DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES

DER FAKULTÄT FÜR CHEMIE UND PHARMAZIE

DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

The Quantitative Protein Interactome in Yeast and Human

Clemens André Michaelis

aus Mainz, Deutschland

2021

Erklärung

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

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 28.10.2021 Clemens André Michaelis

Dissertation eingereicht am 03.11.2021
1. Gutachter: Prof. Dr. Matthias Mann
2. Gutachter: Prof. Dr. Brenda Schulman
Mündliche Prüfung am 30.11.2021

Summary

Cellular function is closely tied to protein-protein interactions. Mapping these on a large scale, therefore, provides fundamental knowledge about the regulation and structure of biological systems. With the onset of proteomics, the use of affinity purification coupled to mass spectrometry (MS) has become the major tool to map protein interactions. Already twenty years ago, researchers endeavored to build interaction maps of model organisms such as yeast. However, previous large-scale interaction studies in *Saccharomyces cerevisiae* date back more than ten years, covered only about half of all genes, and made use of non-quantitative MS and tandem-affinity purification strategies. These approaches were limited by harsh purification protocols and required large amounts of cell lysate. Additionally large false positive and negative rates hampered their use as a fully reliable source for network studies.

Building on recent improvements in sensitivity and speed of MS technology and the introduction of the concept of 'affinity enrichment coupled to MS,' I developed a fast, robust, and highly reproducible workflow for proteome-wide interaction studies. I applied and optimized the approach for a first full screen in *S. cerevisiae*. The workflow starts from only a few hundred µg of proteins per pull-down and is performed entirely in 96-well format, including cell growth, lysis, and affinity enrichment of GFP-tagged proteins. To increase sample throughput and minimize MS idle time between injections, I turned to the high throughput Evosep One liquid chromatography system. This allowed me to obtain data on 60 baits per day. The system is coupled online to a timsTOF Pro mass spectrometer capable of fragmenting over 100 peptides per second using the parallel accumulation – serial fragmentation (PASEF) technology. This combination of miniaturization and standardization ensured high sample throughput, sensitivity, and robustness.

Altogether, I successfully performed over 4150 pull-downs and completed more than 8300 measurements for the yeast interactome using this next-generation workflow, all in less than 20 weeks of mass spectrometer running time. The dataset has a very high success rate for pull-downs. The near-complete coverage of expressed proteins in our study enabled a novel two-dimensional analysis strategy that efficiently scores interactions. We examined well-known protein complexes, which confirmed very high data quality. Although the yeast interactome has been studied by large-scale methods for decades, the majority of interactions were novel compared to known high-quality interaction databases. Among many striking novel discoveries - I found compelling evidence for interactions between the conserved chromatin remodeler SWI/SNF and SPX-domain-containing plasma-transporters. Using the common GFP-tag for quantification of protein abundance confirmed that our workflow covers a wide range of cellular protein abundances down to a few copies per cell. Redefining the yeast interactome with very high data quality and completeness enabled the study of its fundamental network properties that have been controversially discussed over many years. In total, our protein-protein interaction network encompasses about 4,000 proteins connected via about 30,000 interactions. A full browsable web application is accessible at yeast-interactome.org and allows (sub-) network exploration, interactor validation via volcano plots and correlation maps, and sample quality control.

In a collaboration with the CZ Biohub, we set out to implement the mass spectrometry pipeline developed here to an interaction screen with CRISPR GFP-tagged human HEK293T cells. The reduced sample amount allowed us to screen cell cultures grown in 12-well plates for high throughput. The interaction and localization results of 1,311 processed interactomes in biological triplicates can be accessed at opencell.czbiohub.org.

Table of Contents

Abbreviations

AD	activation domain
AP	affinity purification
BD	binding domain
CCS	collisional cross section
DNA	desoxyribonucleic acid
EI	electron ionization
ESI	electrospray ionization
FDR	false discovery rate
GFP	green fluorescent protein
HPLC	high performance liquid chromatography
IEX	ion-exchange chromatography
IMS	ion-mobility spectrometry
IP	immunoprecipitation
LC	liquid chromatography
MALDI	matrix-assisted laser desorption/ionization
mRNA	messenger ribonucleic acid
MS	mass spectrometry
PASEF	parallel accumulation – serial fragmentation
PD	pull-down
SEC	size-exclusion chromatography
TOF	time-of-flight
Y2H	yeast two-hybrid

1 Introduction

1.1 Biological Interaction Networks

“The concept of randomness and coincidence will be obsolete
when people can finally define a formulation of patterned interaction
between all things within the universe.”

-Toba Beta

1.1.1 Interactions Determine Function, Efficiency, and Health

Interactions are fundamental for a tremendous number of known systems. The entirety of all objects in a system and the links that exist between them is called a network. The character and efficiency of a network is defined by its structure and, therefore, by the way the connections are organized. From telecommunication wires around the globe, social networks that depict relationships, the links that connect webpages, to the dynamics underlying global epidemics, knowing their structure helps to understand them. In these cases, they transmit emails or phone calls most efficiently between sender and receiver, help to understand how information and rumors are transmitted, helps the Google search engine algorithm to identify webpages most suited to an inquiry, and are crucial for the identification of transmission routes of a virus in order to prevent further spread, respectively.

Networks also exist on a physical micro scale. Besides technological, social, and informational networks, biochemical networks – an example from the biological world - are among the most important ones. Despite their microscopic nature, biochemical networks do not lack in complexity. The best-studied ones are metabolic, genetic, and protein-protein interaction networks. Metabolic networks describe the biochemical pathways in a cell, whereby chemical compounds are connected by chemical reactions that convert a substrate into a product. For instance, they provide the information of how cells break down nutrition, and how they rebuild and convert cellular building blocks. Genetic regulatory networks capture the dependencies of genes on the level of transcriptional regulation *(*1*)*. Protein-protein interactions are binding events between two or more proteins that accrue in all cells or organisms in large numbers.

These interactions can last longer in stable formations known as protein complexes or can be of short duration which are termed transient interactions. Protein complexes can be seen as a higher order of protein organization. Different protein "building blocks" come together to form larger molecular machines or structural elements that are too complex to be formed by a single protein. Examples of transient interactions include proteins that biochemically modify each other in order to transmit a cellular signal in response to an external stimulus that requires cellular adaption. These types of modifications can alter their activity, cellular or tissue location, induce or inhibit its degradation, or ultimately change their own interaction pattern. Knowing on a global level how proteins interact within a cell is key for understanding how living organisms function. Building a systematic map of networks therefore helps to answer questions in the case of cellular malfunction as to their potential origin and it can help to assign functions to unknown parts. In the context of protein-protein interactions this translates into finding the cause for diseases and into describing functions for uncharacterized proteins by their association with characterized proteins - a phenomenon termed "guilt by association" *(2)*. Furthermore, only if one knows the blueprint of a system, one can repair it or use it to build something new.

