeniser. Crude nuclei and unbroken cells were excluded by centrifugation at 800  $\times$  g for 5 min at 4 °C. For pre clearing, the cell lysates were incubated with 20  $\mu$ L of Protein G sepharose beads (GE Healthcare) for 1 h at 4 °C with rotation. The resulting supernatant was incubated with 2  $\mu$ g of anti Flag antibody in 15  $\mu$ L of Protein G sepharose beads for 10 h at 4 °C with rotation, or 20  $\mu$ L of GFP Trap (ChromoTek) for 1 h at 4 °C without pre-clearing. After the beads were washed four times with ice cold wash buffer (50 mM Tris HCl [pH 7.4], 100 mM NaCl), 50  $\mu$ L of 2x sample buffer (100 mM Tris HCl [pH 6.8], 4% SDS, 20% glycerol, and 0.01% bromophenol blue, without 2-mercaptoethanol) was added to the beads, and the beads were boiled for 3 min (BME<sup>-</sup>). A portion of the samples was treated with 2 mercaptoethanol and boiled for 3 min (BME<sup>+</sup>). These samples were blotted with UBL3 antiserum (1:100 dilution). For the detection of endogenous UBL3 modification, 50 mg of frozen tissue was pulverised (BMS, ShakeMan2) and homogenised with RIPA buffer (50 mM Tris HCl [pH 8.0], 150 mM NaCl, 1% Nonidet P 40, 0.5% sodium deoxycholate, 0.1% SDS). For pre clearing, the cell lysates were incubated with 40  $\mu$ L of Protein G sepharose beads for 1 h at 4 °C with rotation. The resulting supernatants were incubated with anti UBL3 antibody (1:1000 dilution) in 20  $\mu$ L of Protein G sepharose beads for 10 h at 4 °C with rotation.

The method for the immuno isolation of sEVs using anti CD9 or CD63 antibodies was referred to from Kowal J et al.<sup>21</sup>. Briefly, before performing the immuno isolation, anti CD9 (Millipore, MM2/57, 1:10), anti CD63 (BD, H5C6, 1:50) and control mouse polyclonal IgG (Millipore, 12 371, 1:100) incubated with 100  $\mu$ L of protein A magnetic beads (Thermo, 88845) for 16 h at 4 °C with rotation. The resulting protein A magnetic beads were washed three time with ice cold 0.001% Tween/PBS, resuspended in 500  $\mu$ L of ice cold 0.001% Tween/PBS. Two  $\mu$ g purified sEVs were added to the magnetic beads and incubated for 16 h at 4 °C with rotation. Bead bound sEVs were collected and washed three time in 500  $\mu$ L of ice cold 0.001% Tween/PBS.

For the detection of UBL3 inside the sEVs, sEVs were purified from the media of biotinylated tagged UBL3 transfected MDA MB 231 cells, washed with PBS or SDS buffer (50 mM Tris HCl [pH 8.0], 150 mM NaCl, 1% sodium deoxycholate, 1% NP 40, 2% SDS), and then treated with streptavidin beads (Thermo, 29202). The pulldown (PD) and flow through (FT) fractions were analysed by western blotting with streptavidin HRP or anti CD63 antibodies.

**Miscellaneous procedures.** Tissue homogenisation and western blotting analysis were performed using conventional methods<sup>14,31,32</sup>. Isolation of membrane and cytoplasmic protein fractions were performed following the manufacturer's protocols (Fermentas, ProteoJET; Thermo, Mem PER Plus kit). RNA purification, reverse transcription reaction, and real time qPCR analysis (Takara Bio, Thermal Cycle Dice (TP800), Thermal Cycle Dice Real Time System Ver. 5. 11B) were performed by conventional methods<sup>14,31,32</sup>. Briefly, to detect miRNAs, cDNA was generated using an miScript Reverse Transcription Kit (QIAGEN). Expression of miR 16, miR 23 and miR 21 was detected using an miScript SYBR Green PCR Kit (QIAGEN) with the following primers: 5' TAGCAGCACCGTAAATATTGG 3' for miR 16, 5' ATCACATTGCCAGGGATTTC 3' for miR 23 and 5' TAGCTTAT CAGACTGATGTTGA 3' for miR 21.

**Immunocytochemistry.** MDA MB 231 cells were co transfected with GFP UBL3 (wild type or mutant) and iRFP670 (morphological marker) using Lipofectamine 2000. After 24 h, cells were fixed with 4% PFA/PBS for 15 min; quenched with 0.1 M glycine/PBS; permeabilized/block with 0.2% BSA, 2% goat serum, and 0.05% saponin/PBS for 1 h; incubated with primary antibodies for 18 h at 4 °C and secondary antibodies for 1 h; and mounted with Aqua Poly/Mount (Polysciences). Lysosomes were labelled with 50  $\mu$ M LysoTracker Red (Thermo) for 1 h at 37 °C.

Confocal images were acquired with a 63 $\times$  objective lens (NA 1.4) on a confocal laser microscope (Carl Zeiss, LSM780 DUO NLO, ZEN 2012 SP1 Ver. 8. 1. 0. 0). Representative images projected from several 0.5  $\mu$ m interval sections were shown, each with a single confocal image in the inset. Fluorescence intensity and cell morphology were measured with Fiji (Image J). The percentage of UBL3 was calculated as: (%) = (the summation of the GFP UBL3 signal in each organelle area in each confocal section)/(the summation of the GFP UBL3 signal in the total cell area in each confocal section)  $\times$  100.

**Transfer of PKH67-labelled sEVs.** Purified sEVs from MDA MB 231 cells were labelled using a PKH67 Green Fluorescent kit (Sigma, PKH67GL\_1KT). sEVs were incubated with 2  $\mu$ M PKH67 for 5 min, and then washed five times using a 100-kDa filter (Sartorius, VN011142) to remove excess dye. MDA MB 231 cells had been seeded at a density of  $1.2 \times 10^5$  cells per well onto poly-L-lysine coated cover glasses in a 24 well plate, allowed to attach overnight, and then were incubated with PKH67 labelled sEVs (1  $\mu$ g/ml). After 12 h, MDA MB 231 cells were fixed with 4% PFA/PBS for 15 min; permeabilized/block with 5% goat serum and 0.3% Triton X 100/PBS for 20 min; incubated with anti pERK antibodies for 18 h at 4 °C and secondary antibodies for 1 h; and mounted with Aqua Poly/Mount (Poly sciences). Confocal images were acquired with a 63 $\times$  objective lens (NA 1.4) on a confocal laser microscope (Carl Zeiss, LSM 710, ZEN 2009 Ver. 5. 5. 0. 0). Data were analysed by using Fiji (Image J); the values of pERK signal in PKH67 labelled sEV incorporated cells were presented as the mean signal intensity measured over an area of 2.5  $\mu$ m<sup>2</sup> in a region of the cytoplasm adjacent to the nucleus, and then normalised to the average values of pERK in the two neighbouring PKH67 labelled sEV unincorporated cells, in each image.

**Immunoelectron microscopy.** 3xFlag UBL3 (wild type or UBL3A1) or mock transfected MDA MB 231 cells were fixed with 4% PFA/PBS for 30 min at 4 °C, and subsequently fixed with 4% PFA and 0.1% glutaraldehyde (GA)/PBS for 30 min at 4 °C. After being washed with PBS, the cells were fixed with 1% osmium tetroxide/PBS for 30 min, dehydrated through a graded ethanol series, transferred to n-butyl glycidyl ether (QY 1), and embedded in epoxy resin (TAAB, EPON812). Ultrathin sections were cut on an ultramicrotome (Reichert Nissei Ultracut N) at 100 nm thickness, placed on a nickel grid (VECO, 2552), etched, and washed with 5% sodium metaperiodate for 10 min, and washed with PBS. The sections were incubated with anti Flag antibodies (1:100) for 2 h. After washing with PBS, the sections were incubated with 15 nm gold conjugated goat anti mouse IgG antibodies (Cyto diagnostics, 1:50) for 1 h. After being washed with PBS, the sections were fixed in 2.5% GA/PBS for 10 min and washed with Milli Q water to stabilise the antigen-antibody binding. The immunostained sections were further stained with uranyl acetate for 3 min and lead citrate for 1 min, and then observed under a transmission electron microscope (Hitachi, H 7650, H 7650 control Ver. 01. 10, EMIP Ver. 05. 04). The number of gold particles per unit area in the MVBs, plasma membrane, mitochondrion, and nuclear membrane were counted.

**Electron microscope analysis for sEVs.** Ultramicroscopic analysis was performed following the conventional protocol with slight modification<sup>34</sup>. Briefly, the sEVs isolated by the differential ultracentrifugation protocol from the conditioned medium of MDA MB 231 cells were mixed with 4% PFA/PBS in equal amounts, then deposited on electron microscopy (EM) grids (Nissin EM, Excel Support Film). The sEVs on the EM grids were washed with PBS, further fixed for 5 min with 1% GA/0.1M phosphate buffer, and washed eight times every 2 min with ultrapure water. The EM grid was negatively stained with 10% EM stain (Nissin EM) and analysed using a transmission electron microscope.

**Isolation of EVs and RNA analyses.** sEVs were isolated from a conditioned medium according to the conventional protocol with slight modification<sup>34</sup>. Briefly, exosome depleted FBS (centrifuged for 16 h at 100,000  $\times$  g) was used to prepare the conditioned medium. After incubation for 24 h, the culture medium (approximately 90 mL from  $3 \times 10^7$  cells) was collected and centrifuged at 300  $\times$  g for 10 min at 4 °C. The supernatant was centrifuged at 2000  $\times$  g for 20 min at 4 °C. The supernatant was centrifuged again at 10,000  $\times$  g for 30 min at 4 °C. To remove the cellular debris, the supernatant was filtered through a 0.22  $\mu$ m filter (Millipore, Millex GV). The sEV pellet was then harvested by ultracentrifugation at 100,000  $\times$  g (Beckman, SW32Ti rotor) for 70 min at 4 °C. The sEV pellet was washed with 1 mL of PBS and collected by ultracentrifugation at 100,000  $\times$  g (Beckman, TLA 110 rotor) for 60 min at 4 °C. The sEV pellet was again washed with 1 mL of PBS and collected by ultracentrifugation at 100,000  $\times$  g for 60 min at 4 °C, then resuspended in PBS. For the purification of the 2K, 10K and 100K pellets, a differential ultracentrifugation protocol was used<sup>21</sup>. Briefly, exosome depleted FBS was used to prepare the conditioned medium. After incubation for 24 h, the culture medium (approximately 90 mL from  $3 \times 10^7$  cells) was collected and centrifuged at 300  $\times$  g for 10 min at 4 °C. The supernatant was centrifuged at 2000  $\times$  g for 20 min at 4 °C. The resulting pellet was called 2K pellet. The supernatant was centrifuged again at 10,000  $\times$  g for 40 min at 4 °C. The resulting pellet was called 10K pellet. The supernatant was centrifuged by ultracentrifugation at 100,000  $\times$  g (Beckman, SW32Ti rotor) for 90 min at 4 °C. The resulting pellet was called 100K pellet. The 100K pellet was washed with 1 mL of PBS and collected by ultracentrifugation at

100,000  $\times g$  (Beckman, TLA 110 rotor) for 60 min at 4 °C. The 100K pellet was again washed with 1 mL of PBS and collected by ultracentrifugation at 100,000  $\times g$  for 60 min at 4 °C, then resuspended in PBS. 2K and 10K pellets were washed with 30 mL of PBS, centrifuged at the same speed, and then resuspended in PBS. Filtration through 0.22  $\mu m$  filters (Millipore, Millipore GV) was additionally performed before the purification of the 100K pellet. For the mouse serum, after collection, whole blood was allowed to clot by leaving it undisturbed for 30 min at RT. To remove the clot, the whole blood was centrifuged twice at 8000  $\times g$  for 10 min. For the preparation of mouse plasma, after collection, the whole blood was added to a heparin solution (Mochida Pharmaceutical, heparin Na 5000 units per 5 mL), rotated slowly, centrifuged at 9100  $\times g$  for 5 min, and the supernatant was used as the plasma.

In each experiment, the volume of serum from the *Ub3* KO mice was matched to that of its wild type littermates (approximately 350–500  $\mu L$ ). The harvested serum was immediately diluted with an equal volume of PBS and centrifuged at 2000  $\times g$  for 30 min at 4 °C. The supernatant was centrifuged at 12,000  $\times g$  for 45 min at 4 °C. The supernatant was diluted fivefold with PBS and then filtered through a 0.22  $\mu m$  filter. The EV pellet was then harvested by ultracentrifugation at 110,000  $\times g$  (Beckman, SW55Ti rotor) for 70 min at 4 °C. The sEV pellet was washed with 1 mL of PBS and collected by ultracentrifugation at 100,000  $\times g$  (Beckman, TLA 110 rotor) for 60 min at 4 °C, and resuspended in PBS. To quantify the total RNA amounts of serum sEVs, RNA was purified from the whole serum sEVs from each mouse using the miRNeasy Kit (QIAGEN). The purified RNA solution was measured by a Bioanalyzer 2100 system (Agilent, 2100 Expert Ver. B. 02. 08. SI648, RNA 6000 Pico kit). For real time qPCR analysis, the purified RNA solution (40%) was reverse transcribed into complementary DNA (cDNA), and 10% of the resulting cDNA solution for each miRNA was used in real time qPCR analysis.

Protein concentrations of the serum, plasma, and cell lysates were measured by BCA (Thermo). To quantify the protein concentration of the 2K, 10K, and 100K pellets, each pellet was diluted with 2% SDS and measured by Micro BCA (Thermo).

**Measuring total protein amount in the serum sEVs.** Ten percent of the purified serum sEVs from each experiment was boiled with  $\beta$ ME containing an SDS sample buffer and separated by SDS polyacrylamide gel electrophoresis. The resulting gel was fixed in 50% methanol and 10% acetic acid for 30 min and washed in 10% methanol and 7% acetic acid for 30 min. The gel was stained in SYPRO Ruby Gel Stain (Thermo) for 3 h in the dark. The gel was washed in 10% methanol and 8% acetic acid for 30 min, and then washed in ultrapure water for 10 min. The gel was then scanned using a fluorescence scanner (GE Healthcare, Typhoon 9400, Typhoon Scanner Control Ver. 5.0) set at 532 nm. The total protein intensity was analysed using the ImageJ software.

**Proteomics analyses.** MDA MB 231 cells were transfected with plasmids using Lipofectamine 2000. After 24 h, cells were washed and collected by scraping into ice cold PBS, pelleted by centrifugation at 400  $\times g$  for 1 min at 4 °C, and lysed with 1% Triton buffer for 20 min at 4 °C. Crude nuclei and unbroken cells were excluded by centrifugation at 20,000  $\times g$  for 5 min at 4 °C. For pre clearing, the supernatants were incubated twice with 40  $\mu L$  Protein G sepharose beads (GE Healthcare, 50% slurry in lysis buffer) for 1 h at 4 °C. The supernatants were then incubated with 50  $\mu L$  anti Flag M2 affinity gel (Sigma, 50% slurry in lysis buffer) for 16 h at 4 °C. The beads were washed with 1% Triton buffer, washed three times with wash buffer (50 mM Tris HCl [pH 7.5]) to remove detergent, resuspended in 200  $\mu L$  digestion buffer (2 M urea, 50 mM Tris HCl [pH 7.4], 1 mM DTT, 5 mM iodoacetamide, 1  $\mu g$  of trypsin [Promega, V5280]), and incubated by shaking at 1200 rpm for 16 h at 37 °C. The samples were acidified with 22  $\mu L$  of 10% trifluoroacetic acid (TFA). The peptides contained in the supernatant were then taken by centrifugation at 4400  $\times g$  for 30 sec. The peptides were desalting by C18Tip (Thermo, StageTips), dried by speedvac, and then resuspended in mass spectrometry buffer (2% acetonitrile, 0.2% TFA). Peptides were separated on a 50 cm reversed phase column (75  $\mu m$  inner diameter, packed in house with ReproSil Pur C18 AQ 1.9  $\mu m$  resin [Dr. Maisch GmbH]) with a binary buffer system of buffer A (0.1% formic acid (FA)) and buffer B (80% acetonitrile plus 0.1% FA) over a 100 min gradient (5–30% and 30–65% of buffer B for 95 and 5 min, respectively) using the EASY nLC 1000 system (Thermo) with a flow rate of 300 nL per min. Column temperature was maintained at 50 °C. The nLC system was coupled to a Q Exactive HF mass spectrometer (ThermoFisher Scientific), acquiring full scans (300–1650  $m/z$ , maximum injection time 20 ms, resolution 60,000 at 200  $m/z$ ) at a target of 3e6 ions. The 15 most intense ions were isolated and fragmented with higher energy collisional dissociation (HCD) (target 1e5 ions, maximum injection time 25 ms, isolation window 1.4  $m/z$ , NCE 27%, under fill ratio 0.1%) and detected in the Orbitrap (resolution 15,000 at 200  $m/z$ ).

Raw MS files were processed within the MaxQuant environment (Ver. 1.5.9.3) using the MaxLFQ algorithm for label free quantification and the integrated Andromeda search engine with FDR < 0.01 at the protein and peptide levels<sup>35,36</sup>. The search included variable modifications for oxidised methionine (M), acetylation (protein N term), and fixed modifications for carbamidomethyl (C). Up to two missed cleavages were allowed for protease digestion. Peptides with at least

seven amino acids were considered for identification, and 'match between runs' was enabled with a matching time window of 0.7 min to allow the quantification of MSI features which were not identified in each single measurement. Peptides and proteins were identified using a UniProt FASTA database from *Homo sapiens* (2015) containing 21,051 entries.

The freely available software PERSEUS (Ver. 1.5.4.1) was used to perform all statistical and bioinformatics analyses [<http://www.perseus-framework.org>]. The proteins identified only by site modification or found in the decoy reverse database and the contaminants were filtered out before data analysis. MaxLFQ intensities were taken for quantification and transformed into log2 scale. Three biological replicates of each pulldown were grouped, and a minimum of three valid values was required in at least one group. Missing values were imputed based on a normal distribution (width = 0.3, down shift = 1.8). MaxLFQ intensities were first z-scored, and samples were clustered into five clusters according to the Euclidean metric as a distance measure for column and row clustering (distance threshold = 2.8). Annotations were added from Gene Ontology (GO) for enrichment analysis with a Benjamini Hochberg FDR threshold of 0.02. Significance was assessed using ANOVA analysis including a permutation based FDR of 5% and an S0 value of 1. The direction of significance was also assessed by the post hoc test with 5% FDR. For pairwise comparison of groups, two sample Student's t test (two sided) was used including a permutation based FDR of 5% and an S0 value of 1. Fisher exact test was applied by using p value for truncation with threshold value 0.02.

**Measurement of sEVs by nanoparticle tracking analysis.** Nanoparticle tracking analysis (NTA) was carried out using NanoSight LM10HS with a blue laser system (NanoSight, Amesbury) on isolated serum sEV particles diluted 25 fold with PBS<sup>37</sup>. The system focuses a laser beam through a suspension of the particles of interest. They were visualised by light scattering using a conventional optical microscope aligned perpendicularly to the beam axis, which collects light scattered from every particle in the field of view. A 60 s video records all events for further analysis by NTA software. The Brownian motion of each particle was tracked between frames to calculate its size using the Stokes Einstein equation. Each sample was measured three times, and we used an averaged value for the statistical analysis.

**Statistical analysis.** Statistical analyses were carried out using Prism 6.03 (GraphPad Software), Microsoft Excel, and PERSEUS (Ver. 1.5.4.1). The Mann Whitney test or Wilcoxon signed rank test and Kruskal–Wallis/Dunn's multiple-comparisons test were, respectively, applied for comparisons between 2 groups and 3 or more groups except proteomics analyses. We used more than two independent samples for each experiment. Horizontal bars indicate median values for each group.

## Data availability

Proteomics raw datasets are deposited in jPOST (under accession codes [PJST000315](https://jpost.org/post/PJST000315), [PX10007617](https://jpost.org/post/PX10007617)).

Received: 5 June 2018 Accepted: 21 August 2018

Published online: 26 September 2018

## References

1. Raposo, G. & Stoorvogel, W. Extracellular vesicles: exosomes, microvesicles, and friends. *J. Cell. Biol.* **200**, 373–383 (2013).
2. Katzmann, D. J., Odorizzi, G. & Emr, S. D. Receptor downregulation and multivesicular body sorting. *Nat. Rev. Mol. Cell Biol.* **3**, 893–905 (2002).
3. Colombo, M., Raposo, G. & Théry, C. Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. *Annu. Rev. Cell. Dev. Biol.* **20**, 255–289 (2014).
4. Valadi, H. et al. Exosome mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat. Cell Biol.* **9**, 654–659 (2007).
5. Hoshino, A. et al. Tumour exosome integrins determine organotropic metastasis. *Nature* **527**, 329–335 (2015).
6. Rajendran, L. et al. Alzheimer's disease beta amyloid peptides are released in association with exosomes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11172–11177 (2006).
7. Kannert, D. et al. C terminally truncated forms of Tau, but not full length Tau or its C terminal fragments, are released from neurons independently of cell death. *J. Neurosci.* **35**, 10851–10865 (2015).
8. Fevrier, B. et al. Cells release prions in association with exosomes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9683–9688 (2004).
9. Vella, L. J. et al. Packaging of prions into exosomes is associated with a novel pathway of PrP processing. *J. Pathol.* **211**, 582–590 (2007).
10. Lee, H. J., Bae, E. J. & Lee, S. J. Extracellular alpha synuclein a novel and crucial factor in Lewy body diseases. *Nat. Rev. Neurol.* **10**, 92–98 (2014).

## ARTICLE

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-06197-y

11. Reinstein, E. & Ciechanover, A. Narrative review: protein degradation and human diseases: the ubiquitin connection. *Ann. Intern. Med.* **145**, 676–684 (2006).
12. Ageta, H. et al. Regulation of the level of Ves1 IS/Homer 1a proteins by ubiquitin proteasome proteolytic systems. *J. Biol. Chem.* **276**, 15893–15897 (2001).
13. Ageta, H., Kato, A., Fukazawa, Y., Inokuchi, K. & Sugiyama, H. Effects of proteasome inhibitors on the synaptic localization of Ves1 IS/Homer 1a proteins. *Brain Res. Mol. Brain. Res.* **97**, 186–189 (2001).
14. Yao, I. et al. SCRAPPER dependent ubiquitination of active zone protein RIM1 regulates synaptic vesicle release. *Cell* **130**, 943–957 (2007).
15. Welchman, R. L., Gordon, C. & Mayer, R. J. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.* **6**, 599–609 (2005).
16. Yang, H. et al. Transmembrane and ubiquitin-like domain containing protein 1 (Tmub1/IIOPS) facilitates surface expression of GluR2 containing AMPA receptors. *PLoS One* **3**, e2809 (2008).
17. Johnson, E. S. Protein modification by SUMO. *Annu. Rev. Biochem.* **73**, 355–382 (2004).
18. Watson, I. R., Irwin, M. S. & Ohh, M. NEDD8 pathways in cancer, Sine Quibus Non. *Cancer Cell* **19**, 168–176 (2011).
19. Downes, B. P., Saracco, S. A., Lee, S. S., Crowell, D. N. & Vierstra, R. D. MUBs, a family of ubiquitin fold proteins that are plasma membrane anchored by prenylation. *J. Biol. Chem.* **281**, 27145–27157 (2006).
20. Zhang, F. L. & Casey, P. J. Protein prenylation: molecular mechanisms and functional consequences. *Annu. Rev. Biochem.* **65**, 241–269 (1996).
21. Kowal, J. et al. Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E968–E977 (2016).
22. Bobrie, A., Colombo, M., Krumeich, S., Raposo, G. & Théry, C. Diverse subpopulations of vesicles secreted by different intracellular mechanisms are present in exosome preparations obtained by differential ultracentrifugation. *J. Extracell. Vesicles* **1**, <https://doi.org/10.3402/jev.v10.18397> (2012).
23. Ostrowski, M. et al. Rab27a and Rab27b control different steps of the exosome secretion pathway. *Nat. Cell Biol.* **12**, 19–30 (2010).
24. Gonzales, P. A. et al. Large scale proteomics and phosphoproteomics of urinary exosomes. *J. Am. Soc. Nephrol.* **20**, 363–379 (2009).
25. Weinberg, R. A. *The Biology of Cancer*. (Garland Science, Boca Raton, FL, 2007).
26. Demory Beckler, M. et al. Proteomic analysis of exosomes from mutant KRAS colon cancer cells identifies intercellular transfer of mutant KRAS. *Mol. Cell. Proteom.* **12**, 343–355 (2013).
27. Flotho, A. & Melchior, F. Sumoylation: a regulatory protein modification in health and disease. *Annu. Rev. Biochem.* **82**, 357–385 (2013).
28. Al Nedawi, K. et al. Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. *Nat. Cell Biol.* **10**, 619–624 (2008).
29. Peinado, H. et al. Melanoma exosomes educate bone marrow progenitor cells toward a pro metastatic phenotype through MET. *Nat. Med.* **18**, 883–891 (2012).
30. Rodeheffer, M. S., Birsoy, K. & Friedman, J. M. Identification of white adipocyte progenitor cells in vivo. *Cell* **135**, 240–249 (2008).
31. Uezumi, A., Fukada, S., Yamamoto, N., Takeda, S. & Tsuchida, K. Mesenchymal progenitors distinct from satellite cells contribute to ectopic fat cell formation in skeletal muscle. *Nat. Cell Biol.* **12**, 143–152 (2010).
32. Hitachi, K., Nakatani, M. & Tsuchida, K. Myostatin signaling regulates Akt activity via the regulation of miR 486 expression. *Int. J. Biochem. Cell. Biol.* **47**, 93–103 (2014).
33. Dantuma, N. P., Groothuis, T. A., Salomons, F. A. & Neefjes, J. A dynamic ubiquitin equilibrium couples proteasomal activity to chromatin remodeling. *J. Cell. Biol.* **173**, 19–26 (2006).
34. Théry, C., Amigorena, S., Raposo, G. & Clayton, A. Isolation and characterization of exosomes from cell culture supernatants and biological fluids. *Curr. Protoc. Cell. Biol.* <https://doi.org/10.1002/0471143030.cb0322s30> (2006). Unit3 22.
35. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b. range mass accuracies and proteome wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
36. Cox, J. et al. Accurate proteome wide label free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteom.* **13**, 2513–2526 (2014).
37. Yoshioka, Y. et al. Comparative marker analysis of extracellular vesicles in different human cancer types. *J. Extracell. Vesicles* **2**, <https://doi.org/10.3402/jev.v20.20424> (2013).

## Acknowledgements

The authors thank Showbu Sato, T. Hino, and R. Migishima for the generation of the *Ub3* KO mice; Y. Oda and M. Amrai for the phylogenetic tree analysis; M. Kono, G. Niimi, N. Yamamoto, and K. Taniguchi for suggestions and technical support in electron microscope analysis; K. Otsu, Saori Sato, Y. Kato, and N. Kawamura for technical assistance; S. Kamijo, M. Matsuo, and S. Nagao for maintenance of the *UBL3* KO mice; A. Yamaguchi for genotyping the *Ub3* KO mice; and S. Ono and A. Kato for suggestions and discussion. This work was initiated while H.A. was a research scientist at MITILS, and sEV analyses were mainly performed at PHU. This work was partly supported by JSPS KAKENHI (22700346, 26640084, JP16H06280, 17H03945, 18K07209, and 16K08599); an Intramural Research Grant (26-8 and 29-4) for Neurological and Psychiatric Disorders of NCNP; the Ministry of Education, Culture, Sports, Science and Technology (Grant No.15H03898B1); CREST from Japan Agency for Medical Research and Development; AMED (Grant No.1921910520); Innovative Areas-Resource and technical support platforms for promoting research 'Advanced Bioimaging Support'; 'Fluorescence Live imaging' of the Ministry of Education, Culture, Sports, Science and Technology; and the Program for Strategic Research Foundation at Private Universities.

## Author Contributions

H.A., T.O., M.M., M.S., and K.T. contributed to the study design. H.A., N.A.-I., K. Hitachi, O.K., T.O., H.Y., T.K., K. Hatanaka, K.I., Y.Y., N.K., M.N., A.U., T.I., Y.T., H.S., and M.K. participated in data collection and interpretation. H.A., K. Hitachi, T.K., H.S., and K. Hatanaka cloned constructs. K.N. contributed new reagents/analytic tools. K. Hatanaka generated the anti-*UBL3* antibody. H.A., M.S., and K.T. wrote the paper. All authors discussed and commented on the manuscript.

## Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-018-06197-y>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

### **3.2.3. DIA-based systems biology approach unveils novel substrates of the GID E3 ligase in response to a metabolic shift**

Christine R. Langlois <sup>1,a</sup>, Ozge Karayel <sup>2,a</sup>, Andre C. Michaelis <sup>1</sup>, Matthias Mann <sup>2,b</sup> and Brenda A Schulman <sup>1,b</sup>

<sup>1</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>2</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>a</sup> These authors contributed equally <sup>b</sup> Corresponding authors

Manuscript in preparation (2020)

Proteome remodeling has proven to be a vital cellular mechanism in response to stress and changes in environmental conditions. Therefore, understanding how the proteome changes at a global level in response to environmental cues is crucial to uncover the underlying molecular mechanisms that facilitate cellular adaptation. The yeast *Saccharomyces cerevisiae* is a powerful model system for systems-wide investigation of changes to the proteome in response to a variety of stresses and growth conditions. Indeed, large-scale proteomics screens were pioneered in the yeast model and the proteomics community has achieved nearly comprehensive coverage for this organism. Owing to advances in the mass spectrometry technology and new acquisition modes, such as data-independent acquisition (DIA), we anticipate that it would now be possible to obtain high yeast proteome coverage by a straightforward rapid single run approach. In this study, we describe a systems biology approach employing plate-based sample preparation and rapid, single-run data independent mass spectrometry analysis (DIA). Our approach enables quantitative profiling of hundreds of largely covered yeast proteomes in only a few days. To evaluate its capability, we comprehensively and quantitatively analyzed yeast stress response. Furthermore, using our generic and robust methodology and an assay for protein degradation, we identified and validated novel substrates of the GID E3 ligase, which is a major regulator of cellular metabolism during the switch from gluconeogenic to glycolytic conditions.

**DIA-based systems biology approach unveils novel substrates of the GID E3 ligase in response to a metabolic shift**

Christine R. Langlois <sup>1,a</sup>, Ozge Karayel <sup>2,a</sup>, Andre C. Michaelis <sup>1</sup>, Matthias Mann <sup>2,b</sup> and Brenda A Schulman <sup>1,b</sup>

<sup>1</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>2</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>a</sup> These authors contributed equally

<sup>b</sup> Corresponding authors

Correspondence: [mmann@biochem.mpg.de](mailto:mmann@biochem.mpg.de) and [schulman@biochem.mpg.de](mailto:schulman@biochem.mpg.de)

**Keywords:** yeast, mass spectrometry, LC-MS/MS, DIA, proteomics, GID E3 complex

**SUMMARY**

The yeast *Saccharomyces cerevisiae* is a powerful model system for systems-wide biology screens and to evaluate large-scale proteomics methods. The proteomics community has achieved nearly comprehensive coverage for this organism owing to advances in the mass spectrometry technology. However, it remains challenging to scale the technology for rapid and high-throughput analysis of the yeast proteome to investigate biological pathways on a global scale. Here we describe a systems biology approach employing plate-based sample preparation and rapid, single-run data independent mass spectrometry analysis (DIA). Our approach is straightforward, easy to implement and enables quantitative profiling of hundreds of largely covered yeast proteomes in only a few days. We evaluated its capability by characterizing changes in the yeast proteome in response to a variety of stresses. Moreover, using our generic, rapid and robust methodology we identified and validated novel substrates of the GID E3 ligase which is a major regulator of cellular metabolism during the switch from gluconeogenic to glycolytic conditions.

## INTRODUCTION

Proteome remodeling has repeatedly proven to be a vital cellular mechanism in response to stress, changes in environmental conditions, and toxins or pathogens. Cells must both synthesize proteins which enable them to adapt to the new environmental condition, and inactivate and/or degrade proteins which are detrimental or no longer needed. For each environmental perturbation, the proteome must be precisely and distinctly remodeled to ensure healthy and viable cells [1]. Indeed, decreases in proteome integrity are hallmarks of many human diseases, including cancer, Alzheimer's disease, muscular dystrophies, and cystic fibrosis (reviewed in [2]). Despite the importance of cellular stress response, our understanding of how cellular pathways interact during stress adaptation remains rudimentary. Therefore, understanding precisely how the proteome changes at a global level in response to environmental cues is crucial for identifying the underlying molecular mechanisms that facilitate cellular adaptation.

The yeast *Saccharomyces cerevisiae* is a powerful model system that is widely used to probe biological pathways, due to its ease of manipulation and rapid growth compared to mammalian models. In addition, the availability of extensive genetic toolboxes in yeast, including deletion libraries [3, 4], GFP-tagged libraries [5], and over-expression libraries [6] has made yeast a desirable model system for conducting transcriptomics, proteomics, or metabolomics screens. Indeed, systems wide biology screens and large-scale proteomics were both pioneered in the yeast model and developments of more sensitive and fast mass spectrometers have provided nearly comprehensive coverage for this organism [7-9]. For instance, the optimized sample preparation coupled with MS analysis performed on the Orbitrap hybrid mass spectrometer allowed identification of around 4,000 yeast protein over a 70-min LC-MS/MS run [7, 8]. However, the necessity of

technological expertise and long analysis times for the high-quality in-depth yeast proteome measurements has so far precluded the widespread adoption of cutting-edge proteomics workflows in the yeast research community. With further advances in technology and new acquisition modes, such as data-independent acquisition (DIA), we anticipate that it would now be possible to obtain very accurate and high yeast proteome coverage by a straightforward rapid single run approach, enabling researchers to study biological concepts on a global scale.

One ubiquitous mechanism for maintaining proteome integrity is the marking and degradation of proteins that are damaged or no longer needed with ubiquitin. The conjugation of ubiquitin to its targets is catalyzed by E3 ubiquitin ligases, a diverse group of enzymes that recognize and bind target proteins, and facilitate ubiquitin transfer together with an E2, ubiquitin conjugating enzyme. Ubiquitination relies on a variety of cellular signals to direct E3 ligases to their target proteins, and tight regulation of this process is crucial for cellular viability [10]. For instance, during carbon starvation, yeast cells induce expression of the inactive GID (glucose-induced degradation) E3 ligase, which is subsequently activated upon glucose replenishment. Following its activation, the GID E3 ligase targets gluconeogenic enzymes, leading to their degradation, and allowing the yeast cell to proceed with aerobic growth [11-13]. In addition, ubiquitin ligases also serve as crucial regulators in response to oxidative, heavy metal, and protein folded stresses [14-16]. Despite the importance of ubiquitination in response to cellular perturbation, our understanding of the E3-dependent responses to cellular perturbation remains incomplete.

Here, we describe a systems biology approach employing rapid, single-run data independent mass spectrometry analysis to comprehensively map changes to the yeast proteome in response to a variety of yeast stresses. We investigate

growth conditions commonly used in yeast research, including growth media, heat shock, osmotic shock, amino acid starvation, and nitrogen starvation. We demonstrate that our DIA-based approach is sensitive and robust enough to detect quantitative proteome remodeling in response to stress. In addition, we apply our methodology to probe a specific biological question- to identify substrates of an E3 ligase during a metabolic switch. The GID E3 ligase is a major regulator of cellular metabolism during the switch from gluconeogenic to glycolytic conditions [13, 17]. Here we show not only all known substrates of the GID E3 ligase, but also two novel substrates which display distinct patterns regulation.

## RESULTS

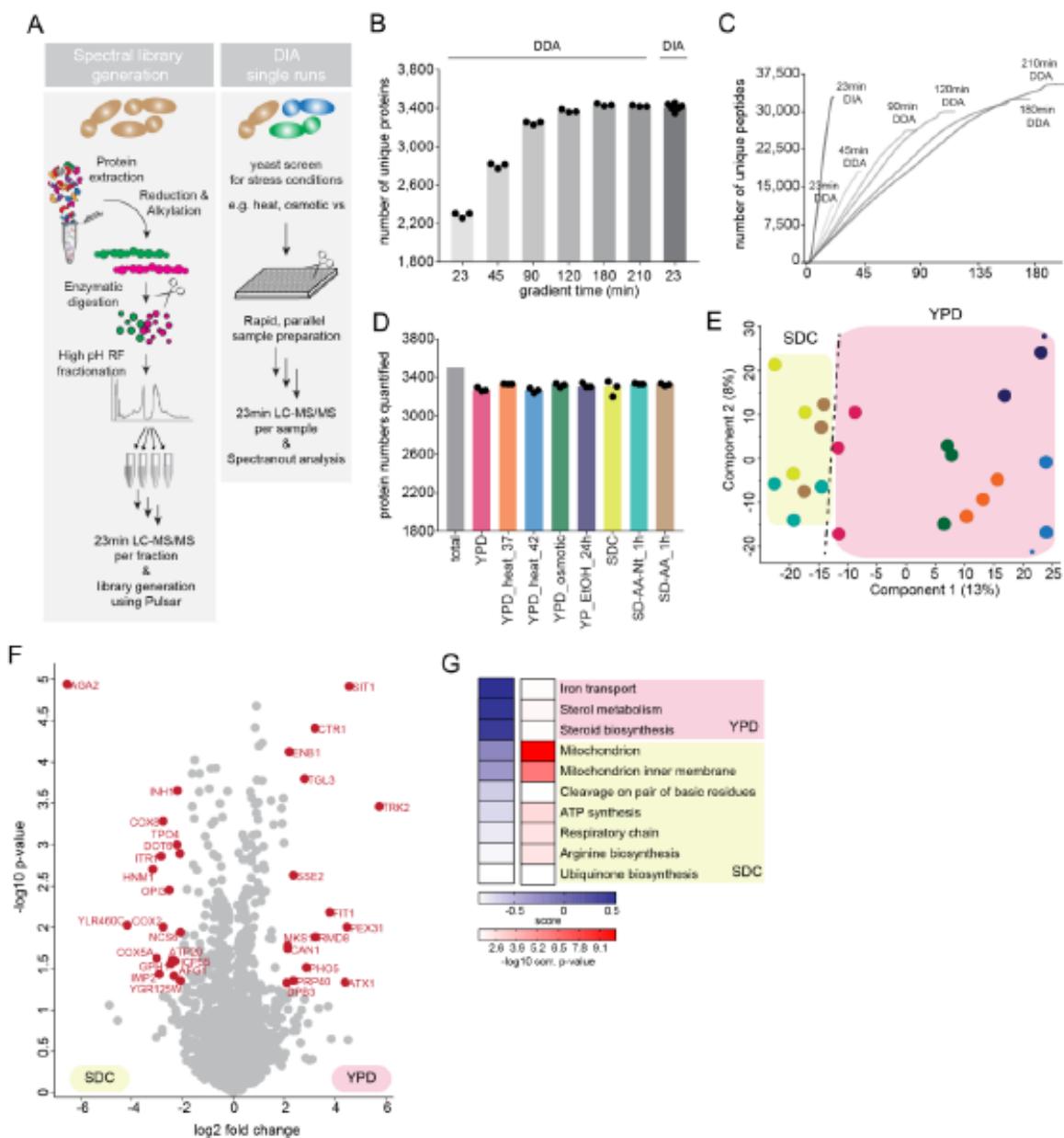
### Streamlined and scalable yeast proteome analysis employing data-independent acquisition

Numerous essential cellular pathways, including the cell cycle, metabolism, proteins quality control and degradation, and signaling pathways such as mitogen-activated protein kinase (MAPK) and target of rapamycin (TOR), are conserved from yeast to humans, making yeast a powerful eukaryotic model system [18-21]. Rapid and robust yeast proteome profiling will allow researchers to systematically and comprehensively probe biological pathways. To establish a fast and scalable single run analysis approach for yeast proteome profiling, we explored data-independent acquisition (DIA) strategy on an Orbitrap mass spectrometer (Figure 1A). Unlike data-dependent acquisition (DDA), a DIA method isolates co-eluting peptide ions together in predefined mass windows, fragmenting and analyzing all ions simultaneously [22], a strategy that has been shown to be crucial to overcome limited capacity of sequential DDA, enabling fast and scalable single-shot analysis workflows yielding substantially higher number of identified proteins with unprecedented quantitative accuracy [23, 24]. To generate a yeast-specific and comprehensive spectral

library necessary for this approach, we have grown yeast under various growth and stress conditions. After extraction and digestion of proteins, we separated peptides obtained from each condition by basic reversed-phase (bRP) chromatography into 8 fractions (Figure 1A). In total, we had 64 fractions (8 fractions x 8 conditions) which were analyzed using a DDA method with 23 min LC gradient and with the Spectronaut software. This method together with LC overhead time took about half an hour, allowing the analysis of 45 samples per day – almost half a 96 well plate. Our library comprised more than 74,103 precursors which mapped into 4,712 unique proteins, covering 87% of whole expressed yeast proteome according to the previous report that computationally combined the vast data resource of 21 different large-scale proteome datasets [25].

Combining with our in-depth spectral library, 23 min DIA method on average identified 33,909 peptides and 3,413 distinct proteins in single measurements of six replicates (Q-value less than 1% at protein and precursor levels, Figure 1B-C). This implies that about 72% of proteins in the deep yeast spectral library were matched into the single runs. Note that the library covers many proteome state and the single runs only one, so the degree of proteome completeness is likely much higher than 73%. Measurements were highly reproducible with Pearson coefficients greater than 0.92 between replicates (Figure S1A) and CVs < 20% for 68% of all common proteins between the six replicates. Single run data-dependent acquisition strategy with the same LC gradient quantified only 11,883 peptides and 2,289 distinct proteins on average (Figure 1B-C). To directly contrast the performance of 23 min DIA method to the DDA method we analyzed the same sample with increasing gradient lengths. We determined that we could reach the same depth using DDA method only with at least 180 min long LC gradients (33,425 peptides and 3,435 proteins) (Figure 1B-C). Thus, the DIA method allows us to obtain comparable

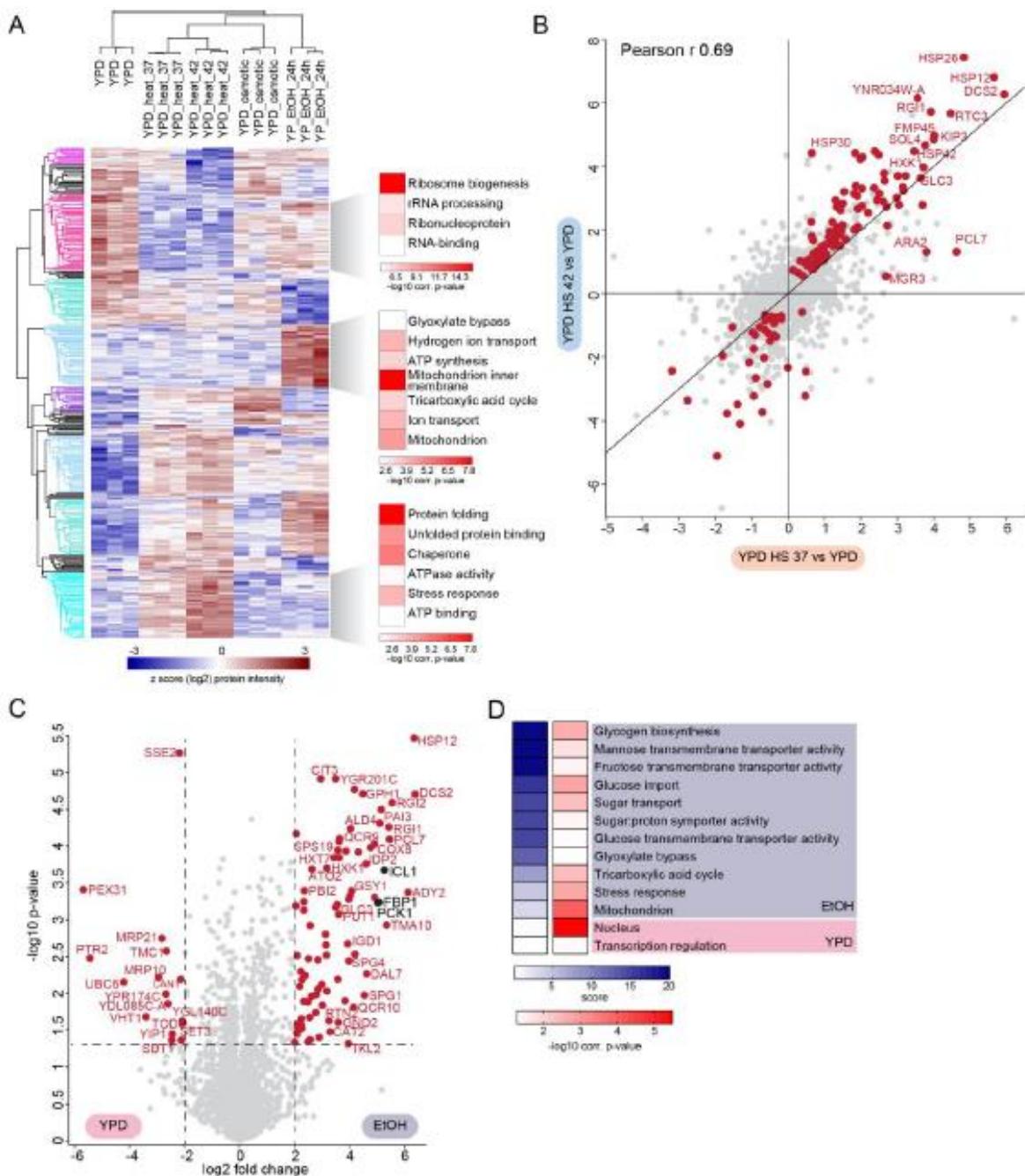
coverage to DDA in high throughput and in-depth fashion in almost one-tenth of the analysis time.



**Figure 1. Fast and scalable yeast proteome analysis using data-independent acquisition**

A. Experimental workflow for yeast spectral library construction (left panel) and fast, single run DIA-based analysis of yeast proteome (right panel). B. Number of identified proteins using DDA with varying LC gradient lengths compared to 23min DIA. C. Cumulative number of identified unique yeast peptides over time using DDA with varying LC gradient lengths and 23min DIA methods. D. Number of quantified proteins in stress conditions. E. PCA of stress conditions along with their biological replicates based on their proteomic expression profiles. F. Volcano plot of the (-log10) p-values vs. the log2 protein abundance differences between yeast grown in YPD vs. SC. The proteins marked in red change significantly (T-test, FDR < 5%). G. GO-term enrichment in the YPD vs. SC fold change

dimension (1D enrichment, FDR < 5%). Terms with positive enrichment scores are enriched in YPD over SC and vice versa.



**Figure 2. Large-scale and quantitative analysis of yeast proteomes under different stresses**

A. Heat map of z-scored protein abundances ( $\log_2$  DIA intensities) of the differentially expressed proteins (ANOVA,  $FDR < 0.01$ ) after hierarchical clustering of stress conditions performed in YPD. Fisher exact test was performed to identify significantly enriched GO-terms in the most prominent profiles ( $FDR < 5\%$ ). B. Correlation of  $\log_2$  fold-changes of the all quantified proteins in HS experiment. The proteins that change significantly in either  $37^{\circ}\text{C}$  or  $42^{\circ}\text{C}$  compared to  $30^{\circ}\text{C}$  YPD control are colored in red (T-test,  $FDR < 5\%$ ). C. Volcano plot of the  $(-\log_{10})$  p-values vs. the  $\log_2$  protein abundance

differences between glucose starvation (ethanol) vs. YPD. Red dots indicate significantly different proteins, determined based on  $p$  value  $< 0.05$  and at least 4-fold change on both directions. D. GO-term enrichment in the ethanol vs. YPD fold change dimension (1D enrichment, FDR  $< 5\%$ ). Terms with positive enrichment scores are enriched in stress condition over YPD control and vice versa.

#### Large-scale and quantitative analysis of yeast stress response in half a day

Using our DIA-based systems biology approach, we next comprehensively and quantitatively analyze proteome changes in response to various stresses in yeast. Each condition was processed in three biological replicates and – after tryptic digestion – the peptides were analyzed in single runs using our rapid DIA method. We quantified 3,506 distinct proteins in total (Figure 1D). Reproducibility was high, with Pearson correlations  $> 0.93$  between the three biological replicates (Figure S1B). Strikingly, over 90% of all detected proteins were consistently quantified at varying levels across all conditions (Supplementary Table 1). Principal component analysis (PCA) demonstrated that the first component segregates the conditions, mainly growth media and accounts for 13% of the variability (Figure 1E).

We first looked more closely at the differences in protein expression during growth in YPD (Yeast extract peptone dextrose; rich media) and SC (Synthetic complete media), the two most common growth media used in yeast research. During growth in YPD, the three most significant proteins (Sit1, Ctr1, and Enb1) are regulators of copper and iron transport (Figure 1F (right), 1G), consistent with the fact that copper and iron are limiting factors for the growth of yeast at more alkaline pH [26]. Conversely, during growth in SDC, many mitochondrial proteins were upregulated compared to YPD (Figure 1F (left), 1G), including the cytochrome c oxidase subunits Cox8, Cox2, and Cox5a, the mitochondrial ATP synthase, Atp20, and the mitochondrial aminopeptidase, Icp55. Because yeast cannot synthesize mitochondria *de novo*, they instead reproduce through fission and must be inherited by daughter cells during cell division [27]. The

upregulation of many mitochondrial proteins is thus consistent with the faster growth rate of our yeast strains in SDC, compared to YPD. Because the choice of media is often considered crucial in experimental design, knowledge of how the proteins in a pathway of interest might be differentially regulated in different media will provide an important resource for yeast researchers.

Next, we investigated proteome changes in yeast grown under various stress conditions. Here, we focused on those commonly utilized in yeast research: heat shock, osmotic shock, carbon starvation, amino acid starvation, and nitrogen starvation. As expected, each produced a discrete stress response, resulting in synthesis or degradation of a distinct set of proteins (Figure 2A, S2A-F). For example, we found that yeast cells grown under heat shock induce expression of chaperones and stress response proteins, a well-characterized response that allows the cell to quickly recover from global heat-induced protein misfolding [28, 29]. Importantly, our DIA method also reveals that the heat shock response is 'dose dependent', with higher induction of the stress response at 42°C compared to 37°C (Figure 2A (teal cluster), 2B).

In addition to temperature, carbon source is a crucial determinant of yeast growth. We compared the proteomes of yeast grown in the aerobic carbon source, glucose (YPD), with the non-fermentable carbon source ethanol. Yeast will preferentially metabolize aerobic carbon sources, such as glucose, when they are present in the media. When only non-fermentable carbon sources, such as ethanol, are present, yeast cells will metabolize them through several pathways, including gluconeogenesis to generate glucose, and conversion of ethanol into pyruvate to allow for ATP generation in the

mitochondria via the tricarboxylic acid cycle [30-32]. Consistent with this, we observe a general upregulation of mitochondrial proteins, and those involved in the tricarboxylic acid cycle during growth in ethanol (Figure 2A, light blue cluster). In addition, many proteins involved in carbon metabolism are differentially regulated in glucose and ethanol containing media. For example, we see more than 16-fold upregulation of the gluconeogenic enzymes Fbp1, Pck1, and Icl1 (Figure 2C). In the absence of glucose, both Hxt7, a glucose transporter, and Hxk1, a hexokinase, are upregulated (Figure 2C, D), allowing the cell to quickly import and metabolize any glucose in the environment. These results are consistent with the idea that yeast have “anticipatory” programming, which not only allows them to adapt to current stressor, but also facilitates a rapid response to future shifts in environmental conditions [33]. Moreover, here we not only identify the proteins that specifically alter in response to a shift in environmental condition but also accurately determine their fold change, giving valuable insight into the protein content under different stress and growth conditions.

Taken together, our results indicate that our fast and robust DIA-based approach can reliably and quantitatively retrieve the known differences and even reveal new and biologically meaningful regulation of protein expression, and thereby provide a comprehensive resource for yeast researchers and a valuable platform to support future studies and quantitative biology.

#### Determining how the yeast proteome changes globally under glucose starvation and recovery

To gain better insights into how yeast regulate metabolism in response to carbon source, we next expanded our analysis to investigate the changes in the yeast proteome during glucose starvation and glucose recovery. Yeast cultures were first grown to logarithmic phase in glucose, then switched to media containing ethanol as a non-fermentable carbon source.

Following 19 hours growth in ethanol, glucose was replenished and the yeast were allowed to recover for 30 minutes or 2 hours (Figure 3A). Our rapid and robust DIA workflow quantified 3,602 distinct proteins in total. PCA analysis demonstrated that the first component segregates the growth conditions, with growth in glucose being largely separated from the ethanol and recovery conditions (Figure 3B, S3A). To further investigate the regulation of metabolism during growth in alternate carbon sources, we compared the proteome changes with that in the transcriptome. PCA analysis of the transcriptome also shows that the first component separates the growth conditions. Interestingly, in this case the cells growing in ethanol are largely separated from the glucose (never starved) and glucose recovery conditions (Figure 3C, S3B), suggesting that under these conditions, yeast cells rapidly remodel their gene expression first through rapid changes in transcription, which facilitates production of new proteins, followed by turnover of existing proteins that are no longer required.

Several regulatory mechanisms have been identified which contribute to carbohydrate metabolism, including allosteric regulation, reversible enzyme inactivation through covalent modifications, and irreversible loss of enzyme activity through proteolysis (reviewed in [34]). Importantly, we observed that protein turnover during glucose recovery occurs rapidly and in less than one cell division, suggesting an active mechanism of protein degradation. One such mechanism that has been well-characterized by our group and others is the ubiquitination and degradation of gluconeogenic enzymes by the GID (glucose-induced degradation) E3 ubiquitin ligase. The GID ligase is present at low levels in all growth conditions, however during growth in ethanol, most of the GID subunits are induced leading to the formation of an inactive anticipatory complex: GID<sup>Ant</sup>. Following glucose replenishment, the substrate receptor, Gid4, is rapidly induced and joins the complex [11], facilitating the recognition of the gluconeogenic

proteins Fbp1, Mdh2, Icl1, and Pck1 via the Pro N-end pathway (Figure 3D) [12]. Indeed, our DIA analysis confirmed that most components of the GID E3 ligase are upregulated around 4-fold during growth in ethanol, with the exception of Gid4, which is rapidly and transiently upregulated around 2-fold within in 30 minutes of glucose replenishment (Figure 3E).

Intriguingly, PCA analysis of individual proteins revealed that the known substrates of the GID E3 ubiquitin ligase, Fbp1, Pck1, Icl1 and, to a lesser extent, Mdh2 are major contributors to the segregation based on growth condition (Figure 3F). While the GID E3 ligase is known to be an important contributor to the regulation of yeast metabolism during the switch from gluconeogenic to glycolytic conditions, and is thought to have additional substrates, the lack of an obvious phenotype in GID mutants has made identification of further substrates problematic. Thus, we applied our rapid and robust DIA-based methodology to search for novel substrates of the GID E3 ligase.

#### Identifying novel GID ligase substrates during recovery from carbon starvation

We compared the transcriptomes and proteomes of wildtype yeast to yeast containing either a Gid4 (substrate receptor) deletion or a Gid2 mutant (Gid2<sup>K365A</sup>) which inactivates the RING domain and thereby inhibits substrate ubiquitination [11], grown under the glucose starvation and recovery conditions described previously. Each condition was measured in triplicates using the rapid DIA method, resulting in the quantification of 3,602 distinct proteins in total (Figure 4A). Importantly, there were no GID-dependent differences in mRNA levels following glucose replenishment (Figure S4A), demonstrating that the GID E3 ligase does not regulate any aspects of protein synthesis, but rather regulates the fate of existing proteins. To confirm that our DIA-based approach would be able to recognize bona fide GID substrates, we first examined the expression patterns of the known substrates

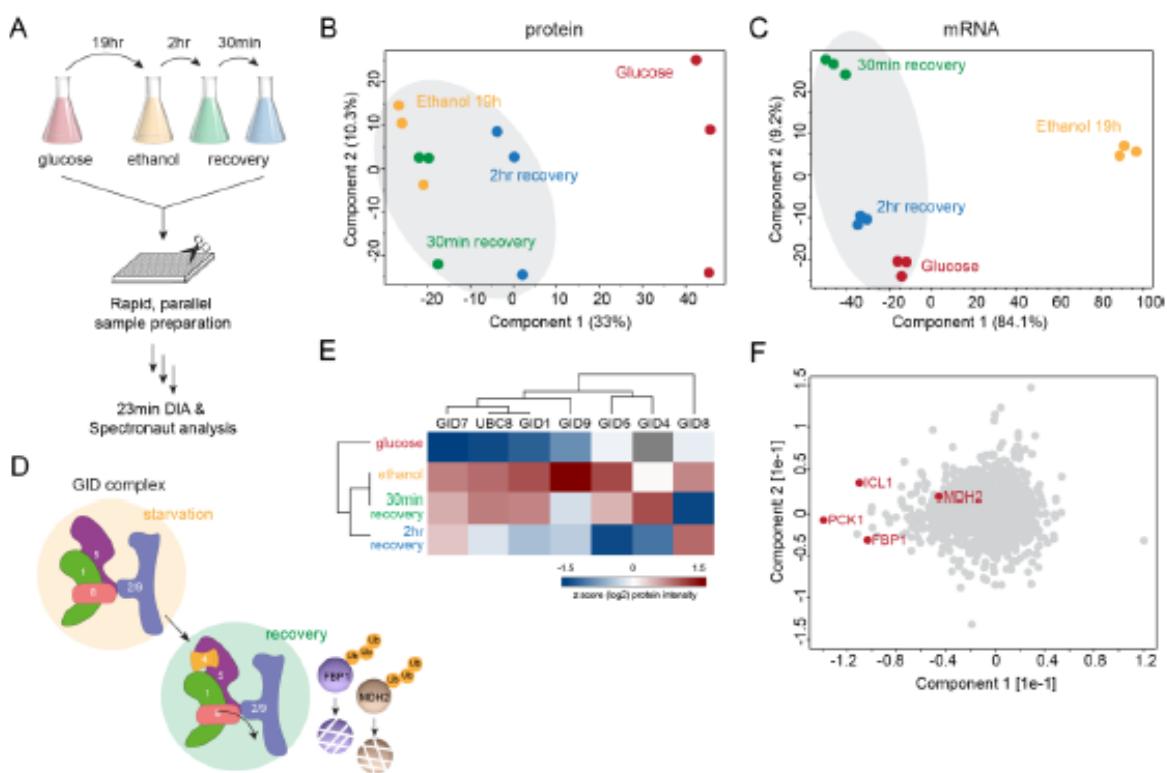
Fbp1 and Mdh2. Indeed, in wildtype cells, Fbp1 and Mdh2 proteins are induced during growth in ethanol and then rapidly turned over within two hours of glucose recovery, with Fbp1 and Mdh2 showing around 8-fold and 5.7-fold reduction, respectively, in protein levels in wildtype yeast. As expected, both proteins are also stabilized in the Gid4-deleted and gid2-mutant cells (Figure 4B, C), confirming that we can robustly identify changes in expression of known substrates.

To identify novel substrates, we searched for proteins with an expression profile similar to the known substrates based on the following criteria: 1) the protein should be expressed more highly in ethanol than glucose, 2) the protein levels should decrease during glucose replenishment, and 3) after two hours of glucose replenishment, the protein should have a higher expression level in the Gid4-deleted and/or gid2-mutant cells, compared to wildtype (Figure S4B). Application of these criteria provided a list of 31 proteins, including all four known GID substrates known so far (Fbp1, Mdh2, Pck1, and Icl1) (Figure S4C). To further narrow the candidate proteins, we required a proline in the second or third position, a genetic and structural requirement of all known substrates [11, 12, 35]. This provided us with a list of 7 proteins: the 4 known substrates, the transcription factor Azf1, and the metabolic enzymes Aro10 and Acs1 (Figure 4D). Interestingly, Azf1 has been implicated in regulation of GID4 transcription [36], suggesting its upregulation in the GID deficient cells may be a cellular compensation mechanism. However, because we did not observe any GID-dependent changes to mRNA expression (Figure S4A), we eliminated Azf1 from further analysis. Acs1 was significantly stabilized in both the Gid4-deleted and gid2-mutant cells, whereas Aro10 was only significantly stabilized in the gid2-mutant.

In order to validate Aro10 and Acs1 as novel substrates *in vivo*, we used the promoter reference technique [37], a transcription

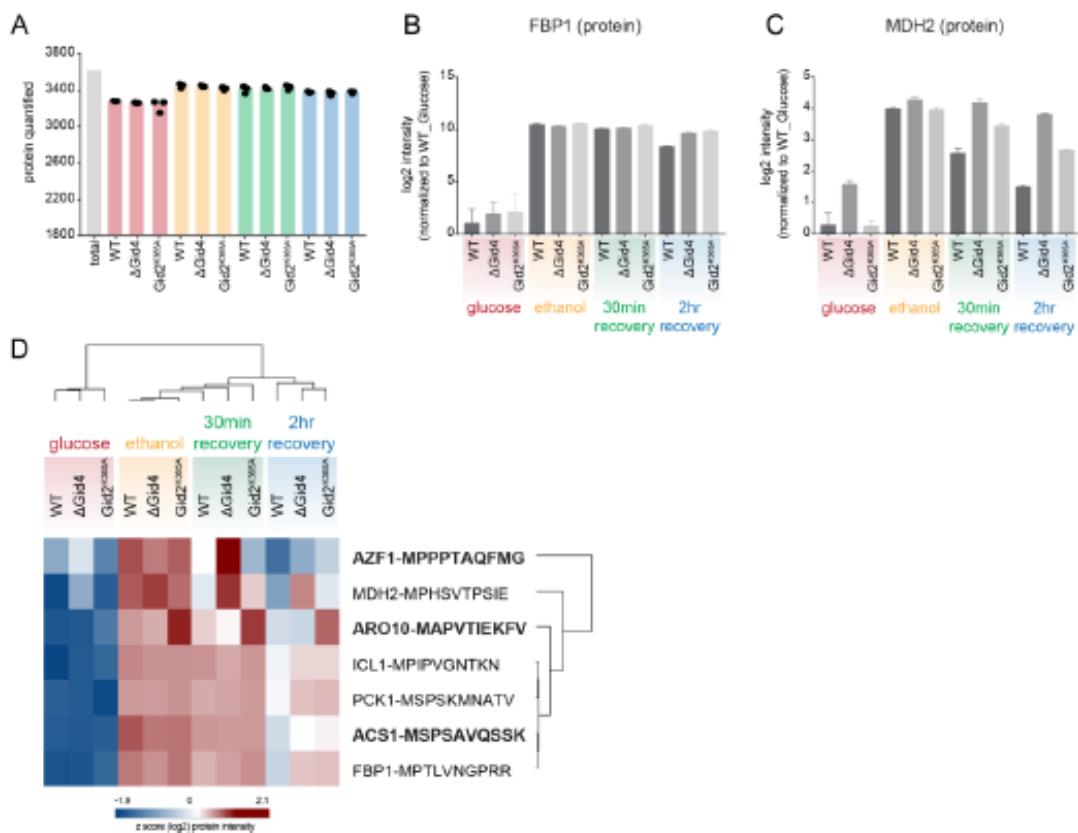
independent method to examine protein turnover. In this method, yeast cells are transformed with a plasmid expressing the test substrate and the control protein DHFR from identical promoters. The transcribed products carry tetracycline-binding RNA aptamers which inhibit protein expression at the level of translation following the addition of tetracycline to the media. Thus, the fate of the existing protein can be monitored. Importantly, this method only terminates synthesis of our test proteins, thus the induction of Gid4 and

activation of the GID complex is not impaired. In agreement with our proteomic findings, the Acs1 protein is completely stabilized in both GID2- and GID4-deleted cells (Figure 5B, D), while Aro10 protein is stabilized in GID2-deleted but not GID4-deleted cells (Figure 5C, E), confirming that Acs1 and Aro10 protein levels are regulated by the GID ligase during recovery from carbon starvation. This demonstrates that Acs1 and Aro10 are novel GID substrates in yeast.



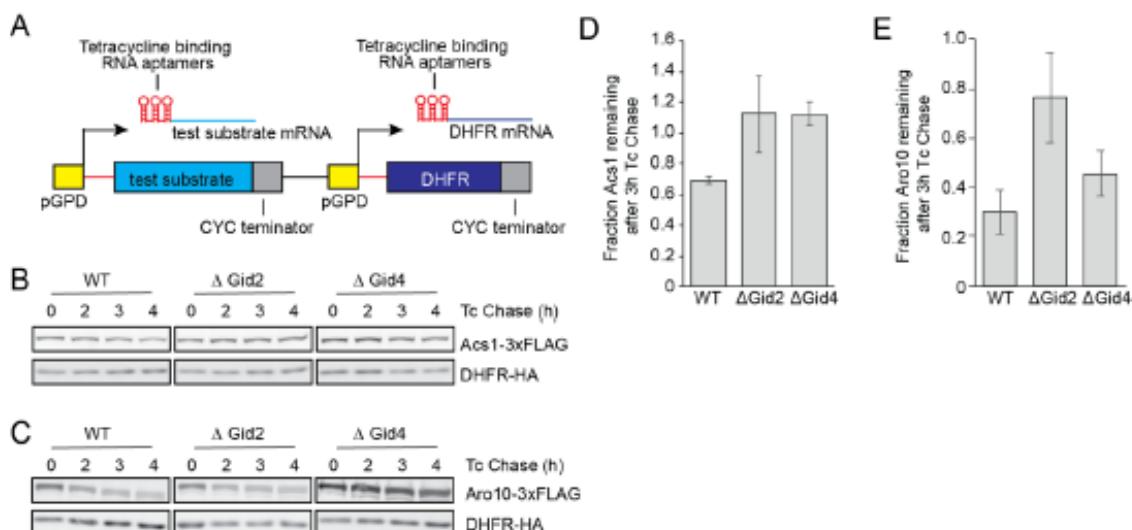
**Figure 3. Global proteome changes of yeast under glucose starvation and recovery**

A. Application of our rapid and robust DIA workflow for yeast proteome profiling under glucose starvation and recovery. B-C. PCA plot of growth conditions along with their biological replicates based on their protein expression (B) and mRNA abundance (C) profiles. D. The GID E3 ubiquitin ligase is a key regulator of the switch from gluconeogenic to glycolytic growth as it degrades the gluconeogenic enzymes Fbp1, Mdh2, Icl1, and Pck1. E. Heat map of z-scored protein abundances ( $\log_2$ ) of the GID complex subunits under glucose starvation and recovery in WT yeast cells. F. PCA plot of proteins during glucose starvation and recovery. Proteins marked in red represent the known GID complex substrates.



**Figure 4: Our rapid and robust DIA-based approach identifies novel GID substrates under recovery after glucose starvation**

A. Number of quantified proteins in WT,  $\Delta$ Gid4 and Gid2<sup>K365A</sup> during glucose starvation and recovery. B-C. Bar graph shows abundances (log2) of Fbp1 (B) and Mdh2 (C) proteins that are normalized to WT Glucose (never starved) condition in WT,  $\Delta$ Gid4 and Gid2<sup>K365A</sup> during glucose starvation and recovery. D. Heat map of z-scored protein abundances (log2) of the proteins which have the criteria of GID substrates.



**Figure 5. *In vivo* validation of novel GID substrates**

A. Schematic of promoter reference technique. B. Representative western blot for Acs1 degradation. C. Representative western blot for Aro10 degradation. D-E. Quantification of Acs1 (D) and Aro10 (E) degradation, based on at least 4 independent replicates. Bars represent mean values, and error bars represent standard deviation

## DISCUSSION

Here we describe a straightforward, streamlined and reproducible systems biology approach for yeast proteome profiling using data-independent acquisition (DIA) to analyze biological pathways much faster and with greater depth. Our minimalistic workflow employs plate-based preparation of digested yeast cell lysate and requires only a few micrograms of yeast as input, and no labeling or special equipment, making it especially amenable for application in non-specialized research groups. Despite its simplicity, our sample preparation workflow combined with the DIA technology provides a sensitive and scalable approach to quantitatively profile hundreds of largely covered yeast proteomes in only a few days. We still anticipate that our approach can be made more robust and streamlined using the latest advances in nanoflow liquid chromatography, especially the new LC system, Evosep, as it significantly reduces the overhead time between sample pick up and start of MS measurement by using pre-formed gradients that already contain the sample [38].

The ability of cells to adapt to stress or changes in environmental conditions relies on extensive proteome remodeling. Understanding these changes provides broad insight into the molecular mechanisms underlying many processes, resolution of heat stress, and adaptation to nutrient availability, and regulation of cell division. Therefore, we applied our DIA-based approach to profile protein level response of yeast to several stress and growth conditions, which also

demonstrated the robustness and specificity of our approach, and provided an in-depth

resource identifying proteomics changes during common stress treatments in yeast, expanding the knowledge on stress mediators regulated at the protein level.

In addition to identifying proteome changes during stress, we have also applied our DIA-based systems biology approach to identify novel proteins that are regulated by the GID E3 ubiquitin ligase. During growth on non-fermentable carbon sources, yeast cells require glucose production through gluconeogenesis. However, in the presence of glucose, this energetically costly pathway is futile and must be deactivated. The GID E3 ubiquitin ligase is a key regulator in this switch [17, 32, 39]. The GID ligase has been shown to be required for the degradation of the gluconeogenic enzymes Fbp1, Mdh2, Icl1, and Pck1 [13]. GID is an eight-protein complex, which is conserved throughout eukaryotes, including mammalian cells, where it is termed the CTLH complex [40]. It is expressed in all growth conditions tested and consists of an inactive anticipatory complex (GID<sup>Ant</sup>), which is activated by binding of the substrate receptor Gid4 [11, 12, 39]. Despite the importance of the GID complex in metabolic regulation, identification of additional substrates has been hindered by the lack of an obvious phenotype, and the variable kinetics of protein degradation. Our DIA workflow identified two novel proteins that are regulated by the GID complex: Acs1 and Aro10, further highlighting the importance and need of such quantitative proteome datasets to provide basis for functional studies.

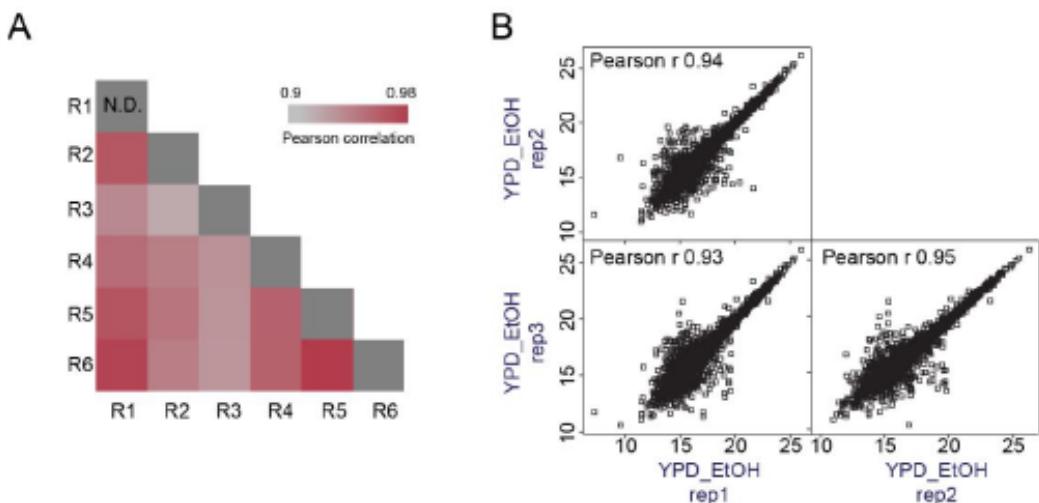
Interestingly, both Acs1 and Aro10, while not considered gluconeogenic enzymes, are important regulators of metabolism and cellular respiration during anaerobic growth. Acs1 encode one of two isoforms of yeast Acetyl-coA synthetase, which catalyzes the formation of acetyl-coA from acetate and coA. Acs1 has a much higher affinity for acetate than its isoform, Acs2, making it more desirable for acetyl-coA production when acetate is limiting, as is the case during growth on non-fermentable carbon sources. During growth in aerobic carbon sources, such as glucose, acetyl-coA is generated from pyruvate following glycolysis and thus, the main energy flux does not require Acs function [41]. Therefore, during growth on glucose, Acs1 expression is suppressed and existing Acs1 protein must be inactivated and degraded. Here, we have shown that the GID E3 ligase is required for Acs1 degradation, through both its RING-containing subunit, Gid2, and its substrate receptor Gid4.

Aro10 is a phenylpyruvate decarboxylase that catalyzes an irreversible step in the Erlich pathway, which breaks down amino acids into fusel acids and fusel alcohols. During growth in non-fermentable carbon sources, the Erlich pathway provides a more energetically favorable means of NADH regeneration. However, following the replenishment of glucose, NADH is regenerated through glycolysis, and thus Aro10 function is no longer required [42, 43]. Here, we show that Aro10 turnover is dependent on the catalytic activity of the GID complex, via its RING-domain, Gid2. Intriguingly, Aro10 is the first protein

identified, where turnover is dependent on Gid2, but not on the substrate receptor Gid4, suggesting an alternate mode of recognition. Indeed, an additional substrate receptor, Gid10, has recently been identified [44], raising the possibility that Aro10 may be the first substrate identified in this novel recognition pathway. Alternatively, Aro10 degradation may be facilitated by a yet uncharacterized mechanism.

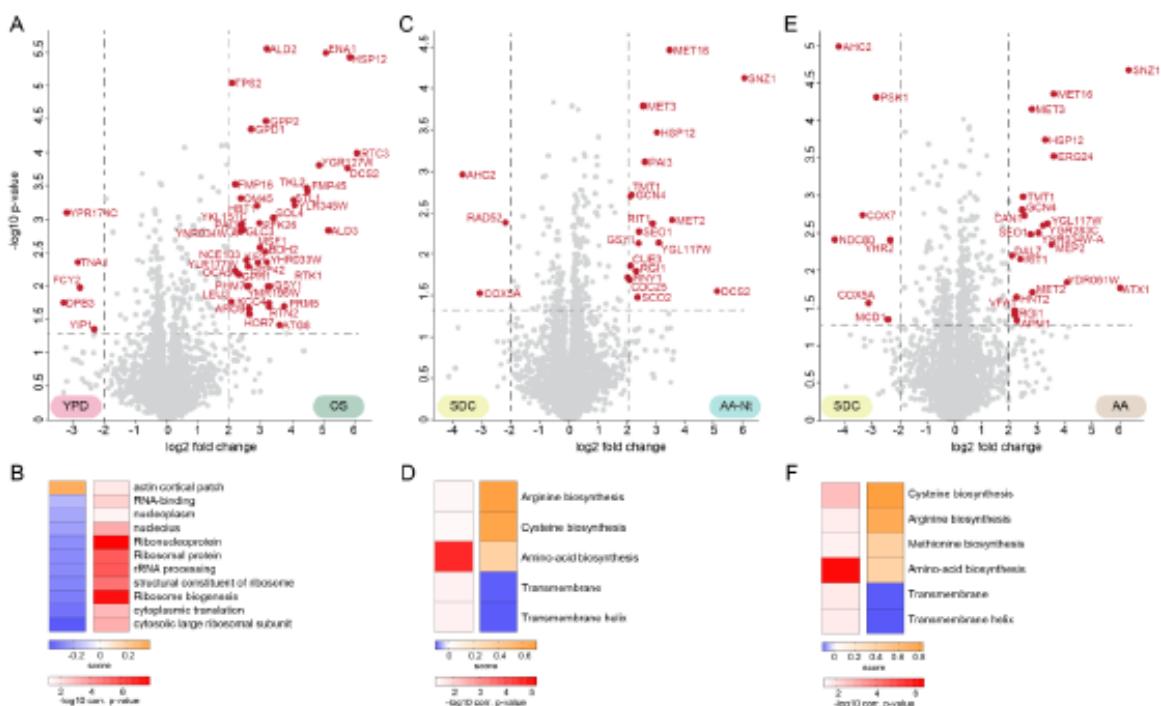
We have identified two novel metabolic steps, entrance into the tricarboxylic acid cycle and the Erlich pathway, which are regulated by the GID E3 ligase during the switch from gluconeogenic to glycolytic growth conditions. Furthermore, we have provided in-depth resource on stress mediators regulated at the protein level, which, we believe, will further allow yeast researchers to probe numerous biological pathways of interest. Thus, the speed and reproducibility of the DIA-based approach presented here allows researchers to probe complex biological pathways and identify novel regulatory mechanisms. Our approach should be applicable to other organisms, generating high-quality quantitative proteome datasets which are required to explain biological processes on a system-wide level. Given that the expressed human proteome (around 15,479, <https://www.proteomicsdb.org>) is around three times larger than the expressed yeast proteome (5,391, [25]), with only three fold increase in sensitivity, it is possible that much of the human proteome will also be analyzable by fast single run approaches.

SUPPLEMENTARY FIGURES



**Supplementary Figure 1. Reproducibility of our workflow**

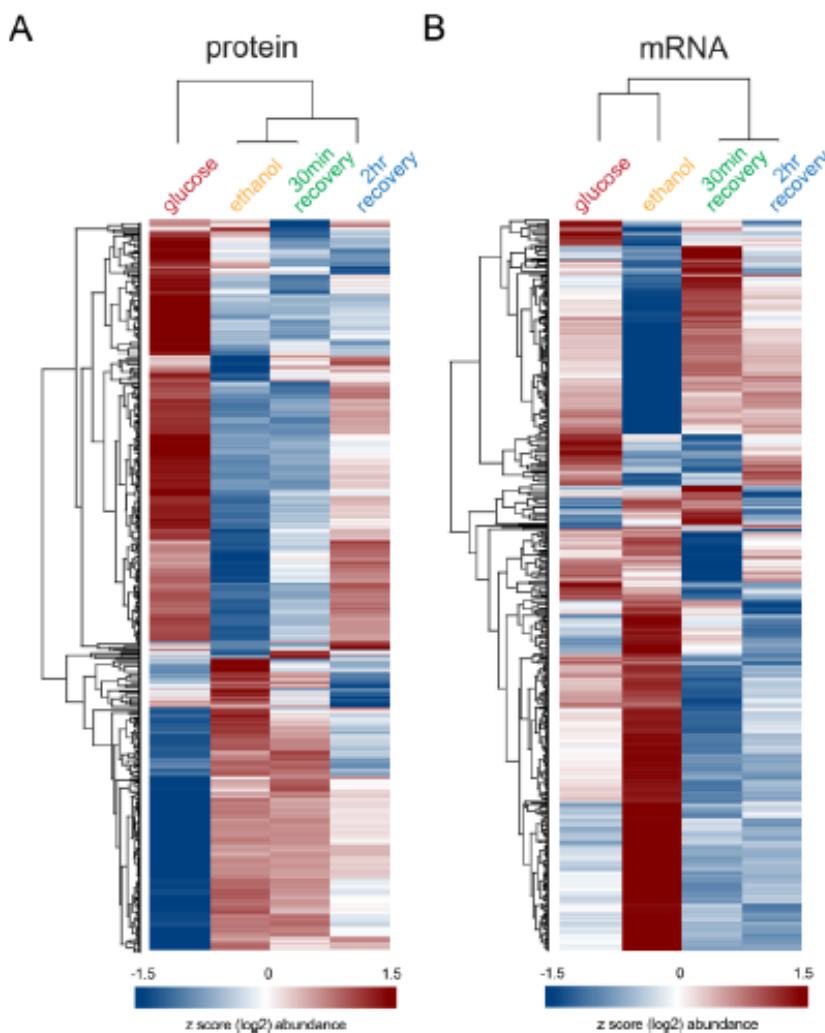
A. Correlation based clustering illustrating the reproducibility between workflow replicates. High (0.98) and lower (0.9) Pearson correlations are denoted in red and grey, respectively. B. The correlation plots illustrating the reproducibility between biological replicates in the yeast stress experiment. Pearson correlation coefficients are shown in the upper left corner.



**Supplementary Figure 2. Mapping changes to the yeast proteome in response to osmotic shock, and amino acid and nitrogen depletion**

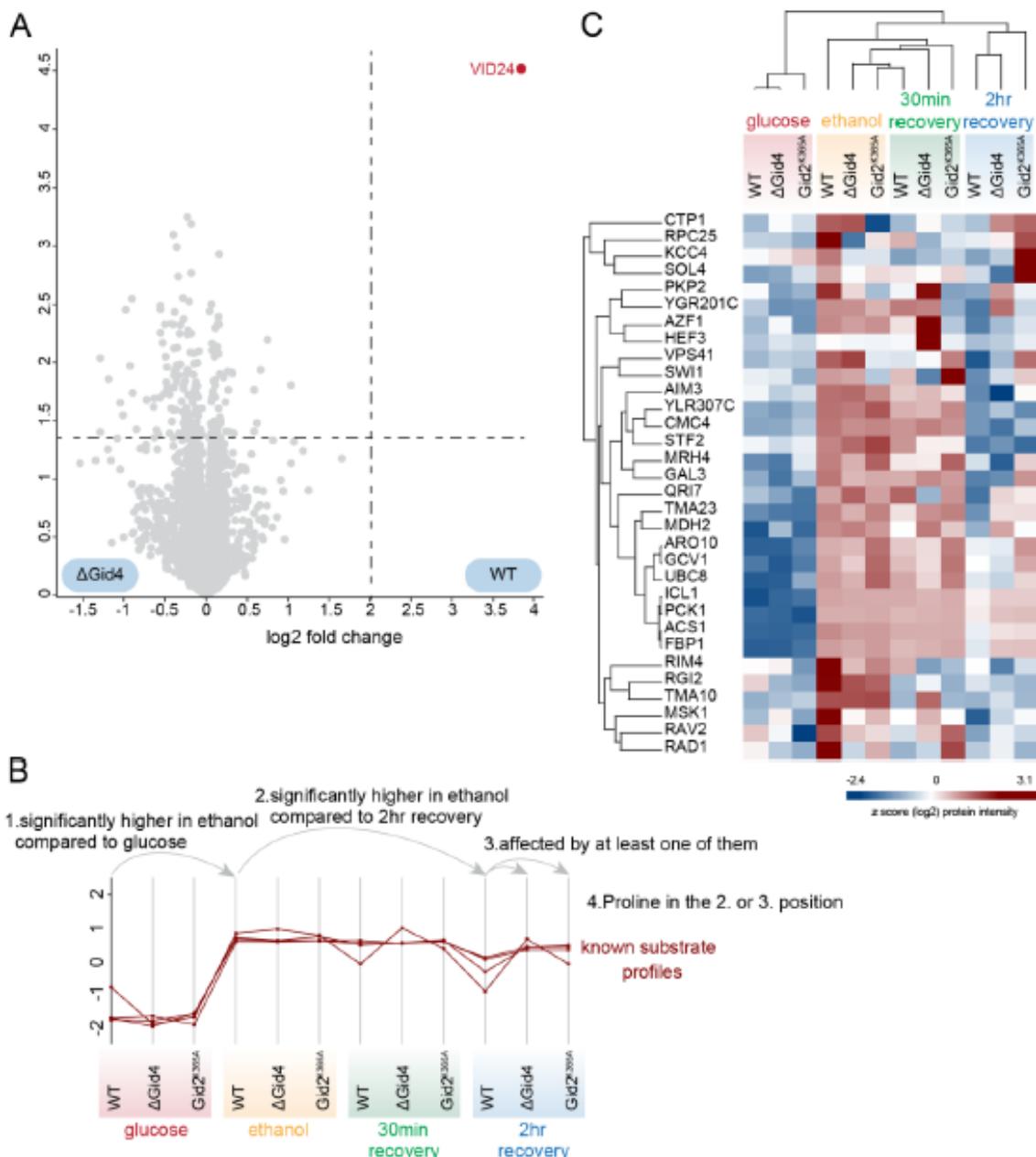
Volcano plot of the (-log<sub>10</sub>) p-values vs. the log<sub>2</sub> protein abundance differences between OS vs. YPD (A), AA-Nt vs. SDC (C), and AA vs. SDC (E). The significant proteins (red dots) are determined based

on p-value  $< 0.05$  and at least 4-fold change on both sides. GO-term enrichment in the OS vs. YPD (B), AA-Nt vs. SDC (D), and AA vs. SDC (F). fold change dimension (1D enrichment, FDR  $< 5\%$ ). Terms with positive enrichment scores are enriched in stress condition over YPD or SDC control and vice versa.



**Supplementary Figure 3. Global proteome and transcriptome changes of yeast under glucose starvation and recovery**

A-B. Heat map of z-scored and differentially regulated proteins (A) and mRNAs (B) ( $\log_2$ ) in wildtype yeast during glucose starvation and recovery.



**Supplementary Figure 4. Identifying GID ligase targets during recovery from carbon starvation**

A. Volcano plot of the (-log10) p-values vs. the log2 mRNA abundance differences between WT vs. Gid4 (substrate receptor) deletion. GID4 was the only significant hit based on p-value < 0.05 and at least 4-fold change on both sides. B. The criteria of the known GID substrates based on their protein profiles: (1) the protein is expressed significantly higher in ethanol compared to glucose and (2) 2hr recovery, (3) decreased abundance of the protein during recovery is dependent on the GID complex, and (4) having Proline (N-degron) in position 2 or 3. C. Heat map of z-scored potential GID substrates which meet the first three conditions in panel B during glucose starvation and recovery.

## METHODS

### Yeast Strains and Growth Conditions

All yeast strains used in this study are derivatives of BY4741 and are listed in Table 1. For rich conditions, yeast cultures were grown in YPD (1% yeast extract, 2% peptone, 2% glucose) or SD (0.67% yeast nitrogen base without amino acids, 2% glucose, containing the appropriate amino acid supplements) media. Unless otherwise specified, yeast were grown at 30°C. For heat shock conditions, yeast cultures were grown in YPD to an OD<sub>600</sub> of 1.0 and then shifted to the indicated temperature for 1 hour. For osmotic shock conditions, yeast cells were grown in YPD to an OD<sub>600</sub> of 1.0, pelleted at 3,000rpm for 3min, and resuspended at an OD<sub>600</sub> of 1.0 in pre-warmed YPD + 0.5M NaCl. For glucose starvation, yeast cells were grown in YPD to an OD<sub>600</sub> of 1.0-2.0, pelleted at 3,000rpm for 3 min, washed once with YPE (1% yeast extract, 2% peptone, 2% ethanol), resuspended in pre-warmed YPE at an OD<sub>600</sub> of 1.0, and grown at 30°C for 19 hours. For glucose recovery, yeast cells were pelleted after 19 hours growth in YPE, and resuspended to an OD<sub>600</sub> of 1.0 in YPD, and allowed to grow at 30°C for 30 minutes or 2 hours.

For proteomics analysis, 50 ODs of cells were pelleted at 3,000rpm for 3 minutes, flash frozen in liquid nitrogen, and stored at -80°C until lysis. For transcriptomes analysis, 10 ODs of yeast were pelleted, flash frozen in liquid nitrogen, and stored at -80°C. Samples were shipped to Novogene (Hong Kong) for RNAseq analysis.

### Protein Degradation Assays (promoter reference technique)

Protein degradation assays using the promoter reference technique were done as previously described [37]. Cells were transformed with plasmid expressing a test substrate and DHFR from identical promoters containing tetracycline-repressible RNA-binding elements. Yeast cells were then grown in SC media lacking histidine, starved in SE (2%

ethanol) media lacking histidine for 19 hours, and then allowed to recover for the indicated times in SC media lacking histidine. At each time point, 1.0 ODs of yeast cells were pelleted, flash frozen in liquid nitrogen, and stored at -80°C until lysis.

For lysis, yeast cells were resuspended in 0.8mL of 0.2M NaOH, followed by incubation on ice for 20 min, and then pelleted at 11,200xg for 1 min. The supernatant was removed and the pellet resuspended in 50uL HU buffer, and incubated at 70°C for 10 min. The lysate was precleared by centrifugation at 11,200xg for 5 min, and then loaded onto a 12% SDS-PAGE gel. Protein samples were transferred to nitrocellulose membrane, and then visualized by western blot using  $\alpha$ FLAG (Sigma, F1804) and  $\alpha$ HA (Sigma H6908) primary antibodies, and Dylight 633 goat anti-Mouse (Invitrogen 35512) and Dylight 488 goat anti-rabbit (Invitrogen 35552) secondary antibodies. Membranes were imaged on a typhoon scanner (Amersham). Bands were quantified with ImageStudio software (Licor).

### Sample preparation for MS analysis

SDC lysis buffer (1 % SDC and 100 mM Tris pH 8.4) were added to the frozen cell pellets to achieve a protein concentration of ~2–3 mg per mL. Lysates were immediately heat-treated for 5 min at 95 °C to facilitate lysis and to inactivate endogenous proteases and transferred to a 96-well plate. Lysates were next homogenized with sonication. Protein concentrations were estimated by tryptophan assay [45] and then all samples were diluted to equal protein concentrations in a 96-well plate. To reduce and alkylate proteins, samples were incubated for 5 min at 45°C with CAA and TCEP, final concentrations of 10 mM and 40 mM, respectively. Samples were digested overnight at 37°C using trypsin (1:100 w/w, Sigma-Aldrich) and LysC (1/100 w/w, Wako). The following day, peptide material was desalted using SDB-RPS StageTips (Empore) [45]. Briefly, samples were diluted with 1% TFA in isopropanol to a final volume of 200  $\mu$ l and

loaded onto StageTips, subsequently washed with 200  $\mu$ l of 1% TFA in isopropanol and 200  $\mu$ l 0.2% TFA/ 2% ACN. Peptides were eluted with 80  $\mu$ l of 1.25% Ammonium hydroxide (NH4OH)/ 80% ACN and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus). Samples were resuspended in buffer A\* (0.2% TFA/ 2% ACN) prior to LC-MS/MS analysis. Peptide concentrations were measured optically at 280nm (Nanodrop 2000, Thermo Scientific) and subsequently equalized using buffer A\*. 300ng peptide was subjected to LC-MS/MS analysis.

To generate the spectral library for DIA measurements cells were lysed in SDC buffer, followed by sonication, protein quantification, reduction, and alkylation and desalting using SDB-RPS cartridges (see above). Around 8 or 30  $\mu$ g of peptides were fractionated into 8 or 24 fractions, respectively, by high pH reversed-phase chromatography as described earlier [46]. Fractions were concatenated automatically by shifting the collection tube during the gradient and subsequently dried in a vacuum centrifuge and resuspended in buffer A\*.

#### LC-MS/MS Measurements

Samples were loaded onto a 20 cm reversed phase column (75  $\mu$ m inner diameter, packed in house with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin [Dr. Maisch GmbH]). The column temperature was maintained at 60°C using a homemade column oven. A binary buffer system, consisting of buffer A (0.1% formic acid (FA)) and buffer B (80% ACN plus 0.1% FA), was used for peptide separation, at a flow rate of 450 nL/min. An EASY-nLC 1200 system (Thermo Fisher Scientific), directly coupled online with the mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) via a nano-electrospray source, was employed for nano-flow liquid chromatography. We used a gradient starting at 5% buffer B, increased to 35% in 18.5 minutes, 95% in a min and stayed at 95% for 3.5 minutes. The mass spectrometer was operated in Top10 data-dependent mode (DDA) with a full scan range of 300-1650 m/z

at 60,000 resolution with an automatic gain control (AGC) target of 3e6 and a maximum fill time of 20ms. Precursor ions were isolated with a width of 1.4 m/z and fragmented by higher-energy collisional dissociation (HCD) (NCE 27%). Fragment scans were performed at a resolution of 15,000, an AGC of 1e5 and a maximum injection time of 60 ms. Dynamic exclusion was enabled and set to 30 s. For DIA measurements full MS resolution was set to 120,000 with a full scan range of 300-1650 m/z, a maximum fill time of 60 ms and an automatic gain control (AGC) target of 3e6. One full scan was followed by 12 windows with a resolution of 30,000 in profile mode. Precursor ions were fragmented by stepped higher-energy collisional dissociation (HCD) (NCE 25.5, 27.30%).

**Data processing and bioinformatics analysis**  
Spectronaut version 13 (Biognosys) was used to generate the spectral libraries from DDA runs by combining files of respective fractionations using the yeast FASTA file (Swissprot, 2018). For the generation of the proteome library default settings were left unchanged. DIA files were analyzed using the proteome library with default settings and enabled cross run normalization. The Perseus software package versions 1.6.0.7 and 1.6.0.9 and GraphPad Prism version 7.03 were used for the data analysis [47]. Protein intensities and mRNA abundances were log2-transformed for further analysis. The data sets were filtered to make sure that identified proteins and mRNAs showed expression or intensity in all biological triplicates of at least one condition and the missing values were subsequently replaced by random numbers that were drawn from a normal distribution (width=0.3 and down shift=1.8). PCA analysis of stress and growth conditions and biological replicates was performed as previously described in [48]. Multi-sample test (ANOVA) for determining if any of the means of stress and growth conditions were significantly different from each other was applied to both mRNA and protein data sets. For truncation, we used permutation-based FDR

which was set to 0.05 in conjunction with an S0-parameter of 0.1. For hierarchical clustering of significant genes and proteins, median protein or transcript abundances of biological replicates were z-scored and clustered using Euclidean as a distance measure for row clustering. GO annotations were matched to the proteome data

based on Uniprot identifiers. Annotation term enrichment was performed with either Fisher exact test or 1D tool in Perseus. Annotation terms were filtered for terms with an FDR of 5% after Benjamini-Hochberg correction.

**Table 1:** Strains used in this study

Strain	Genotype	Reference
<b>BY4741</b>	<i>MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0</i>	Euroscarf
<b>CRLY12</b>	<i>BY4741 GID4::KANMX</i>	This study
<b>CRLY30</b>	<i>BY4741 GID2::KANMX</i>	This study
<b>CRLY131</b>	<i>BY4741 gid2::3xFLAG-GID2K365A</i>	Qiao, 2020

**Table 2:** Plasmids used in this study

Plasmid		Reference
<b>CRLP47</b>	pRS313-P <sub>TDH3</sub> (modified)-Aro10 <sub>3xFlag</sub> -CYC-p <sub>TDH3</sub> (modified)- <sub>Flag</sub> DHFR <sub>ha</sub> -CYC	This study
<b>CRLP48</b>	pRS313-P <sub>TDH3</sub> (modified)-Acs1 <sub>3xFlag</sub> -CYC-p <sub>TDH3</sub> (modified)- <sub>Flag</sub> DHFR <sub>ha</sub> -CYC	This study

## ACKNOWLEDGEMENT

This work was supported by the Max-Planck Society for the Advancement of Science. We thank Arno Alpi, Florian Meier, Igor Paron, Christian Deiml, Johannes B. Mueller, Viola Beier and all the members of the departments of Proteomics and Signal Transduction and Molecular Machines and Signaling at Max-Planck-Institute of Biochemistry for their assistances and helpful discussions.

## AUTHOR CONTRIBUTIONS

CRL and OK performed experiments and analyzed the data. ACM optimized the DIA method with short LC gradient. BAS and MM coordinated and supervised. OK, CRL, MM and BSS designed the study and wrote the paper.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## REFERENCES

1. Sui, X., et al., *Widespread remodeling of proteome solubility in response to different protein homeostasis stresses*. 2020. **117**(5): p. 2422-2431.
2. Hanna, J., et al., *Protein degradation and the pathologic basis of disease*. 2019. **189**(1): p. 94-103.
3. Giaever, G., et al., *Functional profiling of the *Saccharomyces cerevisiae* genome*. 2002. **418**(6896): p. 387-391.
4. Steinmetz, L.M., et al., *Systematic screen for human disease genes in yeast*. 2002. **31**(4): p. 400-404.
5. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast*. 2003. **425**(6959): p. 737-741.
6. Jones, G.M., et al., *A systematic library for comprehensive overexpression screens in *Saccharomyces cerevisiae**. 2008. **5**(3): p. 239-241.
7. Hebert, A.S., et al., *The one hour yeast proteome*. 2014. **13**(1): p. 339-347.
8. Richards, A.L., et al., *One-hour proteome analysis in yeast*. 2015. **10**(5): p. 701-714.

9. Nagaraj, N., et al., *System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap*. 2012. **11**(3).
10. Varshavsky, A.J.A.r.o.b., *The ubiquitin system, an immense realm*. 2012. **81**: p. 167-176.
11. Qiao, S., et al., *Interconversion between anticipatory and active GID E3 ubiquitin ligase conformations via metabolically driven substrate receptor assembly*. 2020. **77**(1): p. 150-163. e9.
12. Chen, S.-J., et al., *An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes*. 2017. **355**(6323).
13. Hämerle, M., et al., *Proteins of newly isolated mutants and the amino-terminal proline are essential for ubiquitin-proteasome-catalyzed catabolite degradation of fructose-1, 6-bisphosphatase of *Saccharomyces cerevisiae**. 1998. **273**(39): p. 25000-25005.
14. Kobayashi, A., et al., *Oxidative and electrophilic stresses activate Nrf2 through inhibition of ubiquitination activity of Keap1*. 2006. **26**(1): p. 221-229.
15. Kaiser, P., et al., *The yeast ubiquitin ligase SCF Met30: connecting environmental and intracellular conditions to cell division*. 2006. **1**(1): p. 1-8.
16. Zhang, D.D., M.J.M. Hannink, and c. biology, *Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress*. 2003. **23**(22): p. 8137-8151.
17. Regelmann, J., et al., *Catabolite degradation of fructose-1, 6-bisphosphatase in the yeast *Saccharomyces cerevisiae*: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways*. 2003. **14**(4): p. 1652-1663.
18. Petranovic, D., et al., *Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism*. 2010. **10**(8): p. 1046-1059.
19. Brodsky, J.L. and W.R.J.C.o.i.c.b. Skach, *Protein folding and quality control in the endoplasmic reticulum: Recent lessons from yeast and mammalian cell systems*. 2011. **23**(4): p. 464-475.
20. Widmann, C., et al., *Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human*. 1999. **79**(1): p. 143-180.
21. De Virgilio, C., R.J.T.i.j.o.b. Loewith, and c. biology, *The TOR signalling network from yeast to man*. 2006. **38**(9): p. 1476-1481.
22. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial*. 2018. **14**(8): p. e8126.
23. Bekker-Jensen, D.B., et al., *Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries*. 2020. **11**(1): p. 1-12.
24. Winter, S.V., et al., *Urinary proteome profiling for stratifying patients with familial Parkinsons disease*. 2020.
25. Ho, B., A. Baryshnikova, and G.W.J.C.s. Brown, *Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome*. 2018. **6**(2): p. 192-205. e3.
26. Serrano, R., et al., *Copper and iron are the limiting factors for growth of the yeast *Saccharomyces cerevisiae* in an alkaline environment*. 2004. **279**(19): p. 19698-19704.
27. Bereiter-Hahn, J., *Behavior of mitochondria in the living cell*, in *International review of cytology*. 1990, Elsevier. p. 1-63.
28. Gomez-Pastor, R., E.T. Burchfiel, and D.J.J.N.R.M.C.B. Thiele, *Regulation of heat shock transcription factors and their roles in physiology and disease*. 2018. **19**(1): p. 4.
29. Vabulas, R.M., et al., *Protein folding in the cytoplasm and the heat shock response*. 2010. **2**(12): p. a004390.
30. Carlson, M.J.C.o.i.m., *Glucose repression in yeast*. 1999. **2**(2): p. 202-207.
31. Schüller, H.-J.J.C.g., *Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae**. 2003. **43**(3): p. 139-160.
32. Gancedo, J.M.J.M. and m.b. reviews, *Yeast carbon catabolite repression*. 1998. **62**(2): p. 334-361.
33. Tagkopoulos, I., Y.-C. Liu, and S.J.s. Tavazoie, *Predictive behavior within microbial genetic networks*. 2008. **320**(5881): p. 1313-1317.
34. Barnett, J.A. and K.D.J.Y. Entian, *A history of research on yeasts 9: regulation of sugar metabolism 1*. 2005. **22**(11): p. 835-894.

35. Dong, C., et al., *Molecular basis of GID4-mediated recognition of degrons for the Pro/N-end rule pathway*. 2018. 14(5): p. 466-473.
36. Slattery, M.G., D. Liko, and W.J.E.C. Heideman, *The function and properties of the *Azfl* transcriptional regulator change with growth conditions in *Saccharomyces cerevisiae**. 2006. 5(2): p. 313-320.
37. Oh, J.-H., S.-J. Chen, and A.J.J.o.B.C. Varshavsky, *A reference-based protein degradation assay without global translation inhibitors*. 2017. 292(52): p. 21457-21465.
38. Bache, N., et al., *A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics*. 2018. 17(11): p. 2284-2296.
39. Santt, O., et al., *The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism*. 2008. 19(8): p. 3323-3333.
40. Francis, O., F. Han, and J.C.J.P.o. Adams, *Molecular phylogeny of a RING E3 ubiquitin ligase, conserved in eukaryotic cells and dominated by homologous components, the muskelin/RanBPM/CTLH complex*. 2013. 8(10): p. e75217.
41. Van Den Berg, M.A., et al., *The two acetyl-coenzyme A synthetases of *Saccharomyces cerevisiae* differ with respect to kinetic properties and transcriptional regulation*. 1996. 271(46): p. 28953-28959.
42. Hazelwood, L.A., et al., *The Ehrlich pathway for fusel alcohol production: a century of research on *Saccharomyces cerevisiae* metabolism*. 2008. 74(8): p. 2259-2266.
43. Pires, E.J., et al., *Yeast: the soul of beer's aroma—a review of flavour-active esters and higher alcohols produced by the brewing yeast*. 2014. 98(5): p. 1937-1949.
44. Melnykov, A., S.-J. Chen, and A.J.P.o.t.N.A.o.S. Varshavsky, *Gid10 as an alternative N-recognin of the Pro/N-degron pathway*. 2019. 116(32): p. 15914-15923.
45. Kulak, N.A., et al., *Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells*. 2014. 11(3): p. 319.
46. Kulak, N.A., et al., *Loss-less nano-fractionator for high sensitivity, high coverage proteomics*. 2017. 16(4): p. 694-705.
47. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data*. 2016. 13(9): p. 731.
48. Deeb, S.J., et al., *Machine learning-based classification of diffuse large B-cell lymphoma patients by their protein expression profiles*. 2015. 14(11): p. 2947-2960.

### 3.2.4. Interconversion between Anticipatory and Active GID E3 Ubiquitin Ligase Conformations via Metabolically Driven Substrate Receptor Assembly

Shuai Qiao<sup>a</sup>, Christine R. Langlois<sup>a\*</sup>, Jakub Chrstowicz<sup>a\*</sup>, Dawafuti Sherpa<sup>a\*</sup>, Ozge Karayel<sup>b</sup>, Fynn M. Hansen<sup>b</sup>, Viola Beier<sup>a</sup>, Susanne von Gronau<sup>a</sup>, Daniel Bollschweiler<sup>c</sup>, Tillmann Schäfer<sup>c</sup>, Arno F. Alpi<sup>a</sup>, Matthias Mann<sup>b</sup>, J. Rajan Prabu<sup>a</sup>, Brenda A. Schulman<sup>a</sup>

<sup>a</sup>Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, 82152, Germany

<sup>b</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, 82152, Germany

<sup>c</sup>Cryo-EM Facility, Max Planck Institute of Biochemistry, Martinsried, 82152, Germany

Published in *Molecular Cell* (2020)

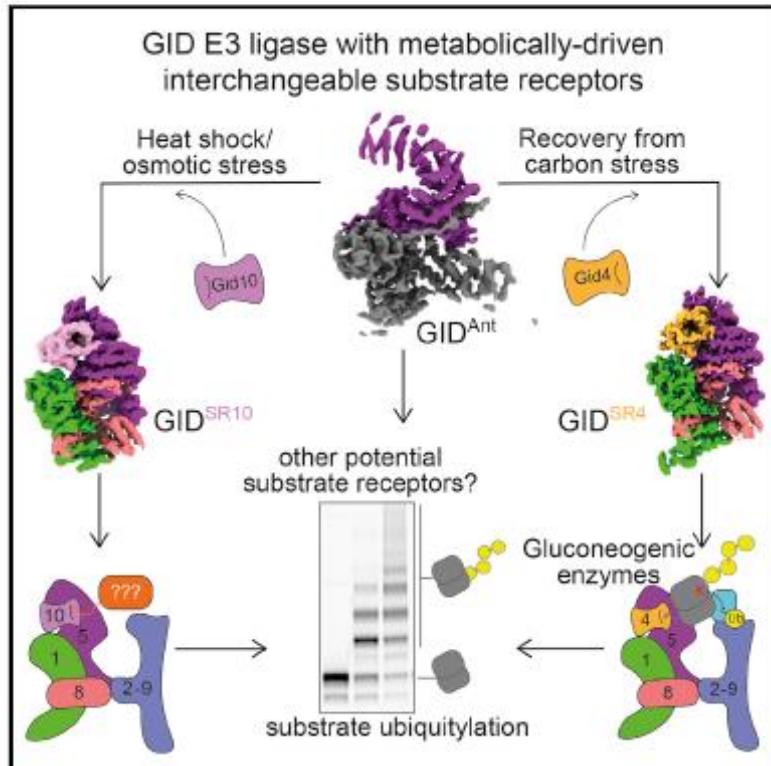
One of the major roles of the ubiquitin system is to respond to environmental changes. A very good example for this is the response to carbon catabolite repression which induces gluconeogenesis in the budding yeast *Saccharomyces cerevisiae* as described a long time ago [165]. The highly conserved Gid E3 ligase complex plays an essential role in glucose-induced degradation (Gid) of Fbp1 –a gluconeogenesis enzyme- in yeast [166-169]. In this study, researchers from the Schulman group aimed to decipher the structure and molecular mechanisms of the GID complex. Their findings revealed mechanisms of stress anticipation and that the Gid is regulated through assembly with interchangeable substrate receptors induced by distinct environmental perturbations.

To assist their study, I performed quantitative proteomic analysis of tagged Gid8 immunoprecipitates to identify the components of the Gid E3 ligase complex. We found that Gid7 has substantially lower abundance relative to other Gid subunits, and Gid7 deletion did not appreciably affect GID assembly. Additionally, we mapped lysine sites that are ubiquitylated *in vitro* by mass spectrometry of Mdh2, a well-known substrate of the complex. Ensuing structural models showed that the identified sites were required respectively for Mdh2 to engage Gid4 simultaneously and to approach the Gid2 RING activated Ubc8~Ub intermediate.

# Molecular Cell

## Interconversion between Anticipatory and Active GID E3 Ubiquitin Ligase Conformations via Metabolically Driven Substrate Receptor Assembly

### Graphical Abstract



### Authors

Shuai Qiao, Christine R. Langlois, Jakub Chrstowicz, ..., Matthias Mann, J. Rajan Prabu, Brenda A. Schulman

### Correspondence

schulman@biochem.mpg.de

### In Brief

Stunning structural intricacies of the yeast GID system, now revealed more than 25 years since the discovery of glucose-induced degradation and of a protein's N terminus serving as a degron, provide a blueprint for a family of multisubunit E3s regulated through assembly with interchangeable substrate receptors induced by distinct environmental perturbations.

### Highlights

- Genetics, biochemistry, proteomics, and cryo-EM define GID E3 ligase regulation
- Carbon stress induces assembly of an inactive anticipatory GID<sup>Ant</sup> complex
- Environmental perturbations trigger substrate receptor assembly into active GID E3s
- Structural model of N-degron substrate ubiquitylation by multisubunit RING-RING E3

# Interconversion between Anticipatory and Active GID E3 Ubiquitin Ligase Conformations via Metabolically Driven Substrate Receptor Assembly

Shuai Qiao,<sup>1</sup> Christine R. Langlois,<sup>1,4</sup> Jakub Chrstowicz,<sup>1,4</sup> Dawafuti Sherpa,<sup>1,4</sup> Ozge Karayel,<sup>2</sup> Fynn M. Hansen,<sup>2</sup> Viola Beier,<sup>1</sup> Susanne von Gronau,<sup>1</sup> Daniel Bölschweiler,<sup>3</sup> Tillman Schäfer,<sup>3</sup> Amo F. Alpi,<sup>1</sup> Matthias Mann,<sup>2</sup> J. Rajan Prabu,<sup>1</sup> and Brenda A. Schulman<sup>1,5,\*</sup>

<sup>1</sup>Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>2</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>3</sup>Cryo-EM Facility, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

<sup>4</sup>These authors contributed equally

<sup>5</sup>Lead Contact

\*Correspondence: schulman@biochem.mpg.de

<https://doi.org/10.1016/j.molcel.2019.10.009>

## SUMMARY

Cells respond to environmental changes by toggling metabolic pathways, preparing for homeostasis, and anticipating future stresses. For example, in *Saccharomyces cerevisiae*, carbon stress-induced gluconeogenesis is terminated upon glucose availability, a process that involves the multiprotein E3 ligase GID<sup>SR4</sup> recruiting N termini and catalyzing ubiquitylation of gluconeogenic enzymes. Here, genetics, biochemistry, and cryoelectron microscopy define molecular underpinnings of glucose-induced degradation. Unexpectedly, carbon stress induces an inactive anticipatory complex (GID<sup>Ant</sup>), which awaits a glucose-induced substrate receptor to form the active GID<sup>SR4</sup>. Meanwhile, other environmental perturbations elicit production of an alternative substrate receptor assembling into a related E3 ligase complex. The intricate structure of GID<sup>Ant</sup> enables anticipating and ultimately binding various N-degron-targeting (i.e., “N-end rule”) substrate receptors, while the GID<sup>SR4</sup> E3 forms a clamp-like structure juxtaposing substrate lysines with the ubiquitylation active site. The data reveal evolutionarily conserved GID complexes as a family of multisubunit E3 ubiquitin ligases responsive to extracellular stimuli.

## INTRODUCTION

Eukaryotes use a plethora of mechanisms to cope with environmental perturbations. Much of our understanding of these processes comes from studies on the yeast *S. cerevisiae*, for example, chaperone induction to enable protein folding during heat stress, kinase activation to control osmolarity, and glycolytic or gluconeogenic enzyme expression to switch metabolism. An emerging concept is that cells also have “anticipatory” programs whereby an altered growth condition triggers not only

pathways rescuing cells from immediate dangers but also expression of proteins that could be required for subsequent shifts in conditions (Mitchell et al., 2009; Tagkopoulos et al., 2008). If the anticipated perturbation does occur, cells can more rapidly adapt to the new environment through expression of yet other genes. For example, chaperones are induced at temperatures below those causing global misfolding, thereby increasing proteostasis capacity should a more severe later stress further compromise cellular protein folding (Klaips et al., 2014). Determination of protein fate by ubiquitylation is another major mechanism orchestrating homeostasis (Ciechanover, 2012; Varshavsky, 2012). Ubiquitylation depends on cellular signals directing E3 ligases to particular targets. Yet our understanding of E3-dependent responses to environmental changes remains rudimentary. The questions of if and how E3 ligase structures play roles in cellular anticipation and responses to perturbations in the extracellular milieu are largely unexplored.

The ubiquitin (Ub) system has long been known to regulate yeast carbon catabolite repression (Zaman et al., 2008). Although yeast growing on non-fermentable carbon sources (e.g., ethanol) require gluconeogenic production of glucose, this energetically costly pathway is futile and therefore terminated when sugars are available. This not only involves multifaceted transcriptional responses but also glucose-induced degradation (Gid) of gluconeogenesis enzymes such as fructose-1,6-bisphosphatase (Fbp1), malate dehydrogenase (Mdh2), and isocitrate lyase (Icl1) (Chiang and Schekman, 1991; Chiang and Chiang, 1998; Gancedo, 1998; Hoffman and Chiang, 1996; Schork et al., 1994a, 1994b). The original Gid gene products defined by genetics and biochemistry include the E2 Ub-conjugating enzyme Gid3 (hereafter referred to as Ubc8), the deubiquitylating enzyme Ubp14 (Gid6), and a Gid complex loosely defined by physical interactions of Gid1, Gid2, Gid4, Gid5, Gid7, Gid8, and Gid9 (Braun et al., 2011; Francis et al., 2013; Menssen et al., 2012; Regelmann et al., 2003; Santt et al., 2008; Schüle et al., 2000). Although the Gid2 and Gid9 subunits each harbor RING (really interesting new gene) domains, the other subunits lack sequences associated with ubiquitylation (Braun et al., 2011; Francis et al., 2013; Menssen et al., 2012; Schüle et al., 2000). A recent breakthrough in our understanding

of the GID E3 came from its assignment as an N-degron-targeting E3 (Chen et al., 2017).

N-degron (formerly termed “N-end rule”) and C-degron (collectively referred to as “terminal degron”) E3s recognize substrate N or C termini and regulate vast biology (Varshavsky, 2019). Nonetheless, beyond knowledge of pathways creating, exposing, or cloaking substrate N- or C-degrons, and structures showing their recognition by E3 ligases, there is limited structural information explaining regulation of terminal degron E3s (Brower et al., 2013; Choi et al., 2010; Dong et al., 2018; Hu et al., 2005; Koren et al., 2018; Lin et al., 2018; Matta-Camacho et al., 2010; Rao et al., 2001; Rusnac et al., 2018; Shemorry et al., 2013; Szoradi et al., 2018; Timms et al., 2019; Varshavsky, 2011; Wang et al., 2008). Moreover, Fbp1, Mdh2, and Icl1 each harbor natively exposed GID E3-targeting N-terminal prolines essential for their degradation (Hämmerle et al., 1998). The question of how their ubiquitylation could be regulated was answered by discovery that glucose availability determines expression of Gid4 (Menssen et al., 2018; Santt et al., 2008), which serves as a substrate receptor for the GID E3 by binding to an N-terminal proline (Chen et al., 2017; Dong et al., 2018). A crystal structure of peptide-bound human Gid4 showed the basis for N-terminal proline recognition (Dong et al., 2018). Although the mammalian GID E3 does not appear to regulate gluconeogenic enzymes (Lampert et al., 2018), and its N-degron substrates remain to be identified, numerous studies suggest that it may also act as a central component in cell fate determination essential for some developmental pathways (Han et al., 2016; Javan et al., 2018; Liu and Pfirrmann, 2019; Nguyen et al., 2017; Pfirrmann et al., 2015; Soni et al., 2006).

Here we reveal molecular mechanisms underlying assembly and activity of the largely mysterious GID E3 and provide general insight into ubiquitylation by the large cohort of terminal degron E3s and by those catalyzing ubiquitylation via heterodimeric RING-RING domains. Unexpectedly, our results also reveal mechanisms of stress anticipation and resolution through assembly of an E3 ligase, and that GID is not a singular complex. GID comprises a family of multisubunit E3s regulated through assembly with interchangeable N-degron-binding substrate receptors induced by distinct environmental perturbations.

## RESULTS

### Carbon-Source-Dependent Anticipatory versus Activated GID E3 Ligase Assemblies

As a prelude for developing and validating a recombinant system, we investigated properties of endogenous Gid proteins. The potential of Gid proteins to stably coassemble with one another *in vivo* was examined using a suite of yeast strains, each harboring a Gid gene tagged at its endogenous locus and validated for activity. Yeasts were grown in various carbon sources known to determine GID E3 ligase activity (Oh et al., 2017; Regelmann et al., 2003), and lysates were subjected to sucrose gradient fractionation (Figure 1A).

Migration of Gid subunits, and their relative levels in the four conditions (Figures 1A and S1), led to three major conclusions (Figure 1B). First, in carbon recovery conditions that prompt

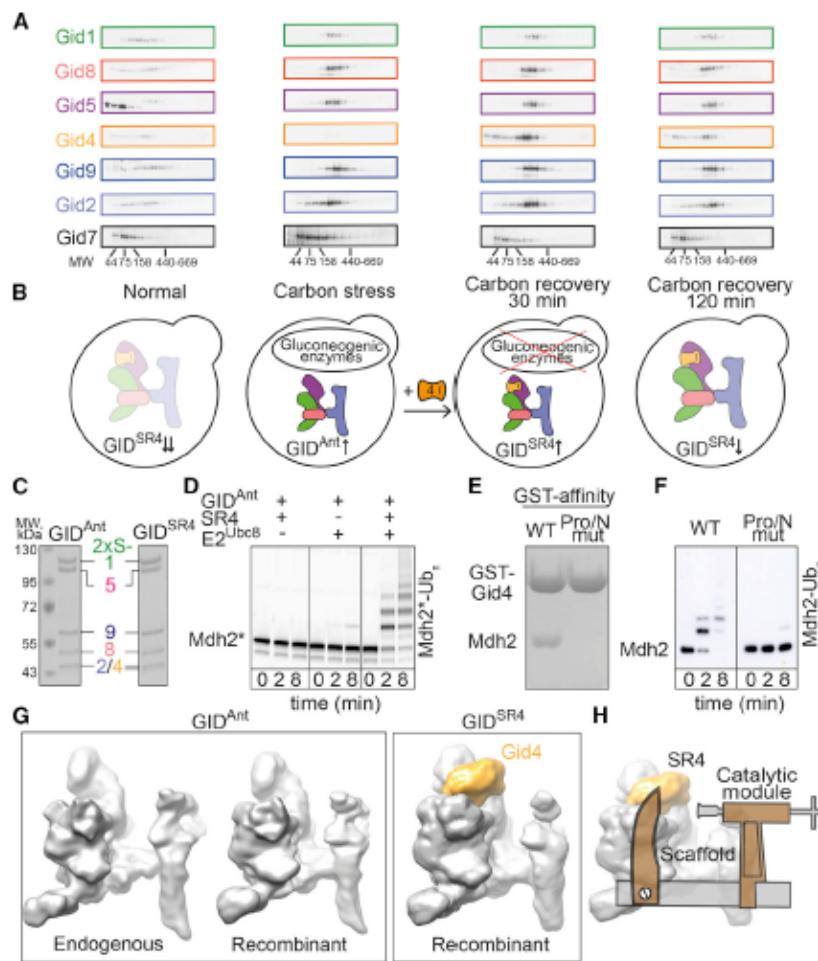
degradation of gluconeogenesis enzymes, Gid1, Gid8, Gid5, Gid4, Gid9, and Gid2 comigrate, suggesting that these subunits form a minimal stable E3 ligase including the substrate receptor Gid4, which we term GID<sup>SR4</sup>. Second, as expected, the relative level of Gid4 is highest during carbon recovery, in agreement with Gid4 expression being the glucose-regulated switch determining E3 activity (Menssen et al., 2018; Santt et al., 2008). Finally and unexpectedly, during carbon stress, the levels of all GID<sup>SR4</sup> subunits except Gid4 increase, and they comigrate in the density gradients. This suggests that during energetically expensive growth on a non-fermentable carbon source, a seemingly unnecessary, inactive complex containing most Gid proteins is produced. This finding can be rationalized by the emerging concept of “anticipatory” programs preparing for a later shift in conditions. Thus, we term the complex containing Gid1, Gid2, Gid5, Gid8, and Gid9 “GID<sup>Ant</sup>,” surmising that when produced during carbon stress GID<sup>Ant</sup> would be benign toward gluconeogenic enzymes but ready and primed for a potential later shift into glucose-containing media, which in turn would rapidly induce Gid4 expression and assembly of the active GID<sup>SR4</sup> E3 ligase.

The distinct migration of Gid7 was corroborated by its substantially lower abundance relative to other Gid subunits observed by quantitative proteomic analysis of a tagged Gid8 immunoprecipitate. Also, Gid7 deletion did not appreciably affect GID assembly (Figures S1A and S1B). Although there may be settings when GID<sup>SR4</sup> and GID<sup>Ant</sup> further include Gid7 *in vivo*, at this point the role of Gid7 remains unknown. Gid7 may bind to a subset of GID complexes, additional factors may contribute to its binding, the interaction may be transient or low affinity, or Gid7 may play alternative roles in regulation.

To mechanistically define regulation, we generated recombinant GID<sup>Ant</sup> and GID<sup>SR4</sup> complexes (Figure 1C) that reconstitute known GID features. First, together with the E2 enzyme Ubc8, GID<sup>SR4</sup>, but not GID<sup>Ant</sup>, catalyzed robust polyubiquitylation of a recombinant gluconeogenic enzyme substrate, Mdh2 (Figure 1D). Second, in accordance with impaired degradation of a Gid substrate upon overexpressing a dominant-negative Ub K48R mutant *in vivo* (Schork et al., 1995), we found that in the context of otherwise lysineless Ub, only K48 supported substantial polyubiquitylation by our recombinant system (Figure S1D). Third, the N-terminal Pro of Mdh2 was required for its binding to Gid4 and ubiquitylation by GID<sup>SR4</sup> (Figures 1E and 1F).

Three-dimensional (3D) reconstructions at 9 Å resolution obtained by cryoelectron microscopy (cryo-EM) further validated our recombinant system. Comparing the EM maps for recombinant GID<sup>Ant</sup> and that purified from yeast cultured in carbon stress conditions revealed a common overall architecture (Figure 1G). Thus, it appears that the native GID<sup>Ant</sup> purified from yeast, at least in terms of subunits overtly visible by cryo-EM at this resolution, is indeed a complex of Gid1, Gid2, Gid5, Gid8, and Gid9.

Prominent additional density correlating with the presence of the substrate receptor subunit Gid4 was readily visible in the map of recombinant GID<sup>SR4</sup> (Figure 1G). The overall structure of the GID<sup>SR4</sup> E3 resembles a clamp, with Gid4 corresponding to one jaw (Figure 1H). A high-resolution structure showed the substrate receptor linked via a scaffold to a catalytic module as described below.



**Figure 1. Regulation and Reconstitution of Yeast GID E3 Ligase Complexes**

(A) Sucrose density gradient fractionation of *S. cerevisiae* lysates from cells harvested from four conditions: stationary growth in glucose-rich medium (normal), ethanol (carbon stress), and switch to glucose-rich medium for 30 and 120 min (carbon recovery 30 and 120 min, respectively). Gid subunits tagged at their endogenous loci were visualized by western blotting. Asterisk indicates a non-specific anti-Flag interaction.

(B) Cartoons representing GID assemblies in different environmental conditions, on the basis of migration patterns of subunits in sucrose density gradients.

(C) Coomassie-stained SDS-PAGE of recombinant GID<sup>Ant</sup> and GID<sup>SR4</sup>.

(D) Fluorescent scan examining ubiquitylation of fluorescently labeled Pro/N-degron substrate Mdh2 (Mdh2\*). Assays test dependence on E2 (Ubc8) and substrate receptor (Gid4). Note: GID<sup>Ant</sup> + Gid4 = GID<sup>SR4</sup>.

(E) Role of substrate N-terminal Pro, tested with wild-type (WT) Mdh2 or N-terminal Pro-to-Ser mutant, in binding GST-tagged substrate receptor Gid4 *in vitro*.