The cell is regulated on several levels, all of them contributing to its phenotype to a different extent. The following chapter describes in more detail the roles and dependencies of these regulations and argues why the study of proteins and their interactions is one of the best available read-outs in systems biology.

1.1.2 The Three Cellular Fundaments of Protein-Protein Interactions

The central dogma of molecular biology depicts the flow of information in almost all cellular systems as two main steps: The first one is transcription and generates a transient copy of the DNA. The emerging molecule from this step, namely mRNA, serves as a template for the second process called translation, which uses the stored information for the assembly of amino acids into proteins. Although the term dogma and the concept of the directed flow of information have been put into perspective *(3)*, it adequately illustrates three central fields in cell biology and medicine: The study of genomes, transcriptomes, and proteomes. With the common -omics suffix that indicates the study of the term's entirety in a particular system like a cell, tissue, organ, or organism, they are called genomics, transcriptomics, and proteomics, respectively. In the last decades, these areas were in the focus of many researchers and kept expanding as technology evolved. The milestone achievement of human genome sequencing *(*4, 5*)* laid the basis for analyzing transcripts and proteomes in a large-scale manner. Due to faster and much cheaper technology, many more organisms followed, providing the sequence information that is an essential precondition for proteomics. The static nature of the genome restricts the information that one can draw from it to make conclusions on the dynamic state of the system. The discrepancy between genotype and phenotype is due to the fact that only some of the genes are actively transcribed at any point in time. Regulation at the translational level is also precise in time and space and restricts the presence of transcripts to a particular phenotype. Proteomics is special in that it deals with the final product of gene expression, thereby overcoming some of the limitations of transcriptomics. It focuses directly on detecting and quantifying the presence of the main functional units in living organisms: proteins. These are the major actors in cellular processes and their direct study more provides the additional information on top of genomics and transcriptomics.

Another level of information that goes even beyond the simple presence of proteins in a system in a certain condition is to study their interactions with another. Many proteins in a cell function in complexes or they fulfill their tasks by interacting with other proteins. This could be due to specific transportation, modification, or degradation of other proteins or for the purpose of building structural units within a cell *(*6*)*. Often there is an architectural reason for proteins to interact with each other, namely complexity. With large and highly sophisticated molecular machines that undergo huge conformational changes - for example in order to catalyze a biochemical reaction – it is necessary to assemble distinct building blocks into a single unit. Gaining access on the powerful information of protein-protein interactions therefore allows the global study on a regulatory, functional, and structural level. The study of interactions, is known as interactomics (**Figure 1**), and it can be achieved by several techniques as described in the next section. One powerful technology involves the use of mass spectrometry. It is identical to the expression proteomics approach except that it uses an additional enrichment step beforehand. While each "-omics" era builds on the knowledge of the previous ones, the presence of proteins and the interactions between them most directly reflects the cellular phenotype.

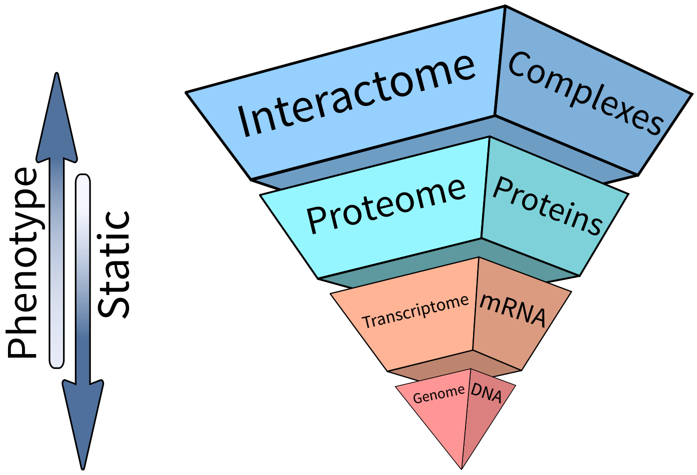


Figure 1. Increasing complexity of the fundaments of protein-protein interactions. The interactome represents an additional layer on top of the proteome that shapes the cellular phenotype.

1.1.3 Identification of Protein-Protein Interactions

A multitude of techniques to study protein-protein interactions have been introduced over the last decades. Label-free techniques include surface plasmon resonance (SPR) spectroscopy, micro-scale thermophoresis (MST), isothermal titration calorimetry (ITC), circular dichroism (CD) spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy. These techniques provide detailed information on the interaction itself, e.g. defining binding constants or the exact site and mode of the interaction. However, they are unsuitable for unbiased and large-scale interaction screens because they require *a priori* knowledge of all potential interacting proteins.

Cell-based bimolecular interaction reporter assays include bioluminescence resonance energy transfer (BRET), the yeast two-hybrid (Y2H) screen, and related split-protein methods like the split-ubiquitin assay *(7)*. The Y2H screen is restricted to the detection of mostly binary interactions: Two potentially interacting proteins are each fused to either the binding domain (BD) or the activation domain (AD) of the transcription factor Gal4. Expressed in yeast, the interaction of both candidate proteins activates Gal4 by bringing together AD and BD domain. Gal4 leads to the transcription of a reporter gene whose read-out corresponds to the interaction of the candidate proteins (**Figure 2B**). In unbiased Y2H interaction screens, large libraries of all potential protein pairs fused to AD and BD need to be generated. One of the limitations of Y2H screens is that interactions can only be detected for soluble proteins that bind each other within the nucleus. In order to detect membrane protein interaction a different assay/tagging of strains - like the split-ubiquitin assay - needs to be deployed *(*8*)*. Y2H is an approach that allows large screens and they have been performed for many organisms like *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and humans *(*9*)*. However, it is only limited to binary interaction studies and prone to false positives and negatives.

There are other assays that are suited for large-scale screens as well, and in contrast to Y2H allow detection of more than one (indirect) interaction in a single experiment. All the following assays have in common that mass spectrometry is used as the final detection method for the identification of interacting proteins. Co-fractionation via gel-filtration - also known as size-exclusion chromatography (SEC) *(7)* or ion-exchange chromatography (IEX) coupled to mass spectrometry - detect proteins that co-elute into different collected fractions based on the size (SEC) or charge (IEX) of the complexes (**Figure 2D**). Proteins in each fraction are identified and quantified via MS. Interactions can be scored based on similar elution behavior using correlation analysis of protein profiles. A main advantage is that there is no need for genetic engineering *(*10*)*, but a current downside is its limited resolution due to the broad elution peak profile over a limited range of typically around 50 fractions *(*11*)*.

Cross-linking coupled to MS uses a chemical linker that covalently bridges proteins that are in close proximity. Cross-linked peptides are then identified via MS and protein-protein interactions inferred from that information (**Figure 2E**). Studies on full proteomes have been conducted in *Escherichia coli* and HeLa cell lysates *(*12*)*.