(F) Ubiquitylation of WT or N-terminal Pro-to-Ser mutant Mdh2-His<sub>6</sub> visualized by western blot with anti-His<sub>6</sub> tag antibody.

(G) Cryo-EM reconstructions at 9 Å resolution of endogenous and recombinant GID<sup>Ant</sup> and recombinant GID<sup>SR4</sup>. Density attributed to substrate receptor Gid4 is yellow.

(H) Clamp-like structure of GID<sup>SR4</sup> assembled from substrate receptor Gid4 (SR4), scaffold, and catalytic modules.

See also Figure S1 and Tables S1 and S2.

### Modular GID E3 Ligase Assembly

Refinement of the cryo-EM data for GID<sup>SR4</sup> yielded a 3D reconstruction at 3.8 Å resolution (Tables 1, 2, S1, and S2; Figures S2–S6). Atomic coordinates for Gid4, Gid5, Gid8, and much of Gid1 and Gid9 were generated by a combinatorial approach involving cryo-EM maps of many variant complexes and automated and manual model building (Table 2; Figures S2–S6). Additional predicted domains from Gid1, Gid2, and Gid9 were approximately docked into lower resolution density (Figures S5B, S6A, and S6B). The multidomain nature of Gid proteins enabled structure validation through (1) testing effects of deleting specific subunits or domains on cryo-EM reconstructions; (2) strong correlations upon superimposing structures of human Gid4 substrate-binding and Gid1 Spla and Ryanodine receptor (SPRY) domains (1.0 and 0.73 root-mean-square deviation [RMSD], respectively) (Dong et al., 2018; Li et al., 2011); (3) visualizing predicted armadillo repeats in Gid5 and LisH-CTLH-CRA domains in Gid1, Gid8, and Gid9; and (4) concordance between the structure and effects of mutations observed in prior studies of GID E3 assembly *in vivo* (Braun et al., 2011; Menssen et al., 2012; Santt et al., 2008).

Overall, the EM data reveal that the GID E3 is organized around three structural and functional modules (Figure 2): the

scaffold—Gid1, Gid5, and Gid8, tightly interacting in a manner that outwardly projects protein interaction domains from each subunit; the substrate receptor—Gid4; and the catalytic module—the Gid2-Gid9 subcomplex. Details of this assembly and how it drives ubiquitylation of N-degron substrates are described below.

### The Scaffold

The foundation of GID<sup>SR4</sup> is an interdigitated assembly of Gid1, Gid8, and Gid5 (Figures 2A and S6C–S6F). Gid1 and Gid8 together form a heterodimeric trefoil-shaped structure. At the vertex, Gid1's LisH and C-terminal segment of the CRA domain (LisH-CRA<sup>C</sup>), and adjacent elements, embrace paralogous regions from Gid8, rationalizing why Gid1 and Gid8 stabilize each other *in vivo* (Menssen et al., 2012). The three lobes of the trefoil are formed by (1) Gid1's SPRY domain, (2) Gid1's CTLH and N-terminal segment of the CRA domain (CRA<sup>N</sup>), and (3) Gid8's CTLH-CRA<sup>N</sup> domain and adjacent sequences (Figure S6C). The distal ends of the CTLH-CRA<sup>N</sup> domains from both Gid1 and Gid8 radiate away from the core, while a continuous V-shaped surface between Gid1's SPRY and Gid8's CTLH domain engages an extended complementary surface

**Table 1. Cryo-EM Data Collection, 3D Reconstruction, and Map Refinement**

	GID <sup>SR4</sup>	GID <sup>Ant</sup>	Endogenous GID <sup>Ant</sup>	GID <sup>SR4</sup> ΔRINGs	GID <sup>SR4</sup>	GID <sup>Ant</sup>	GID <sup>Scaffold</sup> Plus SR <sup>Gid4</sup>	GID <sup>Scaffold</sup> Plus SR <sup>Gid10</sup>	GID <sup>Scaffold</sup>	GID <sup>SR4</sup> Minus Gid2/ΔGid <sup>RING</sup>
EMDB code	N/A	N/A	10331	10332	10327	10326	10330	10329	10328	10333
Microscope/ detector	Arctica/ Falcon III	Arctica/ Falcon III	Arctica/ Falcon III	Arctica/ Falcon III	Krios/K2	Krios/ K2	Krios/K3	Krios/K3	Krios/K3	Krios/K3
Particles	387,982	388,646	192,972	378,602	615,139	994,904	1,106,310	2,132,595	2,132,595	1,645,121
Pixel size	1.612	1.997	1.612	1.612	1.06	1.06	1.09	1.09	1.09	1.09
Defocus range (μm)	1.5–3.5	1.5–3.5	1.5–3.5	1.5–3.5	1.1–3.2	1.1–3.2	1.1–3.2	1.1–3.2	1.1–3.2	1.1–3.2
Voltage (kV)	200	200	200	200	300	300	300	300	300	300
Electron dose (e/Å <sup>2</sup> /s)	21.3	23.07	22.4	21.4	6.675	7.88	13.57	13.8	13.8	13.62
Exposure time (s)	3	3	3	3	8	7	4	4	4	4
Map resolution (Å) <sup>a</sup>	5.1	8.3	9.3	7.3	3.8	3.7	3.4	3.8	3.8	3.2
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	4.5–11.1	6.0–24.0	7.9–25.6	6.0–23.1	3.5–15.8	3.5–11.9	3.1–7.2	3.5–6.5	3.5–7.6	3.0–8.3
Map sharpening B-factor (Å <sup>2</sup> )	–246	–301	–618	–372	–92	–119	–114	–109	–99	–80

See also Figures S2, S3, S4, and S6.

<sup>a</sup>According to the Fourier shell correlation (FSC) cutoff criterion of 0.143 defined in Rosenthal and Henderson (2003).

from Gid5. Gid5's armadillo repeats stack in tandem in a continuous solenoid of roughly one and a half superhelical turns, with the N-terminal domain (NTD) filling the groove between Gid1 and Gid8, and a C-terminal domain (CTD) radiating outward (Figures 2A, S6C, and S6E). The scaffold is further buttressed by loops from all three proteins extending distances up to  $\approx 70$  Å to engage one another.

A protein interaction domain from Gid5 recruits the substrate receptor Gid4, and Gid8 binds the catalytic module Gid2–Gid9 (Figures S6C and S6F). Weak density corresponding to Gid1's CTLH domain also projects outward. Although the structural role of Gid1's CTLH domain is presently unknown, we speculate it binds Gid7 on the basis of its mutation specifically impairing this interaction *in vivo* (Menssen et al., 2012).

#### Scaffold Binding to Substrate Receptor Gid4 Generates GID<sup>SR4</sup>

The substrate receptor, which recruits proteins for ubiquitylation, is an essential E3 ligase element. A prior structure showed that human Gid4's substrate-binding domain is largely a  $\beta$ -barrel, with a funnel-shaped opening at one end binding to short peptides via their N-terminal Pro (Figure 3A) (Dong et al., 2018). The structure of GID<sup>SR4</sup> shows, in turn, how this substrate-binding domain is incorporated into an active E3 ligase (Figures 2B and 3). Gid4's C-terminal eight residues anchor the interaction, by extending into a channel in the concave surface of Gid5<sup>CTD</sup> (Figures 3B and 3C). Here, successive Gid4 side chains protrude in opposite directions. Aromatic residues on one side fill pockets between Gid5 armadillo repeats. Those on the other side contribute to a composite Gid4/Gid5 interface with an aliphatic

stripe across Gid4's barrel. Indeed, mutation of key Gid4 and Gid5 contact residues impaired GID<sup>SR4</sup>-catalyzed substrate ubiquitylation *in vitro* (Figures 3D and 3E). *In vivo*, the Gid5 mutations substantially impaired degradation of the gluconeogenic enzyme Fbp1, as did mutation of Gid4's C-terminal anchor. Although individual conservative amino acid substitutions in Gid4's substrate-binding domain did not have a measurable effect, introduction of a bulky residue or multiple Ala mutations caused substantially impaired glucose-induced degradation of Fbp1 (Figures 3D and 3E).

Additionally, weaker EM density showed Gid4 residues 91–116, upstream of the substrate-binding domain, meandering over 65 Å to loosely wrap around to the convex face of Gid5. Also a loop from Gid1's SPRY domain contacts a peripheral helical portion of Gid4's substrate-binding domain (Figure 3F). However, these residues are neither conserved nor essential for GID<sup>SR4</sup> activity *in vitro* or *in vivo*, suggesting auxiliary roles (Figure 3F).

#### Dynamic Gid5 CTD in Anticipation of a Substrate Receptor

To understand the structure of the GID complex expressed during carbon stress, EM data for recombinant GID<sup>Ant</sup> were refined to yield a 3D reconstruction at 3.7 Å resolution (Figures 4, S2, and S4A; Tables 1 and S1). Comparison with the map of GID<sup>SR4</sup> showed a striking difference in the density corresponding to Gid5's substrate-receptor-binding CTD, which is blurred in GID<sup>Ant</sup>. The Gid5 armadillo repeats are visible, but poor density precluded refinement to high resolution (Figure 4A). Thus, we speculate that anticipation is manifested by conformational

**Table 2. Model Refinement and Validation Statistics**

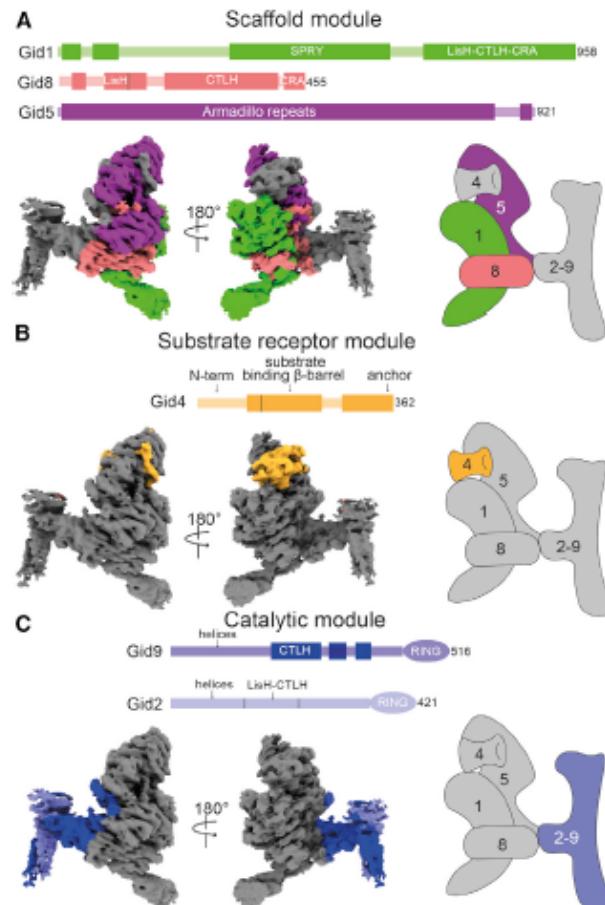
Map	GID <sup>SR4</sup> Minus Gid2/ΔGid9 <sup>RING</sup>
Refinement	
Model Composition	
Non-hydrogen atoms	16,071
Protein residues	2,031
Resolution	3.1
FSC map versus model@0.143 <sup>a</sup>	
RMS Deviations	
Bond lengths (Å)	0.005
Bond angles (Å)	0.960
Validation	
Molprobity score/percentile	1.71
Clashscore/percentile	5.77
Rotamer outliers (%)	0.23
Ramachandran Plot	
% favored	94.15
% allowed	5.8
% outliers	0.05
See also Figures S5A and S6.	
<sup>a</sup> According to the map-versus-model correlation coefficient definitions in Afonine et al. (2018).	

dynamics of the Gid5 CTD armadillo repeats prior to capturing and curling around a substrate receptor subsequently available upon change in environmental conditions.

Because GID<sup>Ant</sup> and GID<sup>SR4</sup> are structurally similar beyond Gid5's CTD and its associated Gid4 (Figure 4B), we hypothesized that GID complexes may display intrinsic catalytic activity irrespective of ability to recruit substrate. To test this, we used an assay that monitors substrate-independent activation of E2~Ub intermediates (Petroski and Deshaies, 2005; Wenzel et al., 2011). First, the reactive Ubc8~Ub intermediate (the symbol ~ refers to thioester linkage) was generated enzymatically, and this reaction was quenched. Next, lysine was added simultaneously with various versions of GID E3s. Ub transfer from Ubc8, presumably to unanchored lysine, was monitored by both disappearance of Ubc8~Ub and appearance of free Ubc8 in SDS-PAGE. Although the Ubc8~Ub intermediate was relatively stable on its own over time, GID<sup>Ant</sup> stimulated its rapid discharge with little effect of titrating a version of Gid4 suitable for substrate recruitment (Figures 1D and 4C). Thus, GID<sup>Ant</sup> is intrinsically competent at activating Ub transfer even without a recruited N-degron substrate or its receptor.

#### A Family of Related GID E3s

The concept of a multiprotein E3 ligase that facultatively associates with a substrate receptor is conceptually reminiscent of cullin-RING and anaphase-promoting complex E3 families. However, these E3s use sets of interchangeable substrate receptors for distinct regulation (Afieri et al., 2017; Lydeard et al., 2013; Watson et al., 2019). Thus, we hypothesized that other GID substrate receptors may exist, and we identified the ORF YGR066C as encoding a protein displaying homology to



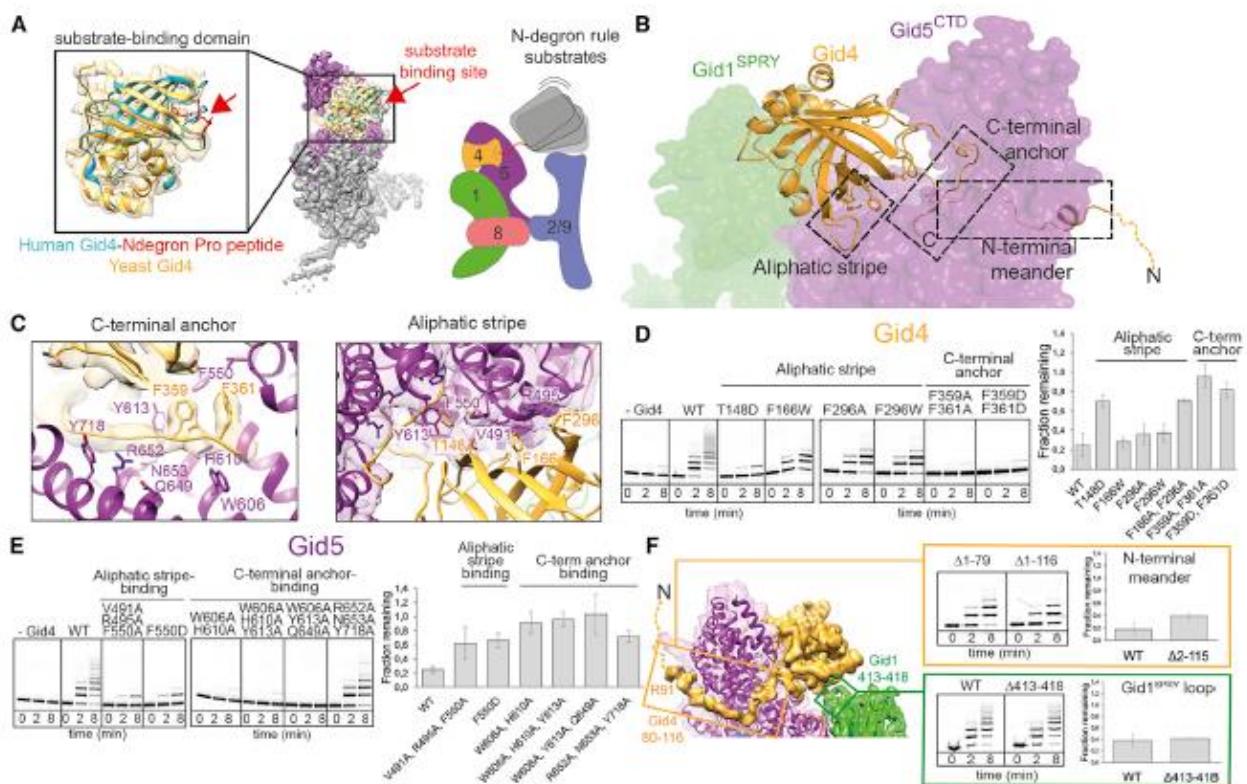
**Figure 2. GID<sup>SR4</sup> E3 Ligase Modular Architecture**

Each panel shows a different module as a domain schematic (top), two views of cryo-EM density (bottom left), and a cartoon (bottom right). Subunits within a module are color-coded, with others in gray. Darker boxes in the domain schematic represent regions of the density map into which an atomic model was built.

- (A) Scaffold module comprising Gid1 (green), Gid8 (salmon), and Gid5 (purple).
- (B) Substrate receptor module consisting of Gid4 (orange).
- (C) Catalytic module composed of Gid2 (light blue) and Gid9 (dark blue).

See also Figures S2–S6 and Tables S1 and S2.

Gid4, including the Gid5-binding hydrophobic stripe and C-terminal anchor (Figure S4C). While our manuscript was under consideration, YGR066C was published as a GID E3 substrate receptor and renamed "Gid10" (Melnikov et al., 2019). We have adopted this nomenclature and had already independently performed several experiments suggesting that Gid10 is a bona fide alternative substrate receptor for a GID E3. First, bacterially expressed Gid10 binds our recombinant GID<sup>Ant</sup> (Figure 5A). Second, Gid10 confers onto GID<sup>Ant</sup> *in vitro* ubiquitylation activity toward an N-degron substrate, albeit with far lower efficiency than Gid4 (Figure 5B). Third, a 3.8-Å-resolution cryo-EM reconstruction of Gid10 bound to the scaffold module showed an overall similar structure to the Gid4-bound complex, including homologous placement of Gid10's C-terminal anchor and a β-barrel



**Figure 3. Formation of GID<sup>SPRY</sup> E3 Ligase through Incorporation of Substrate Receptor Gid4**

(A) Overlay of scaffold-bound *S. cerevisiae* Gid4 substrate-binding domain with crystal structure of human Gid4 bound to N-terminal Pro peptide (PDB: 6CDC) (left), showing potential substrate-binding site with a red arrow. Cartoon of N-degron substrate binding by GID<sup>SPRY</sup> (right).

(B) Overview of Gid4 elements binding to GID scaffold. Gid4 (yellow cartoon) binds Gid5<sup>CTD</sup> (purple surface) via a C-terminal anchor (C, C terminus), an aliphatic stripe and an N-terminal meander. Disordered residues connecting to N terminus (N) shown as dotted line.

(C) Close-up of Gid4 (yellow) C-terminal anchor and aliphatic stripe interactions with Gid5 (purple). Residues mutated in (D) and (E) are represented as sticks.

(D) Assays testing importance of Gid4 residues in aliphatic stripe and C-terminal anchor on *in vitro* Mdh2 ubiquitylation and *in vivo* Fbp1 degradation (quantified as fraction from time 0 remaining after switching from carbon stress to carbon recovery).

(E) Assays testing importance of Gid5<sup>CTD</sup> residues that interact with Gid4 aliphatic stripe and C-terminal anchor on *in vitro* Mdh2 ubiquitylation and *in vivo* Fbp1 degradation.

(F) Left: structure and EM density map depicting auxiliary interactions between Gid5<sup>CTD</sup> and Gid4 N-terminal meander (residues 80–116) and Gid1 SPRY domain loop (residues 413–418) with peripheral helical insertion in Gid4. Right: assays testing if these elements are not essential for *in vitro* ubiquitylation of Mdh2 and *in vivo* Fbp1 degradation.

(D, E, and F) Error bars represent SD ( $n \geq 3$ ).

See also Figures S5 and S6 and Table S2.

domain poised to bind N-degron substrates (Figures 5C, 5D, S2, and S4B; Tables 1 and S1). Indeed, deletion of Gid10's C-terminal anchor abrogates Gid10-dependent ubiquitylation of the recombinant substrate Mdh2 (Figure 5B).

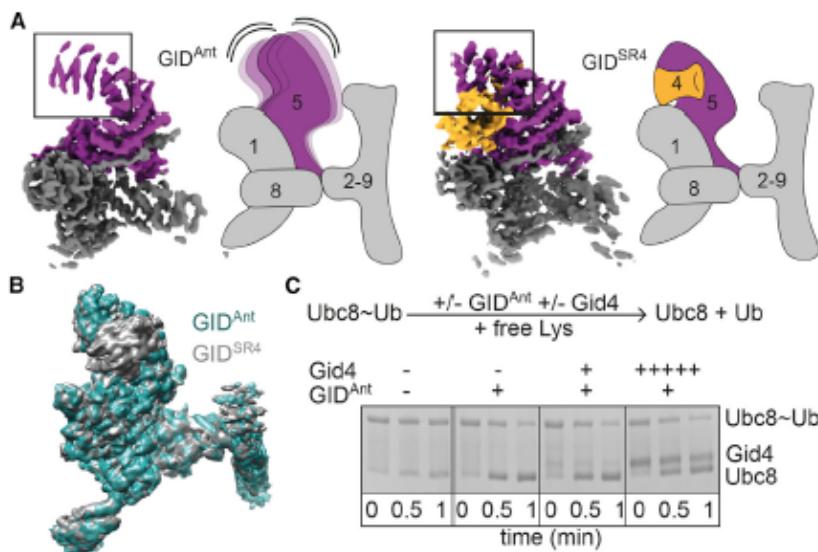
Comparing EM maps with the two substrate receptors in detail shows a potential for the slightly different placement of Gid10 and Gid4 relative to the scaffold (Figures 5C and 5D). This raises the possibility that orientation of substrate-binding domains may underlie mechanisms regulating substrate degradation under different cellular conditions.

Although deletion of the Gid10 gene in yeast did not affect degradation of known Gid substrates after carbon source switching (data not shown and Melnykov et al., 2019), prior transcriptomics, along with our analyses of protein levels, do not

imply Gid10 expression under these conditions. Rather, Gid10 mRNA is expressed during various stresses, including high salinity and heat shock (Gasch et al., 2000; Wanichthanarak et al., 2014). Indeed, we observed Gid10 protein induction under these conditions, presumably leading to its incorporation into an alternative GID<sup>SPRY</sup> E3 complex (Figures 5E and 5F; Melnykov et al., 2019).

#### Embedding of a RING-RING Catalytic Module within Multisubunit E3 Ligase

Most E3 ligases depend on one or more RING domains binding to the E2 and the Ub in a thioester-linked E2~Ub intermediate, thereby stabilizing a "closed conformation" that activates discharge of Ub from the E2 active site (Dou et al., 2012;



**Figure 4. Structural Anticipation by GID<sup>Ant</sup>**

(A) Cryo-EM maps and cartoons showing GID<sup>Ant</sup> and GID<sup>SR4</sup> with Gid5 purple and Gid4 yellow. Black boxes highlight relatively weaker Gid5<sup>CTD</sup> density in GID<sup>Ant</sup>, which we presume represents conformational flexibility in the absence of substrate receptor.

(B) Superposition of cryo-EM maps for GID<sup>SR4</sup> (gray) and GID<sup>Ant</sup> (aqua) at low contour.

(C) Assay testing substrate-independent E3 activity. First, the Ubc8~Ub intermediate is generated enzymatically and this reaction is quenched. Next, free lysine was added. Reactivity probed by loss of Ubc8~Ub and appearance of Ub was tested without an E3, or with GID<sup>Ant</sup> alone, or with addition of equimolar and 5× excess of Gid4. See also Figures S2 and S4A and Table S1.

To validate the locations of the RINGs, we examined mutant versions of GID<sup>SR4</sup> lacking these domains by cryo-EM. Refinement of the data led to two major

Plechanová et al., 2012; Pruneda et al., 2012). Thus, we sought to identify the structural locations and functional roles of the Gid2 and Gid9 RINGs. Having already placed the scaffold and substrate receptor modules, we attributed the remaining density to the catalytic module. This adopts a T-shaped structure, where the base of the T connects the catalytic domain to the scaffold (Figures 2, 6A, 6B, and S6). Here, Gid9's CTLH-CRA<sup>N</sup> domain heterodimerizes with that from Gid8 in a manner resembling a pillar affixed to a base.

The top of the T appears to comprise multiple heterodimeric Gid2-Gid9 subdomains. The relatively poor resolution of this region may suggest mobility of the Gid2-Gid9 subdomains with respect to one another and relative to the scaffold. Although it was not possible to determine which elements derive from Gid2 or Gid9, the density was sufficiently visible at low contour to approximately localize predicted domains (Figures 6B, 6C, S5B, and S6B). One side of the top of the T is a four-stranded coiled coil, which we speculate corresponds to helices predicted at the N termini of Gid2 and Gid9 (Kelley et al., 2015).

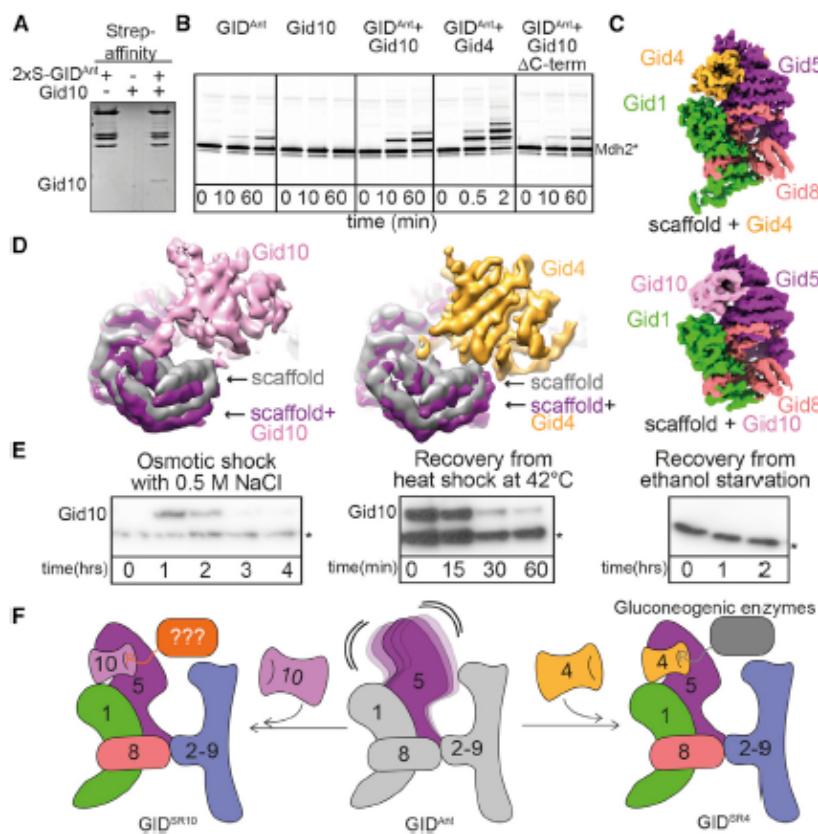
Significantly, the structure of the catalytic core appears to place the Gid2 and Gid9 RING domains in a canonical RING-RING dimer assembly in the clamp-like structure of GID<sup>SR4</sup>, forming the second "jaw" that faces Gid4 (Figures 1H and 6C). We arrived at this conclusion after considering that the remainder of the T-structure consists of two subdomains, and then roughly attributing the unassigned Gid2-Gid9 features. The subdomain at the extreme edge of the complex can be fitted with a homology model of the Gid2 and Gid9 RING domains superimposed on a canonical RING-RING dimer assembly found in many E3 ligases (Kelley et al., 2015). Notably, the notion that the RINGs heterodimerize is consistent with prior mutations of zinc ligands within either protein, which presumably lead to RING misfolding, decreasing Gid2-Gid9 interactions and eliminating glucose-induced substrate degradation *in vivo* (Braun et al., 2011; Regelmann et al., 2003).

classes. One, indeed, showed selective elimination of the density we attribute to a Gid2-Gid9 RING-RING dimer, while the second class superimposed with the map obtained for a sample lacking the entire Gid2 subunit as well as the Gid9 RING domain (Table S1). This is consistent with heterodimeric assembly of the RING domains contributing to Gid2 incorporation into a GID E3 *in vivo* (Braun et al., 2011). Last, we speculate that the remaining density at the T-junction corresponds to a heterodimeric assembly comprising the LisH-CRA<sup>C</sup> domains from Gid2 and Gid9 and/or the ensuing CTLH domain from Gid2, which would match the size of this subdomain (Figure S6B). Moreover, this hypothesis is consistent with the relative orientation of Gid9's CTLH-CRA<sup>N</sup> domain, which is inserted between the LisH and CRA<sup>C</sup> elements in the sequence of Gid9.

#### Model of the Catalytic Center Suggests that the Heterodimeric RING Activates a Single Ubc8~Ub Facing Substrate

As a first step toward structurally modeling GID<sup>SR4</sup>-catalyzed ubiquitylation, each RING domain docked into the EM density was superimposed with a prior structure of an isolated RING-E2~Ub complex (Dou et al., 2012; Plechanová et al., 2012; Pruneda et al., 2012), and then the docked E2 was replaced with Ubc8. Even with uncertain position of the Gid2-Gid9 RING-RING dimer, the structural modeling suggested that only one of the two RING domains would place Ubc8 to face the Gid5-bound substrate receptor.

To test if the Gid2 and/or Gid9 RING primarily binds Ubc8 or plays a supporting role in activating the Ubc8~Ub intermediate, residues were selected for mutation on the basis of homology to three hallmark elements: (1) a hydrophobic surface that binds E2 loops conserved in Ubc8; (2) potential "linchpin" residues, which can be located on either side of the domain, but irrespective of location insert between the E2 and its thioester-linked Ub to stabilize the noncovalent interface between them; and (3) a non-RING priming element flanking a RING sequence that functions



**Figure 5. A Family of Multisubunit GID E3s with Swappable Substrate Receptors**

(A) Streptactin pull-down of GID<sup>Art</sup> testing binding of Gid10.

(B) Ubiquitylation assay testing potential of Gid10 to act as a substrate receptor for the GID<sup>SR4</sup> substrate Mdh2.

(C) Cryo-EM maps of Gid4- and Gid10-bound GID scaffold.

(D) Close-up views of overlays of maps of GID scaffold alone and bound to Gid10 or Gid4.

(E) Western blots showing expression of Gid10, tagged at the endogenous locus, under different environmental conditions. Asterisk indicates a protein interacting non-specifically with anti-Flag antibodies.

(F) Model for family of GID E3s with interchangeable substrate receptors.

See also Figures S2, S3B, S4B, and S4C and Table S1.

in *trans* to allosterically stabilize the closed conformation of the E2-Ub intermediate bound primarily to the opposite RING in a dimer (Figure 6D) (Brown et al., 2014; Dou et al., 2012; Kelley et al., 2015; Plechanovová et al., 2012; Pruneda et al., 2012; Scott et al., 2014; Zheng et al., 2000). Effects on GID<sup>SR4</sup> E3 ligase activity *in vivo* were tested by introducing mutations into tagged versions of Gid2 and Gid9 expressed from their endogenous loci (Figure 6E). Effects of point mutations in predicted E2-binding and linchpin residues of Gid2 mirrored effects of wholesale deletion of Gid2's RING domain on glucose-induced degradation of Fbp1, while there was a relatively minimal effect of mutating Gid2's candidate non-RING priming element. The crucial role for the Gid2 RING's E2-binding site was also confirmed for GID<sup>SR4</sup> E3 ligase activity *in vitro*. In contrast, the opposite pattern was observed for the Gid9 mutants, where only the candidate non-RING priming element significantly abrogated activity. The results suggest that Gid2's RING binds and activates the Ubc8-Ub intermediate, assisted by a non-RING priming element from Gid9, to face the substrate receptor (Figures 6E and 6F).

#### Model of GID<sup>SR4</sup> Ubiquitylating an N-Degron Gluconeogenic Enzyme

The substrate-binding site on Gid4 is  $\approx 50$  Å away from the modeled catalytic center. Although the relatively weak EM density corresponding to the catalytic domain (Figure 6C) suggests flexibility, perhaps for conformational changes during catalysis, it is also possible that the large gap accommodates substrates

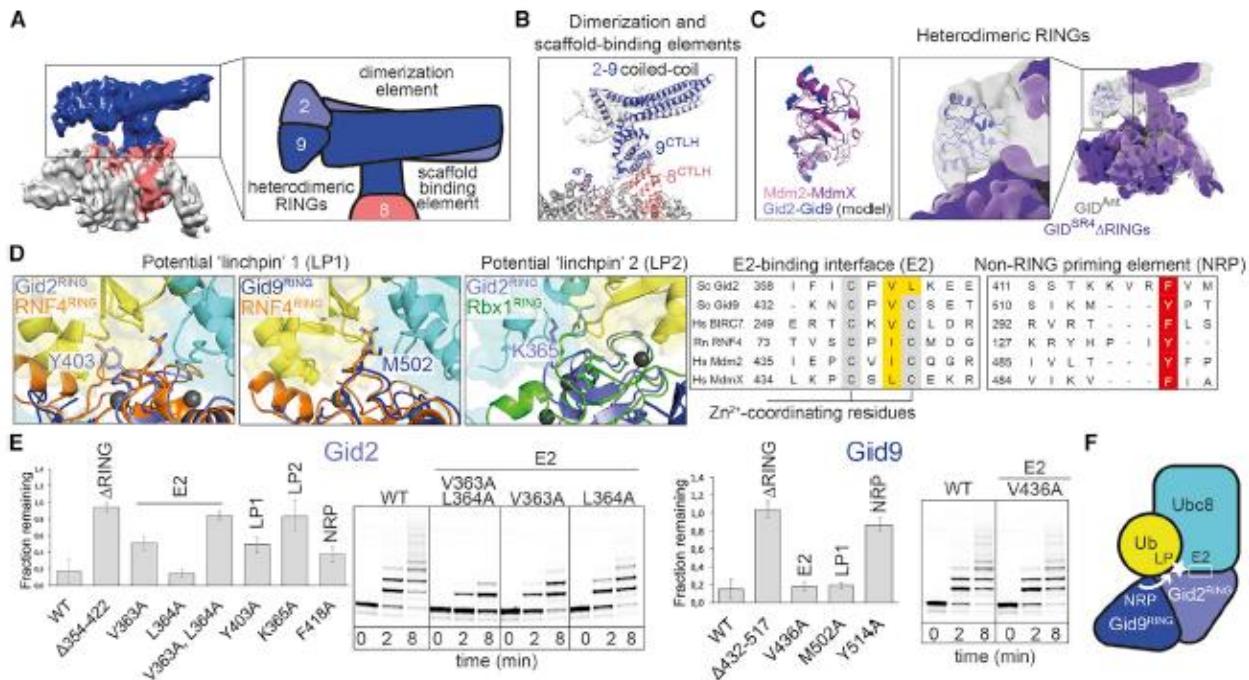
of GID<sup>SR4</sup>, which are large oligomeric enzymes. To model ubiquitylation, the substrate Mdh2 was selected because (1) robust *in vitro* ubiquitylation of bacterially expressed Mdh2 demonstrated that post-translational modifications are not required for its N-degron-based substrate targeting (Figures 1D–1F), and (2) as an  $\approx 80$  kDa homodimer with 34 lysines, Mdh2 is the smallest and structurally

most simplistic of known GID<sup>SR4</sup> substrates (Figure 7). To place Mdh2, the N-terminal four residues were modeled on the basis of the prior structure of the human Gid4 substrate-binding domain bound to a 4-mer peptide. Next, an Mdh2 model was manually rotated while roughly constraining the location of the N-terminal-most ordered residue (L14) proximal to the substrate-binding site on Gid4 (Dong et al., 2018; Kelley et al., 2015).

Overall, the model suggests that some, but not all, Mdh2 lysines would preferentially access the ubiquitylation active site (Figure 7A). To test this, we used mass spectrometry to map sites that are ubiquitylated *in vitro* (Figure S7). Notably, the top sites include a cluster of K254, K256, and K259, as well as K330, and to a lesser extent K360 and K361, for which the 10-residues between Mdh2's N terminus and globular domain would easily accommodate the  $\approx 10$ ,  $\approx 20$ , and  $\approx 15$  Å required, respectively, for Mdh2 to simultaneously engage Gid4 and approach the Gid2 RING activated Ubc8-Ub intermediate. Furthermore, substituting these residues with arginines severely impaired Mdh2 ubiquitylation (Figure 7B). Thus, although detailed knowledge awaits further structural studies, the EM data presented here enable the generation of a geometrically reasonable model for N-degron substrate ubiquitylation (Figure 7C).

#### DISCUSSION

The cryo-EM reconstructions reported here reveal E3 ligase assemblies that vary in response to extracellular stimuli



**Figure 6. GID Catalytic Module**

(A) Left: T-shaped Gid2-Gid9 catalytic module in low contour EM map of GID<sup>Ant</sup>. Right: catalytic module elements shown in cartoon: scaffold-binding domain interacting with Gid8 (salmon), dimerization region, and heterodimeric RINGs.

(B) Homology models of catalytic module elements fitted into EM map generated by focused refinement and signal subtraction. The atomic models of Gid8 and Gid9 CTLL-CRA<sup>N</sup> domains are shown, as is an approximated coiled-coil docked in additional density.

(C) The Gid2-Gid9 RING-RING domain was modeled in triangular density at the tip of the T-shaped catalytic module, as follows: (1) RING-RING domain was generated by superimposing homology models of Gid2 and Gid9 RINGs onto MDM2-MDMX structure (PDB: 2V1E) (left). (2) Model of Gid2-Gid9 RING-RING domain was docked into map of GID<sup>Ant</sup> (center). (3) Density attributed to the Gid2-Gid9 RING-RING domain was not visible in EM map of GID<sup>SP4</sup> with the RINGs deleted (solid violet map, right).

(D) Left: candidate Gid2 and Gid9 RING "linchpins" (LP) identified by superimposing their homology models with crystal structures of RNF4 (PDB: 4AP4) and RBX1 (PDB: 4P50) bearing lynchpin residues R181 and R46, respectively. Corresponding Gid2 and Gid9 residues are shown as sticks. Right: sequence alignments of Gid2 and Gid9 with well-characterized RING domains identified potential E2-binding (E2, yellow) and non-RING priming element (NRP, red) residues.

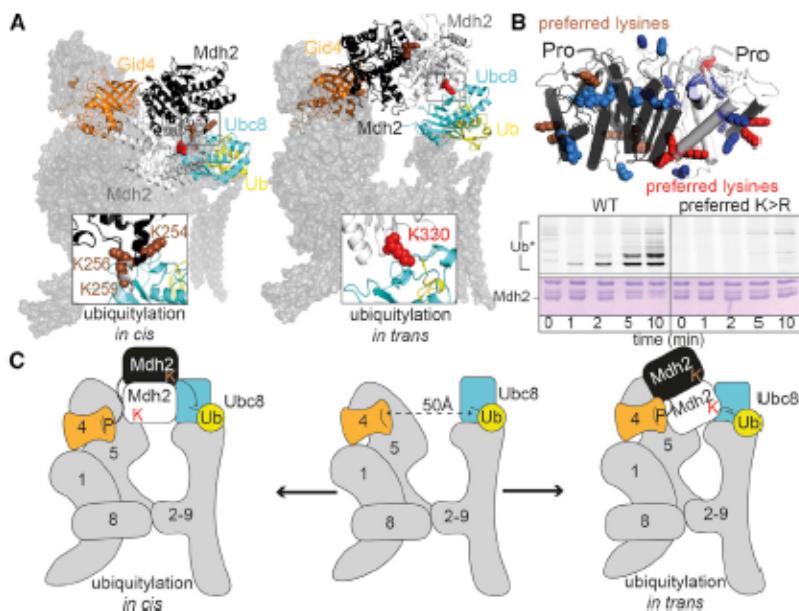
(E) Assays testing effects of Gid2 and Gid9 mutations on GID E3 activity, Fbp1 degradation *in vivo*, and Mdh2 ubiquitylation *in vitro*. Error bars represent SD ( $n \geq 3$ ).

(F) Cartoon summarizing model for Ubc8-Ub activation by Gid2-Gid9 RING-RING domains on the basis of mutational analysis shown in (E). See also Figures S3C, S4A, S5, and S6 and Table S1.

(Figures 1, 4, 5, and S1), molecular mechanisms underlying their regulation (Figures 1, 2, 3, 4, 5, and 6), and a framework for GID E3 ligase-dependent ubiquitylation (Figure 7). The structural data also provide broad insights into major families of E3 ligases, namely, those recognizing terminal degrons and those displaying RING-RING catalytic domains. The modular multiprotein GID<sup>SP4</sup> E3 assembly displays clamp-like properties, established by a central scaffold connecting the two jaws: a variable substrate receptor and the catalytic domain. The structure enables binding of a substrate's unfolded N terminus to Gid4, to direct lysines from a folded domain into the ubiquitylation active site. From the perspective of the other side of the complex, it seems that the RING-RING dimer is the culmination of an intricate heterodimeric Gid2-Gid9 assembly that activates the Ubc8-Ub intermediate facing the N-degron substrate. Some other E3s, for example,

BRCA1-BARD1 or HDM2-HDMX, that contain heterodimeric RING-RING domains, may likewise rely on complex, interconnected assemblies to couple a single active site with a substrate for ubiquitylation.

Although it has long been recognized that GID<sup>SP4</sup> is part of the yeast response to environmental conditions via glucose-induced degradation of gluconeogenic enzymes (Santit et al., 2008), our data indicate that generation of a GID<sup>Ant</sup> complex also occurs in response to an extracellular stimulus: carbon stress (Figure 1A). As GID<sup>Ant</sup> would be inactive toward recruited substrates in the absence of a substrate receptor, we hypothesize that production of this complex allows cells to adapt more rapidly to potential later changes in the extracellular milieu. Our data raise the possibility that carbon stress may prepare cells for a potential return to nutrient-rich conditions. We also cannot rule out the possibility that GID<sup>Ant</sup> could be coupled to a yet unknown substrate



**Figure 7. Model of GID<sup>SR4</sup>-Catalyzed Ubiquitylation of N-Degron Substrate Mdh2**

(A) Structural models for substrate ubiquitylation by GID<sup>SR4</sup>, with a homology model for Mdh2 (protomers in black and gray) placed with its N-terminal Pro binding Gid4 and candidate lysine targets in the active site of a modeled Ubc8-Ub intermediate.

(B) Homology model of Mdh2 dimer showing preferred target lysines identified by mass spectrometry in brown and red and other lysines in blue. Assay testing effects of Arg replacement for preferred target lysines on Mdh2 ubiquitylation. Top: fluorescent scan detecting ubiquitin. Bottom: loading controls showing Mdh2, Gid2 (2), and Gid4 (4).

(C) Cartoon representing the structural models for GID<sup>SR4</sup> ubiquitylation. The clamp-like structure enables multiple configurations for ubiquitylation of preferred lysines from a folded gluconeogenic enzyme substrate.

See also Figure S7.

receptor to allow ubiquitylation of a distinct set of proteins during carbon stress.

We were puzzled by the apparently counterintuitive carbon stress-induced production of GID<sup>Ant</sup> in anticipation of relief from starvation. It is conceptually appealing to envisage microbial anticipatory signaling cascades in terms of one stress serving as a signal for cells to cross-prepare for a looming new stress (Mitchell et al., 2009; Tagkopoulos et al., 2008). Our discovery that recombinant GID<sup>Ant</sup> also binds the Gid4-like protein YGR066C/Gid10, which is induced under several distinct stress conditions (Figure 5; Melnykov et al., 2019), offers a potential mechanism for how GID<sup>Ant</sup> could act as a multifaceted hub integrating responses to various extracellular stimuli. We speculate that carbon stress-induced production of GID<sup>Ant</sup> may enable cells to prepare for ensuing osmotic stress or heat stress through the production of a Gid10-associated GID<sup>SR10</sup> E3 ligase. It seems plausible that Gid10's substrates could be regulators of glycerol or salt intake, protein synthesis, or general stress responses. Mechanistically, it seems likely that substrate selectivity will be influenced not only by protein expression changes under different metabolic conditions but also by subtle differences in the  $\beta$ -barrel domains of Gid4 and Gid10 and their orientations relative to the scaffold (Figure 5). Future studies will be required to identify substrates of a GID<sup>SR10</sup> E3, to visualize substrate ubiquitylation in action, and to understand cross-talk between GID<sup>Ant</sup> assembly, association with multiple substrate receptors, and coupling responses to varying extracellular stimuli.

What does an E3 ligase in "anticipation" look like? Our cryo-EM reconstruction of GID<sup>Ant</sup> suggested motion of Gid5's substrate receptor-binding CTD (Figure 4). Dynamic opening and closing of Gid5 could enable binding, release, and exchange of the substrate receptor. At this point, the molecular stimuli and structural mechanisms underlying substrate receptor dissociation or exchange remain unknown, although Gid4 turnover has

been shown to depend on GID E3-dependent ubiquitylation (Menssen et al., 2018).

We speculate that in humans, a GID<sup>Ant</sup> complex will parallel its yeast counterpart and act as a hub integrating various signals, presumably through many additional binding partners, in turn leading to cell fate determination. Although binding to Gid4 likely generates a human GID<sup>SR4</sup> E3 targeting substrates with N-terminal prolines (Dong et al., 2018), the functions of other partner proteins, including Gid7, remain elusive, and may regulate activity of GID<sup>SR4</sup>, or perhaps form alternative assemblies with GID<sup>Ant</sup> or Gid subunits. Indeed, two Gid7 homologs, along with many other proteins, have been shown to co-precipitate with human Gid subunits (Boldt et al., 2016; Huttlin et al., 2017; Lampert et al., 2018). In addition, the human GID E3 ligase has been reported to ubiquitylate a substrate that does not bear an N-terminal proline (Lampert et al., 2018), despite the apparent requirement of a proline to bind human Gid4 (Dong et al., 2018). This could potentially be reconciled on the basis of our discovery that the GID E3 ligase is not a singular complex but a family of E3 ligases with different substrate receptors (Figure 5). Additional human Gid subunits could substitute for Gid4, modulate substrate specificity, or localize the GID complex (Boldt et al., 2016; Lampert et al., 2018).

The concept of multiple GID E3 assemblies responding to different environmental stimuli is reminiscent of other multiprotein E3 ligases (e.g., cullin RING-ligases) and hubs such as mTOR that integrate signaling with various downstream functions required in certain cellular contexts (González and Hall, 2017; Lydeard et al., 2013; Saxton and Sabatini, 2017). Regulation of these assemblies through interchangeable receptors provides a framework for investigating the GID family. Are there cellular exchange factors that promote swapping Gid4 and Gid10 (Pierce et al., 2013), or inhibitory factors (Duda et al., 2012; Lyapina et al., 2001)? Is GID regulated by modifications or metabolites (González and Hall, 2017; Saxton and Sabatini,

2017)? Does substrate binding play a role in substrate-receptor selection (Emberley et al., 2012; Enchev et al., 2012)? And most curiously, are there other presently unknown substrate receptors? Although future studies will be required to unveil the molecular pathways and mechanisms underlying these complexities, the stunning structural intricacies of the seemingly simple yeast GID N-degron-targeting system, now revealed more than 25 years since the discovery of glucose-induced degradation, provide a blueprint for understanding this important family of multi-subunit E3 ligases.

## STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **LEAD CONTACT AND MATERIALS AVAILABILITY**
- **METHOD DETAILS**
  - Yeast strains and growth conditions
  - Fbp1 degradation assays
  - Affinity Enrichment Mass Spectrometry (AE-MS)
  - Protein digestion of *in vitro* ubiquitylation assays
  - LC-MS/MS sample preparation
  - LC-MS/MS Measurements
  - LC-MS/MS raw data processing
  - Density Fractionation by sucrose gradients
  - Plasmids preparation and Mutagenesis
  - Protein expression and purification for cryo-EM
  - Cryo EM sample preparation and Imaging
  - Data processing
  - Model building and refinement
  - Protein expression and purification for biochemical assays
  - Biochemical assays
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
- **DATA AND CODE AVAILABILITY**

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.molcel.2019.10.009>.

## ACKNOWLEDGMENTS

We thank J. Frye and support from ALSAC/St. Jude for cloning Gid ORFs; A. Varshavsky for plasmids for promoter reference assays in yeast; J. Kellermann for assistance in all work in the Schulman lab; N. Nagaraj, V. Sanchez Cabaliero, N. Krombholz, and A. Wehner for intact mass spectrometry; I. Paron, C. Deiml, and J. Mueller for mass spectrometry maintenance; and M. Strauss, D. Scott, J.W. Harper, and members of the Schulman lab for discussions. This work was supported by the Max Planck Society.

## AUTHOR CONTRIBUTIONS

S.Q. established recombinant GID and prepared cryo-EM samples. S.Q., D.B., T.S., and J.R.P. collected cryo-EM data. S.Q. and J.R.P. determined, refined, and validated atomic models. C.R.L. and V.B. generated yeast strains. C.R.L., V.B., and S.Q. performed yeast biochemistry. S.Q., J.C., and D.S. performed *in vitro* biochemistry. O.K. and F.M.H. performed mass spectrometry supervised by M.M. S.v.G. produced baculoviruses and insect cells expressing

GID complexes. S.Q., J.C., D.S., C.R.L., and B.A.S. wrote the paper. A.F.A. and B.A.S. coordinated and supervised.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 2, 2019

Revised: September 4, 2019

Accepted: October 8, 2019

Published: November 7, 2019

## REFERENCES

Adams, P.D., Afonine, P.V., Bunkóczki, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 66, 213–221.

Afonine, P.V., Poon, B.K., Read, R.J., Sobolev, O.V., Terwilliger, T.C., Urzhumtsev, A., and Adams, P.D. (2018). Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D Struct. Biol.* 74, 531–544.

Alfieri, C., Zhang, S., and Barford, D. (2017). Visualizing the complex functions and mechanisms of the anaphase promoting complex/cyclosome (APC/C). *Open Biol.* 7, 170204.

Boldt, K., van Reeuwijk, J., Lu, Q., Koutrompas, K., Nguyen, T.M., Texier, Y., van Beersum, S.E., Horn, N., Miller, J.R., Mans, D.A., et al.; UK10K Rare Diseases Group (2016). An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat. Commun.* 7, 11491.

Braun, B., Pfirrmann, T., Menssen, R., Hofmann, K., Scheel, H., and Wolf, D.H. (2011). Gid9, a second RING finger protein contributes to the ubiquitin ligase activity of the Gid complex required for catabolite degradation. *FEBS Lett.* 585, 3856–3861.

Brower, C.S., Piatkov, K.I., and Varshavsky, A. (2013). Neurodegeneration-associated protein fragments as short-lived substrates of the N-end rule pathway. *Mol. Cell* 50, 161–171.

Brown, N.G., Watson, E.R., Weissmann, F., Jarvis, M.A., VanderLinden, R., Grace, C.R.R., Frye, J.J., Qiao, R., Dube, P., Petzold, G., et al. (2014). Mechanism of polyubiquitination by human anaphase-promoting complex: RING repurposing for ubiquitin chain assembly. *Mol. Cell* 56, 246–260.

Burnley, T., Palmer, C.M., and Winn, M. (2017). Recent developments in the CCP-EM software suite. *Acta Crystallogr. D Struct. Biol.* 73, 469–477.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66, 12–21.

Chen, S.J., Wu, X., Wadas, B., Oh, J.H., and Varshavsky, A. (2017). An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes. *Science* 355, eaal3655.

Chiang, M.C., and Chiang, H.L. (1998). Vid24p, a novel protein localized to the fructose-1,6-bisphosphatase-containing vesicles, regulates targeting of fructose-1,6-bisphosphatase from the vesicles to the vacuole for degradation. *J. Cell Biol.* 140, 1347–1356.

Chiang, H.L., and Schekman, R. (1991). Regulated import and degradation of a cytosolic protein in the yeast vacuole. *Nature* 350, 313–318.

Choi, W.S., Jeong, B.C., Joo, Y.J., Lee, M.R., Kim, J., Eck, M.J., and Song, H.K. (2010). Structural basis for the recognition of N-end rule substrates by the UBR box of ubiquitin ligases. *Nat. Struct. Mol. Biol.* 17, 1175–1181.

Ciechanover, A. (2012). Intracellular protein degradation: from a vague idea thru the lysosome and the ubiquitin-proteasome system and onto human diseases and drug targeting. *Biochim. Biophys. Acta* 1824, 3–13.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* 62, 1002–1011.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltens, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10, 1794–1805.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 13, 2513–2526.

DiMaio, F., Echols, N., Headd, J.J., Terwilliger, T.C., Adams, P.D., and Baker, D. (2013). Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* 10, 1102–1104.

Dong, C., Zhang, H., Li, L., Tempel, W., Loppnau, P., and Min, J. (2018). Molecular basis of GID4-mediated recognition of degrons for the Pro/N-end rule pathway. *Nat. Chem. Biol.* 14, 466–473.

Dou, H., Buetow, L., Sibbet, G.J., Cameron, K., and Huang, D.T. (2012). BIRC7-E2 ubiquitin conjugate structure reveals the mechanism of ubiquitin transfer by a RING dimer. *Nat. Struct. Mol. Biol.* 19, 876–883.

Duda, D.M., Olszewski, J.L., Tron, A.E., Hammel, M., Lambert, L.J., Waddell, M.B., Mittag, T., DeCaprio, J.A., and Schulman, B.A. (2012). Structure of a globulin-RBX1-CUL1 complex: inhibition of a RING E3 ligase through masking of its E2-binding surface. *Mol. Cell* 47, 371–382.

Emberley, E.D., Mosadeghi, R., and Deshaies, R.J. (2012). Deconjugation of Nedd8 from Cul1 is directly regulated by Skp1-F-box and substrate, and the COP9 signalosome inhibits deneddylated SCF by a noncatalytic mechanism. *J. Biol. Chem.* 287, 29679–29689.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* 60, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* 66, 486–501.

Enchev, R.I., Scott, D.C., da Fonseca, P.C., Schreiber, A., Monda, J.K., Schulman, B.A., Peter, M., and Morris, E.P. (2012). Structural basis for a reciprocal regulation between SCF and CSN. *Cell Rep.* 2, 616–627.

Fernandez-Leiro, R., and Scheres, S.H.W. (2017). A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D Struct. Biol.* 73, 496–502.

Francis, O., Han, F., and Adams, J.C. (2013). Molecular phylogeny of a RING E3 ubiquitin ligase, conserved in eukaryotic cells and dominated by homologous components, the muskelin/RanBPM/CTLH complex. *PLoS ONE* 8, e75217.

Gancedo, J.M. (1998). Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev.* 62, 334–361.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345.

Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., and Ferrin, T.E. (2018). UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* 27, 14–25.

González, A., and Hall, M.N. (2017). Nutrient sensing and TOR signaling in yeast and mammals. *EMBO J.* 36, 397–408.

Guimaraes, C.P., Witte, M.D., Theile, C.S., Bozkurt, G., Kundrat, L., Blom, A.E.M., and Ploegh, H.L. (2013). Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. *Nat. Protoc.* 8, 1787–1799.

Hämmerle, M., Bauer, J., Rose, M., Szallies, A., Thumm, M., Düsterhus, S., Mecke, D., Entian, K.D., and Wolf, D.H. (1998). Proteins of newly isolated mutants and the amino-terminal proline are essential for ubiquitin-proteasome-catalyzed catabolite degradation of fructose-1,6-bisphosphatase of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 273, 25000–25005.

Han, T., Yang, C.S., Chang, K.Y., Zhang, D., Imam, F.B., and Rana, T.M. (2016). Identification of novel genes and networks governing hematopoietic stem cell development. *EMBO Rep.* 17, 1814–1828.

Hoffman, M., and Chiang, H.L. (1996). Isolation of degradation-deficient mutants defective in the targeting of fructose-1,6-bisphosphatase into the vacuole for degradation in *Saccharomyces cerevisiae*. *Genetics* 143, 1555–1566.

Hu, R.G., Sheng, J., Qi, X., Xu, Z., Takahashi, T.T., and Varshavsky, A. (2005). The N-end rule pathway as a nitric oxide sensor controlling the levels of multiple regulators. *Nature* 437, 981–986.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509.

Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., and Knop, M. (2004). A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* 21, 947–962.

Javan, G.T., Salhotra, A., Finley, S.J., and Soni, S. (2018). Erythroblast macrophage protein (Emp): Past, present, and future. *Eur. J. Haematol.* 100, 3–9.

Kaiser, S.E., Riley, B.E., Shaler, T.A., Trevino, R.S., Becker, C.H., Schulman, H., and Kopito, R.R. (2011). Protein standard absolute quantification (PSAQ) method for the measurement of cellular ubiquitin pools. *Nat. Methods* 8, 691–696.

Keilhauer, E.C., Hein, M.Y., and Mann, M. (2015). Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* 14, 120–135.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858.

Klaips, C.L., Hochstrasser, M.L., Langlois, C.R., and Serio, T.R. (2014). Spatial quality control bypasses cell-based limitations on proteostasis to promote prion curing. *eLife* 3, e04288.

Knop, M., Siegers, K., Pereira, G., Zachariae, W., Winsor, B., Nasmyth, K., and Schiebel, E. (1999). Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. *Yeast* 15 (10B), 963–972.

Koren, I., Timms, R.T., Kula, T., Xu, Q., Li, M.Z., and Elledge, S.J. (2018). The eukaryotic proteome is shaped by E3 ubiquitin ligases targeting C-terminal degrons. *Cell* 173, 1622–1635.e14.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* 11, 319–324.

Lampert, F., Stafa, D., Goga, A., Soste, M.V., Gilberto, S., Olieric, N., Picotti, P., Stoffel, M., and Peter, M. (2018). The multi-subunit GID/CTLH E3 ubiquitin ligase promotes cell proliferation and targets the transcription factor Hbp1 for degradation. *eLife* 7, e35528.

Li, J., Li, Z., Ruan, J., Xu, C., Tong, Y., Pan, P.W., Tempel, W., Crombet, L., Min, J., and Zang, J. (2011). Structural basis for specific binding of human MPP8 chromodomain to histone H3 methylated at lysine 9. *PLoS ONE* 6, e25104.

Lin, H.C., Yeh, C.W., Chen, Y.F., Lee, T.T., Hsieh, P.Y., Rusnac, D.V., Lin, S.Y., Elledge, S.J., Zheng, N., and Yen, H.S. (2018). C-terminal end-directed protein elimination by CRL2 ubiquitin ligases. *Mol. Cell* 70, 602–613.e3.

Liu, H., and Pfirrmann, T. (2019). The Gid-complex: an emerging player in the ubiquitin ligase league. *Biol. Chem.* 400, 1429–1441.

Lyapina, S., Cope, G., Shevchenko, A., Serino, G., Tsuge, T., Zhou, C., Wolf, D.A., Wei, N., Shevchenko, A., and Deshaies, R.J. (2001). Promotion of NEDD-CUL1 conjugate cleavage by COP9 signalosome. *Science* 292, 1382–1385.

Lydeard, J.R., Schulman, B.A., and Harper, J.W. (2013). Building and remodelling Cullin-RING E3 ubiquitin ligases. *EMBO Rep.* 14, 1050–1061.

Mastronarde, D. (2003). SerialEM: a program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position. *Microsc. Microanal.* 9, 1182–1183.

Matta-Camacho, E., Kozlov, G., Li, F.F., and Gehring, K. (2010). Structural basis of substrate recognition and specificity in the N-end rule pathway. *Nat. Struct. Mol. Biol.* 17, 1182–1187.

Melnykov, A., Chen, S.J., and Varshavsky, A. (2019). Gid10 as an alternative N-recognin of the Pro/N-degron pathway. *Proc. Natl. Acad. Sci. U S A* 116, 15914–15923.

Menssen, R., Schweiggert, J., Schreiner, J., Kusevic, D., Reuther, J., Braun, B., and Wolf, D.H. (2012). Exploring the topology of the Gid complex, the E3 ubiquitin ligase involved in catabolite-induced degradation of gluconeogenic enzymes. *J. Biol. Chem.* 287, 25602–25614.

Menssen, R., Bui, K., and Wolf, D.H. (2018). Regulation of the Gid ubiquitin ligase recognition subunit Gid4. *FEBS Lett.* 592, 3286–3294.

Mitchell, A., Romano, G.H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O., and Pilpel, Y. (2009). Adaptive prediction of environmental changes by microorganisms. *Nature* 460, 220–224.

Nguyen, A.T., Prado, M.A., Schmidt, P.J., Sendamarai, A.K., Wilson-Grady, J.T., Min, M., Campagna, D.R., Tian, G., Shi, Y., Dederer, V., et al. (2017). UBE2O remodels the proteome during terminal erythroid differentiation. *Science* 357, eaan0218.

Oh, J.H., Chen, S.J., and Varshavsky, A. (2017). A reference-based protein degradation assay without global translation inhibitors. *J. Biol. Chem.* 292, 21457–21465.

Petroski, M.D., and Deshaies, R.J. (2005). Mechanism of lysine 48-linked ubiquitin-chain synthesis by the cullin-RING ubiquitin-ligase complex SCF-Cdc34. *Cell* 123, 1107–1120.

Petersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.

Pfirrmann, T., Villavicencio-Lorini, P., Subudhi, A.K., Menssen, R., Wolf, D.H., and Hollemann, T. (2015). RMND5 from *Xenopus laevis* is an E3 ubiquitin-ligase and functions in early embryonic forebrain development. *PLoS ONE* 10, e0120342.

Pierce, N.W., Lee, J.E., Liu, X., Sweredoski, M.J., Graham, R.L., Larimore, E.A., Rome, M., Zheng, N., Clurman, B.E., Hess, S., et al. (2013). Cand1 promotes assembly of new SCF complexes through dynamic exchange of F box proteins. *Cell* 153, 206–215.

Plechanovová, A., Jaffray, E.G., Tatham, M.H., Naismith, J.H., and Hay, R.T. (2012). Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. *Nature* 489, 115–120.

Pruneda, J.N., Littlefield, P.J., Sosa, S.E., Nordquist, K.A., Chazin, W.J., Brzovic, P.S., and Klevit, R.E. (2012). Structure of an E3:E2-Ub complex reveals an allosteric mechanism shared among RING/U-box ligases. *Mol. Cell* 47, 933–942.

Rao, H., Uhlmann, F., Nasmyth, K., and Varshavsky, A. (2001). Degradation of a cohesin subunit by the N-end rule pathway is essential for chromosome stability. *Nature* 410, 955–959.

Regelmann, J., Schüle, T., Josupeit, F.S., Horak, J., Rose, M., Entian, K.D., Thumm, M., and Wolf, D.H. (2003). Catabolite degradation of fructose-1,6-bisphosphatase in the yeast *Saccharomyces cerevisiae*: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol. Biol. Cell* 14, 1652–1663.

Rosenthal, P.B., and Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* 333, 721–745.

Rusnac, D.V., Lin, H.C., Canzani, D., Tien, K.X., Hinds, T.R., Tsue, A.F., Bush, M.F., Yen, H.S., and Zheng, N. (2018). Recognition of the diglycine C-end degron by CRL2(KLHDC2) ubiquitin ligase. *Mol. Cell* 72, 813–822.e4.

Santt, O., Pfirrmann, T., Braun, B., Juretschke, J., Kimmig, P., Scheel, H., Hofmann, K., Thumm, M., and Wolf, D.H. (2008). The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism. *Mol. Biol. Cell* 19, 3323–3333.

Saxton, R.A., and Sabatini, D.M. (2017). mTOR signaling in growth, metabolism, and disease. *Cell* 168, 960–976.

Scheres, S.H. (2012). A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* 415, 406–418.

Schork, S.M., Bee, G., Thumm, M., and Wolf, D.H. (1994a). Catabolite inactivation of fructose-1,6-bisphosphatase in yeast is mediated by the proteasome. *FEBS Lett.* 349, 270–274.

Schork, S.M., Bee, G., Thumm, M., and Wolf, D.H. (1994b). Site of catabolite inactivation. *Nature* 369, 283–284.

Schork, S.M., Thumm, M., and Wolf, D.H. (1995). Catabolite inactivation of fructose-1,6-bisphosphatase of *Saccharomyces cerevisiae*. Degradation occurs via the ubiquitin pathway. *J. Biol. Chem.* 270, 26446–26450.

Schüle, T., Rose, M., Entian, K.D., Thumm, M., and Wolf, D.H. (2000). Ubc8p functions in catabolite degradation of fructose-1, 6-bisphosphatase in yeast. *EMBO J.* 19, 2161–2167.

Scott, D.C., Sviderskaya, V.O., Monda, J.K., Lydeard, J.R., Cho, S.E., Harper, J.W., and Schulman, B.A. (2014). Structure of a RING E3 trapped in action reveals ligation mechanism for the ubiquitin-like protein NEDD8. *Cell* 157, 1671–1684.

Shemorry, A., Hwang, C.S., and Varshavsky, A. (2013). Control of protein quality and stoichiometries by N-terminal acetylation and the N-end rule pathway. *Mol. Cell* 50, 540–551.

Soni, S., Bala, S., Gwynn, B., Sahr, K.E., Peters, L.L., and Hanspal, M. (2006). Absence of erythroblast macrophage protein (Emp) leads to failure of erythroblast nuclear extrusion. *J. Biol. Chem.* 281, 20181–20189.

Storici, F., and Resnick, M.A. (2006). The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods Enzymol.* 409, 329–345.

Szoradi, T., Schaeff, K., Garcia-Rivera, E.M., Itzhak, D.N., Schmidt, R.M., Bircham, P.W., Leiss, K., Diaz-Miyar, J., Chen, V.K., Muzzey, D., et al. (2018). SHRED is a regulatory cascade that reprograms Ubr1 substrate specificity for enhanced protein quality control during stress. *Mol. Cell* 70, 1025–1037.e5.

Tagkopoulos, I., Liu, Y.C., and Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. *Science* 320, 1313–1317.

Timms, R.T., Zhang, Z., Rhee, D.Y., Harper, J.W., Koren, I., and Elledge, S.J. (2019). A glycine-specific N-degron pathway mediates the quality control of protein N-myristylation. *Science* 365, eaaw4912.

Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., and Cox, J. (2015). Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* 15, 1453–1456.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (proteo)omics data. *Nat. Methods* 13, 731–740.

Varshavsky, A. (2011). The N-end rule pathway and regulation by proteolysis. *Protein Sci.* 20, 1298–1345.

Varshavsky, A. (2012). The ubiquitin system, an immense realm. *Annu. Rev. Biochem.* 81, 167–176.

Varshavsky, A. (2019). N-degron and C-degron pathways of protein degradation. *Proc. Natl. Acad. Sci. U S A* 116, 358–366.

Wang, K.H., Roman-Hernandez, G., Grant, R.A., Sauer, R.T., and Baker, T.A. (2008). The molecular basis of N-end rule recognition. *Mol. Cell* 32, 406–414.

Wanichthanarak, K., Nookaew, I., and Petranovic, D. (2014). yStreX: yeast stress expression database. *Database (Oxford)* 2014, bau068.

Watson, E.R., Brown, N.G., Peters, J.M., Stark, H., and Schulman, B.A. (2019). Posing the APC/C E3 Ubiquitin Ligase to Orchestrate Cell Division. *Trends Cell Biol.* 29, 117–134.

Weissmann, F., Petzold, G., VanderLinden, R., Huis In 't Veld, P.J., Brown, N.G., Lampert, F., Westermann, S., Stark, H., Schulman, B.A., and Peters, J.M. (2016). biGBac enables rapid gene assembly for the expression of large

### 3.2.5. Linkage-specific ubiquitin chain formation depends on a lysine hydrocarbon ruler

Joanna Liwocha<sup>1,8</sup>, David T. Krist<sup>1,2,8</sup>, Gerbrand J. van der Heden van Noort<sup>3,8</sup>, Fynn M. Hansen<sup>4</sup>, Vinh H. Truong<sup>5</sup>, Ozge Karayel<sup>4</sup>, Nicholas Purser<sup>6</sup>, Daniel Houston<sup>6</sup>, Nicole Burton<sup>6</sup>, Mark Bostock<sup>7</sup>, Michael Sattler<sup>7</sup>, Matthias Mann<sup>4</sup>, Joseph S. Harrison<sup>5</sup>, Gary Kleiger<sup>6,\*</sup>, Huib Ovaar<sup>3,\*</sup>, Brenda A. Schulman<sup>1,\*</sup>

<sup>1</sup> Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, 82152, Germany

<sup>2</sup> Present address: Carle Illinois College of Medicine, Champaign, IL, 61820, USA <sup>3</sup> Oncode Institute and Department of Cell and Chemical Biology, Chemical Immunology, Leiden University Medical Centre, Leiden, the Netherlands

<sup>4</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany

<sup>5</sup> Department of Chemistry, University of the Pacific, Stockton, California, 95211, USA <sup>6</sup> Department of Chemistry and Biochemistry, University of Nevada, Las Vegas, Las Vegas, NV, 89154, USA <sup>7</sup> Biomolecular NMR and Center for Integrated Protein Science Munich at Department Chemie, Technical University of Munich, Garching & Germany Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany <sup>8</sup> Co-first author

Under revision in *Nature Chemical Biology* (2020)

Ubiquitin proteins can potentially be conjugated through one of their eight amino groups (seven lysines and the N-terminus), giving rise to a plethora of chain topologies. These different chains can be recognized by downstream readers of the ubiquitin code that determines the fates of modified proteins. Therefore, it is important to uncover features underlying the generation of specific chain linkages. Structural studies have shown how different ubiquitin chains vary in compactness and rigidity depending upon the amine through which they are linked [170]. In the course of chain formation, the lysines in previously attached ubiquitin serve as acceptors for the next ubiquitin molecule forming poly-Ub chains, however little is known about the role of the acceptor Lys in UB chain formation. Researchers from the Schulman group aimed at investigating the potential roles of acceptor lysine architecture on ubiquitylation. We contributed to this study by establishing a targeted MS strategy to quantify the distribution of Ub chain linkage types. We showed that the location of the ubiquitin lysine mutation determines the distribution of di-ubiquitin linkage types.

## Linkage-specific ubiquitin chain formation depends on a lysine hydrocarbon ruler

Joanna Liwocha<sup>1,8</sup>, David T. Krist<sup>1,2,8</sup>, Gerbrand J. van der Heden van Noort<sup>3,8</sup>, Fynn M. Hansen<sup>4</sup>, Vinh H. Truong<sup>5</sup>, Ozge Karayel<sup>4</sup>, Nicholas Purser<sup>6</sup>, Daniel Houston<sup>6</sup>, Nicole Burton<sup>6</sup>, Mark Bostock<sup>7</sup>, Michael Sattler<sup>7</sup>, Matthias Mann<sup>4</sup>, Joseph S. Harrison<sup>5</sup>, Gary Kleiger<sup>6,\*</sup>, Huib Ovaar<sup>3,\*</sup>, Brenda A. Schulman<sup>1,\*</sup>

<sup>1</sup>Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, Martinsried, 82152, Germany

<sup>2</sup>Present address: Carle Illinois College of Medicine, Champaign, IL, 61820, USA

<sup>3</sup>Oncode Institute and Department of Cell and Chemical Biology, Chemical Immunology, Leiden University Medical Centre, Leiden, the Netherlands.

<sup>4</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany

<sup>5</sup>Department of Chemistry, University of the Pacific, Stockton, California, 95211, USA

<sup>6</sup>Department of Chemistry and Biochemistry, University of Nevada, Las Vegas, Las Vegas, NV, 89154, USA

<sup>7</sup>Biomolecular NMR and Center for Integrated Protein Science Munich at Department Chemie, Technical University of Munich, Garching & Germany Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany

<sup>8</sup>Co-first author

\*Corresponding author: gary.kleiger@unlv.edu (G.K), H.Ovaar@lumc.nl (H.O, deceased) schulman@biochem.mpg.de (B.A.S)

### Abstract

Virtually all aspects of cell biology are regulated by a ubiquitin code where distinct ubiquitin chain architectures guide the binding events and itineraries of modified substrates. Various combinations of E2 and E3 enzymes accomplish chain formation by forging isopeptide bonds between the C-terminus of their transiently-linked donor ubiquitin and a specific nucleophilic amino acid on the acceptor ubiquitin, yet it is unknown whether the fundamental feature of most acceptors - the lysine side-chain - affects catalysis. Here, use of synthetic ubiquitins with non-natural acceptor site replacements reveals that the aliphatic side-chain specifying reactive amine geometry is a determinant of the ubiquitin code, through unanticipated and complex reliance of many distinct ubiquitin carrying enzymes on a canonical acceptor lysine.

### Introduction

Ubiquitin (UB) chains are a major post-translational modification controlling protein function in eukaryotic cells<sup>1-3</sup>. Numerous, distinct chains result from the linkage of UB's C-terminus to one of eight amino group acceptors (seven lysines and the N-terminus) on another UB. The different UB chains, which vary in compactness and rigidity, formulate a "UB code" read by cognate UB-chain-specific binding domains that control the fates of modified proteins<sup>1-3</sup>. Studies of endogenous and recombinant proteins have shed light on this code, for example showing that chains linked through UB's lysine 48

(K48) often direct protein targets for proteasomal degradation<sup>1-3</sup>. Chains linked through UB's lysine 63 (K63) often regulate the assembly of multi-subunit complexes and diverse biological processes such as membrane protein trafficking and DNA repair<sup>1-3</sup>. Furthermore, structural studies have shown how specificity is dictated by the spacing between hydrophobic patches displayed from the different UB molecules linked in various chain types<sup>4</sup>. In some cases, the actual isopeptide linkages between UBs, and surrounding residues, also determine recognition of specific UB chains<sup>4,5</sup>.

Recently, chemical biology approaches have also elucidated principles governing important aspects of UB biology<sup>6,7</sup>. Indeed, the chemical synthesis of UB chains with defined linkages and chemically unique properties have illuminated mechanisms underlying protein degradation by the proteasome<sup>8,9</sup> and revealed the potential of hundreds of UB-binding domains to partner with their cognate chain types<sup>10</sup>. Synthetic UBs have also aided in the discovery of small molecules that bind K48-linked chains<sup>11</sup>.

Despite this progress in deciphering how the code is “read” by downstream machineries recognizing UB chains, the mechanisms underlying generation of specific UB chain linkages remain incompletely understood. Chains are forged by combinations of UB-conjugating enzymes (E2s) and UB ligases (E3s)<sup>12,13</sup>. In humans, various pairings amongst  $\approx 30$  E2s and  $\approx 600$  E3s mediate the linkage of UB to selected target proteins, and determine their fates through the generation of UB chains with specific linkages. Different E2s and E3s employ distinct enzymatic mechanisms to achieve polyubiquitylation<sup>12,13</sup>. For example, some E2 enzymes can generate UB chains themselves, whereby after enzymatic linkage of UB's C-terminus to an E2 catalytic cysteine, UB is transferred from the resultant E2-UB intermediate (~ refers to thioester bond) to a lysine on an “acceptor UB”<sup>13</sup>. The preferred UB acceptor lysine may be intrinsic to an E2, and/or may be influenced by an E2 partner protein<sup>14</sup>. In some cases, UB transfer from the E2 is stimulated by the hallmark “RING” domain in many E3s<sup>12,13</sup>. An E2 may also transfer UB to a catalytic cysteine of some E3 classes, such as in a “HECT” catalytic domain, from which UB is ultimately linked to another UB to generate a chain<sup>12,13</sup>.

Many prior studies have identified E2 or E3 residues critical for catalysis, presentation of an acceptor UB by accessory domains, and roles of acceptor UB residues surrounding the targeted lysine<sup>14-20</sup>. However, whether features of a UB's target lysine beyond its nucleophilic primary amino group - such as the distance between the primary amine and the UB polypeptide backbone - influence UB chain formation remains unknown. Within different classes of UB carrying enzymes (e.g. E2 or HECT E3) catalytic domains adopt highly similar structures that have been shown to catalyze covalent bond formation between donor UB and disparate free amino acids - lysine, cysteine, serine, and threonine - acceptors<sup>21-25</sup>. Because substrates of the UB system are often ubiquitylated and degraded even after mutation of their preferred lysines, one has the impression that targeting by at least some enzymes - and certainly across classes of enzymes - is relatively lax. This contrasts, however, from classes of protein interaction domains or histone modifying enzyme families that strictly depend on lysines for specific salt-bridge geometries or substrate targeting<sup>26,27</sup>.

Thus, to investigate if acceptor lysine side-chain features beyond the primary amino group influence UB chain formation, we employed a series of synthetic UBs harboring replacements for K11, K48 or K63 with progressively shorter or longer aliphatic side chains, and tested their reactivities with a broad set of ubiquitylating enzymes. Our results demonstrate that the geometry between the polypeptide backbone and the primary amine strongly influences chain formation for numerous diverse UB chain forming enzymes. Thus, the lysine aliphatic side-chain itself helps to establish the UB code.

### A K63-specific E2 enzyme displays exquisite sensitivity to acceptor UB lysine architecture

The simplest activity of an E2 is the ability to transfer UB to a nucleophilic amino acid free in solution. For some E2s, there is a correlation between the discharge onto a free amino acid acceptor (e.g. lysine, cysteine, threonine, etc.) and the preferred residue-type modified in the context of a protein target<sup>22,25</sup>. We examined reactivity of the well-characterized, human heterodimeric UB-chain forming E2 UBE2N/UBE2V1 complex<sup>19,28,29</sup>. UBE2N/UBE2V1 is distinguished by the partnering of a canonical E2 subunit (UBE2N) with the catalytically inactive dedicated E2-like UBE2V1. UBE2N/UBE2V1 generates nearly pure Lys63-linked UB chains owing to the fact that Lys63 on the acceptor UB is guided by the UBE2V1 subunit towards the thioester linkage between UBE2N's active site cysteine and the donor UB's C-terminus<sup>16,28</sup>. The rate of this UB chain formation is accelerated in the presence of the isolated dimeric RING domain of an E3 (hereafter referred to as RNF4) that stabilizes the active conformation of the donor UB thioester-bonded to the UBE2N active site<sup>30</sup>. Moreover, coupling with both UBE2V1 and RNF4 stimulates intrinsic reactivity of the UBE2N-ubiquitin intermediate, as monitored by ubiquitin discharge to free lysine<sup>29</sup>, albeit with far less efficiency than to an acceptor UB's Lys63.

We examined the ability of RNF4-UBE2V1-activated UBE2N to transfer a donor UB to various free amino acid acceptors using a pulse-chase assay format (Fig. 1a). First, UBE2N was charged with fluorescent donor UB in the pulse reaction using E1 enzyme. After the pulse reaction was quenched, the resultant UBE2N-UB intermediate was incubated with RNF4, UBE2V1 and various unlabeled amino acids. We initially tested L-lysine (four methylene units in the side-chain, referred to here as C4) and two controls: L-serine which is not known to accept UB from RNF4-UEV1-UBE2N, and N<sub>ε</sub>-Acetyl-L-lysine wherein the epsilon amino group is blocked. As expected, L-lysine had high reactivity when compared to the other amino acids (Figure 1b). Next, we verified that the N<sub>ε</sub>-amine of lysine is the preferred acceptor by observing robust reactivity with N<sub>ε</sub>-Acetyl-L-lysine, wherein the alpha amino is blocked but the epsilon amino group is available. With this established, we then tested a range of lysine analogues that differ according to their side-chain length. Here, the C1, C3 and C5 analogues (L-2,3-diaminopropionic acid, N<sub>ε</sub>-Acetyl-L-ornithine, and L-homolysine, respectively) demonstrated robust reactivity (Fig. 1b,c and Extended Data Fig. 1a,b), indicating the absence of an absolute requirement for length of the aliphatic chain between the backbone and the nucleophilic amino group of lysine analogues free in solution.

Since various amino acids with a side-chain primary amino group were capable of preferentially accepting E2-bound donor UB relative to those lacking such a chemical group, we next wondered generally how L-lysine architecture within the context of an acceptor UB would affect UBE2N/UBE2V1 catalytic activity (Fig. 2a). Accordingly, solid-phase peptide synthesis was employed to generate UBs with Lys63 analogs that differ in length by the number of methylene groups - one, two, three, four or five - between the alpha carbon and side-chain amino group: L-2,3-diaminopropionic acid (Dap, referred to here as <sup>K63</sup>UB<sub>C1</sub> for one methylene group in the analog replacing native Lys63), L-2,4-diaminobutyric acid (Dab, referred to here as <sup>K63</sup>UB<sub>C2</sub> for two methylene groups in the analog replacing Lys63), L-ornithine (Om, referred to here as <sup>K63</sup>UB<sub>C3</sub> for three methylene groups in the analog replacing Lys63), L-lysine (Lys, referred to here as <sup>K63</sup>UB<sub>C4</sub> for the four methylene groups in the native acceptor), and L-homolysine (hLys, referred to here as <sup>K63</sup>UB<sub>C5</sub> for five methylene groups in the analog replacing Lys63) (Fig. 2b).

UBE2N/UBE2V1 activity was once again measured using a pulse-chase assay, with the difference being that the acceptor is now UB, and product formation between donor and acceptor UBs results in a di-UB chain. Remarkably, unlike in the discharge to a free amino acid, the removal or addition of only a single

methylene from a canonical Lys63 side-chain now greatly reduces product formation of intrinsic E2-catalyzed free UB chain formation. Moreover, the striking preference for the native lysine persisted upon adding the RNF4 E3 enzyme RING domain that accelerated the reaction (Fig. 2c and Extended Data Fig. 2a).

#### Native lysine is preferred for K48-chain linkage on E3-bound substrates

Since the recruitment of acceptor UB by UBE2N is unique in that it depends on its partner UBE2V1 to bind and orient the acceptor UB<sup>29</sup>, UBE2V1's grip on the acceptor UB may limit the ability of the reactive amine to reposition in the active site upon the addition or removal of a methylene group. Thus, we wondered how changes to the lysine architecture would affect other E2s that rely on their own surfaces to orient the acceptor UB. Hence, two K48 linkage-specific E2s, UBE2G1 and UBE2R2, were assayed in the presence of the K<sup>48</sup>UB<sub>C1-C5</sub> suite<sup>31,32,33</sup>. Again significant di-UB product is only observed with the K<sup>48</sup>UB<sub>C4</sub> acceptor– for the E2s alone, as well as for UBE2R2 and UBE2G1 reactions stimulated by cullin-RING ligase E3s CRL1 or CRL4, respectively<sup>31-35</sup> (Fig. 2d and Extended Data Fig. 2b).

We considered that the reactivity of the acceptor UB may differ substantially when attached to an E3-bound substrate<sup>36,37</sup>. CRLs are multiprotein complexes, wherein a specific receptor protein recruits a substrate degron motif. A well-studied CRL1 substrate receptor is FBW7, a tumor suppressor protein that recruits phosphorylated targets including the cell cycle regulator Cyclin E<sup>38</sup>. A well-studied CRL4 receptor is CCRN, for which chemotherapeutic thalidomide analogs, such as pomalidomide, induce recognition of neosubstrates including transcription factors in the Ikaros family. The activities of CRL1<sup>FBW7</sup> and CRL4<sup>CCRN</sup> can be reconstituted with truncated substrate peptides harboring the degron motif and an acceptor site. To generate such substrates with our synthetic UBs suitable for assaying di-UB chain formation, sortase-mediated transpeptidation was employed to append degron peptides from both CRL1- and CRL4-based substrates (a cyclin E phosphopeptide and IKZF1 zinc finger, respectively) to each UB from the K<sup>48</sup>UB<sub>C1-C5</sub> suite. To test how lysine architecture affects CRL-mediated UB chain formation, these substrates were reacted in the presence of the relevant E3s and UBE2R2 or UBE2G1, respectively<sup>34,35</sup>. Even in this context, only the native lysine supported substantial UB-chain elongation onto CRL-bound substrates (Figure 2d and Extended Data Fig. 2b).

As controls, the suite of K<sup>63</sup>UB<sub>C1-C5</sub> analogs served as acceptors with the Lys48-specific E2 UBE2G1, demonstrating proper folding for all of the synthetic UBs harboring Lys63 substitutions (Fig. 2e and Extended Data Fig. 2c). Similarly, UBE2N/UBE2V1 also produces nearly equivalent amounts of di-UB chains with UB mutants harboring the lysine analogs on the non-acceptor position 48 (K<sup>48</sup>UB<sub>C1-C5</sub>, Fig. 2f and Extended Data Fig. 2d). Moreover, 1D and 2D proton NMR spectra for recombinant UB (aka C4-bio), synthetic UB (aka C4), and K<sup>48</sup>UB<sub>C5</sub> showed good dispersion and were superimposable except for a few resonances, presumably reflecting overall minor impact of the sequence differences between the three forms of UB (Extended Data Fig. 3). Thus, taken together, the data show that K63- and K48-specific E2 enzymes that utilize distinct modes of acceptor UB recruitment display exquisite specificity for the attacking lysine architecture in the context of an acceptor UB.

#### NEDD4-family HECT E3s discharge UB predominantly onto an acceptor UB lysine but not to side-chain analogs

Next, we examined if this remarkable preference for native lysine is preserved in the HECT E3 ligase family, where UB is first transferred from E2 to the HECT domain catalytic cysteine and then discharged onto the substrate lysine<sup>39</sup>. We assayed NEDD4-1 HECT domain<sup>40</sup>, the eponymous member of the NEDD4 family of HECT ligases, and a truncated version of its yeast ortholog Rsp5p, that harbors both

a substrate-binding WW domain and the catalytic domain. Both are well-known to form K63-linked chains<sup>40,41</sup>. Robust di-UB formation is only observed with native lysine acceptor <sup>K63</sup>UBc4, while greatly reduced di-UB formation is observed with <sup>K63</sup>UBc1-c3 or <sup>K63</sup>UBc5 (Fig. 3a and Extended Data Fig. 4a,c). Similar to the reactions assayed above, discharge is maintained when Rsp5p and NEDD4-1 are assayed in the presence of the <sup>K48</sup>UBc1-c5 analogues, which preserve the native lysine at position 63 (Fig. 3b and Extended Data Fig. 4a,c).

Next, we wondered if lysine length influences HECT reactions in the context of a substrate. We again turned to a system where a well-characterized degron motif can serve to recruit a substrate. Sna4p, a substrate of Rsp5p, displays a WW-domain binding PPPY motif<sup>42</sup>. We used sortase-mediated transpeptidation to attach each of the <sup>K63</sup>UBc4-c5 series to a truncated Sna4p peptide harboring the degron motif. Efficient discharge was observed only for native lysine but not a K63R mutant, confirming site-specific di-UB formation in this system. Notably, the Lys63 C5 UB was defective even when recruited via the substrate degron (Fig. 3c and Extended Data 4b).

#### Lysine side-chain architecture impacts ratios of less preferred UB chains produced by the promiscuous E2 UBE2D3

Given the results so far, we pondered whether there may be exceptions to linkage specific ubiquitylation systems relying on native lysine (C4) acceptors. The E2 UBE2S, which generates Lys11 di-UB linkages, seemed like an intriguing candidate, as it is thought that UBE2S catalysis is complemented by acceptor UB side-chains<sup>17</sup>. UBE2S displays weak di-ubiquitin chain synthesis activity on its own, owing to a high Km for the acceptor<sup>17</sup>. This is overcome by fusing a UB-binding domain to the C-terminus to UBE2S, or with the multiprotein E3 enzyme Anaphase-Promoting Complex/Cyclosome (APC/C) whose RING domain recruits the acceptor UB<sup>43,44</sup>. Results from our qualitative assay suggest that this mechanism is less sensitive to lysine side-chain length, as di-UB formation occurred with a suite of <sup>K11</sup>UBc2-c5 acceptors, although <sup>K11</sup>UBc1 was relatively inactive (Extended Data Fig. 5). Notably, a prior study showed that a specific UBE2S glutamate (E34) is important for orienting the Lys11 acceptor<sup>17</sup>, although an E34D replacement is fully functional. We tested whether there is a correlation with acceptor side-chain length in our assays. An E34D replacement did not overtly increase sensitivity to acceptor lysine substitutions, again indicating greater tolerance by UBE2S (Extended Data Fig. 5c).

Motivated by the observation that not all enzymes are sensitive to lysine architecture, we next sought to characterize the UB chain forming enzyme UBE2D3 (aka UBCH5C) as its activity is particularly perplexing. UBE2D3 is promiscuous and thought to be nonspecific: UBE2D3 collaborates with a large swath of E3s, can transfer UB to numerous protein substrates as well as to active site cysteines in HECT- and RBR-family E3s, generates a range of UB chain linkages including Lys11, Lys48 and Lys63 acceptors, and forms branched UB chains in multiple turnover, polyubiquitylation conditions<sup>13,45-48</sup>. We thus established pulse-chase assays to examine the di-UB products of UBE2D3-UB, which revealed preferential targeting to Lys11 and Lys63 according to absolute quantitation by mass spectrometry (Fig. 4a). As such, we were intrigued by the observation that when a <sup>K48</sup>UBc5 acceptor is added to UBE2D3-UB, the SDS-PAGE mobility of the di-UB products differed from those formed with a native UB acceptor. Because different UB chain linkages could impact compactness and other properties influencing electrophoretic migration, the result hinted at distinct di-UB products (Fig. 4b).

We developed a targeted mass spectrometry strategy to quantify the distribution of the UB chain linkage types formed with the native lysines. Although the method does not detect chains linked to the unnatural amino acid, it quantifies the relative UB linkages to the remaining lysines in the in reactions with <sup>K11</sup>UBc5, <sup>K48</sup>UBc5, or <sup>K63</sup>UBc5 compared to reactions with C4 acceptor UB (see methods). The addition of <sup>K11</sup>UBc5 or <sup>K63</sup>UBc5 acceptors to UBE2D3-UB result in a distribution of the measured di-UB linkage

types similar to the one observed with UB<sub>c4</sub> (Fig. 4c,d and Extended Data Fig. 6). However, adding an extra methylene group to the side-chain at Lys48 does alter the distribution of di-UB species formed. The change between preferred acceptors could be accounted for by two observations. First, there is a relative redistribution/redistribution from Lys63- to Lys11-linkages. Second, although the di-UBs linked via Lys27, Lys29, and Lys33 remain a minor proportion of the total, the utilization of these non-preferred acceptors increased compared to UB<sub>c4</sub>. Thus, the location on UB is a determinant of the requirement for a UB<sub>c4</sub> acceptor for a multifunctional ubiquitylating enzyme.

#### Molecular Dynamics simulations suggest pleiotropic effects of an additional lysine side-chain methylene

To gain insights into potential structural effects of adding a methylene group to the acceptor side-chain, we performed Molecular Dynamics (MD) simulations on native UB<sub>c4</sub> or UB<sub>c5</sub> at position 11, 48, or 63. Two independent simulations were performed over 50 ns. For each variant, the overall structure of the UB globular domain (excluding residues 71-76) was preserved, with 1.618 Å (native), 1.271 Å (residue 11), 1.209 Å (residue 48), and 1.494 Å (residue 63) average C $\alpha$  RMSD across the simulations. Nonetheless, during these simulations, we observed differences at the C5 site when compared with the native Lys side-chain: (1) The additional methylene group increases the potential range of distances between the amino acid alpha carbon and side-chain amine. Examining the distributions of this distance shows limited overlap between the relative side-chain amine position in the native Lys and C5 at all three positions, with the mean distance consistently larger for C5 (Fig. 5a). This would effectively impart a greater radius to the C5 side-chain primary amine when considering the backbone as the axis of rotation. (2) The additional methylene group expanded the number of potential rotamers from 81 to 273. While only a subset of rotamers were observed across the MD simulations, more were accessed by the C5 side-chain in every case (C4:C5 ratios: residue 11 56:82; residue 48 65:117; residue 63 43:96). (3) The dynamics for Chi angles also differ. In particular, Chi4 oscillates for C5 fluctuates more frequently between the three rotamer bins. This would cause more rapid fluctuation of the side-chain amine position (Fig. 5b and Extended Data Fig. 7a). (4) With C5 at positions 48 and 63, there were also subtle, but reproducible, increases in fluctuations in phi and psi angles and an increased number of allowable phi/psi combinations observed in the simulations (C4:C5 ratios: residue 11 185:175; residue 48 138:169; residue 63 73:90) (Fig. 5c,d). Collectively, between backbone and rotamer combinations we typically observed more states accessible to the C5 residue, with a noticeable increase, of over 1000 additional states, occurring at positions 48 and 63 (C4:C5 ratios: position 11 2942:3016; position 48 2942:4261; position 63 1188:2561).

We further probed the effect of the C5 side-chain by modeling the UBE2N-donor UB/UBE2V1/acceptor UB intermediate, building on the crystal structure of a chemically-stabilized mimic of this complex, and on prior studies of UB discharge from UBE2N<sup>49,50</sup>. We used Rosetta to construct a starting model, where the deprotonated acceptor side-chain amine group is in a tetrahedral approach in relation to the donor UB G76 alpha carbon and we selected a starting model based on the lowest RMSD to UBE2V1 position. Three independent MD simulations (25 ns each) probed the position of native acceptor K63 or the C5 substitution relative to the UBE2N-UB active site. On average, the MD trajectories for the native acceptor K63 hovered in the vicinity of the thioester. By contrast, its C5 counterpart departed the active site within the first 20% of the MD simulation, and did not return, suggesting that the additional conformational states afforded to K63 C5 could frustrate prolonged positioning of near-transition state intermediates of nucleophilic attack (Fig. 5e and Extended Data Fig. 7b).

### Diverse effects of acceptor lysine architecture on kinetic parameters of UB chain formation by linkage specific enzymes

Quantitative biochemical experiments were performed to estimate kinetic constants and illuminate mechanistic roles for  $\text{UB}_{\text{C}4}$ . Upon substantially increasing reaction time as well as protein levels under steady-state conditions, di-UB formation was now detectable for  $\text{UBE2N/UBE2V1}$  with  $\text{K}^{63}\text{UB}_{\text{C}5}$  acceptor. Here, the value of  $k_{\text{cat}}$  was 17-fold lower for  $\text{K}^{63}\text{UB}_{\text{C}5}$  in comparison with  $\text{K}^{63}\text{UB}_{\text{C}4}$  (Table 1), consistent with the striking results from the pulse-chase assays (Fig. 2).

While defects in enzyme activity can manifest themselves through a variety of perturbations, the failure to activate the acceptor lysine amine or weakening of the binding affinity of the acceptor UB for the E2 are quite common<sup>16,17,51</sup>. Based on a pioneering investigation of the related modification sumoylation, E2s are thought to catalyze ubiquitylation at least in part through the active site complementing the acceptor lysine to achieve  $pK_a$  suppression<sup>51</sup>. More specifically, several residues within the E2 active site considerably suppress the substrate lysine  $pK_a$  through stabilizing interactions that compensate for the energetic penalty of desolvation<sup>19,49,51</sup>. As such,  $\text{UBE2N/UBE2V1}$  activity (with a K92R mutation to decrease autoubiquitylation at high pH<sup>20</sup>) was measured in the presence of  $\text{K}^{63}\text{UB}_{\text{C}4}$  or  $\text{K}^{63}\text{UB}_{\text{C}5}$  across varying pHs (Table 1, Extended Data Fig. 8 a-c). While caution should be taken when interpreting apparent  $pK_a$  values, since both  $k_{\text{cat}}$  and  $K_M$  may display pH dependencies of their own<sup>52</sup>, the data fit best to a model where a single ionizing species is responsible for the pH dependency of  $k_{\text{obs}}$  (Extended Data Fig. 8c). Surprisingly, both C4 and C5 UBs share similar apparent  $pK_a$  values (8.9 and 9.0, respectively; Table 1), suggesting that the observed defect in  $k_{\text{cat}}$  may be due to factors other than desolvation of the acceptor amine group. The estimated  $K_M$  values of  $\text{K}^{63}\text{UB}_{\text{C}4}$  or  $\text{K}^{63}\text{UB}_{\text{C}5}$  for  $\text{UBE2N/UBE2V1}$  were within 2-fold, suggesting that both UBs have similar affinity for the  $\text{UBE2N/UBE2V1}$  complex (Table 1 and Extended Data Fig. 8d,e). Thus, defective catalysis seemingly arises from other effects of the additional methylene in the acceptor UB side-chain.

We also considered that the presence of an E3 may affect acceptor UB lysine specificity as E3s commonly restrict the position of the donor UB-E2 thioester<sup>14</sup>. Indeed, the RING domain of RNF4 greatly impacted  $\text{UBE2N/UBE2V1}$ -catalyzed di-UB formation by both lowering the  $K_M$  of acceptor UB for E2 and by increasing  $k_{\text{cat}}$  by approximately 10-fold (Table 1 and Extended Data Fig. 8f,g). Interestingly, only modest effects were observed for both  $K_M$  (~2.5-fold) as well as  $k_{\text{cat}}$  (~4-fold) in the presence of  $\text{K}^{63}\text{UB}_{\text{C}5}$ . Taken in combination, these effects are not greater than those observed in the absence of E3.

To determine whether the impact of  $\text{K}^{63}\text{UB}_{\text{C}5}$  on  $\text{UBE2N/UBE2V1}$  may apply generally across E2s, kinetic parameters were next estimated for  $\text{UBE2R2}$ . Here,  $\text{K}^{48}\text{UB}_{\text{C}4}$  and  $\text{K}^{48}\text{UB}_{\text{C}5}$  have apparent  $pK_a$  values of 6.6 and 7.3, respectively (Table 1 and Extended Data Fig. 9a,b). While statistically significant, this modest difference is insufficient to explain the nearly 100-fold difference in the rate of di-UB formation when reactions containing  $\text{K}^{48}\text{UB}_{\text{C}4}$  or  $\text{K}^{48}\text{UB}_{\text{C}5}$  were performed at pH 9.7 (Table 1). A 4-fold difference was observed when comparing the estimated  $K_M$  values of  $\text{K}^{48}\text{UB}_{\text{C}4}$  or  $\text{K}^{48}\text{UB}_{\text{C}5}$ , whereas the  $k_{\text{cat}}$  value from reactions containing  $\text{K}^{48}\text{UB}_{\text{C}5}$  was 13-fold lower than for those containing  $\text{K}^{48}\text{UB}_{\text{C}4}$  (Table 1 and Extended Data Fig. 9c,d). Thus, for  $\text{UBE2R2}$ , defects in  $pK_a$  suppression and acceptor UB binding are insufficient, even in combination, to explain the observed losses in the rates of di-UB formation in the presence of C5. Note that apparent  $pK_a$  values could not be performed for  $\text{UBE2R2}$  in the presence of its CRL1 E3 partner due to lack of activity at alkaline pH.

Finally, kinetic experiments were performed on the HECT E3 Rsp5p in the presence of either  $\text{K}^{63}\text{UB}_{\text{C}4}$  or  $\text{K}^{63}\text{UB}_{\text{C}5}$ . While we were also unable to estimate the apparent  $pK_a$  for these reactions due to loss of

enzyme activity at higher pH values, the  $K_M$  of acceptor UB for Rsp5p was estimated for both  $^{K63}\text{UB}_{C4}$  and  $^{K63}\text{UB}_{C5}$  (Table 1 and Extended Data Fig. 9e,f). Remarkably, a 16-fold difference in the  $K_M$  of acceptor UB for Rsp5p was observed, whereas only a ~2.5-fold difference in  $k_{cat}$  occurred when acceptor UB was saturating for Rsp5p. Overall, the kinetic results unveiled a remarkable spectrum of effects of the lysine side-chain ruler on ubiquitin-carrying enzyme activities.

## Discussion

The UB system, with its myriad components that collectively form a UB code oversees a tremendous number of distinct signal transduction events. To do so without producing undesirable cross-talk requires exquisite specificity of the E2 and E3 enzymes that “write” the UB code. Indeed, there has been great focus on how specificity is determined over large scales – tens to hundreds of angstroms – in terms of how substrates are placed relative to UB carrying enzymes<sup>53</sup>. Major effort has also been applied to deciphering the combined catalytic roles of many conserved residues beyond active site cysteines in UB carrying enzymes<sup>15,17,19,20,29</sup>. The roles of acceptor residues have been also studied in terms of different nucleophilic atoms within natural protein amino acid side-chains. UB carrying enzymes display remarkable ranges of reactivities with distinct acceptors within various contexts, including discharge of their thioester-bonded UB to nonspecific amines such as hydroxylamine, free lysine amino acids or even free cysteine or threonine<sup>13,22,25,54</sup> (or Dap, Orn and hLys as shown for UBE2N/UBE2V1 (Fig. 1b,c)). Moreover, UB can be forcibly enzymatically linked to serine or lysine replacements of E2 or E3 catalytic cysteines, as well as to such residues in the vicinity of the active site to enable structural biology<sup>24,55,56</sup>. Thus, in light of the previous findings, it was not surprising that an E2 conjugating enzyme, UBE2S, could tolerate  $^{K11}\text{UB}_{C2-C5}$  acceptors<sup>17</sup>.

Unexpectedly, however, our data shows that many different UB chain forming enzymes are strikingly sensitive to the lysine side-chain hydrocarbon linker at the 1-2 Ångstrom length scale as determined by a single methylene. Biochemical assays show that  $\text{UB}_{C5}$  can affect the  $K_M$ ,  $k_{cat}$ , and  $pK_a$  of the amine (Table 1). Meanwhile, molecular dynamics simulations unveiled pleiotropic structural effects, including additional degrees of freedom, more flexible side chains, and more dynamics in the backbone (Fig. 5). It might stand to reason that side chains that are too short simply could not span the distance between the acceptor UB backbone and UB carrying enzyme active site. However, the fact that the  $\text{UB}_{C5}$  analogs impacted most enzymes tested – and had an effect upon replacing Lys48 in utility of acceptor sites by the seemingly “promiscuous” UBE2D3<sup>48</sup> – further indicates that the role of the acceptor side-chain is more than one of simply spanning the distance between the amine and the active site.

For both UBE2N/UBE2V1 and UBE2R2, the mild effects on apparent  $pK_a$  and/or  $K_M$  are insufficient to explain the defects in  $k_{cat}$  observed upon acceptor lysine substitution with the C5 amino acid (Table 1). The MD simulations pointed toward several possible features of the lysine side chain length that may be optimal for E2-catalyzed UB chain formation. For example, for UB chain forming configurations where substrate binding and/or lysine positioning are rate-limiting, it seems that the increased entropy afforded by an extra methylene group in the targeted side chain could decrease the frequency of catalytic encounter (Fig. 5). Interestingly, this notion would provide a mechanism distinct from that of another E2, UBE2W, for which a confluence of disorder between a flexible substrate N-terminus and a noncanonical E2 C-terminus guides ubiquitylation to a substrate’s N-terminal amine<sup>57</sup>. Rather than demanding disorder, these systems appear to favor a calibrated reach by the nucleophile that also must have restrained degrees of freedom.

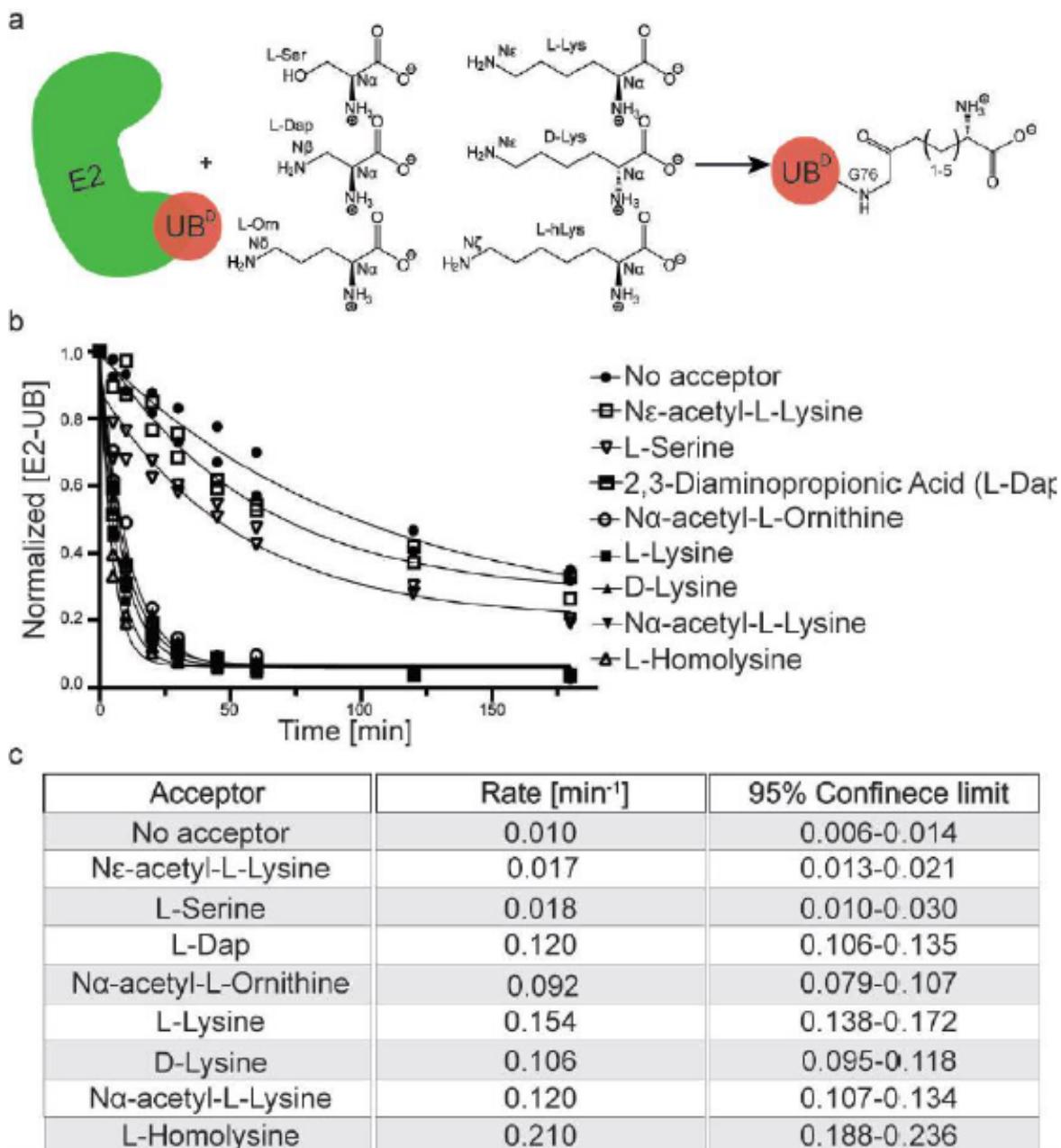
In addition to entropic effects on the side chain, the hydrocarbon linker length could also affect catalysis in cases where chemistry is rate limiting. For example, as shown for UBE2N (without UBE2V1-RING E3 partners), there is a precise “hole” fitting the lysine amine, and attack on the thioester carbonyl is

rate-limiting<sup>50</sup>. The acceptor UB lysine itself may be optimal not only for accessing the amine hole, but also for the chemistry of ubiquitylation. Indeed, this would be consistent with UB discharge to free side-chain amine acceptors irrespective of hydrocarbon length, and a dramatic impact on  $k_{cat}$  in the context of acceptor UB. One would also predict little impact on  $K_M$  in such a case, although lack of an effect on  $K_M$  may also reflect that the additional methylene does not impact acceptor UB recruitment to an auxiliary UB-binding domain.

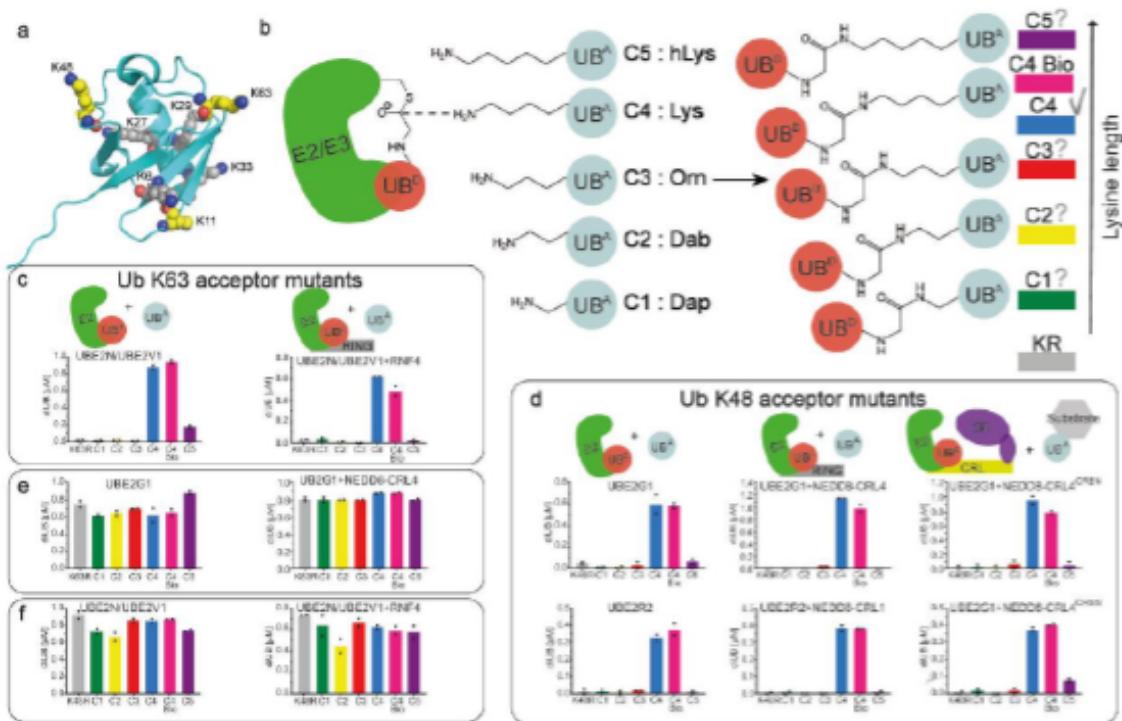
Notably, the impact of acceptor side-chain length on the HECT E3 Rsp5p represents the opposite extreme. The predominant effect on  $K_M$  implies a role of the acceptor Lys itself in productive binding to the E3. It is possible that local interactions – awaiting elucidation by future structural studies – dominate acceptor UB recruitment<sup>58</sup>. It is also possible that placement of the acceptor lysine in the HECT~UB active site allosterically stabilizes the conformation that binds the acceptor UB<sup>59</sup>, or otherwise contributes to productive interactions.

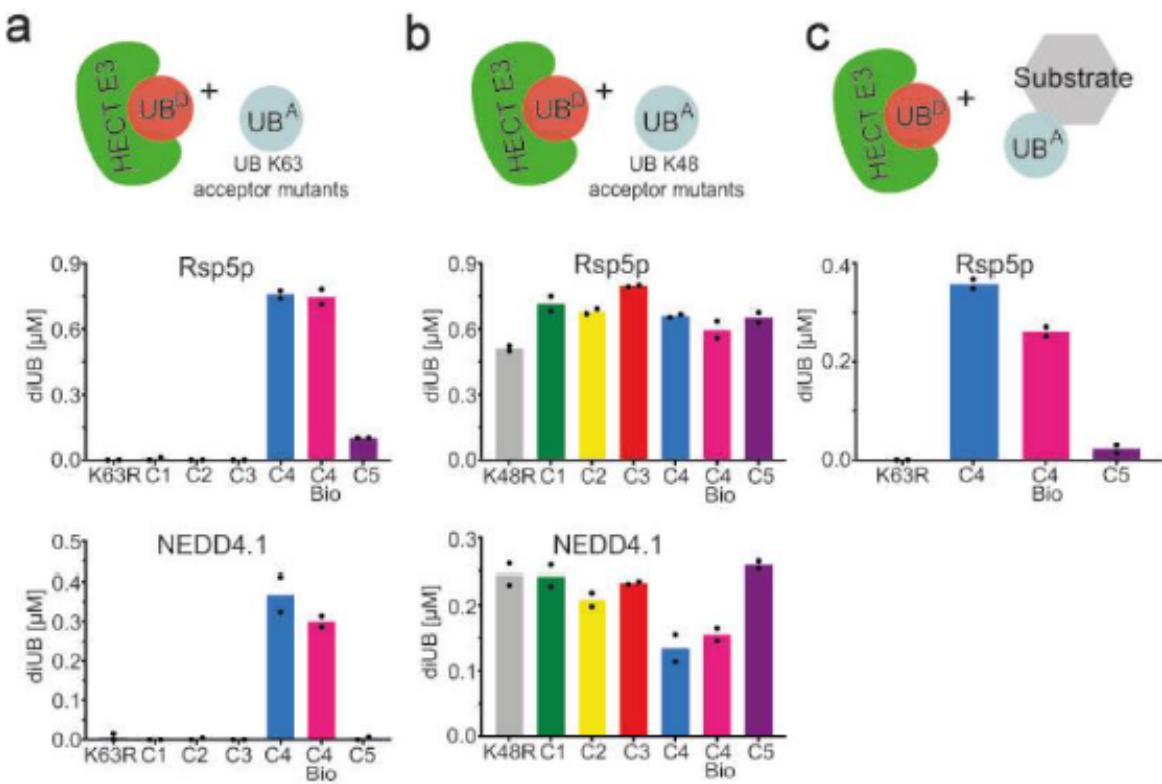
While our study relied on installing side-chain chemical variants, it seems likely that in the cellular milieu, many natural factors – including linkage within a chain and binding to protein partners – could influence presentation of acceptor lysines resulting in specificity with E2 and E3 enzymes. We also speculate that such features may be manifested within the structures of the protein substrates themselves, thereby contributing to lysine targeting during direct substrate ubiquitylation. While strong preferences for lysine side-chains may contribute to the robustness of some ubiquitylation pathways that elicit proteasomal degradation even when preferred targeting sites are unavailable, such features may also influence not only successes but also failures of targeted protein degradation strategies that rely on small molecules to direct proteins of therapeutic interest to ubiquitylating enzymes<sup>60,61</sup>.

Finally, it is noteworthy that E2~UB and (E3) HECT~UB active sites are structurally distinct. Thus, the demand for native lysine C4 acceptor geometry in the course of chain building seems to be a general property of UB acceptors that could extend across many of the hundreds of E2/E3 ligation systems.

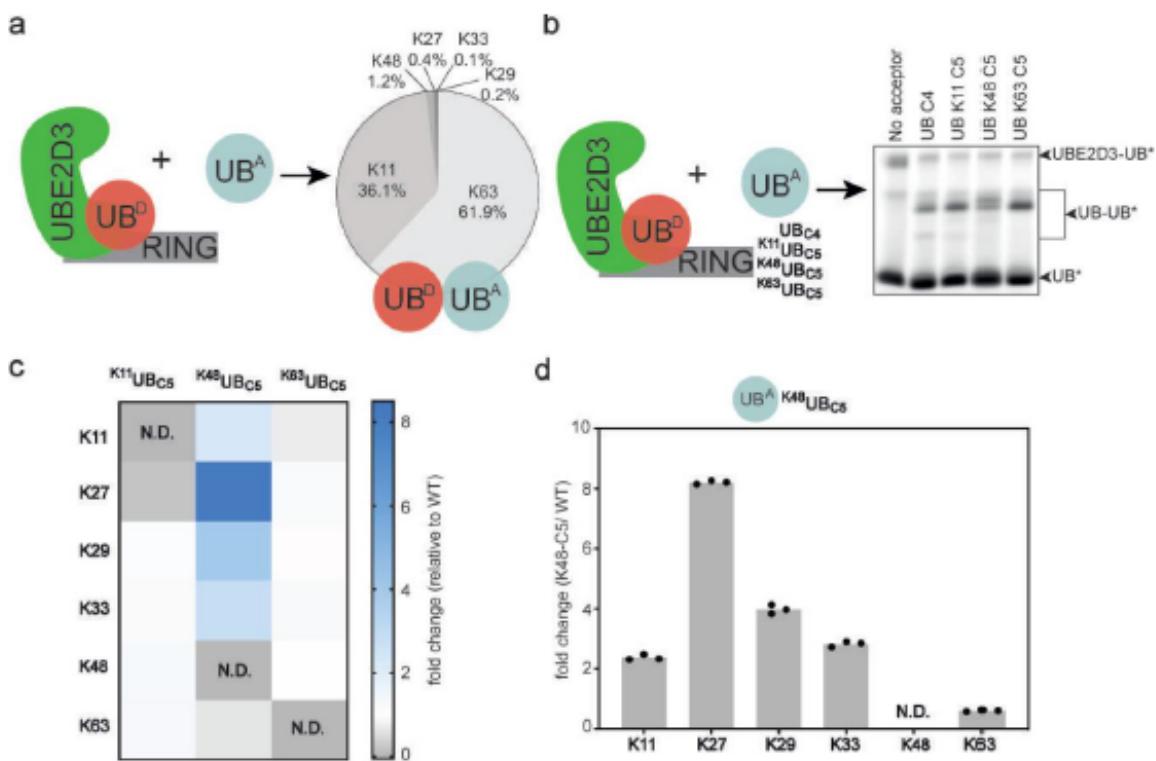


**Figure 1 | UBE2N-UB/UBE2V1/RNF4 RING E3 reacts preferentially with free amino acids harboring amine acceptors and various side-chain hydrocarbon linkers.** (a) Cartoon of experimental scheme, monitoring reactivity of E2-UB (D refers to the “donor” fluorescent UB to be transferred from E2) toward various free amino acid nucleophiles. (b) Time-course of fluorescent UB discharge from UBE2N-UB/UBE2V1/RNF4 RING E3 to the indicated amino acids, normalized to starting signal of fluorescent UB thioester-bonded to UBE2N. (c) Summary of rates in (b) calculated according to a one-phase nonlinear fit of exponential decay.



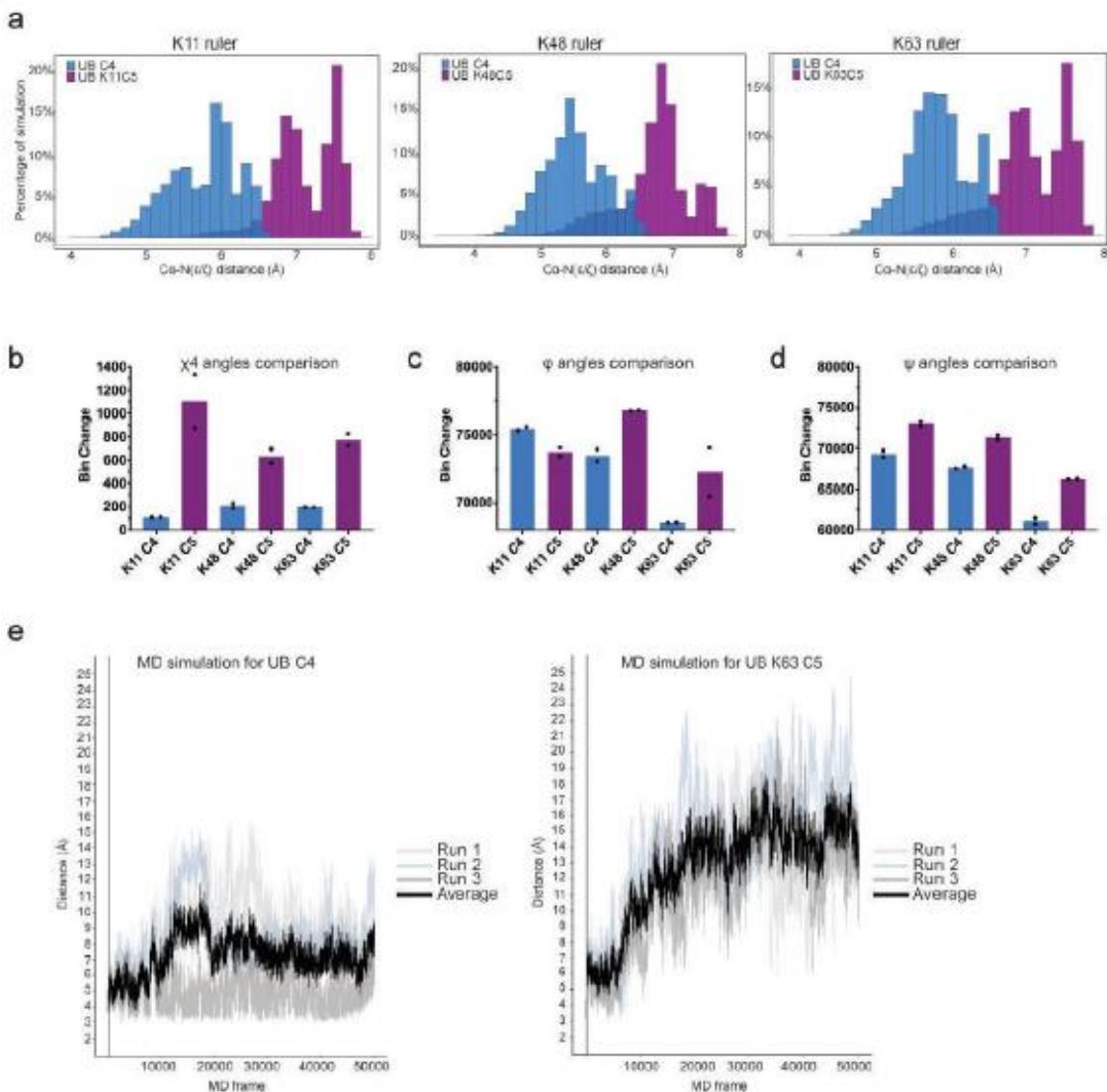


**Figure 3 | K63 chain forming HECT E3 ligases show strong preferences for a native lysine acceptor on ubiquitin.** With an active site that is distinct from E2 conjugating enzymes and unlike any of the other E3s tested, a HECT E3 catalytic cysteine forms a thioester intermediate en route to forming an isopeptide bond with a K63 UB acceptor. (a) The K63 chain-forming HECT E3s NEDD4.1 and Rsp5p strongly prefer the <sup>K63</sup>UB<sub>C4</sub> acceptor, and (b) react equally with the <sup>K48</sup>UB<sub>C1-C5</sub> series. (c) Sortase-mediated transpeptidation was used to fuse the C-terminus of <sup>K63</sup>UB<sub>C4</sub>, <sup>K63</sup>UB<sub>C5</sub>, or <sup>K63</sup>RUB to a truncated peptide of the Rsp5p substrate Sna4p, which harbors a degron motif. Added to preformed E2~UB\*, the <sup>K63</sup>UB<sub>C4</sub>- or <sup>K63</sup>UB<sub>C5</sub>-Sna4p conjugates reveal striking preference for native lysine in the generation of a Lys63-linked polyubiquitin chain by HECT ligase. All bar graphs show the average of two replicates.



**Figure 4 |** The location of homolysine on acceptor UB impacts ratios of less preferred chain-types produced by UBE2D3.