The labeling of proteins that are in close proximity to a protein of interest is a rather recent and promising development. Proximity labeling uses enzyme fusion proteins in which the protein of interest is linked to either a peroxidase named APEX *(*13*)*, or a biotin-ligase known as BioID *(*14*)* and its enhanced version called TurboID *(*15*)*. These enzymes catalyze the biotinylation reaction of proteins in close proximity, that are not necessarily physically interacting (**Figure 2C**). Biotinylated proteins are purified using streptavidin beads and identified via MS *(*16*)*.

The most widely used technology is affinity purification (AP) coupled to MS. AP comes in two main flavors: Immunoprecipitation (IP) and pull-downs (PDs). IP uses immobilized antibodies to capture a specific protein via its antigen-binding site from a cell lysate or any other amenable biological sample. After washing off the unspecific proteins, the purified protein remains. If the purified protein of interest - termed “bait” - is involved in protein-protein interactions, those interacting proteins – termed “preys” – can be detected in the downstream MS analysis as well

(**Figure 2A**). This co-enrichment of interactors is therefore called a co-IP. Since a large-scale screen would require for each IP to generate a corresponding antibody, one can generate tagged proteins (“baits”) that all have the same binding properties to the same antibody. This tag can be a specific sequence of amino acids that is genetically fused to the C- or N-terminus of a bait, or it can be an intact protein like the green fluorescent protein (GFP). In either case the peptides or proteins that are used as tags have a well-characterized stability and antigen properties for available antibodies. This strategy enables the use of the same, generic immobilized antibody for separately executed experiments. In comparison to IPs, pull-downs similarly use an immobilized affinity matrix to capture bait proteins with the only difference being that they do not use the immune-system-derived antibodies. Examples are Ni2+ embedded matrices that enrich His-tagged proteins or immobilized streptavidin that enriches biotinylated proteins *(*17, 18*)*.

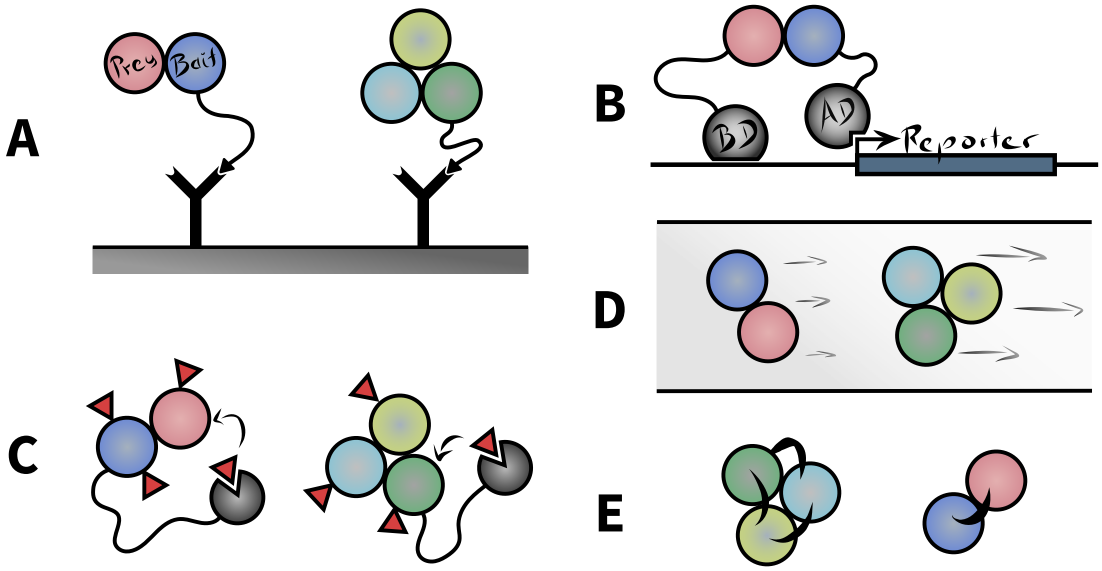


Figure 2. Different methods for studying protein-protein interactions. (A) Affinity purification coupled to mass spectrometry. **(B)** Yeast two-hybrid screening. **(C)** Proximity labelling (APEX, BioID/TurboID). **(D)** Co-fractionation coupled to mass spectrometry (SEC/IEX-MS). **(E)** Cross-linking coupled to mass spectrometry.

In the studies described in this thesis, an endogenous GFP-tagged library in *S. cerevisia*e and human HEK293T cells were used. An additional advantage of GFP tagging is that it can be used for cellular localization screens. While a GFP-library for yeast has been generated and used for a global protein localization study already *(*19*)*, it had not been used for global interaction screen yet. CRISPR-editing nowadays allows the generation of similar endogenous tagged libraries in many systems. The length-limitation of CRISPR-tagging for GFP can be circumvented elegantly by using only β-strand 11 of GFP as the tag. By expressing the remaining part of GFP (β-strand 1-10) in cells the complete tag is reconstituted as a fully functional protein *(*20*)*. This split-GFP strategy enables large-scale CRISPR-based screens that allow fluorescence microscopy localization as well as AP-MS interaction detection studies with the same cell line.

1.1.4 Protein-Protein Interaction Networks

Protein-protein interaction networks are the sum of all known interactions between proteins, or a detected subset of these. They resemble a map of all proteins and their interactions. In general, a network representation consists of nodes that are connected via edges that can be symbolized as circles and lines, respectively. In protein-protein interactions networks, proteins are represented as nodes and interactions as edges (**Figure 3**). The number of neighbors each node has is the "node-degree". Interactions in a network can be – dependent on the underlying data - directed or un-directed. The former is usually depicted by an arrowhead that indicates the direction. Directed interactions can for instance be citation networks or dependency networks of programming packages that always point to the original source, thereby maintaining an important piece of information. When two proteins interact, both participate in an equal way. That is why from a graph theory point of view those networks should be treated as un-directed. Nevertheless, an edge can still be used to visualize further information, for instance for the direction in which an experiment was conducted. In the context of AP-MS, an arrow can indicate which of the protein is the bait and which the prey (pointing from bait to prey).