(a) UBE2D3~UB activated by a RING E3 (RNF4) can catalyze the synthesis of diverse diUB linkages. The diUB linkage type distribution, based on absolute quantitation via mass spectrometry, is represented in a pie chart. The reaction was performed in the presence of wild-type donor and acceptor UBs. (b) We charged UBE2D3 with fluorescent UB (UBE2D3~UB\*), and then added the E3 RNF4 RING domain and the following UB acceptors: <sup>K11</sup>UB<sub>C5</sub>, <sup>K48</sup>UB<sub>C5</sub>, or <sup>K63</sup>UB<sub>C5</sub>. SDS-PAGE detecting UB\* shows diverse mobilities of UB-UB\* products for the chemically variant acceptors, indicating production of different chain types. (c) Relative fold changes of diUB linkage types (rows) produced by reactions similar to those in (b) with <sup>K11</sup>UB<sub>C5</sub>, <sup>K48</sup>UB<sub>C5</sub>, or <sup>K63</sup>UB<sub>C5</sub> acceptors (columns) compared to reactions with UB<sub>C4</sub> as an acceptor. Notice how a C5 linker at UB position 48 distorts the distribution of chain-types. Since our protocol calculates the ratios of analogous peptides between UB<sub>C4</sub> and UB<sub>C5</sub> reactions, the ratios for peptides bearing a C5 linker could not be calculated and are labeled as 'not defined' (N.D.). (d) Relative chain-type abundance comparison between reactions using <sup>K48</sup>UB<sub>C5</sub> and UB<sub>C4</sub> as acceptors. Notably, the amount of Lys11-linked diUB, a main product of the wild-type reaction, more than doubled, while amounts of Lys27, Lys29, and Lys33-linked diUB also significantly increased. Bars represent the mean of three independent replicates (dots).



**Figure 5 | Molecular Dynamics simulations reveal unique geometries for lysine residues on ubiquitin harboring an extra methylene group in the side-chain.** Simulations were completed for UB<sub>C4</sub> and UB<sub>C5</sub> at positions 11, 48, and 63, as well as for the UBE2N~donor UB/UBE2V1/acceptor UB complex. (a) Distribution of acceptor amine to donor C $\alpha$  distances for C4 versus C5 in the simulations. (b) Chi4 interconversion rate between 120-degree rotamer bins in the simulations. Distribution of Lysine side chain Chi4 angles according to 120-degree rotamer bins for C4 versus C5. (c) Phi and (d) Psi interconversion rate between 10 degree bins. (e) 3 independent 25 ns trajectories (colored differently) of the UBE2N~UB acceptor UB<sub>C4</sub> vs <sup>K63</sup>UB<sub>C5</sub> complex. The distance between the acceptor UB's side-chain nitrogen atom and donor UB's carbonyl carbon is calculated for each frame of the trajectories and plotted (in light gray, slate gray, and dark gray) along with an average of the three runs (in black).

$k_{cat}$ , $pK_a^{app}$ and $K_M$ for ubiquitin carrying enzymes with native vs. homolysine acceptor UBs										
E2/E3	UB	Lys	$pK_a^{app}$	95% CL	$k_{obs}(hr^{-1})$ (top pH)	95% CL $K_M(10^{-6}M)$	95% CL $k_{cat}(hr^{-1})$	95% CL		
UBE2N/V1	C4 <sub>bio</sub>	K63				190	142-258	6.1	5.5-6.9	
UBE2N/V1	C4	K63	8.9	8.7-9.0	15.8	14.7-17.1	398	293-553	3.4	2.9-4.0
UBE2N/V1	C5	K63	9.0	8.9-9.2	0.6	0.5-0.7	284	185-456	0.2	0.17-0.26
UBE2N/V1 + RNF4	C4	K63				23	15-34	39.1	35.3-43.2	
UBE2N/V1 + RNF4	C5	K63				58	47-72	9.3	8.7-9.9	
UBE2R2	C4	K48	6.6	6.5-6.8	2.67	2.56-2.78	528	294-939	15.8	13.2-19.2
UBE2R2	C5	K48	7.3	7.0-7.5	0.028	0.025-0.030	1940	1331-2874	1.1	1.0-1.3
RSP5	C4	K63				21.0	17.1-25.7	1.11	1.06-1.17	
RSP5	C5	K63				334.9	269.2-421.2	0.44	0.40-0.49	

**Table 1|**

Kinetic parameters apparent  $pK_a$ , the rate of di-UB formation,  $k_{obs}$ , for UBE2N/UBE2V1 and UBE2R2 at pH 10.1 or 9.7, respectively, the  $K_M$  of acceptor UB for E2 or E3, and  $k_{cat}$  for various ubiquitin carrying enzymes. The 95% confidence limits (CL) are provided as estimates for the error of measurement.  $R^2$  values for the fit of the models to the data have been provided in Extended Data Figures 8 and 9.

#### Acknowledgements

We thank D. Scott, S. Uebel, S. Pettera, V. Sanchez, K. Baek, D. Horn-Ghetko and S. Kostrhon for assistance, reagents, and helpful discussions. We also thank C. Talavera-Ormeño and P. Hekking for assistance in peptide synthesis. BAS has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789016-NEDD8Activate). Also, N.P., D.H., N.B., and G.K. were supported by a grant from the National Institutes of Health (R15GM117555-02). G.J.v.d.H.v.N was supported by grants from NWO (VIDI and Off-Road). HO was supported by a VICI grant from the Netherlands Foundation for Scientific Research (NWO).

#### Author Contributions

Syntheses of UB analogues were designed and executed by G.J.v.d.H.v.N. JL performed all biochemical assays. Kinetics experiments were done by G.K., N.P., D.H., and N.B. Mass spectrometry experiments were designed and conducted by M.M., F.M.H. and O.K. MD simulations were performed by V.T. and J.H. NMR was done by M.B. and M.S. The manuscript was prepared by J.L., D.T.K., G.J.v.d.H.v.N., G.K., H.O., and B.A.S. with input from all authors. The project was supervised by B.A.S., G.K., H.O., and D.T.K.

#### Data availability

All raw gels will be supplied as supplementary information upon manuscript acceptance and ms data deposited in PRIDE.

#### Code availability

ROSETTA software can be downloaded from [www.rosettacommons.org](http://www.rosettacommons.org) and is available free to academic users.

#### Competing Interests statement

H.O. is shareholder of UbiqBio. All other authors declare no competing interests.

## References

- 1 Rajalingam, K. & Dikic, I. SnapShot: Expanding the Ubiquitin Code. *Cell* **164**, 1074-1074.e1071 (2016).
- 2 Yau, R. & Rape, M. The increasing complexity of the ubiquitin code. *Nat Cell Biol* **18**, 579-586 (2016).
- 3 Kwon, Y. T. & Ciechanover, A. The Ubiquitin Code in the Ubiquitin-Proteasome System and Autophagy. *Trends Biochem Sci* **42**, 873-886 (2017).
- 4 Komander, D. & Rape, M. The Ubiquitin Code. *Annual Review of Biochemistry* **81**, 203-229 (2012).
- 5 Dikic, I., Wakatsuki, S. & Walters, K. J. Ubiquitin-binding domains — from structures to functions. *Nature Reviews Molecular Cell Biology* **10**, 659-671 (2009).
- 6 Mattern, M., Sutherland, J., Kadimisetty, K., Barrio, R. & Rodriguez, M. S. Using Ubiquitin Binders to Decipher the Ubiquitin Code. *Trends Biochem Sci* **44**, 599-615 (2019).
- 7 Mulder, M. P. C., Witting, K. F. & Ovaa, H. Cracking the Ubiquitin Code: The Ubiquitin Toolbox. *Curr Issues Mol Biol* **37**, 1-20 (2020).
- 8 Du, J. & Strieter, E. R. A fluorescence polarization-based competition assay for measuring interactions between unlabeled ubiquitin chains and UCH37•RPN13. *Anal Biochem* **550**, 84-89 (2018).
- 9 Sun, H. *et al.* Diverse fate of ubiquitin chain moieties: The proximal is degraded with the target, and the distal protects the proximal from removal and recycles. *Proceedings of the National Academy of Sciences* **116**, 7805-7812 (2019).
- 10 Zhang, X. *et al.* An Interaction Landscape of Ubiquitin Signaling. *Mol Cell* **65**, 941-955.e948 (2017).
- 11 Nawatha, M. *et al.* De novo macrocyclic peptides that specifically modulate Lys48-linked ubiquitin chains. *Nat Chem* **11**, 644-652 (2019).
- 12 Berndsen, C. E. & Wolberger, C. New insights into ubiquitin E3 ligase mechanism. *Nature structural & molecular biology* **21**, 301-307 (2014).
- 13 Stewart, M. D., Ritterhoff, T., Klevit, R. E. & Brzovic, P. S. E2 enzymes: more than just middle men. *Cell Research* **26**, 423-440 (2016).
- 14 Mattioli, F. & Sixma, T. K. Lysine-targeting specificity in ubiquitin and ubiquitin-like modification pathways. *Nature structural & molecular biology* **21**, 308-316 (2014).
- 15 Victor Bernier-Villamor, D. A. S., Michael J. Matunis, Christopher D. Lima. Structural Basis for E2-Mediated SUMO Conjugation Revealed by a Complex between Ubiquitin-Conjugating Enzyme Ubc9 and RanGAP1. *Cell* **108**, 345-356 (2002).
- 16 Eddins, M. J., Carlile, C. M., Gomez, K. M., Pickart, C. M. & Wolberger, C. Mms2–Ubc13 covalently bound to ubiquitin reveals the structural basis of linkage-specific polyubiquitin chain formation. *Nature Structural & Molecular Biology* **13**, 915-920 (2006).
- 17 Wickliffe, K. E., Lorenz, S., Wemmer, D. E., Kuriyan, J. & Rape, M. The mechanism of linkage-specific ubiquitin chain elongation by a single-subunit E2. *Cell* **144**, 769-781 (2011).
- 18 Smit, J. J. *et al.* The E3 ligase HOIP specifies linear ubiquitin chain assembly through its RING-IBR-RING domain and the unique LDD extension. *The EMBO journal* **31**, 3833-3844 (2012).
- 19 Berndsen, C. E., Wiener, R., Yu, I. W., Ringel, A. E. & Wolberger, C. A conserved asparagine has a structural role in ubiquitin-conjugating enzymes. *Nat Chem Biol* **9**, 154-156 (2013).
- 20 Stieglitz, B. *et al.* Structural basis for ligase-specific conjugation of linear ubiquitin chains by HOIP. *Nature* **503**, 422-426 (2013).
- 21 Shimizu, Y., Okuda-Shimizu, Y. & Hendershot, L. M. Ubiquitylation of an ERAD substrate occurs on multiple types of amino acids. *Mol Cell* **40**, 917-926 (2010).
- 22 Wenzel, D. M., Lissounov, A., Brzovic, P. S. & Klevit, R. E. UBCH7 reactivity profile reveals parkin and HHARI to be RING/HECT hybrids. *Nature* **474**, 105-108 (2011).

23 Buetow, L. & Huang, D. T. Structural insights into the catalysis and regulation of E3 ubiquitin ligases. *Nat Rev Mol Cell Biol* **17**, 626-642 (2016).

24 Zheng, N. & Shabek, N. Ubiquitin Ligases: Structure, Function, and Regulation. *Annu Rev Biochem* **86**, 129-157 (2017).

25 Pao, K. C. *et al.* Activity-based E3 ligase profiling uncovers an E3 ligase with esterification activity. *Nature* **556**, 381-385 (2018).

26 Virdee, S., Macmillan, D. & Waksman, G. Semisynthetic Src SH2 domains demonstrate altered phosphopeptide specificity induced by incorporation of unnatural lysine derivatives. *Chem Biol* **17**, 274-284 (2010).

27 Temimi, A. H. K. A. *et al.* Lysine Possesses the Optimal Chain Length for Histone Lysine Methyltransferase Catalysis. *Scientific Reports* **7**, 16148 (2017).

28 McKenna, S. *et al.* Noncovalent interaction between ubiquitin and the human DNA repair protein Mms2 is required for Ubc13-mediated polyubiquitination. *J Biol Chem* **276**, 40120-40126 (2001).

29 Branigan, E., Plechanovová, A., Jaffray, E. G., Naismith, J. H. & Hay, R. T. Structural basis for the RING-catalyzed synthesis of K63-linked ubiquitin chains. *Nature Structural & Molecular Biology* **22**, 597-602 (2015).

30 Branigan, E., Carlos Penedo, J. & Hay, R. T. Ubiquitin transfer by a RING E3 ligase occurs from a closed E2-ubiquitin conformation. *Nature Communications* **11**, 2846 (2020).

31 Petroski, M. D. & Deshaies, R. J. Mechanism of Lysine 48-Linked Ubiquitin-Chain Synthesis by the Cullin-RING Ubiquitin-Ligase Complex SCF-Cdc34. *Cell* **123**, 1107-1120 (2005).

32 Choi, Y.-S. *et al.* Differential ubiquitin binding by the acidic loops of Ube2g1 and Ube2r1 enzymes distinguishes their Lys-48-ubiquitylation activities. *J Biol Chem* **290**, 2251-2263 (2015).

33 Hill, S., Harrison, J. S., Lewis, S. M., Kuhlman, B. & Kleiger, G. Mechanism of Lysine 48 Selectivity during Polyubiquitin Chain Formation by the Ube2R1/2 Ubiquitin-Conjugating Enzyme. *Mol Cell Biol* **36**, 1720-1732 (2016).

34 Lu, G. *et al.* UBE2G1 governs the destruction of cereblon neomorphic substrates. *eLife* **7**, e40958 (2018).

35 Hill, S. *et al.* Robust cullin-RING ligase function is established by a multiplicity of polyubiquitylation pathways. *eLife* **8**, e51163 (2019).

36 Kronke, J. *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301-305 (2014).

37 Lu, G. *et al.* The myeloma drug lenalidomide promotes the cereblon-dependent destruction of Ikaros proteins. *Science* **343**, 305-309 (2014).

38 Skaar, J. R., Pagan, J. K. & Pagano, M. Mechanisms and function of substrate recruitment by F-box proteins. *Nature Reviews Molecular Cell Biology* **14**, 369-381 (2013).

39 Rotin, D. & Kumar, S. Physiological functions of the HECT family of ubiquitin ligases. *Nature Reviews Molecular Cell Biology* **10**, 398-409 (2009).

40 Maspero, E. *et al.* Structure of a ubiquitin-loaded HECT ligase reveals the molecular basis for catalytic priming. *Nature Structural & Molecular Biology* **20**, 696-701 (2013).

41 Kim, H. C. & Huibregtse, J. M. Polyubiquitination by HECT E3s and the Determinants of Chain Type Specificity. *Molecular and Cellular Biology* **29**, 3307 (2009).

42 Gupta, R. *et al.* Ubiquitination screen using protein microarrays for comprehensive identification of Rsp5 substrates in yeast. *Mol Syst Biol* **3**, 116-116 (2007).

43 Bremm, A., Freund, S. M. V. & Komander, D. Lys11-linked ubiquitin chains adopt compact conformations and are preferentially hydrolyzed by the deubiquitinase Cezanne. *Nature structural & molecular biology* **17**, 939-947 (2010).

44 Brown, N. G. *et al.* Dual RING E3 Architectures Regulate Multiubiquitination and Ubiquitin Chain Elongation by APC/C. *Cell* **165**, 1440-1453 (2016).

45 Saville, M. K. *et al.* Regulation of p53 by the ubiquitin-conjugating enzymes UbcH5B/C in vivo. *J Biol Chem* **279**, 42169-42181 (2004).

46 Wu, K., Kovacev, J. & Pan, Z. Q. Priming and extending: a UbcH5/Cdc34 E2 handoff mechanism for polyubiquitination on a SCF substrate. *Molecular cell* **37**, 784-796 (2010).

47 Wauer, T. *et al.* Ubiquitin Ser65 phosphorylation affects ubiquitin structure, chain assembly and hydrolysis. *The EMBO Journal* **34**, 307-325 (2015).

48 Brzovic, P. S. & Klevit, R. E. Ubiquitin transfer from the E2 perspective: why is UbcH5 so promiscuous? *Cell cycle (Georgetown, Tex)* **5**, 2867-2873 (2006).

49 Markin, C. J. *et al.* Catalytic proficiency of ubiquitin conjugation enzymes: balancing pK(a) suppression, entropy, and electrostatics. *J Am Chem Soc* **132**, 17775-17786 (2010).

50 Jones, W. M., Davis, A. G., Wilson, R. H., Elliott, K. L. & Sumner, I. A conserved asparagine in a ubiquitin-conjugating enzyme positions the substrate for nucleophilic attack. *J Comput Chem* **40**, 1969-1977 (2019).

51 Yunus, A. A. & Lima, C. D. Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway. *Nature Structural & Molecular Biology* **13**, 491-499 (2006).

52 Russell, A. J., Thomas, P. G. & Fersht, A. R. Electrostatic effects on modification of charged groups in the active site cleft of subtilisin by protein engineering. *J Mol Biol* **193**, 803-813 (1987).

53 Rennie, M. L., Chaugule, V. K. & Walden, H. Modes of allosteric regulation of the ubiquitination machinery. *Curr Opin Struct Biol* **62**, 189-196 (2020).

54 Saha, A., Lewis, S., Kleiger, G., Kuhlman, B. & Deshaies, R. J. Essential role for ubiquitin-ubiquitin-conjugating enzyme interaction in ubiquitin discharge from Cdc34 to substrate. *Mol Cell* **42**, 75-83 (2011).

55 Kamadurai, H. B. *et al.* Insights into ubiquitin transfer cascades from a structure of a UbcH5B approximately ubiquitin-HECT(NEDD4L) complex. *Mol Cell* **36**, 1095-1102 (2009).

56 Cappadocia, L. & Lima, C. D. Ubiquitin-like Protein Conjugation: Structures, Chemistry, and Mechanism. *Chem Rev* **118**, 889-918 (2018).

57 Vittal, V. *et al.* Intrinsic disorder drives N-terminal ubiquitination by Ube2w. *Nat Chem Biol* **11**, 83-89 (2015).

58 Wang, M., Cheng, D., Peng, J. & Pickart, C. M. Molecular determinants of polyubiquitin linkage selection by an HECT ubiquitin ligase. *Embo j* **25**, 1710-1719 (2006).

59 Scott, D. C. *et al.* Structure of a RING E3 trapped in action reveals ligation mechanism for the ubiquitin-like protein NEDD8. *Cell* **157**, 1671-1684 (2014).

60 Burslem, G. M. & Crews, C. M. Proteolysis-Targeting Chimeras as Therapeutics and Tools for Biological Discovery. *Cell* (2020).

61 Verma, R., Mohl, D. & Deshaies, R. J. Harnessing the Power of Proteolysis for Targeted Protein Inactivation. *Mol Cell* **77**, 446-460 (2020).

### 3.2.6. Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV

Alexey Stukalov<sup>1\*</sup>, Virginie Girault<sup>1\*</sup>, Vincent Grass<sup>1\*</sup>, Valter Bergant<sup>1\*</sup>, Ozge Karayel<sup>2\*</sup>, Christian Urban<sup>1\*</sup>, Darya A. Haas<sup>1\*</sup>, Yiqi Huang<sup>1</sup>, Lila Oubrahim<sup>1</sup>, Anqi Wang<sup>1</sup>, Sabri M. Hamad<sup>1</sup>, Antonio Piras<sup>1</sup>, Maria Tanzer<sup>2</sup>, Fynn Hansen<sup>2</sup>, Thomas Enghleitner<sup>3</sup>, Maria Reinecke<sup>4, 5</sup>, Teresa M. Lavacca<sup>1</sup>, Rosina Ehmann<sup>6, 7</sup>, Roman Wölfel<sup>6, 7</sup>, Jörg Jores<sup>8</sup>, Bernhard Kuster<sup>4, 5</sup>, Ulrike Protzer<sup>1, 7</sup>, Roland Rad<sup>3</sup>, John Ziebuhr<sup>9</sup>, Volker Thiel<sup>10</sup>, Pietro Scaturro<sup>1,11</sup>, Matthias Mann<sup>2</sup> and Andreas Pichlmair<sup>1, 7, §</sup>

<sup>1</sup> Technical University of Munich, School of Medicine, Institute of Virology, 81675 Munich, Germany, <sup>2</sup> Department of Proteomics and Signal transduction, Max-Planck Institute of Biochemistry, Martinsried/Munich, 82152, Germany, <sup>3</sup> Institute of Molecular Oncology and Functional Genomics and Department of Medicine II, School of Medicine, Technical University of Munich, 81675 Munich, Germany, <sup>4</sup> Chair of Proteomics and Bioanalytics, Technical University of Munich, 85354 Freising, Germany, <sup>5</sup> German Cancer Consortium (DKTK), Munich partner site and German Cancer Research Center, (DKFZ) Heidelberg, Germany, <sup>6</sup> Bundeswehr Institute of Microbiology, 80937 Munich, Germany, <sup>7</sup> German Center for Infection Research (DZIF), Munich partner site, Germany, <sup>8</sup> Institute of Veterinary Bacteriology, Department of Infectious Diseases and Pathobiology, University of Bern, Bern, Switzerland, <sup>9</sup> Justus Liebig University Giessen, Institute of Medical Virology, 35392 Giessen, Germany, <sup>10</sup> Institute of Virology and Immunology (IVI), Bern, Switzerland & Department of Infectious Diseases and Pathobiology, University of Bern, Bern, Switzerland, <sup>11</sup> Systems Arbovirology, Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. \* These authors contributed equally §Corresponding author

Under revision in Nature (2020)

At the end of 2019, several cases of severe pneumonia were reported in Wuhan in eastern China, and early in January 2020, next-generation sequencing revealed a novel coronavirus, later named SARS-CoV-2, as the cause for this disease. High infectivity of SARS-CoV-2 resulted in rapid global spreading. Thus, the coronavirus pandemic, also known as the COVID-19 pandemic, arguably became the most serious global health crisis of our time.

The sudden emergence of SARS-CoV-2 urgently requires an in-depth understanding of the viral infection process at the molecular level. In collaboration with the group of Andreas Pichlmair at the TUM, using state-of-the-art mass spectrometry techniques, we characterized the impact of viral infection on protein abundance, ubiquitination and phosphorylation in a time-resolved manner. Building on my previous experience and technological developments, I led the efforts on PTM analysis of virus infection on the Max Planck side of the collaboration. The information we have gained about the dynamic modulation of multiple host signaling pathways upon virus entry into host cells enabled us to identify unique and common molecular mechanisms of SARS coronaviruses. We hope that our resource data will provide an important basis for further translational research and likely lead to new approaches for the search for antiviral compounds.

## Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV

Alexey Stukalov<sup>1\*</sup>, Virginie Girault<sup>1\*</sup>, Vincent Grass<sup>1\*</sup>, Valter Bergant<sup>1\*</sup>, Ozge Karayel<sup>2\*</sup>, Christian Urban<sup>1\*</sup>, Darya A. Haas<sup>1\*</sup>, Yiqi Huang<sup>1</sup>, Lila Oubrahim<sup>1</sup>, Anqi Wang<sup>1</sup>, Sabri M. Hamad<sup>1</sup>, Antonio Piras<sup>1</sup>, Maria Tanzer<sup>2</sup>, Fynn M. Hansen<sup>2</sup>, Thomas Engleitner<sup>3</sup>, Maria Reinecke<sup>4,5</sup>, Teresa M. Lavacca<sup>1</sup>, Rosina Ehmann<sup>6,7</sup>, Roman Wölfel<sup>6,7</sup>, Jörg Jores<sup>8</sup>, Bernhard Kuster<sup>4,5</sup>, Ulrike Protzer<sup>1,7</sup>, Roland Rad<sup>3</sup>, John Ziebuhr<sup>9</sup>, Volker Thiel<sup>10</sup>, Pietro Scaturro<sup>1,11</sup>, Matthias Mann<sup>2</sup> and Andreas Pichlmair<sup>1,7,§</sup>

<sup>1</sup>Technical University of Munich, School of Medicine, Institute of Virology, 81675 Munich, Germany,

<sup>2</sup>Department of Proteomics and Signal transduction, Max-Planck Institute of Biochemistry,

Martinsried/Munich, 82152, Germany, <sup>3</sup>Institute of Molecular Oncology and Functional Genomics and

Department of Medicine II, School of Medicine, Technical University of Munich, 81675 Munich,

Germany, <sup>4</sup>Chair of Proteomics and Bioanalytics, Technical University of Munich, 85354 Freising,

Germany, <sup>5</sup>German Cancer Consortium (DKTK), Munich partner site and German Cancer Research

Center, (DKFZ) Heidelberg, Germany, <sup>6</sup>Bundeswehr Institute of Microbiology, 80937 Munich,

Germany, <sup>7</sup>German Center for Infection Research (DZIF), Munich partner site, Germany, <sup>8</sup>Institute of

Veterinary Bacteriology, Department of Infectious Diseases and Pathobiology, University of Bern,

Bern, Switzerland, <sup>9</sup>Justus Liebig University Giessen, Institute of Medical Virology, 35392 Giessen,

Germany, <sup>10</sup>Institute of Virology and Immunology (IVI), Bern, Switzerland & Department of Infectious

Diseases and Pathobiology, University of Bern, Bern, Switzerland, <sup>11</sup>Systems Arbovirology, Heinrich

Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany.

\* these authors contributed equally

§Corresponding author:

### Abstract:

The sudden global emergence of SARS-CoV-2 urgently requires an in-depth understanding of molecular functions of viral proteins and their interactions with the host proteome. –Several omics studies have extended our knowledge of COVID-19 pathophysiology, including some focused on a proteomics aspects<sup>1-3</sup>. To understand how SARS-CoV-2 and related coronaviruses manipulate the host we here characterized interactome, proteome and signaling processes in a systems-wide manner. This identified connections between the corresponding cellular events, revealed functional effects of the individual viral proteins and put these findings into the context of host signaling pathways. We investigated the closely related SARS-CoV-2 and SARS-CoV viruses as well as the influence of SARS-CoV-2 on transcriptome, proteome, ubiquitinome and phosphoproteome of a lung-derived human cell line. Projecting these data onto the global network of cellular interactions revealed relationships between the perturbations taking place upon SARS-CoV-2 infection at different layers and identified unique and common molecular mechanisms of SARS coronaviruses. The results highlight the functionality of individual proteins as well as vulnerability hotspots of SARS-CoV-2, which we targeted with clinically approved drugs. We exemplify this by identification of kinase inhibitors as well as MMPase inhibitors with significant antiviral effects against SARS-CoV-2.

**Main text:**

To identify interactions of SARS-CoV-2 and SARS-CoV with cellular proteins, we transduced A549 lung carcinoma cells with lentiviruses expressing individual HA-tagged viral proteins (Figure 1a, Extended data Fig. 1a, Supplementary Table 1). Affinity purification followed by mass spectrometry analysis (AP-MS) and statistical modelling of the MS1-level quantitative data allowed identification of 1484 interactions between 1086 cellular proteins and 22 SARS-CoV-2 and 26 SARS-CoV bait proteins (Figure 1b, Extended data Fig. 1b, Supplementary Table 2). The resulting virus-host interaction network revealed a wide range of cellular activities intercepted by SARS-CoV-2 and SARS-CoV (Figure 1b, Extended data Table 1, Supplementary Table 2). In particular, we discovered that SARS-CoV-2 targets a number of key cellular regulators involved in innate immunity (ORF7b-MAVS, -UNC93B1), stress response components (N-HSPA1A) and DNA damage response mediators (ORF7a-ATM, -ATR) (Figure 1b, Extended data Fig. 1c-e). Overall, SARS-CoV-2 interacts with specific protein complexes contributing to a range of biological processes (Supplementary Table 2). To evaluate the consequences of these interactions on cellular proteostasis, we proceeded with the total proteome analysis of A549 cells expressing the 69 individual viral proteins (Figure 1a, d, Supplementary Table 4). The analysis of the proteome changes induced by each viral protein and consideration of the interactions of this respective protein provided direct insights into their functions. For instance, we confirmed that ORF9b of SARS-CoV-2 leads to a dysregulation of mitochondrial functions (Figure 1d, Supplementary Table 4), as was previously reported for SARS-CoV<sup>4</sup>, correlating with the binding of ORF9b of both viruses to TOMM70 (Figure 1b, Supplementary Table 2)<sup>1</sup>, a known regulator of mitophagy<sup>5</sup>, which was not yet known for SARS-CoV ORF9b. Importantly, this approach identified novel SARS-CoV-2 activities, such as the regulation of proteins involved in cholesterol metabolism by NSP6 (Figure 1d, Supplementary Table 4). Despite the high similarity of SARS-CoV-2 and SARS-CoV<sup>6</sup>, these datasets allow to discriminate the commonalities and differences of both viruses, which may in part explain the characteristics in pathogenicity and transmission capabilities. By comparing the AP-MS data of homologous SARS-CoV-2 and SARS-CoV proteins, we identified significant differences in the enrichment of individual host targets, highlighting potential virus-specific interactions (Figure 1b (edge color), Figure 1c, Extended data Fig. 1f-h). For instance, we could recapitulate the known interaction between SARS-CoV NSP2 and prohibitins (PHB, PHB2)<sup>7</sup>, whereas their enrichment was not observed for NSP2 of SARS-CoV-2 (Extended data Fig. 1g). Alternatively, we found that ORF8 of SARS-CoV-2, but not its SARS-CoV homolog, binds specifically to the TGFB1-LTBP1 complex (Extended data Fig. 1f, h).

To obtain information on the concerted activity of the viral proteins during infection, we infected ACE2-expressing A549 cells (Extended data Fig. 2a, b) with SARS-CoV-2, and profiled the impact of viral infection on mRNA transcription, protein abundance, ubiquitination and phosphorylation in a time-resolved manner (Figure 2a-e, Supplementary Tables 5 - 8, Methods). In line with previous reports<sup>8,9</sup>, we did not observe major upregulation of type-I interferons and related genes at the mRNA level (e.g. IFNB, IFIT3, MX1; Extended data Fig. 2c, d, Supplementary Table 5), suggesting active viral inhibition of this system. In contrast, SARS-CoV-2 upregulated NF-κB and stress responses, as inferred from the induction of IL6, CXCL2 and JUNs and transcription factor enrichment analysis (Extended data Fig. 2c - e, Supplementary Tables 3, 5, Methods). At the proteome level, we found 1053 regulated proteins (Figure 2a-b). Most notably, SARS-CoV-2 infection failed to induce an interferon response pattern indicative for an appropriate cellular antiviral response in this dataset (Extended data Fig. 2f, Supplementary Table 6). We complemented these data with global MS analysis of protein ubiquitination, which revealed 884 sites that were differentially regulated after SARS-CoV-2 infection (Figure 2a-c, Extended data Fig. 2f, Supplementary Table 7). A number of proteins displayed both differential abundance and dynamic ubiquitination patterns in an infection-dependent manner (e.g.

mediators of caveolar-mediated endocytosis signaling) (Figure 2c, Extended data Fig. 2f, Supplementary Tables 3, 6, 7). Notably, EFNB1, POLR2B, TYMS and DHFR showed concomitant ubiquitination and a decrease at the protein level (Figure 2c, Extended data Fig. 2f, Supplementary Tables 6, 7). Moreover, we identified two upregulated ubiquitination sites on ACE2, including one previously unknown (K702) (Figure 2c, Extended data Fig. 2f-i), suggesting an alternative post-translational mechanism of its degradation upon SARS-CoV-2 infection besides the cleavage by matrix metalloproteinases<sup>10,11</sup>. We identified multiple yet undescribed ubiquitination sites on viral proteins, which may be tied to the interactions with several E3 ligases observed in the interactome (e.g. ORF3 and TRIM47, WWP1/2, STUB1; M and TRIM7; NSP13 and RING1) (Extended data Fig. 2j, Supplementary Table 7) and likely indicative of cross-talks between ubiquitination and viral protein functions. Moreover, in the phosphoproteomic analysis, we mapped multiple novel phosphorylation sites on viral proteins (M, N, S and ORF9b), which correspond to known recognition motifs of GSK3, CSNKs, GPCR, AKT, CAMKs, and ERKs (Extended data Fig. 2k, Supplementary Table 8). Of 11,847 total quantified phosphorylation sites, 1483 showed significant changes after SARS-CoV-2 infection (Extended data Fig. 2l, Supplementary Table 8). The regulation of known phosphosites suggests an involvement of central kinases known to modulate key cellular pathways, e.g. EPHA2 – focal adhesion, RPS6Ks – cell survival, CDKs – cell cycle progression, AKT – cell growth, survival and motility, p38, JNK, ERK – stress responses, ATM, and CHEK1/2 – DNA damage response, during virus infection (Extended data Fig. 2l, m, Supplementary Tables 3, 8). Intriguingly, we could also observe an interplay of phosphorylation and ubiquitination on YAP1, a downstream regulatory target of Hippo signalling (Figure 2d), underlining the value of testing different post-translational modifications simultaneously. Combining these datasets, describing different aspects of SARS-CoV-2 infection, allowed us to determine the key pathways perturbed during the infection, such as stress and DNA damage response, regulation of transcription and cell junction organisation, at various levels (Figure 2e, see Methods).

The systematic interactome and proteome profiling of individual viral proteins provided us with the opportunity to gain deeper understanding of their molecular mechanisms. For each viral protein, we mapped the collected data onto the global network of cellular interactions<sup>12</sup> and applied a network diffusion approach<sup>13</sup>. Such analysis identifies short links of known protein-protein interactions, signaling and regulation events that connect the interactors of the viral protein with the proteins affected by its expression (Figure 3a, Extended data Fig. 3a, Supplementary Table 9). The connections predicted using the real data were significantly shorter than for the randomized data, confirming both relevance of the approach and the data quality (Extended data Fig. 3a, b). Amongst many other findings, this approach pointed towards the potential mechanisms of autophagy regulation by ORF3 and NSP6; the modulation of innate immunity by M, ORF3 and ORF7b; and the Integrin-TGF $\beta$ -EGFR-RTK signaling perturbation by ORF8 of SARS-CoV-2 (Figure 3b-d, Supplementary Table 9). Enriching these subnetworks with the SARS-CoV-2 infection-dependent mRNA abundance, protein abundance, phosphorylation and ubiquitination (Figure 3a) allowed us to gain unprecedented insights into the regulatory mechanisms employed by SARS-CoV-2 (Figure 3e, Extended data Fig. 3c, e). For instance, this analysis confirmed a role of NSP6 in autophagy<sup>14</sup> and revealed a significant inhibition of autophagic flux by ORF3 leading to the accumulation of autophagy receptors (SQSTM1, GABARAPL2, NBR1, CALCOCO2, MAP1LC3B, TAX1BP1), also observed in virus-infected cells (SQSTM1, MAP1LC3B) (Figure 3e-i). This inhibition may be explained by the interaction of ORF3 with the HOPS complex (VPS11, -16, -18, -39, -41), which is essential for autophagosome-lysosome fusion, as well as by the differential phosphorylation of regulatory sites of key components (AKT1, AKT1S1, SQSTM1, RPS6). The inhibition of the interferon response observed at transcriptional and proteome levels was similarly explained by the network diffusion analysis (Extended data Fig. 3c), demonstrating that multiple proteins of SARS-CoV-2 are employed in the disruption of antiviral immunity. Additional functional experiments corroborated the inhibition of interferon induction or signaling by ORF3, ORF6, ORF7a,

ORF7b, ORF9b (Extended data Fig. 3d, Supplementary Table 10). Upon virus infection, we observed upregulation of TGF $\beta$  and EGFR pathways, which modulate cell survival, motility and innate immune responses (Extended data Fig. 3e - g). Besides promoting virus replication, activation of these pathways has been implicated in fibrosis<sup>15,16</sup>, one of the hallmarks of COVID-19<sup>17</sup>. Specifically, our network diffusion analysis revealed the connection between the binding of ORF8 and ORF3 to TGF $\beta$ -associated factors (TGFB1, TGFB2, LTBP1, TGFBR2, FURIN, BAMBI) and the virus-induced upregulation of fibrinogens, fibronectin, SERPINE1 and integrin(s) (Extended data Fig. 3e, h)<sup>18</sup>. The phosphorylation of SMAD1/5, and ERK, JNK, p38 cascade activation, as well as an increased expression of MMPs, DUSPs, JUN, and EGR1 are indicative of TGF $\beta$  and EGFR pathway regulation. In turn, they are known to be potentiated by the increased integrin signaling and activation of YAP-dependent transcription<sup>19</sup>, which we observed upon SARS-CoV-2 infection (Extended data Fig. 3e).

Taken together, the viral-host protein-protein interactions and pathway regulations observed at multiple levels identify potential vulnerability points of SARS-CoV-2 that could be targeted by well-characterized selective drugs for antiviral therapies. To test their antiviral efficacy, we established time-lapse fluorescent microscopy of SARS-CoV-2 GFP-reporter virus infection<sup>20</sup>. Inhibition of virus replication by type-I interferon treatment corroborated the necessity for SARS-CoV-2 to block this pathway and confirmed the reliability of this screening approach (Figure 4a)<sup>9,21</sup>. We tested a panel of 48 drugs modulating the pathways perturbed by the virus for their effects on SARS-CoV-2 replication (Figure 4b, Supplementary Table 9). Notably, B-RAF (Sorafenib, Regorafenib, Dabrafenib), JAK1/2 (Baricitinib) and MAPK (SB 239063) inhibitors, among others, led to a significant increase of virus growth in our *in vitro* infection setting (Figure 4b, Extended data Fig. 4, Supplementary Table 9). In contrast, inducers of DNA damage (Tirapazamine, Rabusertib) or the mTOR inhibitor (Rapamycin) led to suppression of the virus. The highest antiviral effect was seen for Gilteritinib (a designated FLT3/AXL inhibitor), Ipatasertib (AKT inhibitor), Prinomastat and Marimastat (matrix metalloproteinases inhibitors) (Figure 4b-e, Extended data Fig. 4, Supplementary Table 9). Remarkably, these compounds profoundly inhibited replication of SARS-CoV-2 while having no or minor influence on cell growth (Extended data Fig. 4, Supplementary Table 9). These inhibitors may perturb host pathways required by the virus or influence viral protein activity through post-translational modifications. Notably, we identified AKT as a potential kinase phosphorylating SARS-CoV-2 protein N (Extended data Fig. 2k), indicating the possibility of a direct influence of Ipatasertib on the viral protein.

This drug screen demonstrates the value of our combined dataset that profiles the infection of SARS-CoV-2 at multiple levels. Further exploration of these rich data by the scientific community and investigating the interplay between different -omics levels will substantially advance our knowledge of coronavirus biology, in particular on the pathogenicity caused by highly virulent strains such as SARS-CoV-2 and SARS-CoV. Moreover, this resource may streamline the search for antiviral compounds and serve as a base for intelligent design of combination therapies that aim at targeting the virus from multiple, synergistic angles, thus potentiating the effect of individual drugs while minimizing side-effects on healthy tissues.

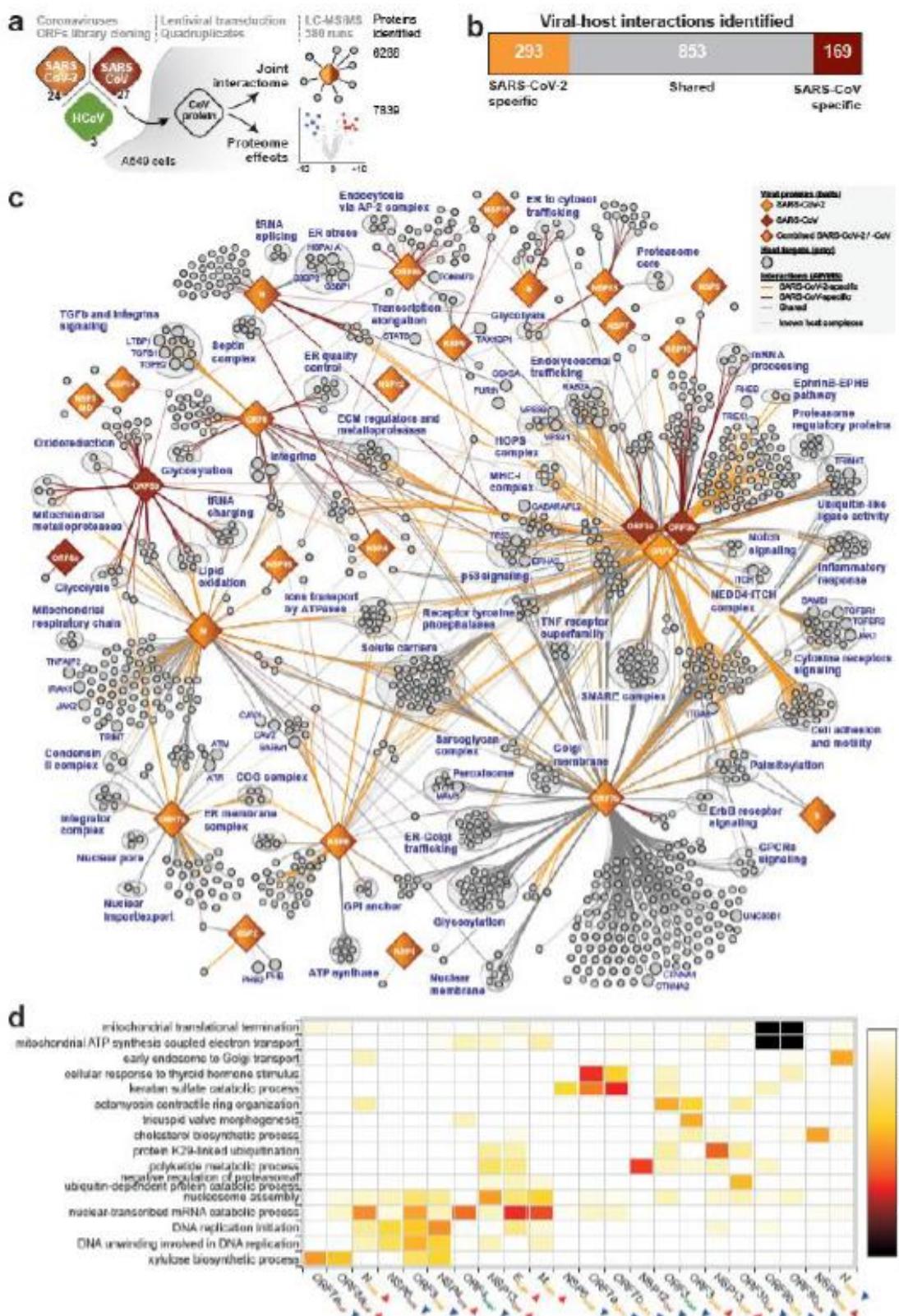
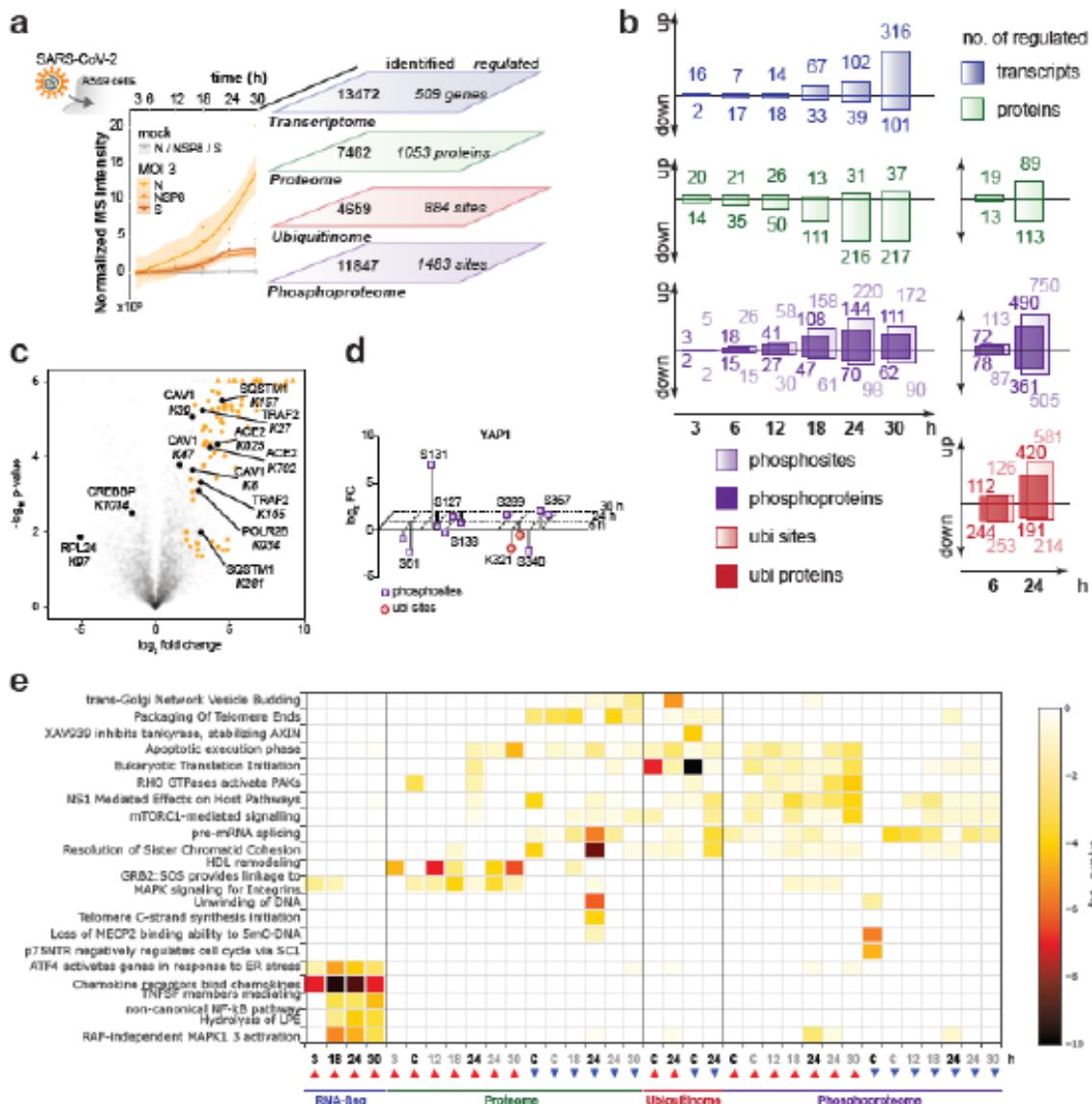


Figure legends:

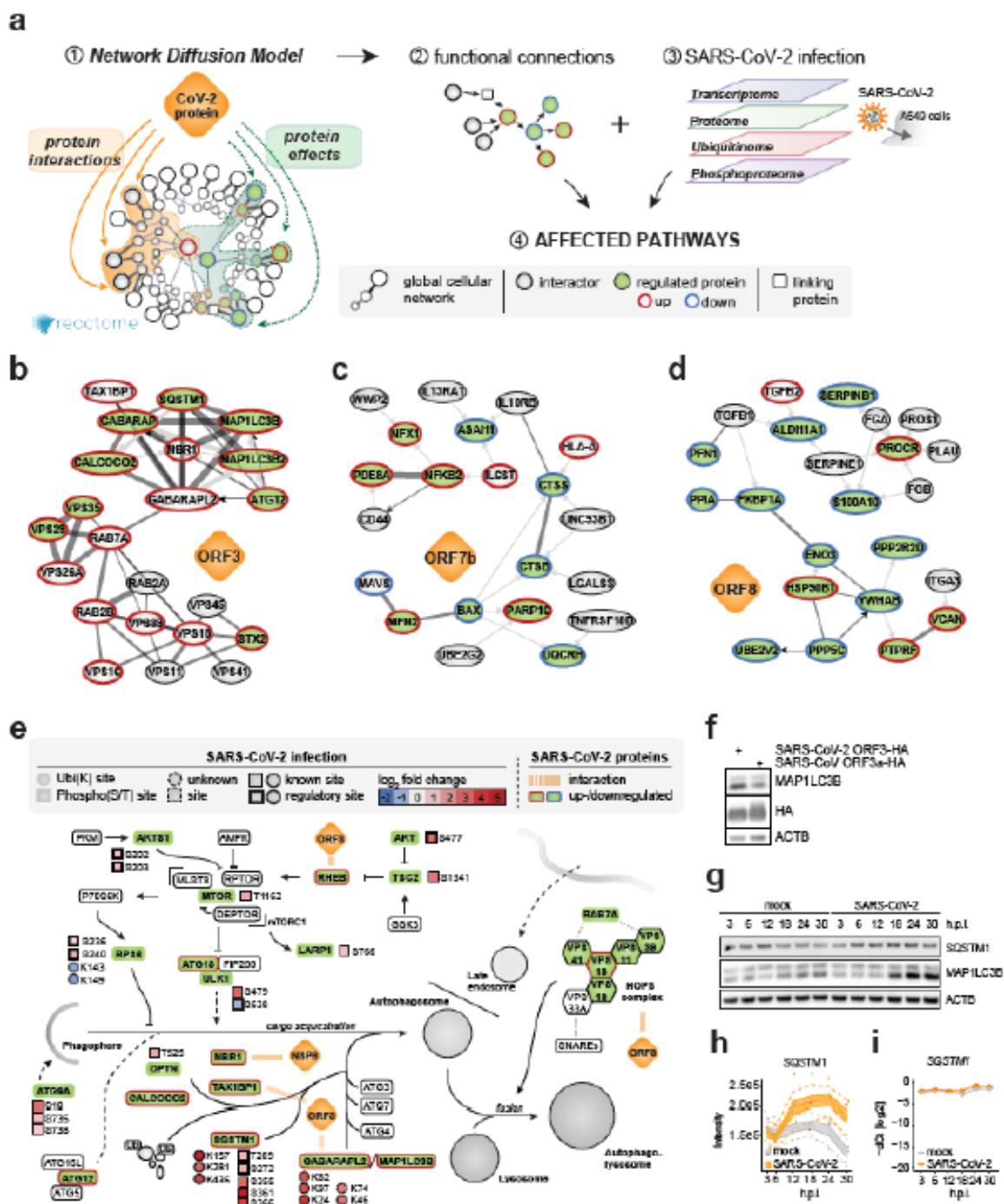
**Figure 1 | Joint analysis of SARS-CoV-2 and SARS-CoV protein-protein virus-host interactomes.**  
 (a) Experimental design to systematically compare the AP-MS interactomes and induced host proteome changes of the homologous SARS-CoV-2 and SARS-CoV viral proteins, with ORF3 homologs of

HCoV-NL63 and HCoV-229E as reference for pan coronavirus specificity. (b) Combined virus-host protein interaction network of SARS-CoV-2 and SARS-CoV measured by affinity-purification coupled to mass spectrometry. Homolog viral proteins are displayed as one node. Shared and virus-specific interactions are denoted by the edge color. (c) The numbers of unique and shared host interactions between the homologous proteins of SARS-CoV-2 and SARS-CoV. (d) Gene Ontology Biological Processes enriched among the cellular proteins that are up- (red arrow) or down- (blue arrow) regulated upon overexpression of individual viral proteins.



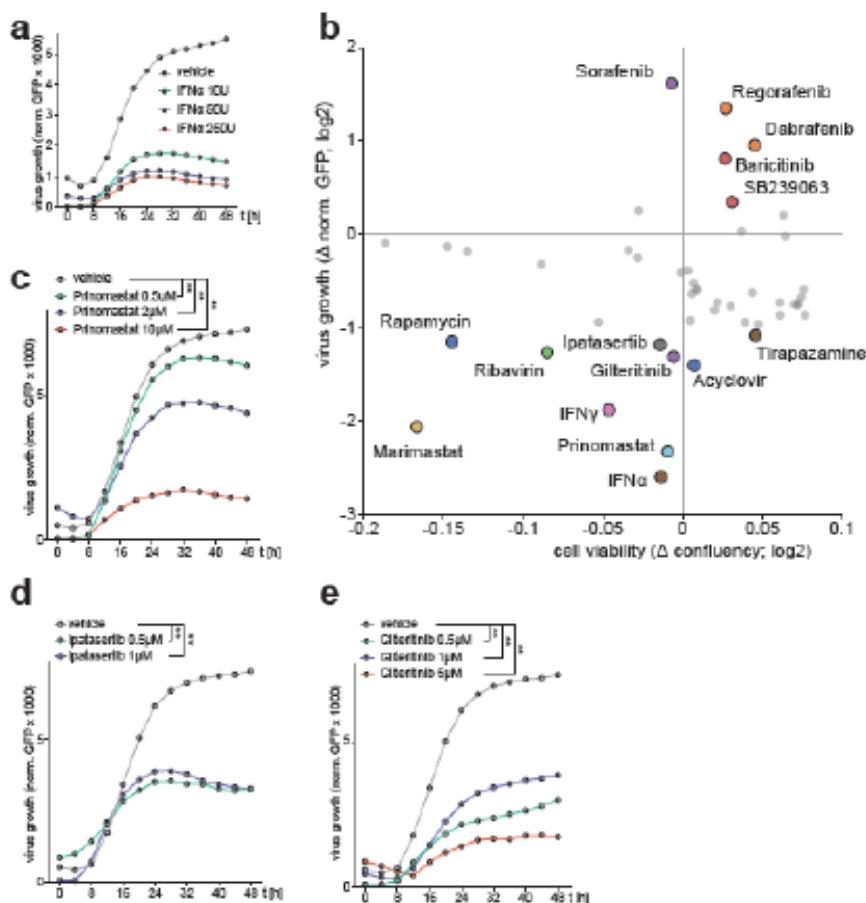
**Figure 2 | Orthogonal profiling of SARS-CoV-2 infection.** (a) Time-resolved profiling of SARS-CoV-2 infection by multiple -omics methods. The plot shows normalized MS intensities of three SARS-CoV-2 viral proteins over time. (b) Numbers of distinct transcripts, proteins, ubiquitination and phosphorylation sites, up- or down-regulated at the indicated time points after infection, as identified using data independent (DIA) or dependent (DDA) acquisition methods. (c) Volcano plot showing ubiquitination sites regulated at 24h after SARS-CoV-2 infection. Viral proteins are marked in orange. Selected significant ubiquitination sites (Student's t-test, two-tailed, permutation-based FDR < 0.05, S0 = 0.1, n = 4) are marked in black. (d) Scatter plot of phosphorylation and ubiquitination sites on Yes-

associated protein (YAP1) regulated upon SARS-CoV-2 infection. Presented are fold changes compared to mock at 6, 24, and 30 hours after infection. S61 and S127 dephosphorylation lead to nuclear translocation, S131 phosphorylation regulates protein stability. Phosphorylation of S289 and S367 is involved in cell cycle regulation. K321 deubiquitination leads to a decrease in YAP1 activation. (e) Reactome pathways enriched in up- (red arrow) or downregulated (blue arrow) transcripts, proteins, ubiquitination- and phosphorylation sites (Fisher's exact test, unadjusted). DIA MS measurements are marked in grey, DDA in black.



**Figure 3 | Network diffusion approach identifies molecular pathways linking protein-protein interactions with downstream changes in the host proteome. (a)** Network diffusion approach to

identify functional connections between the host targets of a viral protein and downstream proteome changes followed by the integration of RNA expression, protein abundance, ubiquitination and phosphorylation changes upon SARS-CoV-2 infection to streamline the identification of affected host pathways. (b-d) Subnetworks of the network diffusion predictions linking host targets of (b) SARS-CoV-2 ORF3 to the accumulation of factors involved in autophagy, (c) ORF7b to the factors involved in innate immunity and (d) ORF8 to the factors involved in TGF $\beta$  signaling. (e) Overview of perturbations to host-cell autophagy, induced by distinct proteins of SARS-CoV-2, derived from the network diffusion model and overlaid with the changes in protein levels, ubiquitination and phosphorylation induced by SARS-CoV-2 infection. (f-i) Western blot of autophagy-associated factors MAP1LC3B-II and SQSTM1 accumulation upon SARS-CoV-2 ORF3 expression in (f) HEK293R1 and (g-i) SARS-CoV-2 infection of A549-ACE2 cells. (h) Profile plot of SQSTM1 MS intensity and (i) line diagram showing *SQSTM1* mRNA level relative to *RPLP0* tested by qRT-PCR upon SARS-CoV-2 infection.



**Figure 4 | SARS-CoV-2-targeted pathways, revealed by multi-omics profiling approach, allow systematic testing of novel antiviral therapies.** (a) A549-ACE2 cells, exposed for 6h to the specified concentrations of interferon alpha and infected with SARS-CoV-2-GFP reporter virus (MOI 3). GFP signal and cell confluence were analysed by live-cell imaging for 48h. Line diagrams show virus growth over time of GFP-positive vs total cell area with indicated mean of four biological replicates. (b) A549-ACE2 cells were treated with the indicated drugs 6h prior to infection with SARS-CoV-2-GFP (MOI 3). Scatter plot represents GFP vs total cell area signal (y-axis) versus cell confluence in uninfected control treatments (x-axis) at 48h after infection. A confluence cutoff of -0.2 log2 fold change was applied to remove cytotoxic compounds. (c-e) as (a) but line diagrams showing virus replication after

(c) Prinomastat, (d) Ipatasertib and (e) Gilteritinib pre-treatment. Asterisks indicate significance to control treatment (Wilcoxon test;  $p$ -value  $\leq 0.01$ ).

## Material and Methods

### Cell lines and reagents

HEK293T, A549, Vero E6 and HEK293R1 cells and their respective culturing conditions were described previously<sup>22</sup>. All cell lines were tested to be mycoplasma-free. Expression constructs for C-terminal HA tagged viral ORFs were synthesised (Twist Bioscience and BioCat) and cloned into pWPI vector as described previously<sup>23</sup> with the following modifications: starting ATG codon was added, internal canonical splicing sites were replaced with synonymous mutations and C-terminal HA-tag, followed by amber stop codon, was added to individual viral open reading frames. C-terminally hemagglutinin(HA)-tagged ACE2 sequence was amplified from an ACE2 expression vector (kindly provided by Stefan Pöhlmann)<sup>24</sup> into the lentiviral vector pWPI-puro. A549 cells were transduced twice, and ACE2-expressing A549 (ACE2-A549) cells were selected with puromycin. Lentiviruses production, transduction of cells and antibiotic selection were performed as described previously<sup>23</sup>. RNA-isolation (Macherey-Nagel NucleoSpin RNA plus), reverse transcription (TaKaRa Bio PrimeScript RT with gDNA eraser) and RT-qPCR (Thermo-Fisher Scientific PowerUp SYBR green) were performed as described previously<sup>25</sup>. RNA-isolation for NGS applications was performed according to manufacturer's protocol (Qiagen RNeasy mini kit, RNase free DNase set). For detection of protein abundance by western blotting, HA-HRP (Sigma-Aldrich), ACTB-HRP (Santa Cruz), ATM, MAP1LC3B, MAVS, HSPA1A, TGF $\beta$  and SQSTM1, phospho-JNK (T183/Y185), JNK, phospho-p38 (T180/Y182), p38 (Cell Signaling), SARS-CoV-2 (Sino Biological) antibodies were used. For AP-MS and AP-WB applications, HA-beads (Sigma-Aldrich and Thermo Fisher Scientific) and Streptactin II beads (IBA Lifesciences) were used. Secondary Abs: HRP and WB imaging was performed as described previously<sup>25</sup>. For the stimulation of cells in the reporter assay, recombinant human interferon- $\alpha$  (IFN- $\alpha$ ) was a kind gift from Peter Stäheli, recombinant human IFN- $\gamma$  were purchased from PeproTech and IVT4 was produced as described before<sup>26</sup>. All compounds tested during the viral inhibitor assay are listed in Suppl. Table 9.

### Virus strains, stock preparation, plaque assay and *in vitro* infection

SARS-CoV-2-MUC-IMB-1 and SARS-CoV-2-GFP strains<sup>20</sup> were produced by infecting Vero E6 cells cultured in DMEM medium (10% FCS, 100  $\mu$ g/ml Streptomycin, 100 IU/ml Penicillin) for 2 days (MOI 0,01). Viral stock was harvested and spun twice (1000g/10min) before storage at -80°C. Titer of viral stock was determined by plaque assay. Confluent monolayers of VeroE6 cells were infected with serial five-fold dilutions of virus supernatants for 1 h at 37 °C. The inoculum was removed and replaced with serum-free MEM (Gibco, Life Technologies) containing 0.5% carboxymethylcellulose (Sigma-Aldrich). Two days post-infection, cells were fixed for 20 minutes at room temperature with formaldehyde directly added to the medium to a final concentration of 5%. Fixed cells were washed extensively with PBS before staining with H2O containing 1% crystal violet and 10% ethanol for 20 minutes. After rinsing with PBS, the number of plaques was counted and the virus titer was calculated.

A549-ACE2 cells were infected with SARS-CoV-2-MUC-IMB-1 strain (MOI 3) for the subsequent experiments. At each time point, the samples were washed once with 1x TBS buffer and harvested in SDC lysis buffer (100 mM Tris HCl pH 8.5; 4% SDC) or 1x SSB lysis buffer (62.5 mM Tris HCl pH 6.8; 2% SDS; 10% glycerol; 50 mM DTT; 0.01% bromophenol blue) or RLT (Qiagen) for proteome-

phosphoproteome-ubiquitinome, western blot, and transcriptome analyses, respectively. The samples were heat-inactivated and frozen at -80°C until further processing, as described in the following sections.