An important finding in network science revealed that most known complex networks have a characteristic of higher-ordered structure that differentiates them from random networks: their node-degree distribution follows a power-law. In simpler terms, these networks have many nodes with a few connections and few nodes with a large number of neighbors. Such networks are called "scale-free" and differ from random networks in which the node-degree is Poisson distributed. "Scale-freeness" in networks is based on (i) the continuous expansion of the network by adding new edges and (ii) the preferential attachment of edges to nodes that are already highly connected *(*21*)*. Due to gene duplication events during evolution, it is thought that protein interaction networks evolved in a similar preferential attachment mode and that protein-protein interaction might follow a power-law. The attributes of scale-free networks explain some interesting features of complex networks, for instance, the “small-world” effect in which two nodes can be reached via only a few edges. These routes are called “shortest paths” and usually pass through highly connected nodes, called “hubs”. Additionally, such networks are robust against random removal of nodes, since the chances of removing a less important one is high. On the other hand, the targeted removal of central hubs, can have dramatic effects on the function of the network *(*22*)*. While previous studies have suggested scale-free properties for protein-protein interaction networks, there seems to be doubt about the quality of the underlying data *(*23*)*.

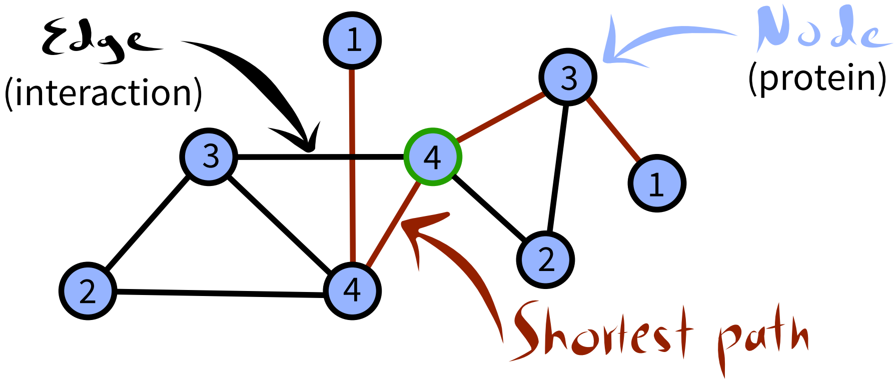


Figure 3. Small network representation: Depicted are eight nodes and ten edges representing proteins connected via detected protein-protein interactions. The numbers indicate the “node-degree” which is equal to the number of its neighbors. The “shortest path” between both nodes with a node-degree of 1 is 4 steps (highlighted in red). Central nodes through which many shortest paths pass have a high “betweenness-centrality” or are called hubs (green).

1.1.5 From non-Quantitative to Quantitative Interaction Screens

The final chapter of the ‘Nature Milestone’ series on mass spectrometry lists the field of interactomics as its latest achievement in the application category *(*24*)*. Indeed, the breakthrough developments in protein ionization and peptide mass fingerprinting *(*25, 26*)* opened the opportunity for large-scale applications in the field of AP-MS. Here I give a short overview of previous large-scale interaction screens in *S. cerevisiae*. This will highlight their remarkable achievements as well as their limitations and will reason why quantitative proteomics can generate interaction data of superior quality compared to previous non-quantitative approaches.

Due to the need of endogenous tagging to establish near-physiological conditions, yeast is an ideal candidate for systems-wide interaction studies. Its natural system of homologous recombination allows the rapid and efficient introduction of tags at specific loci. Almost two decades ago the first two initial AP-MS screens were conducted in yeast *(*27*,* 28*)* that were then followed by two larger-scale studies four years later *(*29*,* 30*)*. The underlying assumption in non-quantitative AP-MS is that all co-purified proteins are specific interactors. Usually, AP samples were separated on gels and sliced bands used for MS identification. This assumption of all co-purified and detected proteins being specific was soon realized to be false. The presence of unspecific binding proteins or contaminants was reduced by the use of tandem affinity purification (TAP) tags, as these allowed more stringent washing in a dual purification step that includes partial tag cleavage *(*31*)*. While those strategies reduced unspecific binding, more stringent washing also caused loss of weaker interactors and needed larger input materials. Generally, the mentioned interaction screens required around 4L of cell culture per pull-down. Altogether this required the processing of about 10 g of yeast pellets per pull-down, involving grinding with dry ice in a coffee grinder *(*29*,* 32*)*. Even then it was necessary to manually remove proteins that commonly appeared in different purifications as unspecific background binders, potentially introducing biases. For example, *Gavin et al.* manually removed dozens of preys and almost all ribosomal subunits *(*33*)*. A database named the "CRAPome" was generated to help exclude those false positives from AP-MS data *(*34*)*. While these milestone studies enabled the understanding of many cellular functions, their limitations clearly reduced data quality *(*33*)*. This is also reflected in the large discrepancies between the two yeast AP-MS interaction datasets that only overlap in 13% of their reported interactions, although they used similar approaches. To overcome this drawback, *Collins et al.* reanalyzed the raw data sets from these two main interaction studies to build a single consensus interactome *(*35*)*. While the resulting data is of higher quality and shaped the interactome landscape it came with the trade-off: size. The combined dataset encompasses about 1,600 proteins, only about one third of the expressed yeast proteome *(*36*)* and much less than the two original studies had reported, leaving the yeast interactome far from complete. Even ongoing studies of the human interactome use non-quantitative approaches, although their unspecific binder correction became more sophisticated *(*37–39*)*.

While the origin of quantitative proteomics dates back to the beginning of this century *(*40*,* 41*)*, it is the recent developments of label-free quantification and normalization methods *(*42*)*, novel approaches of how to group bait samples into a single control group for efficient background identification and concepts of how to use correlation and abundance information that finally allowed scoring for interactions in quantitative acquired MS data (43–45). The basic principle is that the high sequencing speed and sensitivity of mass spectrometers are used to identify and precisely quantify not only a few co-purified proteins, but also "background binders" to a much larger degree (43). The number of detected background proteins can thereby exceed thousands of proteins in a single PD, while only a few specific proteins are present. This is enabled by the precise quantification that allows detection of subtle enrichments of specific proteins in comparison to control samples. The yeast interactome study presented in this thesis, likewise builds on the large number of quantified background binders across all samples' constant background, by applying only very gently washing steps. Those steps do not use mixing, but rather dilute proteins that do not stick on the mobile phase, allowing precise normalization and quantitative interactomics. Together with the highly efficient mass spectrometric read out described next, this forms the basis of a very high-quality interactome.

1.2 Mass Spectrometry-Based Proteomics

"The difficulties which would have to be overcome to make several of the preceding experiments conclusive are so great as to be **almost** insurmountable."