#### **Affinity purification mass spectrometric analyses of SARS-CoV-2, SARS-CoV and HCoV protein expressing A549 cells**

For the determination of SARS-CoV-2, SARS-CoV and partial HCoV interactomes, four replicate affinity purifications were performed for each HA-tagged viral protein. A549 cells ( $6 \times 10^6$  cells per 15-cm dish) were transduced with HA-tagged SARS-CoV-2, SARS-CoV or HCoV protein coding lentivirus and harvested three days post transduction. Cell pellets of two 15-cm dishes were lysed in lysis buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 0.2% (v/v) NP-40, 5% (v/v) glycerol, cOmplete protease inhibitor cocktail (Roche), 0.5% (v/v) 750 U/μl Sm DNase) and sonicated (5 min, 4°C, 30 sec on, 30 sec off, low settings; Bioruptor, Diagenode SA). Following normalization of protein concentrations cleared lysates, virus protein-bound host proteins were enriched by adding 50 μl anti-HA-agarose slurry (Sigma-Aldrich, A2095) with constant agitation for 3h at 4°C. Non-specifically bound proteins were removed by four subsequent washes with lysis buffer followed by three detergent-removal steps with washing buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 5% (v/v) glycerol). Enriched proteins were denatured, reduced, alkylated and digested by addition of 200 μl digestion buffer (0.6 M GdmCl, 1 mM TCEP, 4 mM CAA, 100 mM Tris-HCl pH 8, 0.5 μg LysC (WAKO Chemicals), 0.5 μg trypsin (Promega) at 30°C overnight. Peptide purification on StageTips with three layers of C18 Empore filter discs (3M) and subsequent mass spectrometry analysis was performed as described previously<sup>22,23</sup>. Briefly, purified peptides were loaded onto a 20 cm reverse-phase analytical column (75 μm diameter; ReproSil-Pur C18-AQ 1.9 μm resin; Dr. Maisch) and separated using an EASY-nLC 1200 system (Thermo Fisher Scientific) with a 90 min gradient (80% acetonitrile, 0.1% formic acid; 5% (80% acetonitrile) to 30% for 65 min, 30% to 95% for 10 min, wash out at 95% for 5 min, readjustment to 5% in 10 min) at a flow rate of 300 nL per min. Eluting peptides were directly analyzed on a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) with data-dependent acquisition including repeating cycles of one MS1 full scan (150-2,000 m/z, R = 60,000 at 200 m/z) followed by 15 MS2 scans of the highest abundant isolated and higher-energy collisional dissociation fragmented peptide precursors.

#### **Proteome analyses of SARS-CoV-2, SARS-CoV and HCoV protein expressing cells**

For the determination of proteome changes in A549 cells expressing SARS-CoV-2, SARS-CoV or HCoV proteins, a fraction of  $1 \times 10^6$  lentivirus-transduced cells from the affinity purification samples were lysed in guanidinium chloride buffer (6 M GdmCl, 10 mM TCEP, 40 mM CAA, 100 mM Tris-HCl pH 8), boiled at 95°C for 8 min and sonicated (10 min, 4°C, 30 sec on, 30 sec off, high settings). Protein concentrations of cleared lysates were normalized to 50 μg and proteins were pre-digested with 1 μg LysC at 37°C for 1h followed by a 1:10 dilution (100 mM Tris-HCl pH 8) and overnight digestion with 1 μg trypsin at 30°C. Peptide purification on StageTips with three layers of C18 Empore filter discs (3M) and subsequent mass spectrometry analysis was performed as described previously<sup>22,23</sup>. Briefly, 300 ng of purified peptides were loaded onto a 50 cm reversed phase column (75 μm inner diameter, packed in house with ReproSil-Pur C18-AQ 1.9 μm resin [Dr. Maisch GmbH]). The column temperature was maintained at 60°C using a homemade column oven. A binary buffer system, consisting of buffer A (0.1% formic acid (FA)) and buffer B (80% ACN, 0.1% FA), was used for peptide separation, at a flow rate of 300 nL/min. An EASY-nLC 1200 system (Thermo Fisher Scientific), directly coupled online with the mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) via a nano-electrospray source, was employed for nano-flow liquid chromatography. Peptides were eluted by a linear 80 min gradient from 5% to 30% buffer B (0.1% v/v formic acid, 80% v/v acetonitrile), followed by a 4 min

increase to 60% B, a further 4 min increase to 95% B, a 4 min plateau phase at 95% B, a 4 min decrease to 5% B and a 4 min wash phase of 5% B. To acquire MS data, the data-independent acquisition (DIA) scan mode operated by the XCalibur software (Thermo Fisher) was used. DIA was performed with one full MS event followed by 33 MS/MS windows in one cycle resulting in a cycle time of 2.7 seconds. The full MS settings included an ion target value of  $3 \times 10^6$  charges in the 300 – 1650 m/z range with a maximum injection time of 60 ms and a resolution of 120,000 at m/z 200. DIA precursor windows ranged from 300.5 m/z (lower boundary of first window) to 1649.5 m/z (upper boundary of 33rd window). MS/MS settings included an ion target value of  $3 \times 10^6$  charges for the precursor window with an Xcalibur-automated maximum injection time and a resolution of 30,000 at m/z 200.

To generate the proteome library for DIA measurements purified peptides from the first replicates and the fourth replicates of all samples were pooled separately and 25  $\mu$ g of peptides from each pool were fractionated into 24 fractions by high pH reversed-phase chromatography as described earlier<sup>27</sup>. During each separation, fractions were concatenated automatically by shifting the collection tube every 120 seconds. In total 48 fractions were dried in a vacuum centrifuge, resuspended in buffer A\* (0.3% TFA/ 2% ACN) and subsequently analyzed by a top12 data-dependent acquisition (DDA) scan mode using the same LC gradient and settings. The mass spectrometer was operated by the XCalibur software (Thermo Fisher). DDA scan settings on full MS level included an ion target value of  $3 \times 10^6$  charges in the 300 – 1650 m/z range with a maximum injection time of 20 ms and a resolution of 60,000 at m/z 200. At the MS/MS level the target value was  $10^5$  charges with a maximum injection time of 60 ms and a resolution of 15,000 at m/z 200. For MS/MS events only, precursor ions with 2-5 charges that were not on the 20 s dynamic exclusion list were isolated in a 1.4 m/z window. Fragmentation was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27eV.

#### Infected proteome-phosphoproteome time-course

Frozen cell lysates of infected A549-ACE2 harvested at 3, 6, 12, 18, 24 and 30h post infection were thawed on ice and boiled for 5 min at 95 degrees. Lysates were transferred to a 96-well plate (Covaris) and sonicated for 5 min. Protein concentrations were estimated by tryptophan assay<sup>28</sup> and protein material was equalized to 200  $\mu$ g per sample. CAA (10 mM) and TCEP (40 mM) along with trypsin (1:100 w/w, Sigma-Aldrich) and LysC (1/100 w/w, Wako) were added to the samples that were digested at 37°C overnight. Peptides were desalted using SDB-RPS cartridges (PreOmics).

Briefly, samples were mixed with 300  $\mu$ l 1% TFA in isopropanol, loaded onto cartridges and washed with 200  $\mu$ l 1% TFA in isopropanol and 200  $\mu$ l 0.2% TFA. Peptides were eluted with 150  $\mu$ l of 1.25% Ammonium hydroxide (NH<sub>4</sub>OH)/ 80% ACN and 10  $\mu$ l aliquots were taken and dried separately for global proteome analysis. The rest was dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus) and resuspended in 105  $\mu$ l of equilibration buffer (1% TFA/ 80% ACN) for phosphopeptide enrichment. The AssayMAP Bravo robot (Agilent) performed the enrichment for phosphopeptides by priming AssayMAP cartridges (packed with 5  $\mu$ l Fe(III)-NTA) with 1 % TFA in 99 % ACN followed by equilibration in equilibration buffer and loading of peptides. Enriched phosphopeptides were eluted with 1 % Ammonium hydroxide, dried in a vacuum centrifuge and resuspended in 1 % FA buffer. Evotips were activated by wetting in 40 ml 1-propanol in Evotipbox and subsequently washed with 100  $\mu$ l of 0.1% FA. Peptides were loaded onto tips which were subsequently washed with 100  $\mu$ l of 0.1 % FA. The tips were then loaded with 100  $\mu$ l 0.1 % FA and centrifuged very shortly.

To generate the proteome library for DIA measurements cells were lysed in 4 % SDC and 100 mM Tris pH 8.4, followed by sonication, protein quantification, reduction, and alkylation and desalting using SDB-RPS cartridges (see above). 100  $\mu$ g of peptides were fractionated into 24 fractions by high pH

reversed-phase chromatography as described earlier<sup>27</sup>. Fractions were concatenated automatically by shifting the collection tube every 120 seconds and subsequently dried in a vacuum centrifuge and resuspended in buffer A\* (0.3% TFA/ 2% ACN).

To generate the library for phosphoproteome DIA measurements A549 cells were treated with 100 ng/ml Calyculin and 2 mM Sodium orthovanadate for 20 min. Cells were lysed and treated as for the proteome library generation. After overnight digestion, peptides were desalted using Sepax Extraction columns. 5.5 mg of desalted peptides were fractionated into 84 fractions on a C18 reversed-phase column (4.6 x 150 mm, 3.5  $\mu$ m bead size) under basic conditions using a Shimadzu UFCL operating at 1 ml/minute. Buffer A (2.5 mM Ammoniumbicarbonat in MQ) and Buffer B (2.5 mM ABC in 80% CAN) were used to separate peptides on a linear gradient of 2.5% B to 44% B for 64 minutes and 44% B to 75% B for 5 minutes before a rapid increase to 100% B which was kept for 5 minutes. Fractions were subsequently concatenated into 24 fractions and lyophilized. Phosphopeptides of these 24 fractions were enriched using the AssayMAP Bravo robot and loaded on Evtips. Another 5.5 mg of desalted peptides were split into 24 samples and enriched for phosphopeptides by the AssayMAP Bravo robot. Eluted phosphopeptides were combined, dried and fractionated into 24 fractions by neutral pH reversed-phase chromatography<sup>27</sup>. Fractions were dried and loaded on Evtips as described above.

#### Infected proteome-phosphoproteome-diGly-proteome (6 and 24hr)

Frozen cell lysates of infected A549-ACE2 harvested at 6 and 24h post infection were thawed on ice and sonicated for 1 min (Branson Sonifier). Protein concentrations were estimated by tryptophan assay<sup>28</sup>. To reduce and alkylate proteins, samples were incubated for 5 min at 45°C with CAA and TCEP, final concentrations of 10 mM and 40 mM, respectively. Samples were digested overnight at 37°C using trypsin (1:100 w/w, Sigma-Aldrich) and LysC (1/100 w/w, Wako).

For proteome analysis, 10  $\mu$ g of peptide material were desalted using SDB-RPS StageTips (Empore) (2). Briefly, samples were diluted with 1% TFA in isopropanol to a final volume of 200  $\mu$ l and loaded onto StageTips, subsequently washed with 200  $\mu$ l of 1% TFA in isopropanol and 200  $\mu$ l 0.2% TFA/ 2% ACN. Peptides were eluted with 60  $\mu$ l of 1.25% Ammonium hydroxide (NH<sub>4</sub>OH)/ 80% ACN and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus). They were resuspended in buffer A\* prior to LC-MS/MS analysis. Peptide concentrations were measured optically at 280nm (Nanodrop 2000, Thermo Scientific) and subsequently equalized using buffer A\*. 500ng peptide was subjected to LC-MS/MS analysis.

To generate phosphoproteome data, the EasyPhos protocol was used for the enrichment of phosphopeptides<sup>29</sup>. Briefly, samples were adjusted with the lysis buffer to a volume of 300  $\mu$ l, transferred into a 96-deep-well plate and mixed with 100  $\mu$ l 48% TFA, 8 mM KH<sub>2</sub>PO<sub>4</sub>. Phosphopeptides were captured by 5 min incubation at 40°C with 5 mg TiO<sub>2</sub> beads. Thereafter, beads were washed 5 times with 5% TFA/ 60% isopropanol, followed by a transfer in 0.1% TFA/ 60% isopropanol into C8 StageTips. Phosphopeptides were eluted twice with 30  $\mu$ l 20% NH<sub>4</sub>OH/ 40% ACN and concentrated for 30 min at 45°C using a SpeedVac centrifuge. Concentrated samples were immediately diluted with 100  $\mu$ l 1% TFA in isopropanol and transferred into SDB-RPS StageTips. Peptides were washed, eluted and dried as described above. Dried peptides were resuspended in 6  $\mu$ l buffer A\* and 5  $\mu$ l was subjected to LC-MS/MS analysis.

For diGly peptide enrichment, samples were four-fold diluted with 1% TFA in isopropanol and loaded onto SDB-RPS cartridges (Strata<sup>TM</sup>-X-C, 30 mg/ 3 ml, Phenomenex Inc), pre-equilibrated with 4 ml 30% MeOH/1% TFA and washed with 4 ml 0.2% TFA. Samples were washed twice with 4 ml 1% TFA

in isopropanol, once with 0.2% TFA/ 2% ACN and eluted twice with 2 ml 1.25% NH<sub>4</sub>OH/ 80% ACN. Eluted peptides were diluted with ddH<sub>2</sub>O to a final ACN concentration of 35%, snap frozen and lyophilized. Lyophilized peptides were reconstituted in IAP buffer (50 mM MOPS, pH 7.2, 10 mM Na<sub>2</sub>HPO<sub>4</sub>, 50 mM NaCl) and the peptide concentration was estimated by tryptophan assay. K-ε-GG remnant containing peptides were enriched using the PTMScan® Ubiquitin Remnant Motif (K-ε-GG) Kit (Cell Signaling Technology). Crosslinking of antibodies to beads and subsequent immunopurification was performed with slight modifications as previously described<sup>30</sup>. Briefly, two vials of crosslinked beads were combined and equally split into 16 tubes (~31 µg of antibody per tube). Equal peptide amounts (800 µg) were added to crosslinked beads and the volume was adjusted with IAP buffer to 1 ml. After 1h of incubation at 4°C and gentle agitation, beads were washed twice with cold IAP and 5 times with cold ddH<sub>2</sub>O. Thereafter, peptides were eluted twice with 50 µl 0.15% TFA. Eluted peptides were desalted and dried as described for proteome analysis with the difference that 0.2% TFA instead of 1%TFA in isopropanol was used for the first wash. Eluted peptides were resuspended in 9 µl buffer A\* and 4 µl was subjected to LC-MS/MS analysis.

#### DDA Measurements

Samples were loaded onto a 50 cm reversed phase column (75 µm inner diameter, packed in house with ReproSil-Pur C18-AQ 1.9 µm resin [Dr. Maisch GmbH]). The column temperature was maintained at 60°C using a homemade column oven. A binary buffer system, consisting of buffer A (0.1% formic acid (FA)) and buffer B (80% ACN plus 0.1% FA), was used for peptide separation, at a flow rate of 300 nL/min. An EASY-nLC 1200 system (Thermo Fisher Scientific), directly coupled online with the mass spectrometer (Q Exactive HF-X, Thermo Fisher Scientific) via a nano-electrospray source, was employed for nano-flow liquid chromatography.

For proteome measurements, we used a gradient starting at 5% buffer B and stepwise increasing to 30% in 95 min, 60% in 5 min and 95% in 5 min. The mass spectrometer was operated in Top12 data-dependent mode (DDA) with a full scan range of 300-1650 m/z at 60,000 resolution with an automatic gain control (AGC) target of 3e6 and a maximum fill time of 20ms. Precursor ions were isolated with a width of 1.4 m/z and fragmented by higher-energy collisional dissociation (HCD) (NCE 27%). Fragment scans were performed at a resolution of 15,000, an AGC of 1e5 and a maximum injection time of 60 ms. Dynamic exclusion was enabled and set to 30 s.

For phosphopeptide samples 5µl were loaded and eluted with a gradient starting at 3% buffer B and stepwise increased to 19% in 60 min, 41% in 30 min, 36% in 39 min and 95% in 5 min. The mass spectrometer was operated in Top10 DDA with a full scan range of 300-1600 m/z at 60,000 resolution with an AGC target of 3e6 and a maximum fill time of 120 ms. Precursor ions were isolated with a width of 1.4 m/z and fragmented by HCD (NCE 28%). Fragment scans were performed at a resolution of 15,000, an AGC of 1e5 and a maximum injection time of 50 ms. Dynamic exclusion was enabled and set to 30 s.

For the analysis of K-ε-GG peptide samples, we use a gradient starting at 3% buffer B and stepwise increased to 7% in 6 min, 20% in 49 min, 36% in 39 min, 45% in 10 min and 95% in 4 min. The mass spectrometer was operated in Top12 DDA with a full scan range of 300-1350 m/z at 60,000 resolution with an AGC target of 3e6 and a maximum fill time of 20ms. Precursor ions were isolated with a width of 1.4 m/z and fragmented by HCD (NCE 28%). Fragment scans were performed at a resolution of 30,000, an AGC of 1e5 and a maximum injection time of 110 ms. Dynamic exclusion was enabled and set to 15 s.

### DIA Measurements

Samples were loaded onto a 15 cm reversed phase column (150  $\mu$ m inner diameter, packed in house with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin [Dr. Maisch GmbH]), which was kept in a homemade column oven at 60°C. Peptides were separated by the EVOsep One LC system using the pre-programmed 44 minutes gradient for proteome samples and the 21 minutes gradient for phosphoproteome samples. The same gradients were used for the acquisition of proteome and phosphoproteome library fractions. The EVOsep One system was coupled to a Q Exactive HF-X Orbitrap (Thermo Fisher Scientific) via a nano-electrospray source.

The proteome and phosphoproteome fractions we used to build the libraries were measured in DDA mode. To acquire proteome fractions the mass spectrometer was operated in Top15 data-dependent mode with a full scan range of 300-1650 m/z at 60,000 resolution, an automatic gain control (AGC) target of 3e6 and a maximum fill time of 20ms. For the generation of the phosphoproteome library the mass spectrometer was operated in Top12 data-dependent mode (DDA) with a full scan range of 300-1650 m/z at 60,000 resolution, an automatic gain control (AGC) target of 3e6 and a maximum fill time of 25ms. For both libraries precursor ions were isolated with a width of 1.4 m/z and fragmented by higher-energy collisional dissociation (HCD) (NCE 27%). Fragment scans were performed at a resolution of 15,000, an AGC of 1e5 and a maximum fill time of 28 ms. Dynamic exclusion was enabled and set to 30 s for the proteome library and 20 s for the phosphoproteome library.

For proteome and phosphoproteome DIA measurements, full MS resolution was set to 60,000 with a full scan range of 300-1650 m/z, a maximum fill time of 60 ms and an automatic gain control (AGC) target of 3e6. One full scan was followed by 32 windows with a resolution of 30,000 and a maximum fill time of 54 ms for proteome measurements and 40 windows for phosphoproteome measurements with a resolution of 15,000 and maximum fill time of 28 ms in profile mode. Precursor ions were fragmented by higher-energy collisional dissociation (HCD) (NCE 27%).

### Data-processing of affinity purification, (phospho)proteome and ubiquitinome LC-MS/MS analyses

Raw MS data files of experiments conducted in DDA mode were processed with MaxQuant (version 1.6.14) using the standard settings and label-free quantification enabled (LFQ min ratio count 1, normalization type none, stabilize large LFQ ratios disabled). For profiling of post-translational modifications, additional variable modifications for ubiquitination (GlyGly(K)) and phosphorylation (Phospho(STY)) were added. Spectra were searched against forward and reverse sequences of the reviewed human proteome including isoforms (UniprotKB, release 10.2019) and SARS-CoV-2, SARS-CoV and HCoV proteins by the built-in Andromeda search engine<sup>31</sup>.

For AP-MS data, the alternative protein group definition was used: only the peptides identified in AP-MS samples could be regarded as protein group-specific, protein groups that differed by the single specific peptide or had less than 25% different specific peptides were merged to extend the set of peptides used for protein group quantitation and reduce the number of protein isoform-specific interactions.

For the experiments conducted in DIA mode, Spectronaut version 13 (Biognosys) was used to generate the proteome and phosphoproteome libraries from DDA runs by combining files of respective fractionations using the human fasta file (Uniprot, 2019, 42,431 entries). For the generation of the proteome library default settings were left unchanged. For the phosphoproteome library generation 2 x 24 files received by both fractionation strategies were combined and phosphorylation at

Serine/Threonine/Tyrosine was added as variable modification to default settings. Maximum number of fragment ions per peptide was increased from 6 to 25. Proteome DIA files were analyzed using the proteome library with default settings and disabled cross run normalization. Phospho DIA files were analyzed using the phosphoproteome library using default settings with disabled PTM localization filter and cross run normalization. To search for viral proteins, we also generated the “hybrid” spectral library by merging DDA proteome library with a direct-DIA library generated from the DIA analysis of DIA proteome samples. For this search, the sequences of viral proteins were added to the human fasta file.

### Bioinformatic analysis

Unless otherwise specified, the bioinformatic analysis was done in R (version 3.6), Julia (version 1.4) and Python (version 3.8) using a collection of in-house scripts (available upon request).

#### Statistical analysis of MS data

For all MS datasets, except DDA phosphoproteome and ubiquitome data, the Bayesian linear random effects models were used to define how the abundances of proteins change between the conditions. To specify and fit the models we employed *msglm* R package (<https://gitlab.lrz.de/tumevi4/msglm>), which depends on *rstan* package (version 2.19)<sup>32</sup> for inferring the posterior distribution of the model parameters. In all the models, the effects corresponding to the experimental conditions have regularized horseshoe+ priors<sup>33</sup>, while the batch effects have normally distributed priors. Laplacian distribution was used to model the instrumental error of MS intensities. For each MS instrument used, the heteroscedastic intensities noise model was calibrated with the technical replicate MS data of the instrument. These data were also used to calibrate the logit-based model of missing MS data (the probability that the MS instrument will fail to identify the protein given its expected abundance in the sample). The model was fit using unnormalized MS intensities data. Instead of transforming the data by normalization, the inferred protein abundances were scaled by the normalization multiplier of each individual MS sample to match the expected MS intensity of that sample. This allows taking the signal-to-noise variation between the samples into account when fitting the model. Due to high computational intensity, the model was applied to each protein group separately. For all the models, 4000 iterations (2000 warmup + 2000 sampling) of the No-U-Turn Markov Chain Monte Carlo were performed in 7 or 8 independent chains, every 4th sample was collected for posterior distribution of the model parameters. For estimating the statistical significance of protein abundance changes between the two experimental conditions, the *P-value* was defined as the probability that a random sample from the posterior distribution of the first condition would be smaller (larger) than a random sample drawn from the second condition. No multiple hypothesis testing corrections were applied, since this is handled by the choice of the model priors.

#### Statistical analysis of AP-MS data and filtering for specific interactions

Given the sparsity of the AP-MS data (each peptide is quantified in a small fraction of experiments), to take advantage of the missing data modeling by *msglm*, the statistical model was applied directly to the MS1 intensities of protein group-specific LC peaks (*evidence.txt* table of MaxQuant output). In R GLM formula language, the model could be specified as

$$\log(\text{Intensity}) \sim 1 + APMS + Bait + Bait:Virus + MS1peak + MSbatch,$$

where *APMS* effect models the average shift of intensities in AP-MS data in comparison to full proteome samples, *Bait* is the average enrichment of a protein in AP-MS experiments of homologous proteins of both SARS-CoV and SARS-CoV-2, and *Bait:Virus* corresponds to the virus-specific changes in protein enrichment. *MS1peak* is the log-ratio between the intensity of a given peak and the total protein abundance (the peak is defined by its peptide sequence, PTMs and the charge; it is assumed that the peak ratios do not depend on experimental conditions<sup>34</sup>), and *MSbatch* accounts for batch-specific

variations of protein intensity. *APMS*, *Bait* and *Bait:Virus* effects were used to reconstruct the batch effect-free abundance of the protein in AP-MS samples.

The modeling provided the enrichment estimates for each protein in each AP experiment. Specific AP-MS interactions had to pass the two tests. In the first test, the enrichment of the candidate protein in a given bait AP was compared against the background, which was dynamically defined for each interaction to contain the data from all other baits, where the abundance of the candidate was within 50%-90% percentile range (excluding top 10% baits from the background allowed the protein to be shared by a few baits in the resulting AP-MS network). The non-targeting control and Gaussia luciferase baits were always preserved in the background. Similarly, to filter out any potential side-effects of very high bait protein expression, the ORF3 homologs were always present in the background of M interactors and vice versa. To rule out the influence of the batch effects, the second test was applied. It was defined similarly to the first one, but the background was constrained to the baits of the same batch, and 40%-80% percentile range was used. In both tests, the protein has to be 4 times enriched against the background (16 times for highly expressed baits: ORF3, M, NSP13, NSP5, NSP6, ORF3a, ORF7b, ORF8b, HCoV ORF4a) with the *P-value*  $\leq 1E-3$ .

Additionally, we excluded the proteins that, in the viral protein expression data, have shown upregulation, and their enrichment in AP-MS data was less than 16 times stronger than observed upregulation effects. Finally, to exclude the carryover of material between the samples sequentially analyzed by MS, we removed the putative interactors, which were also enriched at higher levels in the samples of the preceding bait, or the one before it.

For the analysis of interaction specificity between the homologous viral proteins, we estimated the significance of interaction enrichment difference (corrected by the average difference between the enrichment of the shared interactors to adjust for the bait expression variation). Specific interactions have to be 4 times enriched in comparison to the homolog with *P-value*  $\leq 1E-3$ .

#### Statistical analysis of DIA proteome effects upon viral protein overexpression

The statistical model of the viral protein overexpression data set was similar to AP-MS data, except that protein-level intensities provided by Spectronaut were used. The PCA analysis of the protein intensities has identified that the 2nd principal component is associated with the batch-dependent variations between the samples. To exclude their influence, this principal component was added to the experimental design matrix as an additional batch effect.

As with AP-MS data, the two statistical tests were used to identify the significantly regulated proteins. First, the absolute value of  $\log_2$ -fold change of protein abundance upon overexpression of a given viral protein in comparison to the control samples had to be above 0.25 with *P-value*  $\leq 1E-3$ . Second, the protein had to be significantly regulated (same  $\log_2$ -fold change and *P-value* applied) against the background distribution of its abundance in the selected experiments of the same batch (experiments, where the tested protein abundance was within the 20%-80% percentile range of the whole batch, were dynamically selected for each protein).

#### Statistical analysis of DDA proteome data of virus infection

For DDA proteome data the following linear model was used:

$$\log(\text{Intensity}(t)) \sim \sum_{t_i \leq t} (\text{after}(t_i) + \text{treatment:after}(t_i)),$$

where *after(24h)* effect corresponds to the protein abundance changes in mock-infected samples that happened between 6h and 24h, and *treatment:after(t<sub>i</sub>)* (*t*=6,24) is the effect of interaction between the SARS-CoV-2 infection and the timepoint, and corresponds to the changes in the virus-infected cells (in

comparison to the mock-infected samples) within the first 6h hours after infection, and between 6h and 24h after infection, respectively. The absolute value of  $\log_2$ -fold change between the conditions below 0.25 and the corresponding  $P\text{-value} \leq 1\text{E-}3$  criteria were used to define the significant changes.

#### Statistical analysis of DIA proteome and phosphoproteome data of virus infection

For DIA phosphoproteome data, to convert peptide-level output of Spectronaut into PTM site-level report, the Peptide Collapse Perseus plugin was used ([https://github.com/AlexHgO/Perseus\\_Plugin\\_Peptide\\_Collapse](https://github.com/AlexHgO/Perseus_Plugin_Peptide_Collapse))<sup>35</sup>. Phosphosites with less than 0.75 localization probability were excluded.

For DIA proteome and phosphoproteome datasets, the following linear model was used:

$$\log(\text{Intensity}(t)) \sim \sum_{t_i \leq t} (\text{after}(t_i) + \text{treatment:after}(t_i)),$$

where  $\text{after}(t_i)$  effect corresponds to the protein abundance changes in mock-infected samples that happened between  $t_{i-1}$  and  $t_i$  ( $t=6,12,18,24,30$ ), and  $\text{treatment:after}(t_i)$  ( $t=3,6,12,18,24,30$ ) is the effect of interaction between the SARS-CoV-2 infection and the timepoint. The absolute value of  $\log_2$ -fold change between the conditions below 0.25 and the corresponding  $P\text{-value} \leq 1\text{E-}3$  criteria were used to define the significant changes for proteome data, and  $|\log_2\text{-fold change}| \geq 0.5$ ,  $P\text{-value} \leq 1\text{E-}2$  for phosphoproteome data.

#### Statistical analysis of DDA total proteome, phosphoproteome and ubiquitinome data 6 and 24 hours post SARS-CoV-2 infection of A549-ACE2 cells

The output of MaxQuant was analyzed with Perseus (version 1.6.14.0)<sup>36</sup> and visualized with R (version 3.6.0) and RStudio (version 1.2.1335). For total proteome analysis, detected protein groups within the proteinGroups output table identified as known contaminants, reverse sequence matches, only identified by site or quantified in less than 3 out of 4 replicates in at least one condition were excluded. Following  $\log_2$  transformation, missing values were imputed for each replicate individually by sampling values from a normal distribution calculated from the original data distribution (width =  $0.3 \times \text{s.d.}$ , downshift =  $-1.8 \times \text{s.d.}$ ). Differentially regulated protein groups between mock and SARS-CoV-2 infection at 6 and 24 h.p.i. were identified via two-sided Student's T-tests corrected for multiple hypothesis testing applying a permutation-based FDR ( $S_0 = 0.1$ ;  $\text{FDR} < 0.05$ , 250 randomizations). Protein groups were further removed for statistical testing if not at least one T-test condition contained a minimum of three non-imputed values. For phosphoproteome analysis, phosphosites within the Phospho (STY)Sites output table identified as known contaminants, reverse sequence matches or less than 0.75 localization probability were excluded. Following  $\log_2$  transformation, phosphosite intensities were normalized based on sites that were quantified in at least 90% of all samples to account for technical variations. In detail, the median of phosphosite-specific intensities across samples was subtracted from individual intensities to normalize for different phosphosite abundances (row-wise normalization). Next, the median of normalized phosphosite intensities per sample was used as final normalization factor and subtracted from individual non-normalized phosphosite intensities (column-wise normalization). Phosphosites were further filtered for quantification in at least 3 replicates and missing values were imputed for each replicate individually by sampling values from a normal distribution calculated from the original data distribution (width =  $0.3 \times \text{s.d.}$ , downshift =  $-1.8 \times \text{s.d.}$ ). Differentially regulated phosphosites between mock and SARS-CoV-2 infection at 6 and 24 h.p.i. were identified via two-sided Student's T-tests corrected for multiple hypothesis testing applying a permutation-based FDR ( $S_0 = 0.5$ ;  $\text{FDR} < 0.05$ , 250 randomizations). Phosphosites were further removed from statistical testing if not at least one T-test condition contained a minimum of 3 non-imputed values. For ubiquitinome analysis, ubiquitination sites within the GlyGly (K)Sites output table were processed as mention for phosphosites,

but normalization for technical variation was based on ubiquitination sites quantified in more than 60% of all samples. Differentially regulated ubiquitination sites between mock and SARS-CoV-2 infection at 6 and 24 h.p.i. were identified via two-sided Student's T-tests corrected for multiple hypothesis testing applying a permutation-based FDR ( $S_0 = 0.1$ ; FDR  $< 0.05$ , 250 randomizations) and removed from statistical testing if not at least one T-test condition contained a minimum of 3 non-imputed values. Annotation of detected protein groups, phosphosites and ubiquitination sites with GOBP, -MF, -CC, KEGG, Pfam, GSEA, Keywords and Corum as well as PhosphoSitePlus kinase-substrate relations and regulatory sites (version May 1st 2020)<sup>37</sup> was performed in Perseus.

#### Transcriptomic analysis of SARS-CoV-2 infected A549-ACE2 cells

A549-ACE2 cells used for transcriptional profiling of SARS-CoV-2 infection were cultured and infected as described above. RNA isolation was performed using RNeasy Mini kit (Qiagen) according to the manufacturer's protocol with addition of a heat inactivation step after cell lysis. Library preparation for bulk 3'-sequencing of poly(A)-RNA was done as described previously<sup>38</sup>. Briefly, barcoded cDNA of each sample was generated with a Maxima RT polymerase (Thermo Fisher) using oligo-dT primer containing barcodes, unique molecular identifiers (UMIs) and an adapter. 5' ends of the cDNAs were extended by a template switch oligo (TSO) and after pooling of all samples full-length cDNA was amplified with primers binding to the TSO-site and the adapter. cDNA was fragmented and TruSeq-Adapters ligated with the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina® (NEB) and 3'-end-fragments were finally amplified using primers with Illumina P5 and P7 overhangs. In comparison to Parekh *et al.*<sup>38</sup> the P5 and P7 sites were exchanged to allow sequencing of the cDNA in read1 and barcodes and UMIs in read2 to achieve better cluster recognition. The library was sequenced on a NextSeq 500 (Illumina) with 75 cycles for the cDNA in read1 and 16 cycles for the barcodes and UMIs in read2.

As for the analysis of the transcriptome data, Gencode gene annotations v28 and the human reference genome GRCh38 were derived from the Gencode homepage (EMBL-EBI). Dropseq tool v1.12<sup>39</sup> was used for mapping raw sequencing data to the reference genome. The resulting UMI filtered count matrix was imported into R v3.4.4. CPM (counts per million) values were calculated for the raw data and genes having a mean cpm value less than 1 were removed from the dataset. Prior differential expression analysis with DESeq2 v1.18.1<sup>40</sup>, dispersion of the data was estimated with a parametric fit using a multiplicative model where infection status (MOCK, virus infected) and time were incorporated as covariates in the model matrix. The Wald test was used for determining differentially regulated genes across timepoints in individual infection states and shrunk log2 fold changes were calculated afterwards. Transcripts with low mean normalized count that were flagged by the independent filtering procedure of DESeq2 were removed and those with absolute aapeglm shrunk log2 fold change  $> 0.5$  and the p-value  $< 0.05$  were considered differentially expressed in distinct conditions.

Raw sequencing data are available under the accession number PRJEB38744.

#### qRT-PCR analysis

RNA isolation from SARS-CoV-2 infected A549-ACE2 cells was performed as described above (Qiagen). 500 ng total RNA was used for reverse transcription with PrimeScript RT with gDNA eraser (Takara). For relative transcript quantification PowerUp SYBR Green (Applied Biosystems) was used. Primer sequences can be provided upon request.

### Gene Set Enrichment Analysis

We have used Gene Ontology, Reactome and other EnrichmentMap gene sets of human proteins<sup>41</sup> as well as protein complexes annotations from IntAct Complex Portal (version 2019.11)<sup>42</sup> and CORUM (version 2019)<sup>43</sup>.

To find the nonredundant collection of annotations describing the unique and shared features of multiple experiments in a dataset, we have used Julia package OptEnrichedSetCover.jl (<https://github.com/alyst/OptEnrichedSetCover.jl>), which employs evolutionary multi-objective optimization technique to find a collection of annotation terms that have both significant enrichments in the individual experiments and minimal pairwise overlaps.

For transcription factor enrichment analysis the significantly regulated transcripts were submitted to ChEA3 web-based application<sup>44</sup>. Transcription factor – target gene set libraries from ENCODE were used<sup>45</sup>. Transcriptome, proteome, ubiquitinome and phosphoproteome changes along with unchanged transcripts/proteins/sites were submitted to the core ingenuity pathway analysis (IPA) ([www.ingenuity.com](http://www.ingenuity.com)).

The following cut-offs were used for differentially expressed transcripts: the absolute values of apegIm-shrunk  $\log_2$  fold change  $> 0.5$ , the p-value  $< 0.05$ . Transcripts with low mean normalized count that were flagged by the independent filtering procedure of DESeq2 were removed prior pathway analysis. The following cut-offs were used for differentially expressed proteins or regulated sites: p-value  $< 0.05$  and absolute  $\log_2$  fold change  $\geq 0.5$ . Ingenuity knowledge base was used as a reference dataset, only experimentally observed findings were used for confidence filtering, additionally human species and A549-ATCC cell line filters were set. Input datasets were used to identify the most significant canonical pathways and upstream regulators (in case of transcriptome). Right-tailed Fisher's exact test with Benjamini-Hochberg's correction was used to calculate p-values, which are presented in Supplementary Table 3.

### Prediction of Functional Links between AP-MS and viral protein overexpression data

To systematically detect functional interactions, which may connect the cellular targets of each viral protein with the downstream changes it induces on proteome level, we have used the network diffusion-based HierarchicalHotNet method<sup>13</sup> as implemented in Julia package HierarchicalHotNet.jl (<https://github.com/alyst/HierarchicalHotNet.jl>). Specifically, for network diffusion with restart, we used the ReactomeFI network (version 2019)<sup>12</sup> of cellular functional interactions, reversing the direction of functional interaction (e.g. replacing kinase→substrate interaction with substrate→kinase). The proteins with significant abundance changes upon bait overexpression ( $|\text{median}(\log_2\text{fold change})| \geq 0.25$ ,  $P\text{-value} \leq 1\text{E-}3$  both in the comparison against the controls and against the baits of the same batch) were used as the sources of signal diffusion with weights set to  $\sqrt{|\text{median}(\log_2\text{fold change})| \cdot |\log_{10} P\text{-value}|}$ , and the restart probability was set to 0.4. To find the optimal cutting threshold of the resulting hierarchical tree of strongly connected components (SCCs) of the weighted graph corresponding to the stationary distribution of signal diffusion and to confirm the relevance of predicted functional connections, the same procedure was applied to 1000 random permutations of vertex weights as described in Reyna *et al.*<sup>13</sup> (vertex weights are randomly shuffled between the vertices with similar in- and out-degrees). Since cutting the tree of SCCs at any threshold  $t$  (keeping only the edges with weights above  $t$ ) and collapsing each resulting SCC into a single node produces the directed acyclic graph of connections between SCCs, it allowed efficient enumeration of the paths from the “source” nodes (proteins perturbed by viral protein expression with vertex weight  $w$ ,  $w \geq 1.5$ ) to the “sink” nodes (interactors of the viral protein). We have used this property of the tree to

calculate the average source-to-sink path length at each cutting threshold of the network diffusion weighted graph. At each threshold  $t$ , the average path from source to sink nodes was calculated as:

$$L_{avg}(t) = \left( N_{dis} \cdot (N_{scc} + 1) + \sum_p L_{scc}(p) \right) / (N_{src} \cdot N_{sink}),$$

where  $N_{src}$  is the number of “sources”,  $N_{sink}$  is the number of “sinks”,  $N_{dis}$  is the number of disconnected pairs of sources and sinks,  $N_{scc}$  is the number of SCC at given threshold,  $L_{scc}(p)$  is the number of SCCs that the given path  $p$  from source to sink goes through, and the sum is for all paths from sources to sinks. For the generation of the diffusion network we have selected the  $t_{opt}$  threshold that maximized the difference between the median of  $L_{avg}(t)$  for randomly shuffled data and  $L_{avg}(t)$  for the real data.

### Co-immunoprecipitation and western blot analysis

HEK293T cells were transfected with pWPI plasmid encoding single HA-tagged viral proteins, alone or together with pTO-SII-HA expressing host factor of interest. 48 hours after transfection, cells were washed in PBS, flash frozen in liquid nitrogen and kept at -80°C until further processing. Co-immunoprecipitation experiments were performed as described previously<sup>22,23</sup>. Briefly, cells were lysed in lysis buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 0.2% (v/v) NP-40, 5% (v/v) glycerol, cOmplete protease inhibitor cocktail (Roche), 0.5% (v/v) 750 U/μl Sm DNase) and sonicated (5 min, 4°C, 30 sec on, 30 sec off, low settings; Bioruptor, Diagenode SA). HA or Streptactin beads were added to cleared lysates and samples were incubated for 3h at 4°C under constant rotation. Beads were washed six times in the lysis buffer and resuspended in 1x SDS sample buffer 62.5 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 50 mM DTT, 0.01% bromophenol blue). After boiling for 5 minutes at 95°C, a fraction of the input lysate and elution were loaded on NuPAGE™ Novex™ 4-12% Bis-Tris (Invitrogen), and further submitted to western blotting using Amersham Protran nitrocellulose membranes. Imaging was performed by HRP luminescence (ECL, Perkin Elmer).

SARS-CoV-2 infected A549-ACE2 cell lysates were sonicated (10 min, 4°C, 30 sec on, 30 sec off, low settings; Bioruptor, Diagenode SA). Protein concentration was adjusted based on Pierce660 assay supplemented with ionic detergent compatibility reagent. After boiling for 5 min at 95°C and brief max g centrifugation, the samples were loaded on NuPAGE™ Novex™ 4-12% Bis-Tris (Invitrogen), and blotted onto 0.22 μm Amersham™ Protran® nitrocellulose membranes (Merck). Primary and secondary antibody stainings were performed according to the manufacturer’s recommendations. Imaging was performed by HRP luminescence using Femto kit (ThermoFischer Scientific) or Western Lightning PlusECL kit (Perkin Elmer).

### Reporter Assay and IFN Bioassay

The following reporter constructs were used in this study: pISRE-luc was purchased from Stratagene, EF1-*a*-ren from Engin Gürlevik, pCAGGS-Flag-RIG-I from Chris Basler, pIRF1-GAS-ff-luc, pWPI-SMN1-flag and pWPI-NS5 (ZIKV)-HA was described previously<sup>23,46</sup>.

For the reporter assay, HEK293RI cells were plated in 24-well plates 24 hours prior to transfection. Firefly reporter and Renilla transfection control were transfected together with plasmids expressing viral proteins using polyethylenimine (PEI, Polysciences) for untreated and treated conditions. In 18 hours cells were stimulated for 8 hours with a corresponding inducer and harvested in the passive lysis buffer (Promega). Luminescence of Firefly and Renilla luciferases was measured using dual-luciferase-reporter assay (Promega) according to the manufacturer’s instructions in a microplate reader (Tecan).

Total amounts of IFN- $\alpha/\beta$  in cell supernatants were measured by using 293T cells stably expressing the firefly luciferase gene under the control of the mouse Mx1 promoter (Mx1-luc reporter cells)<sup>47</sup>. Briefly, HEK293RI cells were seeded, transfected with pCAGGS-flag-RIG-I plus viral protein constructs and stimulated as described above. Cell supernatants were harvested in 8 hour. Mx1-luc reporter cells were seeded into 96-well plates in triplicates and were treated 24 hours later with supernatants. At 16 hours post incubation, cells were lysed in the passive lysis buffer (Promega), and luminescence was measured with a microplate reader (Tecan). The assay sensitivity was determined by a standard curve.

#### Viral inhibitors assay

A549-ACE2 cells were seeded into 96-well plates in DMEM medium (10% FCS, 100  $\mu$ g/ml Streptomycin, 100 IU/ml Penicillin) one day before infection. Six hours before infection, or at the time of infection, the medium was replaced with 100 $\mu$ l of DMEM medium containing either the compounds of interest or DMSO as a control. Infection was performed by adding 10 $\mu$ l of SARS-CoV-2-GFP (MOI 3) per well and plates were placed in the IncuCyte S3 Live-Cell Analysis System where whole well real-time images of mock (Phase channel) and infected (GFP and Phase channel) cells were captured every 4h for 48h. Cell viability (mock) and virus growth (mock and infected) were assessed as the cell confluence per well (Phase area) and GFP area normalized on cell confluence per well (GFP area/Phase area) respectively using IncuCyte S3 Software (Essen Bioscience; version 2019B Rev2).

#### Acknowledgements

We thank Stefan Pöhlmann for sharing ACE2 plasmids and Robert Baier for technical assistance, Juan Pancorbo and Johannes Albert-von der Gönna from the Leibniz Supercomputing Centre ([www.lrz.de](http://www.lrz.de)) for technical assistance. We further thank the Karl Max von Bauernfeind - Verein for support for the screening microscope. Work in the author's laboratories was supported by an ERC consolidator grant (ERC-CoG ProDAP, 817798), the German Research Foundation (PI 1084/3, PI 1084/4) and the German Federal Ministry of Education and Research (COVINET) to A.P. This work was also supported by the German Federal Ministry of Education and Research (CLINSPECT-M) to BK. AW was supported by the China Scholarship Council (CSC). The work of J.Z. was supported by the German Research Foundation (SFB1021, A01 and B01; KFO309, P3), the State of Hessen through the LOEWE Program (DRUID, B02) and the German Ministry for Education and Research (COVINET).

#### References

1. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 1–13 (2020) doi:10.1038/s41586-020-2286-9.
2. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 1–8 (2020) doi:10.1038/s41586-020-2332-7.
3. Messner, C. B. *et al.* Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Systems* (2020) doi:10.1016/j.cels.2020.05.012.
4. Shi, C.-S. *et al.* SARS-CoV-2 Open Reading Frame-9b Suppresses Innate Immunity by Targeting Mitochondria and the MAVS/TRAF3/TRAF6 Signalosome. *The Journal of Immunology* 193, 3080–3089 (2014).
5. Heo, J.-M. *et al.* Integrated proteogenetic analysis reveals the landscape of a mitochondrial-autophagosome synapse during PARK2-dependent mitophagy. *Science Advances* 5, eaay4624 (2019).
6. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 395, 565–574 (2020).

7. Cornillez-Ty, C. T., Liao, L., Yates, J. R., Kuhn, P. & Buchmeier, M. J. Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling. *Journal of Virology* **83**, 10314–10318 (2009).
8. Chu, H. *et al.* Comparative replication and immune activation profiles of SARS-CoV-2 and SARS-CoV in human lungs: an ex vivo study with implications for the pathogenesis of COVID-19. *Clin Infect Dis* doi:10.1093/cid/ciaa410.
9. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036-1045.e9 (2020).
10. Heurich, A. *et al.* TMPRSS2 and ADAM17 Cleave ACE2 Differentially and Only Proteolysis by TMPRSS2 Augments Entry Driven by the Severe Acute Respiratory Syndrome Coronavirus Spike Protein. *J Virol* **88**, 1293–1307 (2014).
11. Lambert, D. W. *et al.* Tumor Necrosis Factor- $\alpha$  Convertase (ADAM17) Mediates Regulated Ectodomain Shedding of the Severe-acute Respiratory Syndrome-Coronavirus (SARS-CoV) Receptor, Angiotensin-converting Enzyme-2 (ACE2). *J. Biol. Chem.* **280**, 30113–30119 (2005).
12. Wu, G., Dawson, E., Duong, A., Haw, R. & Stein, L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res* **3**, 146 (2014).
13. Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018).
14. Cottam, E. M., Whelband, M. C. & Wileman, T. Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* **10**, 1426–1441 (2014).
15. Andrianifahanana, M. *et al.* ERBB receptor activation is required for profibrotic responses to transforming growth factor beta. *Cancer Res.* **70**, 7421–7430 (2010).
16. Pittet, J.-F. *et al.* TGF- $\beta$  is a critical mediator of acute lung injury. *J Clin Invest* **107**, 1537–1544 (2001).
17. Mo, X. *et al.* Abnormal pulmonary function in COVID-19 patients at time of hospital discharge. *European Respiratory Journal* (2020) doi:10.1183/13993003.01217-2020.
18. Zavadil, J. *et al.* Genetic programs of epithelial cell plasticity directed by transforming growth factor- $\beta$ . *Proc Natl Acad Sci U S A* **98**, 6686–6691 (2001).
19. Qin, Z., Xia, W., Fisher, G. J., Voorhees, J. J. & Quan, T. YAP/TAZ regulates TGF- $\beta$ /Smad3 signaling by induction of Smad7 via AP-1 in human skin dermal fibroblasts. *Cell Commun. Signal* **16**, 18 (2018).
20. Thao, T. T. N. *et al.* Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. *Nature* **1–8** (2020) doi:10.1038/s41586-020-2294-9.
21. Mantlo, E., Bukreyeva, N., Maruyama, J., Paessler, S. & Huang, C. Antiviral activities of type I interferons to SARS-CoV-2 infection. *Antiviral Res* **179**, 104811 (2020).
22. Hubel, P. *et al.* A protein-interaction network of interferon-stimulated genes extends the innate immune system landscape. *Nat. Immunol.* **20**, 493–502 (2019).
23. Scaturro, P. *et al.* An orthogonal proteomic survey uncovers novel Zika virus host factors. *Nature* **561**, 253–257 (2018).
24. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
25. Gebhardt, A. *et al.* The alternative cap-binding complex is required for antiviral defense in vivo. *PLoS Pathog.* **15**, e1008155 (2019).
26. Goldeck, M., Schlee, M., Hartmann, G. & Hornung, V. Enzymatic synthesis and purification of a defined RIG-I ligand. *Methods Mol. Biol.* **1169**, 15–25 (2014).
27. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol. Cell Proteomics* **16**, 694–705 (2017).

28. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
29. Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nat Protoc* **13**, 1897–1916 (2018).
30. Udeshi, N. D., Mertins, P., Svinkina, T. & Carr, S. A. Large-scale identification of ubiquitination sites by mass spectrometry. *Nat Protoc* **8**, 1950–1960 (2013).
31. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301–2319 (2016).
32. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, 1–32 (2017).
33. Bhadra, A., Datta, J., Polson, N. G. & Willard, B. The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Anal.* **12**, 1105–1131 (2017).
34. Goeminne, L. J. E., Gevaert, K. & Clement, L. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics. *Mol. Cell Proteomics* **15**, 657–668 (2016).
35. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun* **11**, 787 (2020).
36. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
37. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261–D270 (2012).
38. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep* **6**, 25533 (2016).
39. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
41. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
42. Meldal, B. H. M. *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2019).
43. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
44. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
45. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
46. Vogt, C. *et al.* The interferon antagonist ML protein of thogoto virus targets general transcription factor IIB. *J. Virol.* **82**, 11446–11453 (2008).
47. Jorns, C. *et al.* Rapid and simple detection of IFN-neutralizing antibodies in chronic hepatitis C non-responsive to IFN-alpha. *J. Med. Virol.* **78**, 74–82 (2006).

## 4. Discussion and Outlook

In this thesis, I have demonstrated that improvements in the sensitivity, throughput and performance of MS-driven signaling approaches can greatly accelerate systems biology as well as clinical research, revealing rich and complex insights into cellular networks in health and diseases. Looking forward, the sensitive and quantitative proteomics strategies developed in this thesis can be applied to other biological scenarios and particularly in diseases, where sensitive PTM-specific proteomics profiling of primary tissues and individual patient samples may aid precision medicine in the future.

Our optimized EasyPhos platform addresses important issues of in-depth quantitative phosphoproteomics as it requires minimal sample amounts and measurement time without compromising depth of analysis. It makes large-scale and high-throughput signaling studies feasible for samples with low amounts such as typically available from primary cells. Our unbiased system-wide study of erythropoiesis demonstrated the breadth and depth of coverage that can now be achieved by MS-based proteomics and our EasyPhos technology. Here, the analysis of phosphorylation-based signaling in purified and differentiated erythroid precursors at distinct maturation stages generated data that can be mined for hypothesis generating problems related to erythroid biology. It would be also interesting to investigate other post translational protein modifications such as ubiquitination in the same system especially as we already observed distinct regulation of members and targets of the ubiquitin machinery.

Quantitative completeness is an important aspect of large scale PTM proteomics and represents a particular analytical challenge in data dependent analysis (DDA)-based workflows. Recently, the Olsen group showed that converting from a DDA to a DIA workflow dramatically increases the number of phosphorylation sites that can consistently and significantly be quantified in a single run analysis [171]. However, DIA suffers from lower throughput compared to single run TMT-based workflows. The latest advances in nanoflow liquid chromatography, such as the new LC system called Evosep, allow rapid, robust and deep DIA-based proteome and phosphoproteome profiling [172]. It significantly reduces the overhead time between sample pick up and start of MS measurement by using pre-formed gradients that already contain the sample. In our COVID19 study, we generated proteome and phosphoproteome time-course data on infected using the Evosep system and DIA-based label-free quantification. Using these state-of-the-art mass spectrometry and LC techniques, we characterized the impact of viral infection on protein abundance, ubiquitination and phosphorylation in a time-resolved manner by quantifying around 7,500 proteins, 5,000 ubiquitination and 12,000 phosphorylation events.

Given the central importance of ubiquitination, we also investigated the power of DIA for improving data completeness and sensitivity in large scale ubiquitinome analysis. Our sensitive

and robust DIA-based workflow was capable of identifying around 35,000 diGly peptides in single run measurements with unprecedented quantitative accuracy. Importantly, the workflow requires no offline fractionation, making it easy to implement and could be also applied to quantify other PTMs relying on antibody-based enrichment such as lysine acetylation and tyrosine phosphorylation. One disadvantage of our workflow is that its sensitivity is currently limited by the antibody-based enrichment, which requires 0.5-1 mg of protein material. To enable the analysis of systems such as human primary cell culture models where protein material is limited, we need further increase in sensitivity. This could be provided by either scaling down the need for high input amount, e.g. improving antibodies, or eliminating the steps that cause sample loss. For instance, omitting a peptide-clean-up step like in our EasyPhos protocol would further improve throughput and reproducibility of ubiquitinome analysis, making the entire workflow more streamlined. Furthermore, DIA-based ubiquitinome analysis would also likely benefit from the Evosep LC system, but this remains to be investigated in future studies.

One bottleneck for DIA analysis is that its performance with a project-specific spectral library is superior in terms of the coverage and quantification to library-free DIA which is much easier to implement [171, 173]. There are considerable efforts currently being invested into producing prediction tools for MS/MS spectra and retention time to make library-free approaches feasible. DIA analysis without need for specific libraries will surely simplify DIA workflows for PTMs and make them more amenable for the proteomics community [174].

In our quest towards precision medicine, we also showed that PTM-specific proteomics can provide valuable insights into cellular network function in pathophysiological contexts. In a breakthrough for the field, using genetic mouse models, specific inhibitors and EasyPhos workflow, Martin Steger from our group has identified and verified a subset of Rab GTPases as bona fide substrates of Parkinson's disease kinase LRRK2. Among those, Rab10 appears to be a key physiological kinase substrate as we later showed that all known pathogenic forms of LRRK2 enhance this phosphorylation and conversely inhibition of LRRK2 by small molecule inhibitors decrease it. In this thesis, I further aimed to establish the relevance of pRab10 as a PD marker by a targeted phosphoproteomics approach. This led to an accurate and highly-sensitive targeted MS-assay for determining the Rab10 phosphorylation stoichiometry and how it changes in Parkinson's disease. Using stable-isotope labeled spike-in peptides and a differential filling strategy for an Orbitrap analyzer, our assay enabled the quantification down to 50 attomoles, much beyond what is detectable by typical immunoassays and in a much more specific way. We were able to capture very small (around 1-2%) but significant phosphorylation differences between healthy controls, idiopathic PD patients and PD patients with defined genetic cause. We showed that pRab10 stoichiometry when measured precisely using targeted MS-based phosphoproteomics method can serve as a robust target engagement and patient stratification marker in clinical studies. In this study, we used neutrophils to demonstrate the potential of our

assay as we found that these cells contain relatively high levels of both LRRK2 and Rab10. It would be interesting to investigate whether pRab10 can be detected in other tissues such as brain or bodily fluids such as cerebrospinal fluid (CSF). We focused on two different mutation carriers with PD: LRRK2 G2019S and VPS35 D620N and analyzed a small cohort. We believe that analysis of larger cohorts and inclusion of patient samples with higher LRRK2 activity, such as the R1441G/C mutations, should further establish pRab10 as a PD marker.

Furthermore, our targeted mass spectrometric assay for accurate and sensitive measurement of phosphorylation levels is completely generic and not restricted to PD. It could be applied to measure phosphosite occupancies of prominent oncogenic factors to study how their levels influence tumorigenesis. The only limitation of our assay is its limited throughput due to gel-based enrichment of the target protein, especially when applied to very large cohorts. Our group recently developed a 'global targeting approach', bridging the approaches of shotgun and targeted proteomics (MaxQuant.Live platform). In this strategy, tens of thousands of precursor ions can be isolated in targeted manner in real-time, followed by fragmentation in single LC-MS run. It significantly improves sensitivity using elution time prediction algorithm based on the endogenous background population. We envision that the MaxQuant.Live global targeting approach will allow routine targeting of several disease-associated ions in a single MS run and eliminate up-front enrichment step as it significantly improves sensitivity on the MS side. This platform can be readily extended to PTM-specific peptides to measure them globally with very high reproducibility and accuracy. In summary, developments described in my thesis have contributed to a future in which we can elucidate and follow PTM-based processes in cellular biology and pathophysiology in unprecedented scale and detail, an important basis of system biology and precision medicine.

## 5. References

1. Humphrey, S.J., S.B. Azimifar, and M.J.N.b. Mann, *High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics*. 2015. **33**(9): p. 990-995.
2. Steger, M., et al., *Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases*. Elife, 2016. **5**: p. e12813.
3. Nature, E.P.C.J., *An integrated encyclopedia of DNA elements in the human genome*. 2012. **489**(7414): p. 57-74.
4. Smith, L.M., et al., *Proteoform: a single term describing protein complexity*. 2013. **10**(3): p. 186-187.
5. Aebersold, R. and M.J.N. Mann, *Mass spectrometry-based proteomics*. 2003. **422**(6928): p. 198-207.
6. Kim, M.-S., et al., *A draft map of the human proteome*. 2014. **509**(7502): p. 575-581.
7. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome*. 2014. **509**(7502): p. 582-587.
8. Savitski, M.M., et al., *A scalable approach for protein false discovery rate estimation in large proteomic data sets*. 2015. **14**(9): p. 2394-2404.
9. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules*. 1989. **246**(4926): p. 64-71.
10. Tanaka, K., et al., *Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry*. 1988. **2**(8): p. 151-153.
11. Fornelli, L., et al., *Top-down proteomics: where we are, where we are going?* 2018. **175**: p. 3.
12. Hein, M., et al., *Proteomic analysis of cellular systems In Handbook of systems biology: concepts and insights*. 2013, Academic Press.
13. Tsatsiani, L. and A.J.J.T.F.j. Heck, *Proteomics beyond trypsin*. 2015. **282**(14): p. 2612-2626.
14. Shevchenko, A., et al., *In-gel digestion for mass spectrometric characterization of proteins and proteomes*. 2006. **1**(6): p. 2856-2860.
15. Shevchenko, A., et al., *Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels*. 1996. **68**(5): p. 850-858.
16. Wiśniewski, J.R., et al., *Universal sample preparation method for proteome analysis*. 2009. **6**(5): p. 359-362.
17. Rappaport, J., Y. Ishihama, and M.J.A.c. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics*. 2003. **75**(3): p. 663-670.
18. Kulak, N.A., et al., *Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells*. 2014. **11**(3): p. 319.
19. Ding, H., et al., *Urine Proteomics: Evaluation of Different Sample Preparation Workflows for Quantitative, Reproducible, and Improved Depth of Analysis*. 2020. **19**(4): p. 1857-1862.
20. Berger, S.T., et al., *MStern blotting—high throughput polyvinylidene fluoride (PVDF) membrane-based proteomic sample preparation for 96-well plates*. 2015. **14**(10): p. 2814-2823.
21. Hughes, C.S., et al., *Ultrasensitive proteome analysis using paramagnetic bead technology*. 2014. **10**(10): p. 757.
22. Batth, T.S., et al., *Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation*. 2019. **18**(5): p. 1027-1035.
23. Gilar, M., et al., *Orthogonality of separation in two-dimensional liquid chromatography*. 2005. **77**(19): p. 6426-6434.

24. Kulak, N.A., et al., *Loss-less nano-fractionator for high sensitivity, high coverage proteomics*. 2017. **16**(4): p. 694-705.
25. Haag, A.M., *Mass analyzers and mass spectrometers*, in *Modern Proteomics—Sample Preparation, Analysis and Practical Applications*. 2016, Springer. p. 157-169.
26. Grinfeld, D., et al., *Phase-constrained spectrum deconvolution for Fourier transform mass spectrometry*. 2017. **89**(2): p. 1202-1211.
27. Lisacek, F., *Proteome Informatics*. 2019.
28. Meier, F., et al., *BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes*. 2018. **15**: p. 440-448.
29. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial*. 2018. **14**(8): p. e8126.
30. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis*. 2012. **11**(6).
31. Borràs, E. and E.J.P. Sabidó, *What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry*. 2017. **17**(17-18): p. 1700180.
32. Eng, J.K., A.L. McCormack, and J.R.J.J.o.t.a.s.f.m.s. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database*. 1994. **5**(11): p. 976-989.
33. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data*. 1999. **20**(18): p. 3551-3567.
34. Elias, J.E. and S.P.J.N.m. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. 2007. **4**(3): p. 207-214.
35. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue*. 2010. **7**(5): p. 383-385.
36. Zeiler, M., et al., *A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines*. 2012. **11**(3).
37. Dupuis, A., et al., *Protein Standard Absolute Quantification (PSAQ) for improved investigation of staphylococcal food poisoning outbreaks*. 2008. **8**(22): p. 4633-4636.
38. Winiewski, J.R., et al., *A'proteomic ruler'for protein copy number and concentration estimation without spike-in standards*. Molecular & cellular proteomics, 2014: p. mcp. M113. 037309.
39. Wiese, S., et al., *Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research*. 2007. **7**(3): p. 340-350.
40. Ross, P.L., et al., *Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents*. 2004. **3**(12): p. 1154-1169.
41. Thompson, A., et al., *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS*. 2003. **75**(8): p. 1895-1904.
42. Rauniyar, N. and J.R.J.J.o.p.r. Yates III, *Isobaric labeling-based relative quantification in shotgun proteomics*. 2014. **13**(12): p. 5293-5309.
43. Winter, S.V., et al., *EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification*. 2018. **15**(7): p. 527-530.
44. Kang, C., Y. Lee, and J.E.J.W.j.o.g. Lee, *Recent advances in mass spectrometry-based proteomics of gastric cancer*. 2016. **22**(37): p. 8283.
45. Mann, M., et al., *The coming age of complete, accurate, and ubiquitous proteomes*. 2013. **49**(4): p. 583-590.