-J.J. Thomson

1.2.1 A Century of Innovations in Mass Spectrometry

Based on the discovery of Wilhelm Wien in 1898 that beams of charged particles could be deflected by a magnetic field *(*46*)*, Joseph John Thomson constructed the first instrument capable of acquiring a mass spectrum in the early 20th century *(*47*)*. Thomson, who became known as the father of mass spectrometry, built the parabola spectrograph that applied magnetic and electric fields to deflect gaseous ions based on their charge and mass. His observations on the properties of the electron were rewarded with the Nobel Prize of physics in 1906. Thomson's work led to the discovery of atoms and isotopes, and his apparatus laid the basis for the field of mass spectrometry *(*48–51*)*. In the following century, three more Nobel Prizes were awarded for groundbreaking work in the field of mass spectrometry. In 1922, Thomson's former research assistant Francis William Aston who further improved the instrumental setup was recognized for his discovery of isotopes in a large number of non-radioactive elements *(*52–58*)*. In 1989, Wolfgang Paul and Hans G. Dehmelt shared the Nobel Prize for the development of the ion trap technology *(*59*)*. Paul's quadrupole and Dehmelt's magnetron are also known as the Paul and Penning traps, respectively, and evolved versions of either device made their way into most commercial mass spectrometers available today. Transferring an analyte into the mass spectrometer, controlling its movement within the device, and allowing its separation based on the mass-to-charge ratio by applying magnetic or electric fields, requires that the otherwise neutral molecules have to be ionized beforehand. The 'gold standard' ionization method in the first half of the last century that replaced the initial gas discharge experiments was electron ionization (EI) also known as electron impact or bombardment ionization. In EI, an electron stream is generated and focused with magnets onto the analyte for its ionization. Although variations of EI as the field ionization (FI), the field desorption (FD), or the chemical ionization (CI) provided a 'softer' alternative to ionize small organic molecules, they were still too harsh and destructive for large biomolecules *(*26*)*. The breakthrough discovery for proteomics application was the development of ionization methods that are compatible to larger biomolecules like peptides or intact proteins. In 2002 John B. Fenn and Koichi Tanaka shared the Nobel Prize in chemistry for their contributions on the development of electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), respectively (60). With MALDI, the analyte is embedded in a protective matrix that absorbs the energy of a pulsed laser beam that is used to transfer the analyte into the gas phase. With ESI the liquid analyte is guided through a needle to which a high voltage is applied. The development of nano-ESI with flow rates in the low nL/min range makes use of an efficient dispersion of the liquid and causes a dramatic increase in sensitivity (61–64). Both, MALDI and ESI are standard ionization technologies in mass spectrometry for the analysis of larger biomolecules today. The major advantage of ESI over MALDI however is the 'online' use of a liquid chromatography (LC) upfront of the ionization process, making it the preferred choice for reproducible analysis of complex samples in proteomics.

The first century in the field of mass spectrometry was a fascinating one accompanied with great inventions that began to enable its application in medicine, quality control, forensics, food chemistry, biochemistry, and in many other areas of life science (50). Nevertheless, it is only in recent years that these promising developments have come to fruition. Nowadays, the increase in sensitivity of mass spectrometers allows unprecedented depth and analysis of samples of only a few cells and even of a single cell soon (65, 66). At the same time, many scientists, as well as established and newly founded companies, focus on developing solutions to improve up- and downstream processes in mass spectrometry. This includes efficient sample preparation (67), innovative columns and liquid chromatography systems (68, 69), novel data acquisition modes and analysis tools (70–73), as well as next-generation mass spectrometers that outdo one another in terms of sensitivity and resolution (65, 74).

The following sections will give an overview on recent technological developments that were pivotal for this thesis and which are on the brink of becoming standards for high-throughput applications in science, medicine and industry.

1.2.2 Bottom-up Proteomics

In "bottom-up" proteomics proteins are first extracted, denatured, and digested by sequence-specific proteases into peptides. Following the enzymatic cleavage, peptides are separated via liquid chromatography and their masses are analyzed in the mass spectrometer. In order to obtain sufficient information on the peptide sequence, peptides are fragmented inside the mass spectrometer and the resulting fragment masses are obtained as well. Proteins are identified by comparing peptide sequences to an in-silico digested reference database. The "bottom-up" approach is frequently used since it is very powerful due to the ease of handling peptides and the superior analysis possibilities of peptides compared to intact proteins. Intact proteins are used in the counter-part approach named "top-down", in which intact proteins are analyzed without a prior digestion step. As illustrated in **Figure 4**, a classical bottom-up MS-based proteomics workflow can be divided into sample preparation (A), LC-MS/MS analysis (B), and data analysis (C) *(75)*. The steps are described in more detail in the next sections.

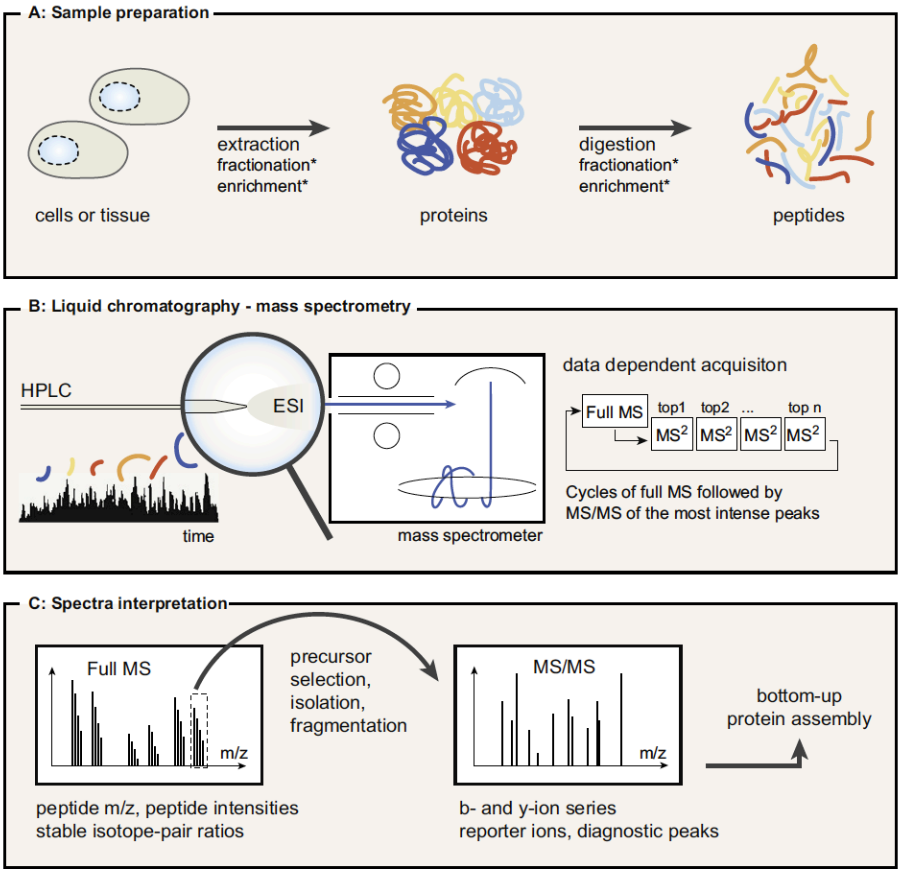


Figure 4. Bottom-up proteomics workflow. Classical steps in proteomics for: **(A)** Sample preparation for the extraction of proteins from cells or any other amenable biological sample, followed by enzymatic digestion. **(B)** High performance liquid chromatography (HPLC)-based separation of peptides and ionization

via electrospray ionization (ESI) followed mass spectrometric analysis (here exemplary shown for an orbitrap analyzer). The mass-to-charge-ratio in data dependent acquisition modes is detected for intact co-eluting peptides (precursors, full-MS) followed by detection of the most abundant fragmented peptides (MS^2). **(C)** The acquired mass spectra are used for database search containing the sequence of all potential proteins of the sample. Figure by *Hein et al. (*75*)*.

1.2.3 Sample preparation methods

The accurate and reliable identification of several thousands of proteins – which in "bottom-up" proteomics is inferred from peptide information – requires a specific sample preparation procedure. The preparation step is crucial in proteomics and the execution is dependent on the sample type. The main goal is to efficiently extract and isolate proteins from a sample of interest without inducing unspecific proteolysis. This sample can, in theory, be anything that contains proteins. In proteomics, primarily biological samples are of interest. These can be tissues, body fluids like plasma, cells from culture, parts of plants, yeast cells, bacteria, or other organisms. The basic steps include sample homogenization, cell lysis and extraction of proteins, protein denaturation, reduction of disulfide bonds, cysteine alkylation, proteolysis, and sample cleaning for complete removal of contaminants like salts or detergents before LC-MS analysis.

For biological samples, homogenization and cell lysis can be achieved by mechanical disruption such as cryogenic grinding or bead-beating. Alternatives are sonication, heating, or the use of chemicals. Additionally, different agents are used to "deactivate" the sample by denaturing all proteins and therefore inhibit all enzymatic activity that could potentially alter the proteome, such as unwanted modifications or unspecific proteolysis. These include detergents like SDS (sodium dodecyl sulfate) and SDC (sodium deoxycholate), organic solvents like ACN (acetonitrile), or chaotropic agents like urea, thiourea, and guanidinium chloride *(*76*,* 77*)*.

In the next step, stable disulfide bonds are disrupted by using reducing agents such as TCEP (tris(2-carboxyethyl)phosphine) or DTT (dithiothreitol). The reformation of disulfide bridges is prevented by the alkylation of cysteines, typically using agents such as IAA (iodoacetamide) or CAA (chloroacetamide). Due to the identical masses of an alkylation side product with the ubiquitin diglycine adduct when using IAA, CAA is preferable in some cases *(*78*)*.

Different sequence-specific enzymes can be used for digestion. Most frequently used is the combination of the enzymes trypsin and LysC. LysC cleaves specifically C-terminally to lysine, while trypsin cleaves C-terminally to lysine and arginine. The sequence specificity is vital for generating peptides with a certain average length while restricting the database search in the last step to peptides with an already known C-terminal amino acid. Other enzymes that can be used include Asp-N, Lys-N, Arg-C, or GluC *(*79*)*.

Before digested peptides are analyzed by mass spectrometry, they need to be purified. This process includes removing all potential damaging agents for the LC, column, or mass spectrometer from the sample. Removed are detergents, salts, chaotropic agents, or other aggregates that might clog, contaminate, or interfere with the LC-MS pipeline. Improper clean-up can also suppress analyte ionization and impurities can deposit on hardware components of the mass spectrometer, thereby decreasing performance or damaging the instrument. The cleaner the sample, the longer the high-performant instrument run time and the more reproducible and reliable the data gets. A cornerstone in this area was the development of a peptide-tip-based purification technique named Stop and Go Extraction tips, short "StageTips". It consists of small discs of retention material inserted into pipet tips that serve as a sample clean-up reservoir. This procedure is easily applicable and has become a standard in proteomics for sample purification and concentration *(*67*,* 80*,* 81*)*.

Peptide sample complexity can be decreased before LC-MS analysis by using fractionation. Fractionation is the separation of a single sample into several less complex peptide mixtures. Measuring several fractions instead of a single sample allows one to analyze more input material and spend more MS time on it. This procedure increases the depth and allows the detection of less abundant peptides. A crucial thing to consider is that fractionation is not just splitting the sample into different vessels but instead uses a chromatographic separation. The separation method should differ from the one that is used later in the LC-MS setup. This orthogonal separation can, for example, be an off-line high pH reverse-phase LC. The eluted fractions are then used for the on-line LC-MS analysis, which usually uses a low pH reverse-phase separation *(*82*,* 83*)*. Our group has developed a 'loss-less nano-spider' fractionator, which automatically concatenates the collected fractions via a rotating valve. This fractionator enables the quantification of around 12,000 proteins from very low-µg starting peptide material *(*84*)*.

1.2.4 High-Throughput Liquid Chromatography

Subsequent to the enrichment and purification steps in bottom-up proteomics, different peptide species are separated from another on a liquid chromatography system which is coupled "online” to a mass spectrometer. In this LC-MS setup, the separation step takes place on the LC column, which is filled with a solid, hydrophobic material (typically octadecylsilane, C18). Peptides are separated based on physiochemical properties - mainly their hydrophobicity - and the separation is caused by the differences in interactions with the moving liquid and stationary solid phase. While the properties of the solid phase remain constant, the hydrophobicity of the mobile phase increased during each run. This mobile phase linear gradient controls the elution of the more hydrophobic peptides from the reversed-phase column. This causes peptides of the same species to co-elute from the column in narrow packages within the range of seconds with bell-shaped like intensities, called chromatographic peaks. This LC-MS setup allows to decrease the sample complexity in a time dimension, by submitting co-eluting peptides of the same species consecutively to the mass spectrometer. At the end of the column peptides are transferred to the gas phase by electrospray ionization (ESI). This happens at the entrance of the mass spectrometer where ionized peptides are then transferred into the vacuum *(*25*,* 85*)*. Ionization of peptides can be aided by the presence of protons from formic acid in the solution. Two factors drastically influence ionization efficiency: the flow rate and the droplet size forming at the site of ESI. Both can be reduced by using a slow stream of liquid and by using (usually long) columns with a small diameter *(*63*)*. While this setup causes an excellent peak separation, it requires high-pressure pumps in order to provide a constant flow through the column. These high pressures can cause pump-, valve and column failures. Another downside is that the long columns work preferentially with long gradients, since sample loading, passing of sample through the column, and washing off the column takes rather long and causes large gap times between runs.