46. Bekker-Jensen, D.B., et al., *Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries*. 2020. **11**(1): p. 1-12.
47. Itzhak, D.N., et al., *Global, quantitative and dynamic mapping of protein subcellular localization*. 2016. **5**: p. e16950.
48. Huttlin, E.L., et al., *Architecture of the human interactome defines protein communities and disease networks*. 2017. **545**(7655): p. 505-509.
49. Hein, M.Y., et al., *A human interactome in three quantitative dimensions organized by stoichiometries and abundances*. 2015. **163**(3): p. 712-723.
50. Bader, J.M., et al., *Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease*. 2020. **16**(6): p. e9356.
51. Niu, L., et al., *Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease*. 2019. **15**(3).
52. Albrechtsen, N.J.W., et al., *Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery*. 2018. **7**(6): p. 601-612. e3.
53. Geyer, P.E., et al., *Proteomics reveals the effects of sustained weight loss on the human plasma proteome*. 2016. **12**(12).
54. Geyer, P.E., et al., *Proteomics reveals the effects of sustained weight loss on the human plasma proteome*. 2016. **12**(12): p. 901.
55. Geyer, P.E., et al., *Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies*. 2019. **11**(11).
56. Geyer, P.E., et al., *Revisiting biomarker discovery by plasma proteomics*. 2017. **13**(9).
57. Wang, Y.-C., S.E. Peterson, and J.F.J.C.r. Loring, *Protein post-translational modifications and regulation of pluripotency in human stem cells*. 2014. **24**(2): p. 143-160.
58. Fuhs, S.R., et al., *Monoclonal 1-and 3-phosphohistidine antibodies: new tools to study histidine phosphorylation*. 2015. **162**(1): p. 198-210.
59. Besant, P., et al., *Focus on phosphoarginine and phospholysine*. 2009. **10**(6): p. 536-550.
60. Chen, M.J., J.E. Dixon, and G.J.S.s. Manning, *Genomics and evolution of protein phosphatases*. 2017. **10**(474).
61. Macek, B., et al., *Global and site-specific quantitative phosphoproteomics: principles and applications*. 2009. **49**: p. 199-221.
62. Mayya, V. and D.K.J.E.r.o.p. Han, *Phosphoproteomics by mass spectrometry: insights, implications, applications and limitations*. 2009. **6**(6): p. 605-618.
63. Tay, R.Y., et al., *Treatment of metastatic colorectal cancer: focus on panitumumab*. 2015. **7**: p. 189.
64. Roskoski Jr, R.J.P.r., *ERK1/2 MAP kinases: structure, function, and regulation*. 2012. **66**(2): p. 105-143.
65. de Oliveira, P.S.L., et al., *Revisiting protein kinase–substrate interactions: Toward therapeutic development*. 2016. **9**(420): p. re3-re3.
66. Miller, M.L., et al., *Linear motif atlas for phosphorylation-dependent signaling*. 2008. **1**(35): p. ra2-ra2.
67. Ardito, F., et al., *The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy*. 2017. **40**(2): p. 271-280.
68. Tonks, N.K.J.N.r.M.c.b., *Protein tyrosine phosphatases: from genes, to function, to disease*. 2006. **7**(11): p. 833-846.
69. Cohen, P.J.E.J.o.B., *The role of protein phosphorylation in human health and disease. Delivered on June 30th 2001 at the FEBS Meeting in Lisbon*. 2001. **268**(19): p. 5001-5010.

70. Steen, H., et al., *Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements*. 2006. **5**(1): p. 172-181.
71. Wilkes, E. and P.R. Cutillas, *Label-free phosphoproteomic approach for kinase signaling analysis*, in *Kinase Signaling Networks*. 2017, Springer. p. 199-217.
72. Jedrychowski, M.P., et al., *Evaluation of HCD-and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics*. 2011. **10**(12).
73. Nagaraj, N., et al., *Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation*. 2010. **9**(12): p. 6786-6794.
74. Quan, L., M.J.M.C. Liu, and Applications, *CID, ETD and HCD fragmentation to study protein post-translational modifications*. 2012.
75. Ochoa, D., et al., *The functional landscape of the human phosphoproteome*. 2020. **38**(3): p. 365-373.
76. Franciosa, G., A. Martinez-Val, and J.V.J.N.B. Olsen, *Deciphering the human phosphoproteome*. 2020. **38**(3): p. 285-286.
77. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse*. 2012. **40**(D1): p. D261-D270.
78. Grimsrud, P.A., et al., *Phosphoproteomics for the masses*. 2010. **5**(1): p. 105-119.
79. Thingholm, T.E. and M.R. Larsen, *The use of titanium dioxide micro-columns to selectively isolate phosphopeptides from proteolytic digests*, in *Phospho-Proteomics*. 2009, Springer. p. 57-66.
80. Thingholm, T.E., et al., *Highly selective enrichment of phosphorylated peptides using titanium dioxide*. 2006. **1**(4): p. 1929.
81. Larsen, M.R., et al., *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns*. 2005. **4**(7): p. 873-886.
82. Jensen, S.S. and M.R.J.R.C.i.M.S.A.I.J.D.t.t.R.D.o.u.t.t.M.R.i.M.S. Larsen, *Evaluation of the impact of some experimental procedures on different phosphopeptide enrichment techniques*. 2007. **21**(22): p. 3635-3645.
83. Bodenmiller, B., et al., *Reproducible isolation of distinct, overlapping segments of the phosphoproteome*. 2007. **4**(3): p. 231-237.
84. Schmelzle, K., et al., *Temporal dynamics of tyrosine phosphorylation in insulin signalling*. 2006. **55**(8): p. 2171-2179.
85. Kim, J.-E. and F.M.J.T.J.o.I. White, *Quantitative analysis of phosphotyrosine signaling networks triggered by CD3 and CD28 costimulation in Jurkat cells*. 2006. **176**(5): p. 2833-2843.
86. Wolf-Yadlin, A., et al., *Effects of HER2 overexpression on cell signaling networks governing proliferation and migration*. 2006. **2**(1): p. 54.
87. Huang, P.H., et al., *Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma*. 2007. **104**(31): p. 12867-12872.
88. Reeve, A., E. Simcox, and D.J.A.r.r. Turnbull, *Ageing and Parkinson's disease: why is advancing age the biggest risk factor?* 2014. **14**: p. 19-30.
89. Klein, C. and A.J.C.S.H.p.i.m. Westenberger, *Genetics of Parkinson's disease*. 2012. **2**(1): p. a00888.
90. Li, Y.I., et al., *Prioritizing Parkinson's disease genes using population-scale transcriptomic data*. 2019. **10**(1): p. 1-10.
91. Chang, D., et al., *A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci*. 2017. **49**(10): p. 1511.

92. Rudenko, I.N. and M.R.J.N. Cookson, *Heterogeneity of leucine-rich repeat kinase 2 mutations: genetics, mechanisms and therapeutic implications*. 2014. **11**(4): p. 738-750.

93. Tolosa, E., et al., *LRRK2 in Parkinson disease: challenges of clinical trials*. 2020: p. 1-11.

94. Hopkins, C.R., *Novel Therapeutic Approaches to the Treatment of Parkinson's Disease: An Overview and Update*. Vol. 18. 2016: Springer.

95. Parisiadou, L., et al., *Phosphorylation of ezrin/radixin/moesin proteins by LRRK2 promotes the rearrangement of actin cytoskeleton in neuronal morphogenesis*. 2009. **29**(44): p. 13971-13980.

96. Jaleel, M., et al., *LRRK2 phosphorylates moesin at threonine-558: characterization of how Parkinson's disease mutants affect kinase activity*. 2007. **405**(2): p. 307-317.

97. Stenmark, H., *Rab GTPases as coordinators of vesicle traffic*. *Nature reviews Molecular cell biology*, 2009. **10**(8): p. 513-525.

98. Taymans, J.-M. and E.J.C.n. Greggio, *LRRK2 kinase inhibition as a therapeutic strategy for Parkinson's disease, where do we stand?* 2016. **14**(3): p. 214-225.

99. West, A.B.J.E.n., *Achieving neuroprotection with LRRK2 kinase inhibitors in Parkinson disease*. 2017. **298**: p. 236-245.

100. Deng, X., et al., *Characterization of a selective inhibitor of the Parkinson's disease kinase LRRK2*. 2011. **7**(4): p. 203.

101. Dzamko, N., et al., *Measurement of LRRK2 and Ser910/935 phosphorylated LRRK2 in peripheral blood mononuclear cells from idiopathic Parkinson's disease patients*. 2013. **3**(2): p. 145-152.

102. Dzamko, N., et al., *Inhibition of LRRK2 kinase activity leads to dephosphorylation of Ser910/Ser935, disruption of 14-3-3 binding and altered cytoplasmic localization*. 2010. **430**(3): p. 405-413.

103. Nichols, R.J., et al., *14-3-3 binding to LRRK2 is disrupted by multiple Parkinson's disease-associated mutations and regulates cytoplasmic localization*. 2010. **430**(3): p. 393-404.

104. Ito, G., et al., *Phos-tag analysis of Rab10 phosphorylation by LRRK2: a powerful assay for assessing kinase function and inhibitors*. 2016: p. BCJ20160557.

105. Steger, M., et al., *Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases*. *Elife*, 2016. **5**.

106. Sheng, Z., et al., *Ser1292 autophosphorylation is an indicator of LRRK2 kinase activity and contributes to the cellular effects of PD mutations*. 2012. **4**(164): p. 164ra161-164ra161.

107. Reynolds, A., et al., *LRRK2 kinase activity and biology are not uniformly predicted by its autophosphorylation and cellular phosphorylation site status*. 2014. **7**: p. 54.

108. Kluss, J.H., et al., *Detection of endogenous S1292 LRRK2 autophosphorylation in mouse tissue as a readout for kinase activity*. 2018. **4**(1): p. 13.

109. Fraser, K.B., et al., *Urinary LRRK2 phosphorylation predicts parkinsonian phenotypes in G2019S LRRK2 carriers*. 2016. **86**(11): p. 994-999.

110. Lis, P., et al., *Development of phospho-specific Rab protein antibodies to monitor in vivo activity of the LRRK2 Parkinson's disease kinase*. *Biochemical Journal*, 2018. **475**(1): p. 1-22.

111. Steger, M., et al., *Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis*. 2017. **6**: p. e31012.

112. Steger, M., et al., *Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases*. 2016. **5**: p. e12813.

113. Collins, F.S. and H. Varmus, *A new initiative on precision medicine*. *N Engl J Med*, 2015. **372**(9): p. 793-5.

114. Emamzadeh, F.N. and A.J.F.i.n. Surguchov, *Parkinson's disease: biomarkers, treatment, and risk factors*. 2018. **12**: p. 612.
115. Emamzadeh, F.N.J.J.o.r.i.m.s.t.o.j.o.l.U.o.M.S., *Alpha-synuclein structure, functions, and interactions*. 2016. **21**.
116. Surguchov, A., *Intracellular dynamics of synucleins: "here, there and everywhere"*, in *International review of cell and molecular biology*. 2015, Elsevier. p. 103-169.
117. Mollenhauer, B., et al.,  *$\alpha$ -Synuclein and tau concentrations in cerebrospinal fluid of patients presenting with parkinsonism: a cohort study*. 2011. **10**(3): p. 230-240.
118. Hong, Z., et al., *DJ-1 and  $\alpha$ -synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease*. 2010. **133**(3): p. 713-726.
119. Li, Q.-X., et al., *Plasma  $\alpha$ -synuclein is decreased in subjects with Parkinson's disease*. 2007. **204**(2): p. 583-588.
120. Foulds, P.G., et al., *A longitudinal study on  $\alpha$ -synuclein in blood plasma as a biomarker for Parkinson's disease*. 2013. **3**: p. 2540.
121. Masuda-Suzukake, M., et al., *Prion-like spreading of pathological  $\alpha$ -synuclein in brain*. 2013. **136**(4): p. 1128-1138.
122. Emamzadeh, F.N.J.J.o.M.N., *Role of Apolipoproteins and  $\alpha$ -Synuclein in Parkinson's Disease*. 2017. **62**(3-4): p. 344-355.
123. Wang, E.S., et al., *Tetranectin and apolipoprotein A-I in cerebrospinal fluid as potential biomarkers for Parkinson's disease*. 2010. **122**(5): p. 350-359.
124. Swanson, C.R., et al., *Plasma apolipoprotein A1 associates with age at onset and motor severity in early Parkinson's disease patients*. 2015. **30**(12): p. 1648-1656.
125. Vitali, C., C.L. Wellington, and L.J.C.r. Calabresi, *HDL and cholesterol handling in the brain*. 2014. **103**(3): p. 405-413.
126. Hagan, J.J., et al., *Orexin A activates locus coeruleus cell firing and increases arousal in the rat*. 1999. **96**(19): p. 10911-10916.
127. Wienecke, M., et al., *Progressive dopamine and hypocretin deficiencies in Parkinson's disease: is there an impact on sleep and wakefulness?* 2012. **21**(6): p. 710-717.
128. Pickart, C.M. and M.J.J.B.e.B.A.-M.C.R. Eddins, *Ubiquitin: structures, functions, mechanisms*. 2004. **1695**(1-3): p. 55-72.
129. Kliza, K. and K.J.F.i.M.B. Husnjak, *Resolving the Complexity of Ubiquitin Networks*. 2020. **7**: p. 21.
130. Zheng, N. and N.J.A.r.o.b. Shabek, *Ubiquitin ligases: structure, function, and regulation*. 2017. **86**: p. 129-157.
131. Komander, D., M.J. Clague, and S.J.N.r.M.c.b. Urbé, *Breaking the chains: structure and function of the deubiquitinases*. 2009. **10**(8): p. 550-563.
132. Swatek, K.N. and D.J.C.r. Komander, *Ubiquitin modifications*. 2016. **26**(4): p. 399-422.
133. Komander, D. and M.J.A.r.o.b. Rape, *The ubiquitin code*. 2012. **81**: p. 203-229.
134. Huang, X. and V.M.J.C.r. Dixit, *Drugging the undruggables: exploring the ubiquitin system for drug development*. 2016. **26**(4): p. 484-498.
135. Hershko, A. and A.J.A.r.o.b. Ciechanover, *The ubiquitin system*. 1998. **67**(1): p. 425-479.
136. Peng, J., et al., *A proteomics approach to understanding protein ubiquitination*. 2003. **21**(8): p. 921-926.
137. Wagner, S.A., et al., *A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles*. 2011. **10**(10).
138. Kim, W., et al., *Systematic and quantitative assessment of the ubiquitin-modified proteome*. 2011. **44**(2): p. 325-340.

139. Akimov, V., et al., *UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites*. 2018. **25**(7): p. 631-640.

140. Nakagawa, T. and K.J.G.t.C. Nakayama, *Protein monoubiquitylation: targets and diverse functions*. 2015. **20**(7): p. 543-562.

141. Clague, M.J. and S.J.C. Urbé, *Ubiquitin: same molecule, different degradation pathways*. 2010. **143**(5): p. 682-685.

142. Yuan, W.-C., et al., *K33-linked polyubiquitination of coronin 7 by Cul3-KLHL20 ubiquitin E3 ligase regulates protein trafficking*. 2014. **54**(4): p. 586-600.

143. Matsumoto, M.L., et al., *K11-linked polyubiquitination in cell cycle control revealed by a K11 linkage-specific antibody*. 2010. **39**(3): p. 477-484.

144. Akutsu, M., I. Dikic, and A.J.J.o.c.s. Bremm, *Ubiquitin chain diversity at a glance*. 2016. **129**(5): p. 875-880.

145. Gatti, M., et al., *RNF168 promotes noncanonical K27 ubiquitination to signal DNA damage*. 2015. **10**(2): p. 226-238.

146. Zhao, J.J.C. and m.l. sciences, *Sumoylation regulates diverse biological processes*. 2007. **64**(23): p. 3017-3033.

147. Zhang, D., D.-E.J.J.o.i. Zhang, and c. research, *Interferon-stimulated gene 15 and the protein ISGylation system*. 2011. **31**(1): p. 119-130.

148. Nielsen, J.M., et al., *Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level*. 2011. **10**(3).

149. Franco, M., et al., *A novel strategy to isolate ubiquitin conjugates reveals wide role for ubiquitination during neural development*. 2011. **10**(5).

150. Xu, G., J.S. Paige, and S.R.J.N.b. Jaffrey, *Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling*. 2010. **28**(8): p. 868-873.

151. Wagner, S.A., et al., *Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues*. Mol Cell Proteomics, 2012. **11**(12): p. 1578-85.

152. Elia, A.E., et al., *Quantitative Proteomic Atlas of Ubiquitination and Acetylation in the DNA Damage Response*. Mol Cell, 2015. **59**(5): p. 867-81.

153. Rose, C.M., et al., *Highly Multiplexed Quantitative Mass Spectrometry Analysis of Ubiquitylomes*. Cell Syst, 2016. **3**(4): p. 395-403 e4.

154. Udeshi, N.D., et al., *Large-scale identification of ubiquitination sites by mass spectrometry*. Nat Protoc, 2013. **8**(10): p. 1950-60.

155. van der Wal, L., et al., *Improvement of ubiquitylation site detection by Orbitrap mass spectrometry*. J Proteomics, 2018. **172**: p. 49-56.

156. Kim, W., et al., *Systematic and quantitative assessment of the ubiquitin-modified proteome*. Mol Cell, 2011. **44**(2): p. 325-40.

157. Akimov, V., et al., *StUbEx PLUS A Modified Stable Tagged Ubiquitin Exchange System for Peptide Level Purification and In-Depth Mapping of Ubiquitination Sites*. 2018. **17**(1): p. 296-304.

158. Fujimuro, M., H. Sawada, and H.J.F.I. Yokosawa, *Production and characterization of monoclonal antibodies specific to multi-ubiquitin chains of polyubiquitinated proteins*. 1994. **349**(2): p. 173-180.

159. Matsumoto, M.L., et al., *Engineering and structural characterization of a linear polyubiquitin-specific antibody*. 2012. **418**(3-4): p. 134-144.

160. Hjerpe, R., et al., *Efficient protection and isolation of ubiquitylated proteins using tandem ubiquitin-binding entities*. 2009. **10**(11): p. 1250-1258.

161. Michel, M.A., et al., *Ubiquitin linkage-specific affimers reveal insights into K6-linked ubiquitin signaling*. 2017. **68**(1): p. 233-246. e5.

162. Crowe, S.O., et al., *Ubiquitin chain enrichment middle-down mass spectrometry enables characterization of branched ubiquitin chains in cellulo*. 2017. **89**(8): p. 4428-4434.

163. Swatek, K.N., et al., *Insights into ubiquitin chain architecture using Ub-clipping*. 2019. **572**(7770): p. 533-537.

164. Kirkpatrick, D.S., S.A. Gerber, and S.P.J.M. Gygi, *The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications*. 2005. **35**(3): p. 265-273.
165. Zaman, S., et al., *How *Saccharomyces* responds to nutrients*. Annu Rev Genet, 2008. **42**: p. 27-81.
166. Schule, T., et al., *Ubc8p functions in catabolite degradation of fructose-1, 6-bisphosphatase in yeast*. EMBO J, 2000. **19**(10): p. 2161-7.
167. Francis, O., F. Han, and J.C. Adams, *Molecular phylogeny of a RING E3 ubiquitin ligase, conserved in eukaryotic cells and dominated by homologous components, the muskelin/RanBPM/CTLH complex*. PLoS One, 2013. **8**(10): p. e75217.
168. Braun, B., et al., *Gid9, a second RING finger protein contributes to the ubiquitin ligase activity of the Gid complex required for catabolite degradation*. FEBS Lett, 2011. **585**(24): p. 3856-61.
169. Menssen, R., et al., *Exploring the topology of the Gid complex, the E3 ubiquitin ligase involved in catabolite-induced degradation of gluconeogenic enzymes*. J Biol Chem, 2012. **287**(30): p. 25602-14.
170. Komander, D. and M. Rape, *The Ubiquitin Code*. Annual Review of Biochemistry, 2012. **81**(1): p. 203-229.
171. Bekker-Jensen, D.B., et al., *Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries*. Nat Commun, 2020. **11**(1): p. 787.
172. Bache, N., et al., *A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics*. 2018. **17**(11): p. 2284-2296.
173. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial*. Mol Syst Biol, 2018. **14**(8): p. e8126.
174. Zhang, T., et al., *High-throughput discovery of genetic determinants of circadian misalignment*. PLoS Genet, 2020. **16**(1): p. e1008577.

## Appendix

### Urinary proteome profiling for stratifying patients with familial Parkinson's disease

Sebastian Virreira Winter<sup>1,a,b</sup>, Ozge Karayel<sup>1,a</sup>, Maximilian T Strauss<sup>1,b</sup>, Shalini Padmanabhan<sup>2</sup>, Matthew Surface<sup>3</sup>, Kalpana Merchant<sup>4</sup>, Roy N. Alcalay<sup>3</sup>, Matthias Mann<sup>1,5,c</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany <sup>2</sup> The Michael J. Fox Foundation for Parkinson's Research, NY, USA <sup>3</sup> Department of Neurology, Columbia University, NY, USA <sup>4</sup> Northwestern University Feinberg School of Medicine, IL, USA <sup>5</sup> Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark <sup>a</sup> These authors contributed equally <sup>b</sup> Current address: OmicEra Diagnostics GmbH, Am Klopferspitz 19, 82152 Planegg, Germany <sup>c</sup> Corresponding author

Under peer review in *EMBO Molecular Medicine*

Parkinson's disease (PD) is a prevalent neurologic disorder which is diagnosed in up to 10 million people worldwide. However, there are currently no treatment strategies and therapeutics that alter the course of the disease. Their development would greatly benefit from specific, sensitive, and non-invasive clinical assays and biomarkers to detect PD early and monitor disease progression and treatment efficacy. Together with Sebastian Virreira Winter and in addition to my other studies, I developed a scalable and sensitive proteomics workflow for urinary proteome profiling by combining high-throughput sample preparation with state-of-the-art MS-based proteomics. Using this workflow and minimal sample volumes, we quantified more than 2,000 proteins in each of more than 200 urine samples from two independent cohorts. The urinary proteome was significantly different between PD patients and healthy controls as well as between LRRK2 G2019S carriers and non-carriers in both cohorts. We observed high overlap of proteins exhibiting significantly perturbed levels in the two independent cohorts. This study demonstrates that valuable information can be inferred from the urinary proteome for neurodegenerative diseases. Strikingly, our data revealed strong lysosomal dysregulation in individuals with the LRRK2 G2019S mutation and urinary proteome profiles. Using machine-learning techniques, this alone can remarkably well classify pathogenic LRRK2 carriers from controls and non-manifesting carriers from those who develop the disease.

## Urinary proteome profiling for stratifying patients with familial Parkinson's disease

Sebastian Virreira Winter<sup>1,a,b</sup>, Ozge Karayel<sup>1,a</sup>, Maximilian T Strauss<sup>1,b</sup>, Shalini Padmanabhan<sup>2</sup>, Matthew Surface<sup>3</sup>, Kalpana Merchant<sup>4</sup>, Roy N. Alcalay<sup>3</sup>, Matthias Mann<sup>1,5,c</sup>

<sup>1</sup> Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

<sup>2</sup> The Michael J. Fox Foundation for Parkinson's Research, NY, USA

<sup>3</sup> Department of Neurology, Columbia University, NY, USA

<sup>4</sup> Northwestern University Feinberg School of Medicine, IL, USA

<sup>5</sup> Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

<sup>a</sup> These authors contributed equally

<sup>b</sup> Current address: OmicEra Diagnostics GmbH, Am Klopferspitz 19, 82152 Planegg, Germany

<sup>c</sup> Corresponding author

Correspondence: [mmann@biochem.mpg.de](mailto:mmann@biochem.mpg.de)

### SUMMARY

The prevalence of Parkinson's disease (PD) is increasing but the development of novel treatment strategies and therapeutics altering the course of the disease would benefit from specific, sensitive and non-invasive biomarkers to detect PD early. Here, we describe a scalable and sensitive mass spectrometry (MS)-based proteomic workflow for urinary proteome profiling. Our workflow enabled the reproducible quantification of more than 2,000 proteins in more than 200 urine samples using minimal volumes from two independent patient cohorts. The urinary proteome was significantly different between PD patients and healthy controls, as well as between *LRRK2* G2019S carriers and non-carriers in both cohorts. Interestingly, our data revealed lysosomal dysregulation in individuals with the *LRRK2* G2019S mutation. When combined with machine learning, the urinary proteome data alone was sufficient to classify mutation status and disease manifestation in mutation carriers remarkably well, identifying VGF, ENPEP and other PD-associated proteins as the most discriminating features. Taken together, our results validate urinary proteomics as a valuable strategy for biomarker discovery and patient stratification in PD.

### KEYWORDS

Parkinson's disease, urine, body fluid, mass spectrometry, LC-MS/MS, DIA, neurodegeneration, proteomics, biomarker

## INTRODUCTION

With a population prevalence of 0.2%, Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease [1]. It is characterized by the progressive loss of dopaminergic neurons and accumulation of  $\alpha$ -synuclein-containing protein aggregates called Lewy bodies in the cytoplasm of the remaining neurons. As a result of dopaminergic neuron loss, PD manifests with motor signs and symptoms including bradykinesia, tremor and rigidity, and these characteristics are used for diagnosing the disease [1-3].

PD is a genetically complex disorder. Most patients do not carry a single pathogenic variant linked to PD, but a subset of about 10% of patients carry an identifiable pathogenic variant in genes such as *SNCA*, *PRKN*, *LRRK2* or *GBA*. For these individuals, the risk of developing the disease increases to 2 to 5% [2]. Among these genes, *LRRK2* is relatively common and causes PD in an autosomal dominant with incomplete penetrance fashion. How *LRRK2* mutations cause PD is unknown, however, several studies have indicated that disease-linked *LRRK2* mutations elevate its kinase activity and contribute to PD pathogenesis [4]. We have previously identified multiple Rab GTPases as endogenous targets of mutant *LRRK2* [5-7]. Furthermore, inhibitors of this kinase have emerged as promising therapeutics for PD and clinical trials have already passed phase 1 [8]. Although idiopathic forms of PD presumably represent a heterogeneous collection of pathogenic mechanisms, *LRRK2*-associated PD and idiopathic PD (iPD) show a similar phenotype in terms of disease symptoms and response to levodopa. The interest in *LRRK2* as a therapeutic target is also fueled by the association between common variants in *LRRK2* and sporadic PD [9]

and the observation that *LRRK2* activity is increased in autopsied brain tissue from iPD patients without a known pathogenic mutation [10]. Thus, it will be important to conduct studies on ante-mortem biospecimens to gain insights into *LRRK2* mutation-induced changes and thereby identify iPD patients who could benefit from *LRRK2*-targeted therapies.

Current treatments, including levodopa - the most effective PD medication, mainly alleviate the motor symptoms rather than slow disease progression or reverse the pathology. Given the growing number of PD patients worldwide, and escalating economic and societal implications, there is an urgent need for disease-modifying therapeutics. The development of new therapeutic strategies requires better insights into the pathophysiologic mechanisms of PD as well as biomarkers to detect the earliest stages of PD before severe motor impairment is evident and irreversible brain damage has already occurred. Although cerebrospinal fluid (CSF) has been frequently used for biomarker studies of brain disorders, recent studies indicate that urine offers another promising clinically viable matrix for PD since it can be frequently and non-invasively collected in large volumes [11]. Importantly, urine contains not only kidney and urinary tract proteins but also filtered plasma proteins originating from distal organs, including the brain [12, 13]. Therefore, urine protein analysis may provide diagnostic and prognostic opportunities for both urogenital and non-renal diseases [14-26]. Recent technical advances in fast and high-throughput sample preparation methods in conjunction with improvements in high-accuracy mass spectrometry (MS)-based proteomics have enabled characterization of the urinary proteome [27-30]. However, to what extent neurodegenerative disorders including PD affect the urinary proteome remains unknown.

Our group has recently employed state-of-the-art MS-based proteomics to obtain proteome profiles of the two body fluids, plasma and CSF, in multiple disease conditions [31-35]. Here, we extend this technology to urinary proteome profiling and provide first evidence that this approach can be used for PD biomarker discovery. More specifically, we focused our analysis on two large patient cohorts, both including healthy control subject, non-manifesting carriers of the frequently found *LRRK2* G2019S mutation, manifesting patients with the same mutational signature, and PD patients without the *LRRK2* mutation. The composition of the cohorts, quality of the data and the depth of proteome coverage allowed us to identify pathogenic *LRRK2*-regulated lysosomal protein signatures that could serve as biomarkers to stratify subjects with pathogenic *LRRK2*. Taken together, our study offers evidence that quantitative MS-based proteomics represents a clinically useful strategy for non-invasive monitoring of disease progression and treatment response as well as patient stratification in PD.

## RESULTS

### Overview of PD cohorts for urinary proteomics

Here, we employ a recently described 'rectangular' biomarker discovery strategy in which as many proteins as possible are measured using shotgun MS-based proteomics for all the individuals in both discovery and validation cohorts [33, 36]. To map proteome changes between individuals with different mutation status and manifestation of disease, we analyzed 235 urine samples from two independent cross-sectional cohorts each comprised of four subject groups: (1) healthy controls (HC, *LRRK2*-/PD-); (2) non-manifesting carriers (NMC) harboring the *LRRK2* G2019S mutation (*LRRK2*+/PD-); (3) idiopathic PD patients (iPD, *LRRK2*-/PD+); and

(4) manifesting PD patients with *LRRK2* G2019S (*LRRK2* PD, *LRRK2*+/PD+) (Figure 1A & Table 1).

The first cohort was recruited at Columbia University Irving Medical Center (hereinafter referred to as 'Columbia cohort' and color-coded with orange) [11, 37]. Participants in the Columbia cohort included 35 HC, 16 NMC, 40 iPD, 28 *LRRK2* PD individuals, and one PD patient with an unknown *LRRK2* status. The cohort included 52 female sex and 68 male sex individuals (Figure 1A and Table 1). The *GBA* (gene that encodes for lysosomal acid glucosylceramidase (GCase)) mutation status was also available for all individuals, with 22 of them harboring a *GBA* variant and 98 the wild type allele. PD+ and PD- subjects were frequency-matched by age with means of  $67.0 \pm 9.3$  and  $64.1 \pm 12.0$  ( $\pm$  SD) years, respectively (Supplementary Figure 1A). Their motor skills were assessed using the Unified Parkinson's Disease Rating Scale part III (UPDRS-III) and cognitive functioning with the Montreal Cognitive Assessment (MoCA) test (Supplementary Figure 1B-C). Genotyping for *LRRK2* G2019S mutation was conducted as previously described [38].

To confirm findings from the Columbia cohort, we additionally analyzed a subset of biobanked urine samples from the Michael J. Fox Foundation for Parkinson's Research (MJFF)-funded *LRRK2* Cohort Consortium (LCC) (hereinafter referred to as 'LCC cohort' and color-coded with blue). We determined urine proteomes for 26 HC, 37 NMC, 29 iPD, and 23 *LRRK2* PD individuals (53 female and 62 male) (Figure 1A and Table 1). In the LCC cohort, individuals in the non-diseased group were somewhat younger ( $53.8 \pm 13.9$ ) than PD patients ( $67 \pm 7.6$ ) (means  $\pm$  SD) years (Supplementary Figure 1D). In addition, LCC sample collection protocols were less stringent than in the

Columbia cohort and UPDRS-III and MoCA scores were not available, indicating that the Columbia cohort is more powerful for our analyses. Both studies were approved by local institutional review boards, and each participant signed an informed consent (See **Supplementary Table 1** for a detailed overview).

**Proteomic characterization of urine samples**  
 For the proteomic profiling of individual urine samples, we developed a high-throughput proteomics workflow building on the PVDF-based sample processing method MStern blotting by the Steen group [27] combined with data-independent acquisition (DIA) LC-MS/MS [39, 40] (Figure 1A). To maximize proteome depth, we generated two cohort-specific hybrid spectral libraries by merging three sub-libraries: (1) a library constructed by data-dependent acquisition (DDA) consisting of 24 fractions of pooled neat urine samples; (2) a DDA library consisting of 8 fractions of extracellular vesicles isolated from pooled neat urine samples; and (3) a direct-DIA library generated from the DIA analysis of all analyzed samples (see Methods). In these hybrid libraries, we identified a total of 4,564 and 5,725 protein groups for the Columbia and LCC cohorts, respectively (Supplementary Figure 1E). Applying this robust workflow, we quantified on average 2,026 (Columbia) and 2,162 (LCC) protein groups per neat urine sample, in single runs of 45 minutes and using less than 100  $\mu$ l of starting material (Supplementary Table 2). Three outlier samples were excluded from further analysis due to low proteome depth (Fig 1B and 1C, Supplementary Table 1). The quantified protein intensities spanned five orders of magnitude in both cohorts and the top ten most abundant proteins contributed about half to the total urinary proteome signal (Fig 1D and 1E). As observed before [17], the molecular weight distribution spanned a wide range with many proteins exceeding 100 kDa. More than 2,000 proteins were in common between the two cohorts. To

best of our knowledge, this study presents the deepest urinary proteome coverage for single-run analysis to date, a promising basis for the discovery of biomarkers.

Data from repeated measurements of individual samples revealed a high reproducibility with more than 90% of proteins having an intra- and inter-plate coefficient of variation (CV) below 50% in both studies and about 60% of proteins with a CV below 20% (Figure 1G, 1H, Supplementary Figure 2). The intra- and inter-plate variability within each cohort was even lower (Supplementary Figure 2C and 2I), while the inter-individual variability was much larger with no protein having a CV below 20% (Supplementary Figure 2F and 2L). Thus, our proteomic quantification accuracy greatly exceeds the biological variability that we seek to measure.

#### Quality assessment of urine samples

Pre-analytical variation caused by inconsistent sample processing and contaminations during sample collection can have a strong impact on the results and may cause the reporting of incorrect biomarkers [34]. To ensure that the observed proteome changes are not caused by artifacts related to sample handling and processing, we assessed each sample for potential quality issues. To this end, we used a previously reported quality marker panel to determine the degree of contamination with erythrocytes [34] (Figure 2A-Band Supplementary Table 3). Insufficient removal of cells and cellular debris from urine leads to an increased detection of intracellular proteins with a high sample-to-sample variability compared to regularly secreted urinary proteins [41]. We therefore generated a second urine-specific quality marker panel to assess the degree of contamination with cells and cellular debris that could originate from aged, inflamed or damaged tissue of the kidneys, bladder or the urinary tract (see Methods). Although urine

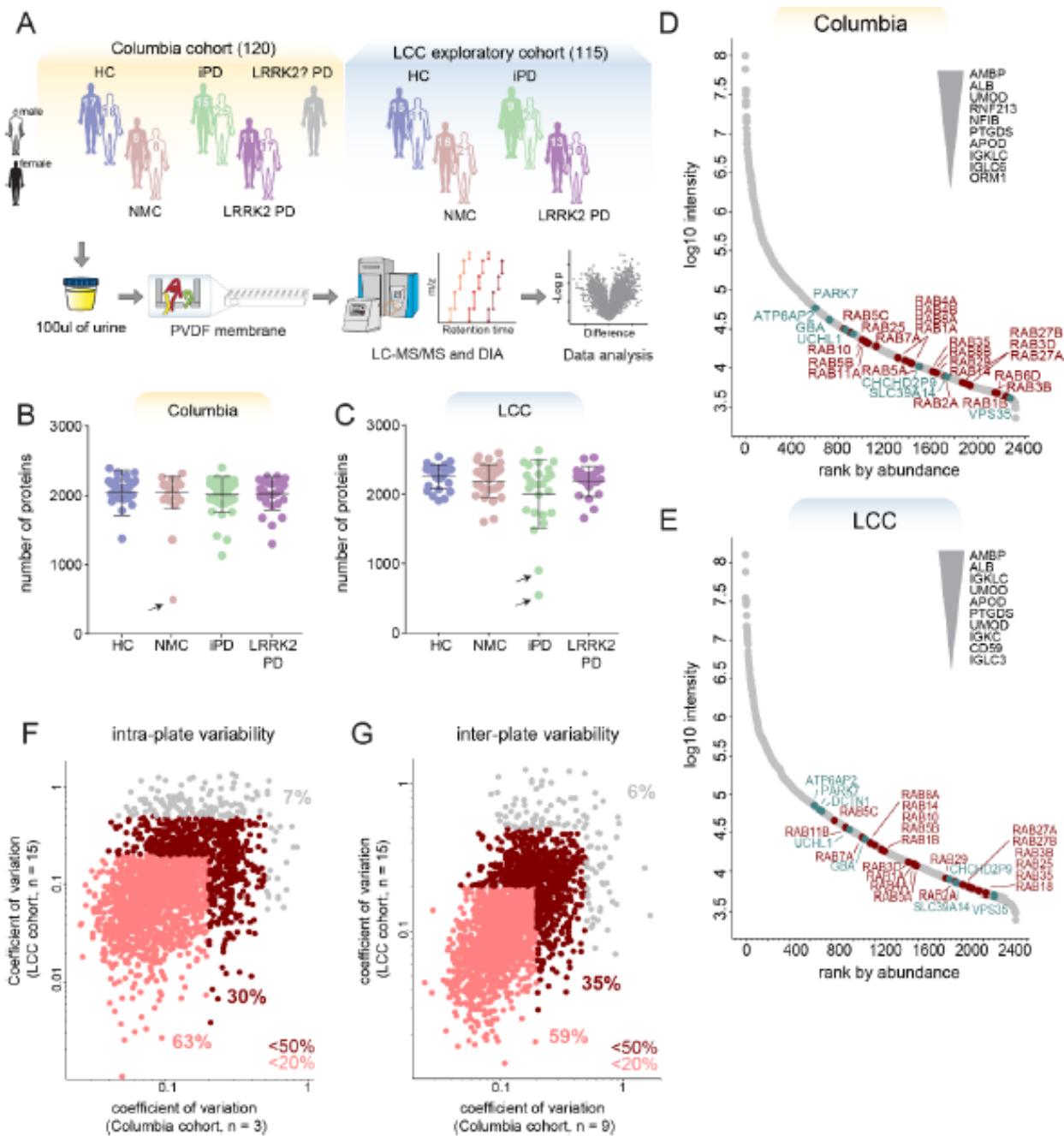
samples from both cohorts were cleared by centrifugation following collection to avoid this systematic bias, our procedure flagged four samples from the Columbia cohort for potential contamination with cellular components (Figure 2A-B). Taken together, 6 samples from the Columbia cohort and 4 samples from the LCC cohort showed increased intensities of contamination markers and were thus excluded from further analyses. In addition, we further excluded one sample from the Columbia cohort, as it clustered far away from all other samples in a principal component analysis (PCA), likely indicating pre-analytical variation.

Next, we generated a global correlation map of the urinary proteome to identify clusters of functional co-regulation as previously reported for plasma proteome profiling [31]. The global correlation map contains pairwise relations of all urinary proteins across 112 samples from the Columbia cohort. Unsupervised hierarchical

clustering of the pairwise Pearson correlation coefficients revealed four main and several small clusters of co-regulated proteins (Figure 2C). The largest of these clusters was chiefly enriched for proteins with the Gene ontology (GO)-term 'extracellular exosome' as well as other significant terms (Supplementary Table 4). We also identified a cluster of highly correlated proteins that was enriched for the GO-terms 'immunoglobulin' and 'B-cell receptor', suggesting that these proteins originate from immune cells. The two further main clusters were enriched for proteins originating from sex-specific tissues such as the prostate and vagina (Figure 2C) [42]. This shows that sex-dependent anatomical differences strongly affect the urinary proteome and thus should be considered as confounding factors. Indeed, a principal component analysis indicated sex as the strongest contributor to the inter-individual variance of the urinary proteome (Figure 2D and 2E).

**Table 1. Demographics of all participants**

Columbia cohort	HC ( <i>LRRK2</i> -/PD-) (n = 35)	NMC ( <i>LRRK2</i> +/PD-) (n = 16)	iPD ( <i>LRRK2</i> -/PD+) (n = 40)	<i>LRRK2</i> PD ( <i>LRRK2</i> +/PD+) (n = 28)
Age at collection, mean (SD)	67.5 (10.3)	56.8 (12.7)	64.9 (9.2)	70.7 (8.7)
Age at onset, mean (SD)	n/a	n/a	57.8 (11)	57.9 (11.3)
Sex (female/male)	17/18	8/8	15/25	11/17
<i>GBA</i> (mut/WT)	7/28	1/15	11/29	3/25
MoCA	27.5 (2)	28.7 (1.1)	26.9 (1.6)	26.3 (4.5)
UPDRS-III	1.1 (1.5)	0.8 (1.1)	17.7 (10)	20.6 (8)
LCC cohort	HC ( <i>LRRK2</i> -/PD-) (n = 26)	NMC ( <i>LRRK2</i> +/PD-) (n = 37)	iPD ( <i>LRRK2</i> -/PD+) (n = 29)	<i>LRRK2</i> PD ( <i>LRRK2</i> +/PD+) (n = 23)
Age at collection, mean (SD)	56.1 (16)	52.1 (12.1)	65 (9.1)	68.4 (5)
Age at onset, mean (SD)	n/a	n/a	58.4 (9.1)	57.7 (7.6)
Sex (female/male)	15/11	16/21	9/20	13/10



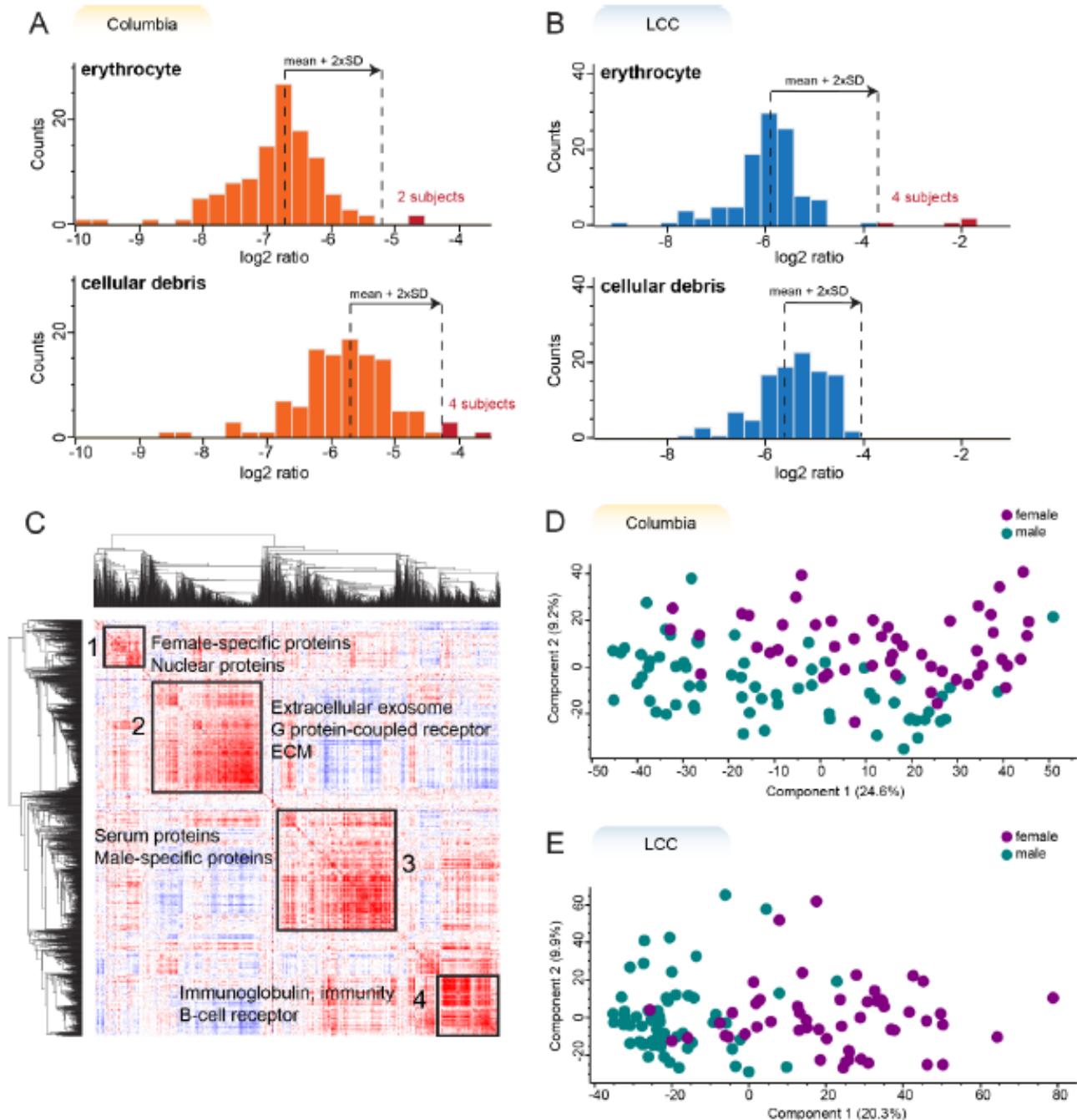
**Figure 1. MS-based proteomic analysis of two independent urinary PD cohorts has high depth and precision**

A) Overview of the two cohorts and the proteomic workflow. Urine samples comprised of four subject groups ((HC, iPD, NMC and LRRK2 PD) were prepared using MStern blotting and analyzed by LC-MS/MS using data-independent acquisition (DIA). The sex and total number of subjects per cohort group is shown.

B-C) Number of proteins identified and quantified with a 1% false discovery rate (FDR) in each sample in the Columbia (B) and LCC (C) cohorts. Bars indicate mean and standard deviation. Arrows point at one subject from the Columbia and two subjects from the LCC cohort that were excluded from further analysis due to low proteome depth.

**D-E)** Proteins identified in the Columbia (D) and LCC (E) cohort were ranked according to their MS signals, which covered more than five orders of magnitude. The top ten most abundant Parkinson-related proteins (green) and Rab GTPases (red) are labeled.

**F-G)** Quantification precision assessed by calculating the intra-plate (F) and inter-plate (G) coefficients of variation (CVs) for the Columbia and LCC cohorts. Proteins with a CV below 20% and 50% in both cohorts are highlighted in light and dark red, respectively and the fractions of proteins above and below these CV thresholds are shown. A total of 2,051 proteins were consistently quantified in both cohorts.



**Figure 2. The large majority of urine samples has high quality and shows sex-specific protein expression**

**A-B)** Histograms of log2 transformed ratios of the summed intensity of the proteins in the respective quality marker panel and the summed intensity of all proteins in Columbia (A) and LCC (B) cohorts. A sample was flagged for potential contamination and removed from further analysis if the ratio differed more than two standard deviations from the mean of all samples within the cohort. The proteins in each quality marker panel are listed in Supplementary Table 3.

**C)** Global correlation map of proteins generated by clustering the Pearson correlation coefficients of all pairwise protein comparisons for the Columbia cohort.

**D-E)** Principal component analysis (PCA) of all subjects based on their urinary proteome profiles. Female subjects are shown in purple and males in green.

**Detection of PD-related proteome alterations in urine**

Although PD primarily manifests in the central nervous system and is characterized by motor impairments, it is known to affect and potentially initiate in peripheral tissues and is associated with non-motor symptoms [43, 44]. Thus, we asked if the disease also causes changes of the urinary proteome, which reflects proteins from both central and peripheral organs. To establish PD-associated changes in urine proteome, we first determined which proteins are differentially present in the urine of PD patients compared to the controls, irrespective of their *LRRK2* status (HC and NMC). To control for confounders, we performed an analysis of covariance (ANCOVA) considering sex, age at sample collection, *LRRK2* status and *GBA* status (only available for the Columbia cohort) as confounding factors. Applying a 5% false discovery rate (FDR) cut off, we identified 361 proteins that displayed significantly different levels in PD patients when compared to controls (HC and NMC) (298 in Columbia cohort and 73 in LCC cohort)

(Supplementary Table 5). The smaller number of significantly different proteins in the LCC cohort as well as the relatively small overlap between the cohorts could be explained by a less stringent sample collection protocol and worse age-matching in the LCC cohort. The log2 fold-

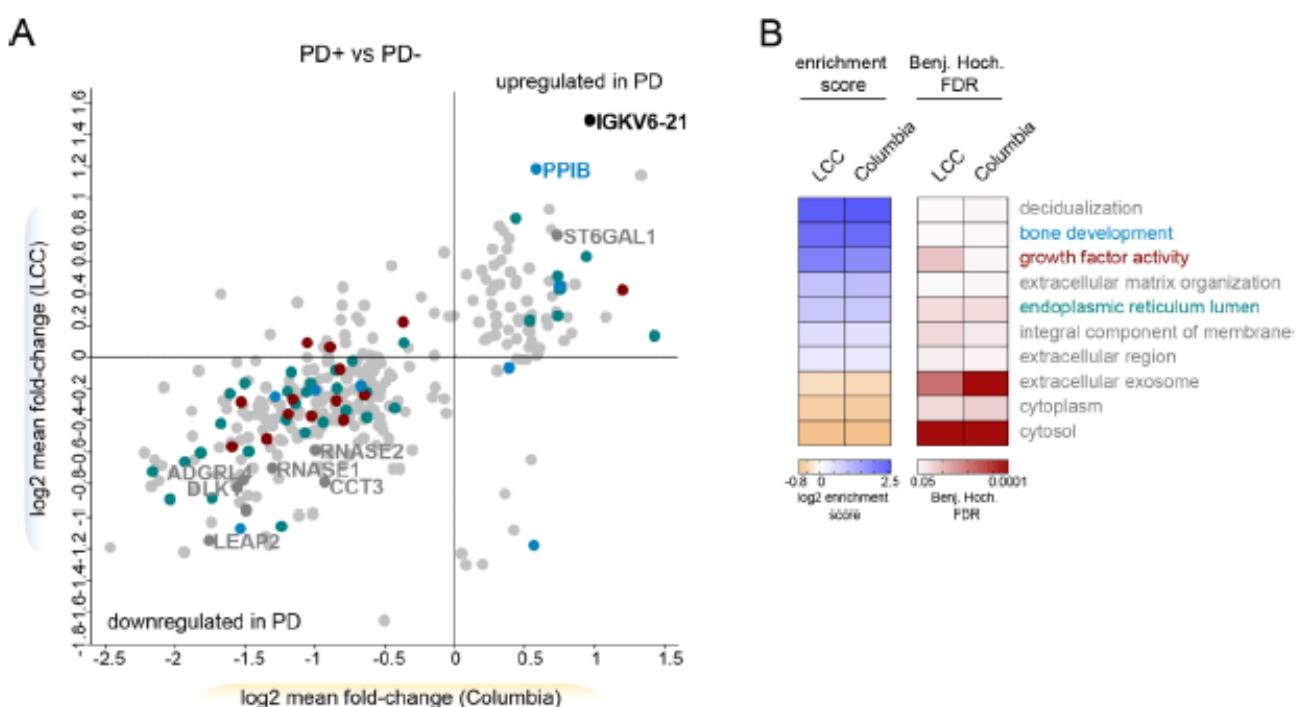
changes between PD patients and non-diseased individuals show a good correlation between the two cohorts (Pearson  $r = 0.65$ ) (Figure 3A), reflecting both reproducibility of the applied proteomic workflow and pathobiological consistency. The mean fold-changes of the 330 PD-associated proteins that were quantified in both cohorts were larger for the Columbia cohort (Columbia: 1.43 (up) & 0.49 (down) vs. LCC: 1.27 (up) & 0.75 (down)). Furthermore, 90% of the PD-associated proteins were detected with at least two peptides and quantified with CVs below 50% (Supplementary Figure 3A).

Protein misfolding is known to be involved in many neurodegenerative conditions including PD [45]. Interestingly, some of the proteins exhibiting the largest differential levels between the urine of controls vs. PD patients include proteins assisting other proteins in folding, such as peptidyl-prolyl cis-trans isomerase B (PPIB) and T-complex protein 1 subunit gamma (CCT3) (Figure 3A). We also identified two of the eight human canonical ribonucleases (RNASE1 and RNASE2) to be PD-associated in both cohorts (Figure 3A). The levels of the four apolipoproteins APOA1, APOA2, APOA4 and APOC1 were also elevated in PD patients (Figure 3A). While they show a similar trend in both cohorts, they reached statistical significance only in the Columbia cohort, corroborating that

this cohort has greater power to detect PD-associated changes.

Next, we analyzed if any GO-terms assigned to the 361 PD-associated proteins were significantly enriched compared to the urinary proteome (**Figure 3B**). This analysis examines if PD affects individual cellular compartments and particular biological signaling networks in urine. The term

'bone development' was significantly enriched, in line with previous findings that PD patients are at increased risk for osteoporosis and osteopenia [46]. In summary, we observed disease-associated protein signatures with a high correlation between the two independent cohorts and identified promising candidates that could serve as biomarkers for PD and provide mechanistic insights into disease pathogenesis.



**Figure 3. PD affects the urinary proteome**

A) Correlations of mean fold-changes of the proteins changing PD-dependently in the Columbia and LCC cohorts. Only proteins quantified in both cohorts are shown (n=330). The colors match to the GO terms shown in (B). Proteins overlapping between the two cohorts are labeled with their name.  
 B) Fisher exact test to identify significantly enriched GO-terms in the PD-associated proteins in urine. All GO-terms that were significant in both cohorts are displayed (FDR < 5%).

**Pathogenic *LRRK2*-dependent changes are linked to lysosomes and glycosphingolipid metabolism**

Encouraged by the observation of disease-dependent proteome changes in urine, we next

asked if the urinary proteome is altered by the presence of the *LRRK2* G2019S mutation. We again applied an ANCOVA analysis with sex, age at sample collection, PD status and *GBA* status (only available for the Columbia cohort) as confounding factors and compared the proteomes

between G2019S and wild type allele carriers. Applying an FDR of 5%, the mutation altered the abundance of 237 proteins (FDR < 5%, Columbia: 166, LCC: 104) (Figure 4A and Supplementary Table 5). A subset of 33 proteins differed significantly in G2019S carriers in both cohorts and all these proteins were upregulated in pathogenic *LRRK2* carriers. A pairwise comparison of the four subject groups (HC, NMC, iPD and LRRK2 PD) using a student's t-test confirmed that the abundance of the overlapping proteins changed in a G2019S-dependent manner but was unaffected by the PD status (Figure 4B).

In total, 227 *LRRK2*-status associated proteins were quantified in both cohorts and the fold-changes of these were similar between the two cohorts (Figure 4C), although the effect sizes were slightly larger for the Columbia cohort (Columbia: 1.43 (up) & 0.76 (down) vs. LCC: 1.39 (up) & 0.89

(down)). Interestingly, one of the proteins exhibiting the largest increase in *LRRK2* G2019S carriers in both cohorts was a phosphatase, the Intestinal-type Alkaline Phosphatase (ALPI). As for the proteins that changed dependent on PD disease status, most *LRRK2*-status associated proteins were detected with at least two peptides and quantified with CVs below 50% (Supplementary Figure 3B).

A GO-term analysis revealed strong enrichment of proteins associated with lysosome-related terms such as 'autolysosome', 'lysosome', 'lysosomal lumen', 'azurophil granule lumen' and 'lysosomal membrane' as well as 'glycosphingolipid metabolic processes' in *LRRK2* G2019S carriers in both cohorts (Figure 4D). Among the proteins associated with the lysosome-related GO-terms were multiple members of the cathepsin family including cathepsins A, B, C, D, H, L, O, S, and Z. The

widely used lysosomal marker proteins, LAMP1 and LAMP2, were also significantly altered in *LRRK2* carriers in the LCC cohort, while LAMP3 was significantly changed in the Columbia cohort. In total, 13 proteins were associated with the GO term 'sphingolipid metabolic process', most of them upregulated in *LRRK2* G2019S carriers. Among them were multiple lysosomal enzymes including GCase (encoded by *GBA*), galactocerebrosidase (GALC), sphingomyelin phosphodiesterase (SMPD1), and the beta-hexosaminidase subunits alpha and beta (HEXA and HEXB).

Heterozygous pathogenic mutations in *GBA* are one of the most common PD risk factors while homozygous loss of function mutations of the same protein cause the lysosomal storage disorder (LSD) Gaucher's disease. Past studies have described increased GCase activity in *LRRK2* deficient mice, and decreased GBA activity in *LRRK2* G2019S carrier neurons [47, 48]. Additionally, it has been reported that PD patients with mutations in both proteins develop symptoms at a younger age compared to patients with only one affected gene [49-51]. However, despite these reports, it remained unclear whether mutations in *GBA* and *LRRK2* contribute to the pathogenesis of PD via common pathways. Participants in the Columbia cohort were sequenced for mutations in *GBA* [38] and 22 individuals were found to carry a pathogenic mutation in this locus. To determine which proteins were changed specifically in carriers of *GBA* variants, we performed an ANCOVA analysis with sex, age at sample collection, PD status and *LRRK2* status as confounding factors. Using an FDR of 5%, we found that levels of 74 proteins were affected by *GBA* (Figure 4E and Supplementary Table 5). Interestingly, only Intercellular adhesion molecule 1 (ICAM1), Adenosylhomocysteinase (AHCY) and Stomatin (STOM) were affected by pathogenic mutations in both *LRRK2* and *GBA*, suggesting that the two

mutations largely affect distinct pathways. Furthermore, the *GBA*- and *LRRK2*-dependent protein fold-changes were poorly correlated (Pearson  $r = 0.21$ ) (Figure 4F) but future well-powered studies on *GBA* cohorts are needed to firm up the data. Of note, most proteins associated with the GO term 'glycosphingolipid metabolic process' were increased in *LRRK2* G2019S carriers but decreased in pathogenic *GBA* carriers, most notably GM2 activator (GM2A).

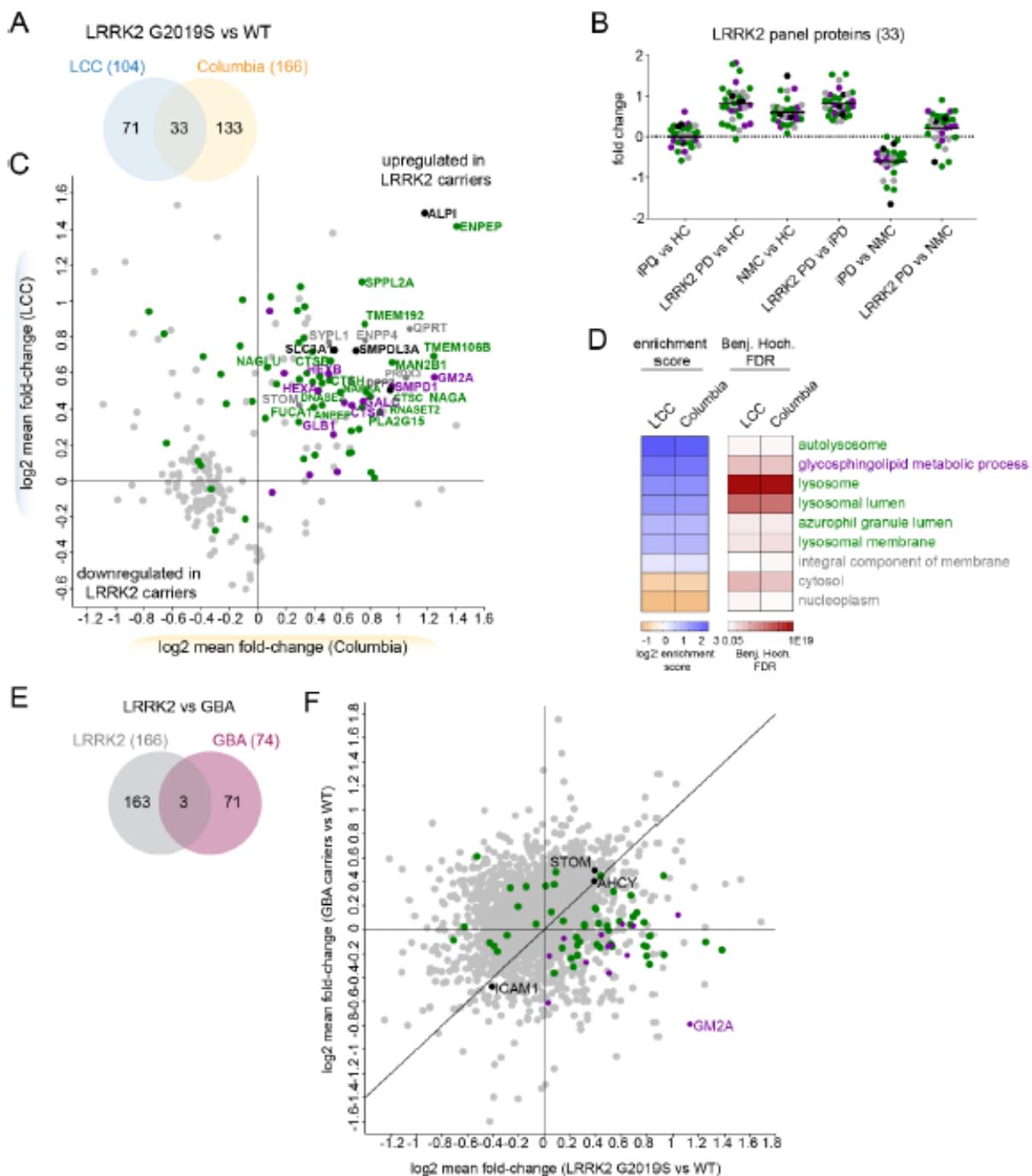
Together, we identified pathogenic mutant *LRRK2*-dependent protein signatures with a high correlation between the two independent cohorts. The *LRRK2* mutational status-dependent changes of the urinary proteome include lysosomal proteins that could serve as biomarkers to stratify subjects with pathogenic *LRRK2*.