To overcome these expensive idle times of the mass spectrometer and to enable high-throughput usage with short gradient runs, a novel concept for liquid chromatography has been developed: The Evosep One. This LC uses mainly low-pressure pumps, runs with short columns, utilizes very short gradients, and drastically reduces gap times to a minimum. This setup promises to be a robust LC for high-throughput projects that need short gradients *(*69*)*. The first difference to conventional LC systems is the direct elution of peptides from the C18 material that is embedded at the bottom of a pipette tip (“StageTip”, see 1.2.3). While upstream peptide enrichment and washing steps remain similar, peptides are not manually eluted *(*67*,* 80*)*, but the loaded tips are placed inside a box and put onto the LC. The availability of a commercial and standardized version known as “Evotips”, helps to reduce handling variability and increases reproducibility. Two low-pressure pumps (A, B, **Figure 5**) elute the peptide sample from the C18 material. This has two major advantages: The partial elution keeps impurities on the disposable Evotips and the gradual elution into the storage loop allows for the peptides to form of pre-gradient. The low-pressure pumps C and D generate an off-set gradient, by diluting down the organic component of the liquid phase. This off-set gradient reduces the interaction of the peptides with mobile phase and down-stream allows a better interaction with the solid phase on the column which contains C18 as well.

Once the sample is on the loop, it is pushed quickly onto the column with a single high-pressure pump (H, **Figure 5**). The moment the sample has left the loop, the pressure and flowrate are reduced and data acquisition starts. Meanwhile the loop is washed and loaded with the next sample. Due to this, on a 21 min gradient the overhead time is only 3 min allowing to process 60 samples a day with this setting. Other gradients reach from 3 to 44 min allowing 300 to 30 samples per day *(69)*. The above stated principles of pre-gradient and off-gradient formation, compensate for the above-mentioned disadvantages of this high-flow system for proteomics. At the same time, its robustness and high-throughput capabilities make it a perfect use-case for the here presented interaction studies that need the reliable processing of many thousands of samples.

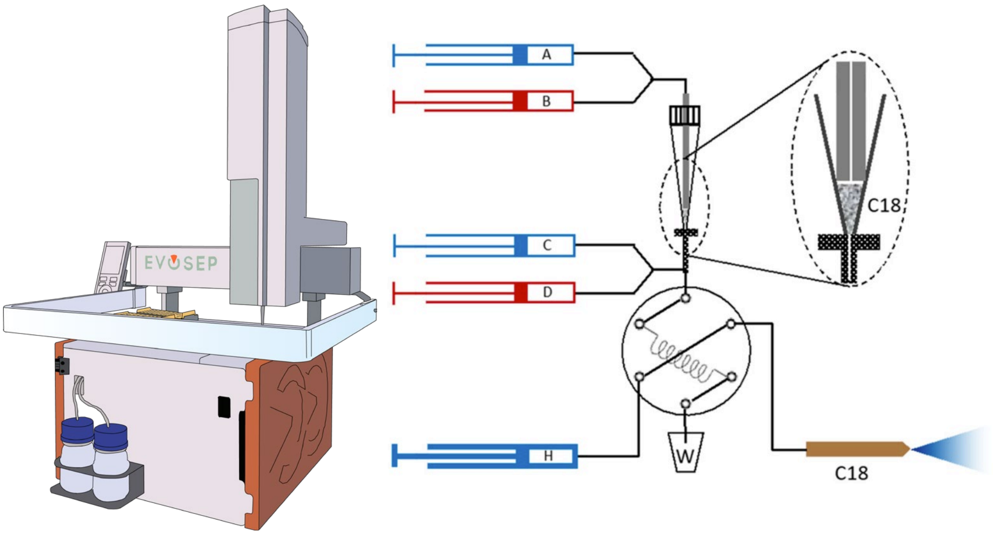


Figure 5. Evosep liquid chromatography system. (Left) Evosep LC system device. **(Right)** Schematic representation depicts the use of 4 low-maintenance low-pressure pumps (A-D) that allow direct sample elution from the C18-material packed pipette tip ("Evotip") via a gradient formed by pump A+B. An off-set gradient is formed by the pumps C+D that allows sharper peptide peaks/separation on LC column with the same C18 material. A single high pressure pump H pushes the pre-formed gradient from the storage loop (lower central circle) onto the column. This allows mass spectrometric data acquisition to proceed while the loop is washed and filled with the next sample. This strongly reduces gap times between runs in which the mass spectrometer would be idle. Figure by *Bache et al. (69)*.

1.2.5 The Mass Spectrometer

Central to the proteomics workflow is the mass spectrometer, a device that detects and quantifies the masses – more precisely the mass-to-charge-ratio (m/z) - of the analyte. The mass spectrometer is composed of three main parts: the ion source, the mass analyzer, and a detector. To avoid collisions of the analyte with gas molecules and to avoid interferences, the mass spectrometer operates under an ultra-high vacuum (down to about 10^{-9} mbar) *(*50*)*. The most frequently used high-performance instruments in proteomics in the last two decades were Orbitrap platform mass spectrometers, that replaced the much slower and impracticable Fourier transform ion cyclotron resonance (FT ICR) analyzer, the time-of-flight (TOF) instruments that had ion transmission deficiencies or ion traps with low mass accuracy. Based on a commonly used proteomics data submission website, 80 % of used machines are now Orbitraps and 8 % TOF based mass spectrometers (accessed on 2021/10/26, excluding other instruments: proteomecentral.proteomexchange.org). Since its first presentation over 20 years ago and the first launch of the LTQ Orbitrap in 2005, Orbitraps quickly became the prevalent instrument type *(*50*,* 86*)*. Alexander Markov, the chief instrumentalist working on the Orbitrap, soon developed a combined ion trap and analyzer which is based on the Kingdon and Knight ion traps, that have their origins in 1923 and 1981, respectively. The problem of capturing stable ions in the Orbitrap analyzer was solved by a principle termed “electrodynamic squeezing”, in which the central electrode potential is increased the moment ions are injected axial to the Orbitrap. As an external pulsed ion source, Makarov designed a bend quadrupole – known as “C-trap”.

An Orbitrap, as the name suggests, measures the mass-to-charge ratio of the analyte by detecting the induced current of the axial oscillating ions along the spindle pole *(*87*)*.

TOF instruments measure the time an analyte needs to pass a defined drift path distance in the vacuum until it hits the detector. This time-of-flight is dependent on its mass-to-charge ratio, causing ions with a smaller m/z to arrive earlier at the detector. A prerequisite for the time-of flight measurement is that the acceleration of each ion set takes place in a precisely defined short time frame. This controlled acceleration was enabled by pulsed ionization methods in the late 1980s that made the combination of MALDI with TOF instruments a perfect match for larger biomolecules. In this setting a continuous acceleration of ions is prevented by controlling the ionization and gas phase transfer with a pulsed laser beam instead *(*86*)*.

One of the major advantages of Orbitrap over MALDI-TOF instruments for proteomics was their capability of operating downstream of HPLC (high performance liquid chromatography) devices. This on-line setup as described in 1.2.4 allows to perform analysis of complex proteomic samples therefore increasing signal to noise ratio. Despite the much higher speed of TOF instruments, in which single spectra can be acquired in less than a ms *(*50*)*, the discussed reasons made Orbitraps the preferred choice.