#### Correlation of proteome profiles with clinical parameters

Given that clinical parameters, including disease severity scores, were available for the Columbia cohort, we were interested in exploring whether any of these clinical parameters correlate with proteomic changes we detected. We were especially interested in the cognitive capabilities of the participants as evaluated using the Montreal Cognitive Assessment (MoCA) test, and the motor performance as assessed using the Unified Parkinson's Disease Rating Scale part III

(UPDRS-III). Within the Columbia cohort, MoCA scores ranged from 8 to 30, on a scale from 0, for severe cognitive impairment, to 30, for no measurable cognitive impairment. We observed that two proteins, Tenascin-R (TNR) and Furin (FURIN), showed a strong negative correlation with the MoCA score in PD patients (TNR Pearson  $r: -0.66$ ; FURIN  $r: -0.65$ ;  $p < 10^{-7}$  for both), mainly driven by *LRRK2* G2019S carriers (TNR  $r: -0.77$ ;

FURIN  $r: -0.78$ ;  $p < 10^{-5}$  for both) (Figure 5A). When similar type of analysis was done with UPDRS-III scores, which ranged from 0 to 38 in the Columbia cohort (on a scale from 0 assigned for normal to 56 for severely affected motor function), we observed that immunoglobulin kappa variable 6-21 (IGKV6-21), was the highest correlated protein in PD patients ( $r: 0.54$ ,  $p < 10^{-5}$ ) (Figure 5C). This protein also exhibited one the highest fold-change in abundance when comparing urine of PD patients with non-diseased individuals (Figure 3B). Of note, the correlation between UPDRS-III scores and levels of IGKV6-12 was mainly driven by iPD patients ( $r: 0.68$ ;  $p < 10^{-5}$ ) and much weaker in *LRRK2* G2019S PD patients ( $r: 0.36$ ; not significant) (Figure 5C). Collectively, this analysis suggests that iPD and *LRRK2* G2019S patients could be stratified based on the differences between MoCA and UPDRS-III score correlations with different urine proteins.



**Figure 4. Pathogenic *LRRK2*-dependent lysosomal dysregulation is reflected in the urinary proteome and distinct from pathogenic *GBA*-induced alterations**

A) Proteins that significantly differ between pathogenic *LRRK2* carriers and controls using an ANCOVA analysis with sex, age, PD status and *GBA* status as confounders and an FDR of 5%.

B) Mean fold-changes for each of the 33 proteins that were *LRRK2*-dependently regulated in both cohorts using a pairwise t-test comparing the four subgroups (HC, NMC, iPD and *LRRK2* PD).

C) Correlation of mean fold-changes of the proteins changing *LRRK2*-dependently in the Columbia and LCC cohorts. Only proteins identified in both cohorts are shown (n=227). The colors match to the GO terms shown in (D). Proteins overlapping between the two cohorts are labeled with their name.

D) Fisher exact test was performed to identify significantly enriched GO-terms in the *LRRK2*-dependently regulated proteins in urine. All GO-terms that were significant in either cohort are displayed (FDR < 5%).

E) Proteins that differ significantly between pathogenic *GBA* carriers and controls or pathogenic *LRRK2* carriers and controls using an ANCOVA analysis with sex, age, PD status and *LRRK2/GBA* status as confounders and an FDR of 5%. In total, 237 proteins were differentially expressed in these two comparisons with 166 and 74 regulated proteins in the *LRRK2* carriers and *GBA* carriers, respectively, only three of which were common between both mutations.

F) Correlation of mean fold-changes of the proteins changing *LRRK2*-dependently and *GBA*-dependently (n=237) in the Columbia cohort. Carriers of pathogenic variants in both *GBA* and *LRRK2* were excluded from the analysis. The colors match to the GO terms shown in (D).

### Machine learning-based classification of urinary proteomes

Finally, we assessed how well machine learning models can discriminate between PD patients and non-diseased individuals, between *LRRK2* G2019S and wildtype allele carriers, and between NMC individuals and *LRRK2* G2019S patients based on the acquired urinary proteome profiles. Since the accuracy of the model largely depends on the number of samples, we combined all samples from the Columbia and LCC cohorts for these analyses. We first selected and ranked which protein features to use in the machine learning model by employing a decision tree. To classify individuals as having PD or not, the decision tree selected the 15 most important features of the PD<sup>-</sup> vs. PD<sup>+</sup> urinary proteomes, with the intensity of PPIB, one of the proteins that displayed the largest difference in abundance when PD samples were compared to the controls (Figure 3A), being on top of the list (Supplementary Figure 5A). Using these proteins, we trained an XGBoost model, a commonly applied algorithm for gradient boosting, a machine learning technique that is used to build robust predictive models based on ensembles of weaker predictions, such as decision trees. Samples were cross-validated by applying a stratified 4-fold split. This was

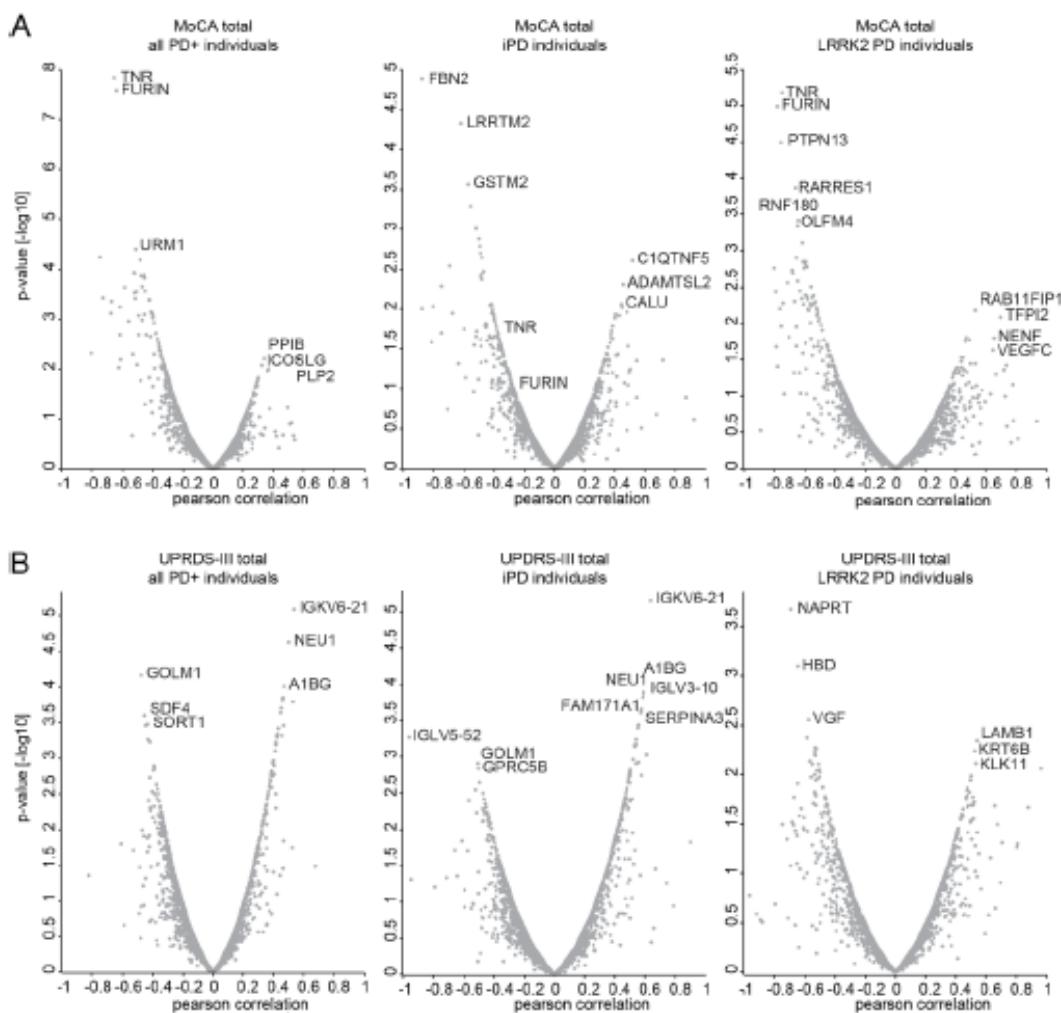
repeated (n=15) with shuffling the dataset to have a total of 60 train/test-splits to achieve a robust estimate of model performance. Each time, we determined a receiver operating characteristic (ROC) curve and found the mean area under the curve (AUC), which is often used to assess the performance of a model, to be 0.84±0.05 (Figure 6A). On average, we correctly classified 91 out of 117 PD patients and 77 out of 106 controls in the test sets (Figure 6B). Accordingly, the machine learning model reached a sensitivity of 78% and a specificity of 73%. When we trained the model on the one cohort and tested it on the other cohort, we obtained AUCs of 0.86 or 0.72, further demonstrating the robustness of the model (Figure 6C).

We next used the same machine learning methods to classify G2019S and wildtype *LRRK2* carriers using the same strategy as described above. The decision tree selected the 15 most important features, with the intensity of ENPEP being the most important one (Figure 4A; Supplementary Figure 5B). Using these proteins and the XGBoost algorithm, we obtained a mean AUC of the ROC curves of 0.87±0.04 (Figure 6D). For the test sets, we could correctly classify 73 out of the 99 *LRRK2* G2019S carriers and 103 out of the 123 wildtype allele carriers, corresponding to a 74% sensitivity and 84% specificity (Figure 6E).

When we trained the model on one of the cohorts and tested it on the other, we obtained AUCs of 0.76 or 0.80 (Figure 6F). Additionally, we trained the model on all individuals with a known *LRRK2* status and classified the sample from the Columbia cohort with an unknown *LRRK2* status with an 87% probability to be wild type *LRRK2*. After we had finished this machine learning modelling, the mutational status of this individual was determined as wild type *LRRK2*, further verifying the machine learning model.

Encouraged by these results, we wanted to see how well machine learning can discriminate *LRRK2*+ PD patients from NMCs that also carry a *LRRK2* mutation and are at increased risk of

developing the disease. Using a decision tree, we selected seven proteins for training the model (Supplementary Figure 5C). Interestingly, VGF, a neurotrophic factor, was identified as the most important feature. When using these features to train a classifier with our cross-validation scheme, the obtained mean AUC of the ROC curve was  $0.94 \pm 0.05$  and the obtained sensitivity and specificity were both 88% (Figure 6G/H). Using samples from only one cohort as a training set and applying the model to the other cohort resulted in AUCs of 0.93 and 0.74 (Figure 6I). Taken together, machine learning allowed us to classify the PD and *LRRK2* states with high specificities and sensitivities.



**Figure 5. Correlations with clinical parameters**

A) Pearson correlation scores and associated p-values [-log10] of all protein intensities with the MoCA total score. Either all PD patients (left), iPD patients (middle) or LRRK2 PD patients (right) were included in the analysis.

B) Pearson correlation scores and associated p-values [-log10] of all protein intensities with the UPDRS-III score. Either all PD patients (left), iPD patients (middle) or LRRK2 PD patients (right) were included.

**Machine learning-based classification of urinary proteomes**

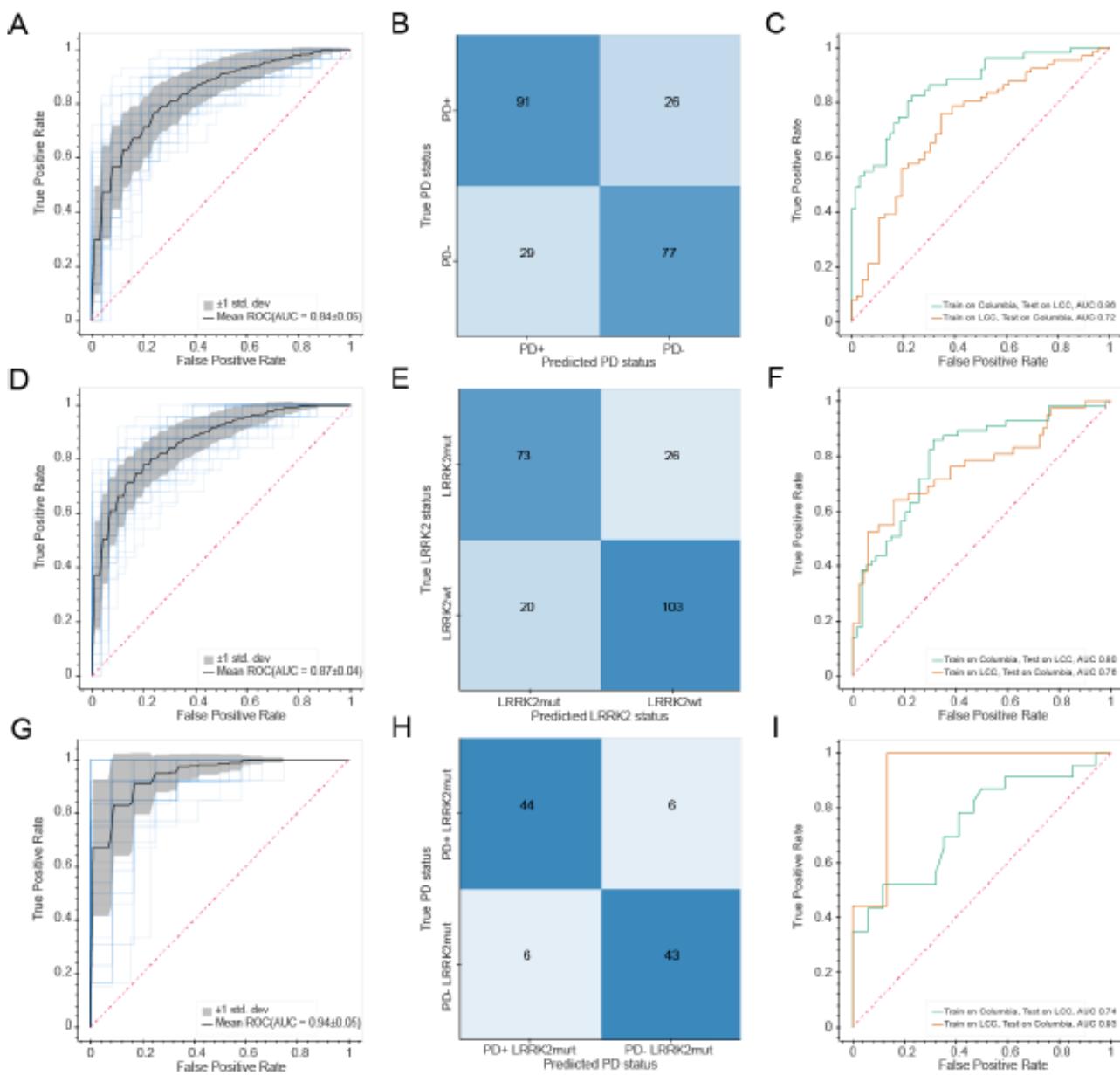
Finally, we assessed how well machine learning models can discriminate between PD patients and non-diseased individuals, between *LRRK2* G2019S and wildtype allele carriers, and between NMC individuals and *LRRK2* G2019S patients based on the acquired urinary proteome profiles. Since the accuracy of the model largely depends on the number of samples, we combined all samples from the Columbia and LCC cohorts for these analyses. We first selected and ranked which protein features to use in the machine learning model by employing a decision tree. To classify individuals as having PD or not, the decision tree selected the 15 most important features of the PD<sup>-</sup> vs. PD<sup>+</sup> urinary proteomes, with the intensity of PPIB, one of the proteins that displayed the largest difference in abundance when PD samples were compared to the controls (Figure 3A), being on top of the list (Supplementary Figure 5A). Using these proteins, we trained an XGBoost model, a commonly applied algorithm for gradient boosting, a machine learning technique that is used to build robust predictive models based on ensembles of weaker predictions, such as decision trees. Samples were cross-validated by applying a stratified 4-fold split. This was repeated (n=15) with shuffling the dataset to have a total of 60 train/test-splits to achieve a robust estimate of model performance. Each time, we determined a receiver operating characteristic (ROC) curve and found the mean area under the curve (AUC), which is often used to assess the performance of a model, to be 0.84±0.05 (Figure

6A). On average, we correctly classified 91 out of 117 PD patients and 77 out of 106 controls in the test sets (Figure 6B). Accordingly, the machine learning model reached a sensitivity of 78% and a specificity of 73%. When we trained the model on the one cohort and tested it on the other cohort, we obtained AUCs of 0.86 or 0.72, further demonstrating the robustness of the model (Figure 6C).

We next used the same machine learning methods to classify G2019S and wildtype *LRRK2* carriers using the same strategy as described above. The decision tree selected the 15 most important features, with the intensity of ENPEP being the most important one (Figure 4A; Supplementary Figure 5B). Using these proteins and the XGBoost algorithm, we obtained a mean AUC of the ROC curves of 0.87±0.04 (Figure 6D). For the test sets, we could correctly classify 73 out of the 99 *LRRK2* G2019S carriers and 103 out of the 123 wildtype allele carriers, corresponding to a 74% sensitivity and 84% specificity (Figure 6E). When we trained the model on one of the cohorts and tested it on the other, we obtained AUCs of 0.76 or 0.80 (Figure 6F). Additionally, we trained the model on all individuals with a known *LRRK2* status and classified the sample from the Columbia cohort with an unknown *LRRK2* status with an 87% probability to be wild type *LRRK2*. After we had finished this machine learning modelling, the mutational status of this individual was determined as wild type *LRRK2*, further verifying the machine learning model.

Encouraged by these results, we wanted to see how well machine learning can discriminate *LRRK2*+ PD patients from NMCs that also carry a *LRRK2* mutation and are at increased risk of developing the disease. Using a decision tree, we selected seven proteins for training the model (Supplementary Figure 5C). Interestingly, VGF, a neurotrophic factor, was identified as the most important feature. When using these features to train a classifier with our cross-

validation scheme, the obtained mean AUC of the ROC curve was  $0.94 \pm 0.05$  and the obtained sensitivity and specificity were both 88% (Figure 6G/H). Using samples from only one cohort as a training set and applying the model to the other cohort resulted in AUCs of 0.93 and 0.74 (Figure 6I). Taken together, machine learning allowed us to classify the PD and *LRRK2* states with high specificities and sensitivities.



**Figure 6. Machine learning-based classification of PD and *LRRK2* status**

A) Receiver operating characteristic (ROC) curve for the XGBoost-based model to classify PD+ vs. PD- individuals. Random performance is indicated by the dotted diagonal line. The gray area represents the standard deviation from the mean ROC curve. The blue lines show the values for a total of 15 repeats with four stratified train-test splits.

B) Confusion matrix showing the model performance for classifying PD+ vs. PD- individuals. Numbers represent the mean number from 15 repeats of cross-validation with four stratified train-test splits.

C) ROC curve for the XGBoost-based model when trained on one cohort and tested on the other cohort. Random performance is indicated by the dotted diagonal line.

D) Same as A) but for classification of *LRRK2* G2019S vs. *LRRK2* WT carriers.

E) Same as B) but for classification of *LRRK2* G2019S vs. *LRRK2* WT carriers.

F) Same as C) but for classification of *LRRK2* G2019S vs. *LRRK2* WT carriers.

G) Same as A) but for classification of PD+ vs. PD- in *LRRK2* G2019S carriers.

H) Same as B) but for classification of PD+ vs. PD- in *LRRK2* G2019S carriers.

I) Same as C) but for classification of PD+ vs. PD- in *LRRK2* G2019S carriers.

## DISCUSSION

The pathophysiology of PD leads to progressive decline of motor function and results in numerous quality of life issues for patients and their families, and inevitably leads to death within 7 to 14 years from the initial diagnosis. The majority of previous PD biomarker discovery and validation efforts have focused on CSF, serum and blood [52]. Additional strategies included targeted monitoring of  $\alpha$ -synuclein levels, given the known relationship between  $\alpha$ -synuclein accumulation and PD progression [53]. To address this problem, we developed a shotgun proteomics workflow for urinary proteome profiling. We chose to focus on urine given the non-invasive nature of obtaining clinical samples, which is a major advantage when developing a strategy that can be used not only for diagnostic and prognostic purposes, but for long-term disease progression and treatment response monitoring. Additionally, instead of focusing on a single biomarker and/or a subset of molecular entities, our shotgun proteomic approach provides a multiparameter global map of the disease state. We previously showed that this strategy can yield powerful, data-driven

descriptors of a disease [33, 35, 54], and we now confirm, for the first time, that this also works for urinary proteome analysis in the context of a complex neurodegenerative disease, such as PD.

Our quantitative shotgun proteomic workflow represents a sensitive and scalable approach for rapid analysis of a large number of samples. Applying this workflow to more than 200 urine samples from two independent cohorts allowed us to precisely quantify on average more than 2,000 proteins per sample while using minimal sample amounts of less than 100  $\mu$ l. Our approach successfully determined proteins with abundances that varied over more than five orders of magnitude, and quantified more than 1,200 proteins with a CV below 20% across the two cohorts, highlighting the high depth and precision of our study. Moreover, the observed variability between samples was much smaller than the biological variability between subjects, further illustrating the quantitative robustness of our workflow.

Another factor contributing to the quality of the urinary proteome dataset reported here, is the composition of the cohorts we analyzed. The

cohorts included two types of controls, the healthy controls as well as asymptomatic individuals that are carriers of PD-associated mutation G2019S *LRRK2*. The cohorts also included PD patients with and without the mutation, and patients of both sexes, thus allowing for different types of comparisons. For example, the global correlation map and PCA analysis showed that the sex of an individual has a dominant effect on the urinary proteome. This is in line with basic physiology and previous reports [55] but highlights the importance of incorporating sex as a confounding factor for statistical analyses. This is further illustrated by the fact that 42 and 12% of the proteins with differential abundance in PD patients vs. controls, as well as 35 and 14% of the proteins that exhibit different abundance in a *LRRK2* mutational status dependent manner also significantly differed between the sexes in the Columbia and LCC cohorts, respectively.

Applying our ‘rectangular’ strategy for biomarker discovery [36], we discovered 361 and 237 significantly altered proteins in PD patients and pathogenic *LRRK2* carriers, respectively. The observed overlap of proteins exhibiting significantly perturbed levels in the two independent cohorts confirms that valuable information can be inferred from the urinary proteome for neurodegenerative diseases. We note the scalability of our workflow, which will allow its application to larger cohorts with more comprehensive genetic and clinical information. This extension of our work will be important to further validate our results and to discover additional biomarker candidates with improved statistical power.

Our data analysis led to several interesting observations that might suggest opportunities for follow up. Here, we will briefly discuss only a small number of such examples. For those interested in more in depth data mining, we made

our datasets available via publicly accessible depository (see Materials and Methods for accession numbers). An interesting insight that emerged from the GO-term analysis of PD patient vs. control proteomes identified the GO-term ‘bone development’ as significantly enriched. The enzyme PPIB was significantly upregulated in PD patients in both cohorts. This cyclophilin assists the folding of type I collagen and can protect cells against MPP<sup>+</sup>-induced cell death in a PD cell culture model [56]. Inhibitors of the closely related family member cyclophilin D (CypD) are considered as therapeutic agents against several neurodegenerative diseases including PD [57]. Most other proteins associated with the GO-term ‘bone development’ were downregulated in PD patients, in line with recent findings that PD patients frequently suffer from osteoporosis and osteopenia [46, 58]. Going forward, it would be important to examine the relationship between PD profession and bone health more closely, as this connection is currently underexplored.

Another enriched term was ‘growth factor activity’, although none of the proteins associated with this term was significantly regulated in both cohorts. Growth factors and particularly neurotrophic factors have gained strong interest as therapeutic agents in Parkinson’s disease but so far have not produced convincing clinical benefits [59]. The neurosecretory protein VGF was strongly decreased in PD patients in both cohorts (Columbia: 0.24, LCC: 0.54) but only reached statistical significance in the better-controlled Columbia cohort. VGF is synthesized as a prohormone and proteolytically processed to various biologically active peptides. In this study, we identified peptides covering most of the VGF sequence, including sequences contained in the neuroendocrine regulatory peptide-1. However, the applied tryptic digestion complicates a direct link to the endogenous hormone peptides. VGF is exclusively synthesized and secreted by neuronal

and neuroendocrine tissues. In the CNS, VGF promotes neurite growth and exhibits neuroprotective activity, while it also regulates energy homeostasis in peripheral tissues. Gene expression of VGF in the cortex [60] and peptides derived from this gene are reduced in post-mortem parietal brain cortex and plasma from PD patients [61, 62]. Furthermore, VGF has been suggested as a biomarker in CSF for Alzheimer's Disease (AD) and Amyotrophic lateral sclerosis (ALS) and its expression was reduced in the CSF of AD and ALS patients compared to controls [63, 64].

We also identified several apolipoproteins - the major proteinaceous constituent of lipoproteins - to be significantly upregulated in PD patients in the Columbia cohort. They have been linked to neurodegenerative disorders including Alzheimer disease, including in our recent proteomic study of CSF [54]. APOE variants were shown to exhibit neuroprotective activity (reviewed in [65]). APOA1 is the major protein component of plasma high-density lipoprotein and its low levels in CSF and plasma have been reported as a potential PD biomarker [66-68]. While APOE and ApoA1 are the most abundant apolipoproteins in the CSF and highly enriched in the brain [69, 70], APOC1 - a less-abundant brain apolipoprotein - was implicated in Alzheimer disease although its regulation and possible role is poorly understood [71].

In another illustrative example, we analyzed proteomic differences between patients with and without a major inherited mutation associated with familial PD, *LRRK2* G2019S. Lysosomal dysregulation and associated  $\alpha$ -synuclein aggregation appear to be a central event in the pathogenesis of PD [72] and *LRRK2*, through its regulation of the endolysosomal pathways, is a key player in this mechanism [73, 74]. Interestingly, the *LRRK2*-dependent signature in the urinary proteome seemed to be more

consistent than the PD-dependent signature, as indicated by the larger overlap of 33 vs. 10 proteins between the two cohorts. This suggests that the genetic mutation of *LRRK2* not only manifests in the central nervous systems but also dysregulates multiple pathways in distal organs such as the bladder and kidney, where *LRRK2* is actually highly expressed [75]. Our results demonstrate that urine of pathogenic *LRRK2* carriers strongly reflects lysosomal dysregulation associated with increase in *LRRK2* activity [72] [72, 73]. This suggests that the genetic mutation of *LRRK2* not only manifests in the central nervous systems but also dysregulates multiple pathways in distal organs such as the bladder and kidney, where *LRRK2* is highly expressed [75]. Additionally, one of the strongest upregulated proteins in *LRRK2* G2019S carriers was the alkaline phosphatase ALPI. We suggest that this enzyme may counteract the hyperactive kinase by as yet undiscovered feedback mechanisms. Interestingly, knockdown of ALPI has been shown to decrease both *LRRK2* levels and activity in cells [76]. We also found several lysosomal proteins including  $\alpha$ - and  $\beta$ -subunit of  $\beta$ -hexosaminidase A (HEXA and HEXB), GM2A and GCase, whose genes are associated with LSDs, to be upregulated in *LRRK2* G2019S carriers in both cohorts. Mutations in many LSD genes have been associated with PD, suggesting common pathogenic mechanisms underly both diseases. The GO-term 'glycosphingolipid metabolic process' was enriched among *LRRK2*-regulated proteins, in agreement with increased interest in understanding how sphingolipids contribute to PD that stems from the fact that several PD-associated genes including *GBA* are linked to their metabolism [77, 78]. Ceramide levels are increased in *LRRK2*-deficient mouse brains and this decrease is regulated by *GBA* [48]. It still remains unclear how the disruption of sphingolipid metabolism may result in PD-associated neurodegeneration or if *LRRK2* directly or indirectly regulates this process. Our

data suggest that pathogenic mutations in *GBA* and *LRRK2* mainly affect distinct regulatory networks, as only three proteins were significantly altered in common by mutations in both genes. However, further studies on larger *GBA* cohorts are needed to confirm and extend our findings.

One of the cohorts we analyzed (Columbia) included information on clinical scores of cognitive and motor performances. This allowed us to correlate proteomic changes to clinical score, thus revealing that TNR and FURIN levels were strongly correlated with higher cognitive impairment. FURIN is a protease and is involved in NMDA-induced neuronal injury [79]. Furthermore, its homologue in the fruit fly, Furin1, has been reported to be a translational target of pathogenic *LRRK2* and to be involved in neurotoxicity [80]. TNR is a neural extracellular matrix protein exclusively expressed in the brain. It is involved in neurogenesis [81] and extracellular matrix aggregates in the brain called perineurial nets [82]. Of note, rare TNR variants have also been associated with familial PD [83]. Interestingly, IGKV6-21 was highly upregulated in PD patients and also strongly correlated with the UPDRS-III score. Although the underlying biology is unclear, the association with both PD risk and severity make this V region a promising biomarker candidate to pursue in future studies.

To extend utility of our datasets, we developed a machine learning model for stratifying PD patients and *LRRK2* G2019S carriers with high sensitivities and specificities. Importantly, the machine learning model excelled in classifying the PD status in *LRRK2* G2019S carriers. This is

of high interest, because although these carriers are at an increased risk of developing PD, there is no predictive marker to determine whether or not and when a mutation carrier develops the disease. Given the performance of the machine learning model, VGF, LTF, CELA3A, TUBB4B, and SOD2 are promising candidates as predictive markers to early indicate disease development.

In summary, we have demonstrated that a distal body fluid like urine contains brain-specific proteins and can inform about the disease and mutation status in a neurodegenerative disease. Our urinary proteomics workflow is relatively straightforward, readily scalable and thus easily applicable to larger and more powerful cohorts. It would be important to also apply it to longitudinal data to confirm increased levels of PPIB and IGKV6-21 in PD patients and VGF as a potential indicator for disease manifestation in *LRRK2* G2019S carriers but also identify new biomarkers for PD risk and disease progression in idiopathic and genetic forms of PD. Our results demonstrate that urinary proteome profiling enables the discovery of better biomarkers, which could have a major impact on important aspects of disease management: (i) a diagnostic biomarker will enable early and objective diagnosis of PD, (ii) a prognostic biomarker will provide information about the progression of the disease, and (iii) predictive and treatment response biomarkers will allow to monitor whether and how the patients respond to a therapy. Reliable biomarkers assessing *LRRK2* activity can also aid with monitoring compliance of *LRRK2* kinase inhibitors and treatment efficacy, early detection of non-manifesting carriers to prevent disease onset and stratify idiopathic PD patients who could benefit from *LRRK2*-based therapies.

## MATERIALS and METHODS

### Study cohorts

In this study, urine samples from two independent cross- sectional cohorts were analyzed. Both studies were approved by local institutional review boards, and each participant signed an informed consent.

The first cohort was recruited at Columbia University Irving Medical Center (Columbia cohort) and its participants donated urine under a MJFF-funded *LRRK2* biomarker project from March 2016 to April 2017. This cohort contained 35 healthy individuals without pathogenic *LRRK2* mutation (HC), 16 non-manifesting carriers of the *LRRK2* G2019S mutation (NMC), 40 idiopathic PD patients without pathogenic *LRRK2* mutation (iPD) and 28 PD patients with the pathogenic *LRRK2* G2019S mutation (LRRK2 PD) and 1 PD patient with an unknown mutation status of *LRRK2*. Motor performance was evaluated using the Unified Parkinson's Disease Rating Scale part III (UPDRS-III), and cognitive functioning was assessed using the Montreal Cognitive Assessment (MoCA) test. Genotyping for *LRRK2* G2019S and *GBA* mutations was conducted as previously described [38].

To confirm findings from the Columbia cohort, urine from a second cohort consisting of 115 bio-banked urine samples from the Michael J. Fox Foundation for Parkinson's Research (MJFF)-funded *LRRK2* Cohort Consortium (LCC) was analyzed. The cohort used in this study was an exploratory subset of a larger cohort and contained 26 healthy individuals without pathogenic *LRRK2* mutation (HC), 37 non-manifesting carriers of the *LRRK2* G2019S mutation (NMC), 29 idiopathic PD patients without pathogenic *LRRK2* mutation (iPD) and 23 PD patients with the pathogenic *LRRK2* G2019S mutation. UPDRS-III and MoCA scores

were not available for subjects from the LCC cohort.

### Quality assessment

To generate the urine-specific quality marker panel, we recruited three volunteers from within the Department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry who kindly donated 10 ml of urine at three different time points during a day and provided a written informed consent, with prior approval of the ethics committee of the Max Planck Society.

Following the collection, urinary samples were centrifuged at 2000g for 10 minutes, supernatants were harvested and pellets were resuspended in 100  $\mu$ l of Urea sample solution. 100  $\mu$ l of each supernatant and the entire 100  $\mu$ l of the resuspended pellets were used for sample preparation as described below. A sample was flagged for potential contamination if the summed intensity of all proteins in the respective quality marker panel differed more than 2 standard deviations from the mean of all samples within the cohort.

### Sample Preparation

The undiluted neat urine as well as the cleared and pelleted urine samples for the urine-specific quality marker panel were prepared using MStern Blot protocol as described previously [27]. Briefly, 100  $\mu$ l of urine was first diluted in 300  $\mu$ l of Urea sample solution (8 M urea in 50 mM ammonium bicarbonate (ABC)) and subsequently mixed with 30  $\mu$ l of 150 mM dithiothreitol (DTT) solution (150 mM DTT, 8 M urea, 50 mM ABC) in a 96-well plate. The resulting solution was incubated for 20 min at room temperature. Reduced cysteine side chains were alkylated by adding 30  $\mu$ L of iodoacetamide (IAA) solution (700 mM IAA, 8 M urea, 50 mM ABC) and incubated for 20 min in the dark. During incubation, each well of the 96-well

PVDF membrane plates (MSIPS4510, Merck Millipore) was activated and equilibrated with 150  $\mu$ l of 70% ethanol/water and urea sample solution, respectively. The urine samples were transferred through the PVDF membranes using a vacuum manifold (MSVMHTS00, Merck Millipore). Adsorbed proteins were washed two times with 150  $\mu$ l of 50 mM ABC. Digestion was performed at 37°C for 2 hours by adding 100  $\mu$ l digestion buffer (5% v/v acetonitrile (ACN)/50 mM ABC) containing 0.35  $\mu$ g per well of each protease trypsin and LysC. After incubation in a humidified incubator, the resulting peptides were collected by applying vacuum and remaining peptides were eluted twice with 75  $\mu$ l of 40%/0.1%/59.9% (v/v) acetonitrile/formic acid/water. The pooled peptide solutions were dried in a vacuum centrifuge.

Peptides resuspended in 0.1% trifluoroacetic acid (TFA) were desalting on C18 StageTips as described in [29]. The StageTips were centrifuged at 1,000g for washing with 0.1% TFA and elution with 80% ACN/0.1% TFA. The eluate was evaporated to dryness using a vacuum centrifuge and peptides were resuspended in 10  $\mu$ l buffer A\* (2% ACN/0.1% TFA and stored at -20°C. Samples were thawed shortly before mass spectrometric analysis and shaken for 2 minutes at 2000rpm (thermomixer C, Eppendorf). Peptide concentrations were measured optically at 280nm (Nanodrop 2000, Thermo Scientific) and subsequently equalized using buffer A\*. 500ng peptide was subjected to LC-MS/MS analysis.

Cohort-specific libraries for data-independent analyses were generated by pooling of 25 randomly selected samples of each cohort. Sample pools were fractionated into 24 fractions each by high pH (pH 10) reversed-phase chromatography as described earlier [84]. Fractions were concatenated automatically by shifting the collection tube every 120 seconds and

subsequently dried in a vacuum centrifuge and resuspended in buffer A\*.

To increase the depth of each library, extracellular vesicles (EV) were isolated from pooled urine samples of each cohort by ultracentrifugation as described earlier (Add Andy West paper here). Briefly, 8.5 ml of 6 urine samples per group (*LRRK2*-/PD-, *LRRK2*+/PD-, *LRRK2*-/PD+ and *LRRK2*+/PD+) were pooled and centrifuged at 10,000g for 30 min at 4 °C and supernatant was transferred and then centrifuged again at 100,000g for 1h at 4 °C. Supernatants were discarded and pellets were washed by adding 30 mL PBS and centrifugation at 100,000g for 1 h at 4 °C. Supernatant was discarded and pellets were resuspended in 100  $\mu$ l of a sodium deoxycholate-based lysis buffer containing chloroacetamide (PreOmics GmbH) and heated to 95°C for 10 min for reduction and alkylation. After cooling to room temperature, 0.75  $\mu$ g of each protease trypsin and 0 LysC were added to each sample and digestion was performed at 37°C overnight. Peptides were desalting with SDB-RPS (styrenedivinylbenzene-reverse phase sulfonate) StageTips. Samples were mixed with 5 volumes of 1% TFA/isopropanol for loading on StageTips and subsequently washed once with 1%TFA/isopropanol and once with 0.2% TFA as described earlier [29]. Peptides were eluted 80%/5% ACN/ammonium hydroxide. The eluate was completely dried using a vacuum centrifuge and resuspended in 0.1% formic acid. Peptides were then separated into 8 fractions by high pH reversed-phase chromatography as described above for the libraries.

To determine coefficients of variation for the Columbia cohort, urine from five donors in triplicates on one plate were subjected to sample preparation (intra-plate) and this was repeated on three different plates (inter-plate). For the LCC cohort, urine from three donors in duplicates on

one plate were subjected to sample preparation (intra-plate). Urine from nine other subjects were prepared on two different plates (inter-plate).

#### LC-MS/MS analysis

LC-MS/MS analysis was performed on an EASY-nLC 1200 coupled to a Q Exactive HF-X Orbitrap mass spectrometer via a nano-electrospray ion source (all Thermo Fisher Scientific). Purified peptides were separated at 60 °C on 50cm columns with an inner diameter of 75µm packed in-house with ReproSil-Pur C18-AQ 1.9µm resin (Dr.Maisch GmbH). Mobile phases A and B were 99.9/0.1% water/formic acid (v/v) and 80/20/0.1% acetonitrile/water/formic acid (v/v/v). For the LCC cohort, the flow rate was constant at 300 nL/min and the initial concentration of 5% B was linearly increased to 30% B within 36 minutes, and then increased further to 95% within 6 min with a 3 min plateau at the end. For the Columbia cohort, the flow rate was constant at 350 nL/min and the initial concentration of 5% B was linearly increased to 30% B within 35 minutes, and then increased further to 95% within 5 min with a 5 min plateau at the end.

MS data was acquired in the data-independent acquisition (DIA) scan mode for single-shot patient samples, using the MaxQuant Live software and spectral processing with phase-constrained spectrum deconvolution (phi-SDM) [85, 86]. Full MS scans were acquired in the range of m/z 300–1,650 at a resolution of 60,000 at m/z 200 and the automatic gain control (AGC) set to 3e6. For the Columbia cohort, additionally two BoxCar scans with 12 isolation windows each and a resolution of 60,000 at m/z 200 were acquired [87]. Full MS events were followed by 33 MS/MS windows (LCC cohort) or 58 MS/MS windows (Columbia cohort) per cycle in the range of m/z 300–1,650 at a resolution of 15,000 at m/z 200. For the LCC cohort, higher-energy collisional dissociation MS/MS scans were

acquired with a stepped normalized collision energy of 25/27.5/30 and ions were accumulated to reach an AGC target value of 3e6 or for a maximum of 30 ms. For the Columbia cohort, higher-energy collisional dissociation MS/MS scans were acquired with a normalized collision energy of 27 and ions were accumulated to reach an AGC target value of 3e6 or for a maximum of 22 ms.

All fractionated samples including EV fractions were acquired with a top12 data-dependent acquisition (DDA) scan mode. Full MS scans were acquired in the range of m/z 300–1,650 at a resolution of 60,000 (Columbia cohort) or 120,000 (LCC cohort) at m/z 200. The automatic gain control (AGC) target was set to 3e6. Higher-energy collisional dissociation MS/MS scans were acquired with a normalized collision energy of 27 at a resolution of 15,000 at m/z 200. Precursor ions with charge states of 2-7 were isolated in a 1.4 Th window and accumulated to reach an AGC target value of 1e5 or for a maximum of 60 ms. Precursors were dynamically excluded for 20 s after the first fragmentation event.

#### Mass spectrometry data processing

The MS data of the fractionated pools (DDA MS data, 24 neat pool urine and 8 EV fractions) and the single shot subject samples (DIA MS data, 165 and 132 samples in Columbia and LCC, respectively) were used to generate a DDA-library and direct-DIA-library, respectively, which were computationally merged into two cohort-specific hybrid libraries using Spectronaut version 13.9.191106.43655 (Biognosys AG). For all experiments except the machine learning, the two cohorts were quantified separately in Spectronaut. A minimum of 3 and a maximum of 10 fragments was required for each peptide in the library. The hybrid spectral libraries were subsequently used to search the MS data of the

single shot patient samples in the Spectronaut software. All searches were performed against the human SwissProt reference proteome of canonical and isoform sequences with 42,431 entries downloaded in July 2019. Searches used carbamidomethylation as fixed modification and acetylation of the protein N-terminus and oxidation of methionines as variable modifications. Trypsin/P proteolytic cleavage rule was used, permitting a maximum of 2 missed cleavages and a minimum peptide length of 7 amino acids. The Q-value cutoffs for both library generation and DIA analyses were set to 0.01. For generation of the global correlation map, the individual protein correlations with clinical parameters, and the machine learning, the Q-value data filtering setting in Spectronaut was set to 'Qvalue' to use every peptide passing the Q-value threshold for the protein group quantification. For all other analyses, the setting was set to 'Qvalue percentile' with a cutoff of 25%, to use only those peptides for the protein quantification that passed the Q-value threshold in at least 25% of all analyzed samples. The 'Qvalue percentile' setting results in a complete data matrix with no missing values, as the noise is quantified and reported if the peptide did not pass the Qvalue threshold.

#### **Bioinformatics data analysis**

The Perseus software package versions 1.6.0.7 and 1.6.1.3 and GraphPad Prism version 7.03 were used for the data analysis [88]. Protein intensities were log2-transformed for further analysis apart from correlation and coefficient of variation analysis. Coefficients of variation (CVs) were calculated in Perseus for all inter-plate and intra-plate combinations of samples, the median values were reported as overall coefficient of variation. The protein CVs of the main study were calculated likewise within cohorts individually. The protein abundance levels were cross-correlated to generate a matrix of correlation coefficients. Unsupervised

hierarchical clustering was performed using Perseus and proteins were clustered based on Pearson correlation scores. For generation of the abundance curves, median protein abundances across all samples within a proteome were used. ANCOVA analysis was performed in python (version 3.7.6) using the pandas (version 1.0.1), numpy (version 1.18.1) and pingouin (version 0.3.4) packages. For the ANCOVA analysis, age at sample collection, *LRRK2* status (only in PD+ vs. PD-), *GBA* status (only Columbia cohort *LRRK2*+ vs. *LRRK2*-), and PD status (only *LRRK2*+ vs. *LRRK2*-) were set as confounding factors. The FDR was set to 5% after Benjamini-Hochberg correction. GO annotations were matched to the proteome data based on Uniprot and Ensemble identifiers. Annotation term enrichment was performed with Fisher exact test in Perseus separately for each cohort. Annotation terms were filtered for terms with an FDR of 5% after Benjamini-Hochberg correction in each cohort. Calculation of Pearson correlation scores and associated p-values of protein intensities to UPDRS-III and MoCA scores was performed in Perseus.

#### **Machine learning**

Data processing and machine learning was performed in Python (Version 3.7.3). Missing values were not imputed and protein intensities were normalized using the StandardScaler method from the scikit-learn package (0.21.3). The XGBoost package (Version 0.90) was used to classify the samples and results were plotted using the bokeh library (2.1.1). Features were selected using a decision tree. Samples from both Columbia and LCC cohorts were used for the model and cross-validated using four stratified training/test splits and 15 repeats were applied. To assess sensitivity and specificity of the model, the results of the test sets were summed and averaged from 15 repeats.

## ACKNOWLEDGEMENTS

Biospecimens used in the analyses presented in this article were obtained from the MJFF-sponsored *LRRK2* Cohort Consortium (LCC). For up-to-date information on the study, visit [www.michaeljfox.org/lcc](http://www.michaeljfox.org/lcc). We thank all members of the Proteomics and Signal Transduction Group at the Max Planck Institute of Biochemistry and the Clinical Proteomics Group at the NNF Center for Protein Research for help and discussions and in particular Jakob Bader, Philipp Geyer, Igor Paron, Christian Deiml and Alexander Strasser for helpful discussions and technical assistance. We further thank Hanno Steen, Dario Alessi and Suzanne Pfeffer and their group members for helpful discussion. We thank employees of the Michael J. Fox Foundation for Parkinson's research for helpful discussions.

## FUNDING

The work carried out in this project was supported by the Max Planck Society for the Advancement of Science and The Michael J. Fox Foundation.

## AUTHOR CONTRIBUTIONS

SVW and OK designed the experiments, performed, analyzed and interpreted all data. MTS helped with the machine learning. SP, MS, KM and RNA were responsible for sample collection and their distribution. SP, KM and RNA helped with interpretation of the results. MM supervised and guided the project, interpreted results and wrote the manuscript with SVW and OK.

## COMPETING INTERESTS

The authors declare no competing interests.

## REFERENCES

1. De Lau, L.M. and M.M.J.T.L.N. Breteler, *Epidemiology of Parkinson's disease*. 2006. **5**(6): p. 525-535.
2. Reeve, A., E. Simcox, and D.J.A.r.r. Turnbull, *Ageing and Parkinson's disease: why is advancing age the biggest risk factor?* 2014. **14**: p. 19-30.
3. Tysnes, O.-B. and A.J.J.o.N.T. Storstein, *Epidemiology of Parkinson's disease*. 2017. **124**(8): p. 901-905.
4. West, A.B.J.M.d., *Ten Years and Counting: Moving Leucine-R ich Repeat K inase 2 Inhibitors to the Clinic*. 2015. **30**(2): p. 180-189.
5. Steger, M., et al., *Systematic proteomic analysis of LRRK2-mediated Rab GTPase phosphorylation establishes a connection to ciliogenesis*. 2017. **6**: p. e31012.
6. Steger, M., et al., *Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases*. 2016. **5**: p. e12813.
7. Karayel, O., et al., *Accurate MS-based Rab10 phosphorylation stoichiometry determination for LRRK2 activity in Parkinson's disease*. 2019: p. 819607.
8. Tolosa, E., et al., *LRRK2 in Parkinson disease: challenges of clinical trials*. 2020: p. 1-11.
9. Nalls, M.A., et al., *Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies*. 2019. **18**(12): p. 1091-1102.

10. Di Maio, R., et al., *LRRK2 activation in idiopathic Parkinson's disease*. 2018. **10**(451): p. eaar5429.
11. Alcalay, R.N., et al., *Higher urine bis (Monoacylglycerol) phosphate levels in LRRK2 G2019S mutation carriers: implications for therapeutic development*. 2020. **35**(1): p. 134-141.
12. Decramer, S., et al., *Urine in clinical proteomics*. 2008. **7**(10): p. 1850-1862.
13. An, M., Y.J.G. Gao, proteomics, and bioinformatics, *Urinary biomarkers of brain diseases*. 2015. **13**(6): p. 345-354.
14. Kentsis, A., et al., *Urine proteomics for profiling of human disease using high accuracy mass spectrometry*. 2009. **3**(9): p. 1052-1061.
15. Kentsis, A., et al., *Urine proteomics for discovery of improved diagnostic markers of Kawasaki disease*. 2013. **5**(2): p. 210-220.
16. Ferrari, E., et al., *Urinary proteomics profiles are useful for detection of cancer biomarkers and changes induced by therapeutic procedures*. 2019. **24**(4): p. 794.
17. Adachi, J., et al., *The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins*. 2006. **7**(9): p. R80.
18. Nguyen, M.T., et al., *Early prediction of acute renal injury using urinary proteomics*. 2005. **25**(4): p. 318-326.
19. Zimmerli, L.U., et al., *Urinary proteomic biomarkers in coronary artery disease*. 2008. **7**(2): p. 290-298.
20. Duangkumpha, K., et al., *Urine proteomics study reveals potential biomarkers for the differential diagnosis of cholangiocarcinoma and periductal fibrosis*. 2019. **14**(8).
21. Kaiser, T., et al., *Proteomics applied to the clinical follow-up of patients after allogeneic hematopoietic stem cell transplantation*. 2004. **104**(2): p. 340-349.
22. Metzger, J., et al., *Urine proteomic analysis differentiates cholangiocarcinoma from primary sclerosing cholangitis and other benign biliary disorders*. 2013. **62**(1): p. 122-130.
23. Tantipaiboonwong, P., et al., *Different techniques for urinary protein analysis of normal and lung cancer patients*. 2005. **5**(4): p. 1140-1149.
24. Ward, D.G., et al., *Proteomic profiling of urine for the detection of colon cancer*. 2008. **6**(1): p. 19.
25. Mischak, H., et al., *Proteomic analysis for the assessment of diabetic renal damage in humans*. 2004. **107**(5): p. 485-495.
26. Nagaraj, N. and M.J.J.o.p.r. Mann, *Quantitative analysis of the intra-and inter-individual variability of the normal urinary proteome*. 2011. **10**(2): p. 637-645.
27. Berger, S.T., et al., *MStern blotting—high throughput polyvinylidene fluoride (PVDF) membrane-based proteomic sample preparation for 96-well plates*. 2015. **14**(10): p. 2814-2823.
28. Batth, T.S., et al., *Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation*. 2019. **18**(5): p. 1027-1035.
29. Kulak, N.A., et al., *Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells*. 2014. **11**(3): p. 319.
30. Ding, H., et al., *Urine Proteomics: Evaluation of Different Sample Preparation Workflows for Quantitative, Reproducible, and Improved Depth of Analysis*. 2020. **19**(4): p. 1857-1862.
31. Albrechtsen, N.J.W., et al., *Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery*. 2018. **7**(6): p. 601-612. e3.
32. Geyer, P.E., et al., *Proteomics reveals the effects of sustained weight loss on the human plasma proteome*. 2016. **12**(12).
33. Geyer, P.E., et al., *Plasma proteome profiling to assess human health and disease*. 2016. **2**(3): p. 185-195.
34. Geyer, P.E., et al., *Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies*. 2019. **11**(11).
35. Niu, L., et al., *Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease*. 2019. **15**(3).
36. Geyer, P.E., et al., *Revisiting biomarker discovery by plasma proteomics*. 2017. **13**(9).

37. Melachroinou, K., et al., *Elevated In Vitro Kinase Activity in Peripheral Blood Mononuclear Cells of Leucine-Rich Repeat Kinase 2 G2019S Carriers: A Novel Enzyme-Linked Immunosorbent Assay-Based Method*.  
38. Alcalay, R.N., et al., *Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations*. 2015. **138**(9): p. 2648-2658.  
39. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial*. 2018. **14**(8): p. e8126.  
40. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis*. 2012. **11**(6).  
41. Guo, Z., et al., *A proteomic analysis of individual and gender variations in normal human urine and cerebrospinal fluid using iTRAQ quantification*. 2015. **10**(7).  
42. Uhlén, M., et al., *Tissue-based map of the human proteome*. 2015. **347**(6220): p. 1260419.  
43. Jain, S.J.P. and r. disorders, *Multi-organ autonomic dysfunction in Parkinson disease*. 2011. **17**(2): p. 77-83.  
44. Poewe, W.J.E.j.o.n., *Non-motor symptoms in Parkinson's disease*. 2008. **15**: p. 14-20.  
45. Cook, C., C. Stetler, and L.J.C.S.H.p.i.m. Petrucelli, *Disruption of protein quality control in Parkinson's disease*. 2012. **2**(5): p. a009423.  
46. Torsney, K.M., et al., *Bone health in Parkinson's disease: a systematic review and meta-analysis*. 2014. **85**(10): p. 1159-1166.  
47. Ysselstein, D., et al., *LRRK2 kinase activity regulates lysosomal glucocerebrosidase in neurons derived from Parkinson's disease patients*. 2019. **10**(1): p. 1-9.  
48. Ferrazza, R., et al., *LRRK2 deficiency impacts ceramide metabolism in brain*. 2016. **478**(3): p. 1141-1146.  
49. Spitz, M., et al., *Association of LRRK2 and GBA mutations in a Brazilian family with Parkinson's disease*. 2015. **21**(7): p. 825-826.  
50. Duran, R., et al., *The glucocerobrosidase E326K variant predisposes to Parkinson's disease, but does not cause Gaucher's disease*. 2013. **28**(2): p. 232-236.  
51. Yahalom, G., et al., *Carriers of both GBA and LRRK2 mutations, compared to carriers of either, in Parkinson's disease: Risk estimates and genotype-phenotype correlations*. 2019. **62**: p. 179-184.  
52. Chen-Plotkin, A.S., et al., *Finding useful biomarkers for Parkinson's disease*. 2018. **10**(454).  
53. Fields, C.R., N. Bengoa-Vergniory, and R.J.F.i.M.N. Wade-Martins, *Targeting alpha-synuclein as a therapy for Parkinson's Disease*. 2019. **12**.  
54. Bader, J.M., et al., *Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease*. 2020. **16**(6): p. e9356.  
55. Shao, C., et al., *Comprehensive analysis of individual variation in the urinary proteome revealed significant gender differences*. 2019. **18**(6): p. 1110-1122.  
56. Oh, Y., et al., *Cyclophilin B protects SH-SY5Y human neuroblastoma cells against MPP+-induced neurotoxicity via JNK pathway*. 2016. **478**(3): p. 1396-1402.  
57. M Fayaz, S., et al., *CypD: the key to the death door*. 2015. **14**(5): p. 654-663.  
58. Handa, K., et al., *Bone loss caused by dopaminergic degeneration and levodopa treatment in parkinson's disease model mice*. 2019. **9**(1): p. 1-16.  
59. Paul, G. and A.M.J.E.J.o.N. Sullivan, *Trophic factors for Parkinson's disease: Where are we and where do we go from here?* 2019. **49**(4): p. 440-452.  
60. Henderson-Smith, A., et al., *Next-generation profiling to identify the molecular etiology of Parkinson dementia*. 2016. **2**(3): p. e75.  
61. Cocco, C., et al., *VGF peptides as novel biomarkers in Parkinson's disease*. 2020. **379**(1): p. 93-107.  
62. Cocco, C., et al., *Distribution of VGF peptides in the human cortex and their selective changes in Parkinson's and Alzheimer's diseases*. 2010. **217**(6): p. 683-693.  
63. Pasinetti, G., et al., *Identification of potential CSF biomarkers in ALS*. 2006. **66**(8): p. 1218-1222.

64. Carrette, O., et al., *A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease*. 2003. 3(8): p. 1486-1494.

65. Emamzadeh, F.N.J.O.M.N., *Role of Apolipoproteins and  $\alpha$ -Synuclein in Parkinson's Disease*. 2017. 62(3-4): p. 344-355.

66. Wang, E.S., et al., *Tetranectin and apolipoprotein A-I in cerebrospinal fluid as potential biomarkers for Parkinson's disease*. 2010. 122(5): p. 350-359.

67. Swanson, C.R., et al., *Plasma apolipoprotein A1 associates with age at onset and motor severity in early Parkinson's disease patients*. 2015. 30(12): p. 1648-1656.

68. Qiang, J.K., et al., *Plasma apolipoprotein A1 as a biomarker for Parkinson disease*. 2013. 74(1): p. 119-127.

69. Borghini, I., et al., *Characterization of subpopulations of lipoprotein particles isolated from human cerebrospinal fluid*. 1995. 1255(2): p. 192-200.

70. Koch, S., et al., *Characterization of four lipoprotein classes in human cerebrospinal fluid*. 2001. 42(7): p. 1143-1151.

71. Petit-Turcotte, C., et al., *Apolipoprotein CI expression in the brain in Alzheimer's disease*. 2001. 8(6): p. 953-963.

72. Alessi, D.R. and E.J.S. Sammler, *LRRK2 kinase in Parkinson's disease*. 2018. 360(6384): p. 36-37.

73. !!! INVALID CITATION !!! [60-62].

74. Henry, A.G., et al., *Pathogenic LRRK2 mutations, through increased kinase activity, produce enlarged lysosomes with reduced degradative capacity and increase ATP13A2 expression*. 2015. 24(21): p. 6013-6028.

75. Biskup, S., et al., *Dynamic and redundant regulation of LRRK2 and LRRK1 expression*. 2007. 8(1): p. 102.

76. Berndsen, K., et al., *PPM1H phosphatase counteracts LRRK2 signaling by selectively dephosphorylating Rab proteins*. 2019. 8.

77. Plotegher, N., et al., *Ceramides in Parkinson's disease: from recent evidence to new hypotheses*. 2019. 13: p. 330.

78. Lin, G., et al., *Sphingolipids in the Pathogenesis of Parkinson's Disease and Parkinsonism*. 2019. 30(2): p. 106-117.

79. Yamada, M., et al., *Furin inhibitor protects against neuronal cell death induced by activated NMDA receptors*. 2018. 8(1): p. 1-9.

80. Maksoud, E., E.H. Liao, and A.P.J.C.r. Haghghi, *A neuron-glial trans-signaling cascade mediates LRRK2-induced neurodegeneration*. 2019. 26(7): p. 1774-1786. e4.

81. Xu, J.-C., et al., *The extracellular matrix glycoprotein tenascin-R regulates neurogenesis during development and in the adult dentate gyrus of mice*. 2014. 127(3): p. 641-652.

82. Morawski, M., et al., *ECM in brain aging and dementia*, in *Progress in brain research*. 2014, Elsevier. p. 207-227.

83. Farlow, J.L., et al., *Whole-exome sequencing in familial Parkinson disease*. 2016. 73(1): p. 68-75.

84. Kulak, N.A., et al., *Loss-less nano-fractionator for high sensitivity, high coverage proteomics*. 2017. 16(4): p. 694-705.

85. Wichmann, C., et al., *MaxQuant Live enables global targeting of more than 25,000 peptides*. 2019. 18(5): p. 982-994.

86. Grinfeld, D., et al., *Phase-constrained spectrum deconvolution for Fourier transform mass spectrometry*. 2017. 89(2): p. 1202-1211.

87. Meier, F., et al., *BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes*. 2018. 15: p. 440-448.

88. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data*. 2016. 13(9): p. 731.

## Acknowledgement

I would like to express my deepest gratitude to all the wonderful people who contributed to this work and made my PhD time an amazing experience.

First of all, I want to thank my supervisor, Prof. Dr. Matthias Mann for his continuous support during my study and his trust, patience, guidance and immense knowledge. Thank you also for introducing me to Prof. Dr. Brenda Schulman who has been an endless source of motivation and enthusiasm for me. I feel much honored that I could be part of your teams. I could not imagine having a better start to my scientific carrier.

I would like to thank the rest of my thesis committee, Prof. Dr. Ulrich Hartl, Dr. Franz Herzog, Dr. Henrik Daub and Prof. Dr. Karl-Peter Hopfner for their critical reading of my thesis and their comments and Prof. Dr. Chuna Ram Choudhary for being part of my TAC committee and all the scientific input.

Prof. Dr. Dario Alessi, Prof. Dr. Suzanne Pfeffer and their research groups, especially Dr. Esther Sammler and Dr. Francesca Tonelli, for being amazing collaboration partners. Thank you for your advices and sharing our dreams. Moreover, I want to acknowledge the financial support from the Michael J. Fox Foundation for the Parkinson's disease projects included in this thesis and Shalini Padmanabhan, Kalpana Merchant from the MJFF and Roy N. Alcalay from Columbia University for their support over the years. I also want to thank Dr. Mitchell J. Weiss, Dr. Peng Xu and Dr. Senthil Velan Bhoopalan for the great collaboration in the erythropoiesis project.

Martin Steger, Sean Humphrey and Sebastian Virreira Winter for being the best supervisors on this planet. I fell so lucky that I met you and worked with you. You have been great inspirations to me.

Igor Paron, Gaby Sowa and Alison Dalfovo for being great people and your endless help.

Florian Meier, Sophia Doll and Philipp Geyer for the best welcoming environment and your great advises. You made me love this group even before I had started working in here.

Matteo Pecoraro, Charo Robles, Natalie Krahmer and Francesca Sacco for your guidance.

Arno Alpi and Christine Langlois from the Schulman Department for all the discussions and scientific input.

Alberto and Isabell for all the bioinformatics help and the great RBC project collaboration. I hope many more will come in the future!

My office mates, André Michaelis and Jakob Bader for answering my millions of questions and listening to all of my complains.

Andreas Brunner, Catherine Vasilopoulou, Marta Murgia, Fabian Coscia, Michael Wierer and everyone in the Mann department for creating such a prosperous atmosphere of amazing science and helpful discussions.

Fynn Hansen and Maria Tanzer for just being there for me. I do not know what I would do without you two. The COVID project was just the beginning of a collaboration that will last for a lifetime😊

Last but definitely not least, my biggest thank goes to my wonderful family, Belis, Sevket and Pastam... You have made me stronger, better and more fulfilled than I could have ever imagined. You have taught me to work hard for the things that I aspire to achieve. There is no way to express my gratefulness to you, Onur. You have been a constant source of support and encouragement during the challenges of graduate school and life. I am truly thankful for having you in my life. Thank you for always being by my side darling!