Recently, two major developments helped to shift momentum back to TOF mass spectrometer: The improvement of orthogonal accelerators and the implementation of ion-mobility spectrometry (IMS). Orthogonal accelerators allow the use of non-natively pulsed ionization methods like ESI with TOF, rendering them LC- compatible. IMS on the other hand can be added to TOF instruments as an additional dimension of separation uncovering new possibilities of speed and sensitivity. This type of instrument and the modes of operation that enable its efficient use is described in the next section.

1.2.6 Trapped-Ion-Mobility Coupled Time-of-Flight Mass Spectrometry

A trapped ion mobility spectrometer (TIMS) separates ions in the gas phase based on their ion mobility. The ion mobility itself is dependent on the ion-neutral collisional cross section (CCS) and the charge of the molecule. In a TIMS device ions are dragged along a constant flow of a gas (e.g. nitrogen from ambient air) and are pushed back by an opposing constant electrical field until both forces reach an equilibrium that keeps the analyte in a fixed position. The dragging force is caused by the impact of colliding gas molecules onto the analyte and is dependent on the molecule's average accessible cross section: the CCS. The counteracting force is dependent on the charge of the molecule. The TIMS device is a development from *Melvin Park* and colleagues from Bruker Daltonics and is inspired by a conventional drift tube in which the analyte is moving and colliding with a resting gas. By using a gas stream instead, the TIMS device shrinks in size down to centimeters compared to meters in length for a drift tube *(*88*)*. The TIMS device traps ions – separated by their ion mobility – and by decreasing the electric field releases them in packages into the mass spectrometer. An updated version – the dual TIMS analyzer (**Figure 6B**) - separates the funnel in three parts: A trapping unit in which arriving ions from the source are accumulated, a transfer region and a second unit that separates and gradually releases the ions by ramping down the electric field. Ions from unit one are transferred to the second unit and the cycle begins anew. This parallel accumulation enables an up to 100% duty cycle of the ions *(*89*)*.

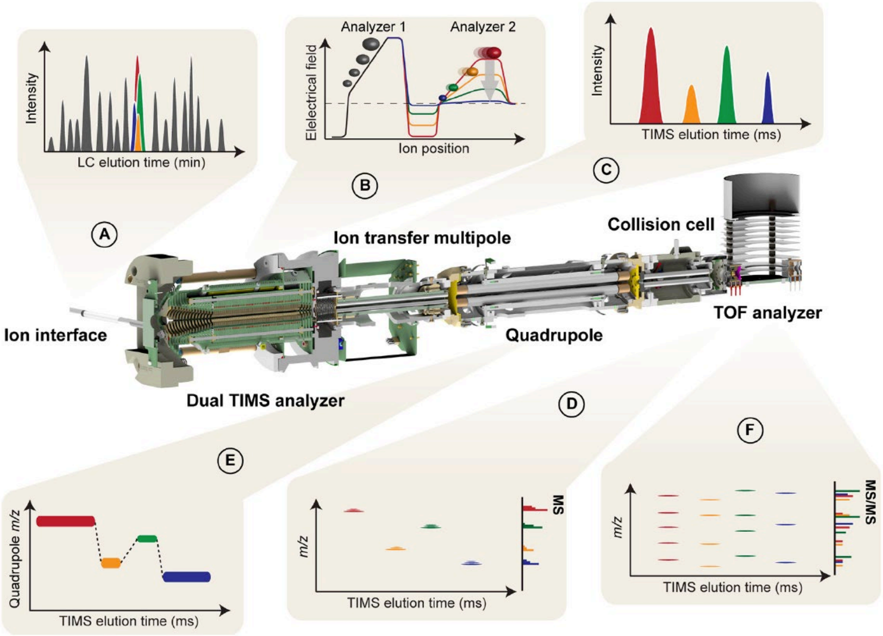


Figure 6. A timsTOF Pro mass spectrometer utilizing the Parallel Accumulation – Serial Fragmentation (PASEF) mode. Shown are the main elements: Trapped ion mobility spectrometry (TIMS) analyzer, quadrupole mass filter, quadrupole collision cell and bottom part of time of flight (TOF) analyzer. The displays depict the timely interplay between the single elements. See text for details. Figure from *Florian Meier et al. (*74*)*.

Another important innovation was made in our lab and is named Parallel Accumulation – Serial Fragmentation. PASEF is a scan mode that utilizes more ions in the same amount of time thereby increasing sequencing capacity about tenfold *(*74, 90*)*. Normally the quadrupole mass filter (**Figure 6E**) is switched in the MS/MS mode to the m/z value for a single in MS1 selected ion, thereby discarding all other ions eluting from the TIMS device. In PASEF mode, the quadrupole is sequentially switched in synchrony to the m/z of several select ions that elute from the TIMS device. This implementation allows about ten PASEF scans per second with a selection of up to 10 or 12 precursors each, resulting in sequencing speeds of > 100 Hz *(*74*)*. Importantly, the TIMS device operates in the millisecond time range and thus fits perfectly in between the peptide elution time from the column (seconds, **Figure 6A**) and the spectra acquisition time in the range of 100 microseconds (**Figure 6F**). Combining the TIMS with a mass spectrometer therefore offers a unique advantage for TOF instruments. The ion mobility dimension adds an additional precursor separation dimension and increases the signal-to-noise ratio by accumulation dense ion packages while drastically multiplying sequencing speed without sample loss. The additional dimension also allows an improvement in the data analysis pipeline: The matching between runs feature, in which identified features can be transferred between runs can benefit from the CCS as an extra dimension *(*91*)*.

I used the timsTOF Pro mass spectrometer, the PASEF scan mode and MBR feature in the interactome studies described in this thesis in order to achieve highly sensitive measurement for low input material and to generate as complete as possible data matrices.

1.3 Aims of the Thesis

The aim of this thesis was to elucidate the entire interactome of the model organism *S. cerevisiae* to understand its network structure and to discover novel biological findings. The project builds on previous work of Eva Keilhauer *(*43*)*, Fabian Hosp *(*45*)* and Marco Hein *(*44*)* who initiated the quest for quantitative interactomics studies in our lab, by establishing new concepts for analysis and testing the limits for input materials.

A major aim of the thesis was to optimize and establish a workflow for affinity-purifications coupled to mass spectrometry in a high throughput and scalable manner for all known to be expressed 4,200 proteins in *S. cerevisiae*. This included the identification of optimal conditions that allow exponential growth – the preferred condition for yeast biologist – as well as to miniaturize and standardize the workflow. One of my major goals was to achieve a workflow in which all steps are in a high-throughput compatible 96-well format. Therefore, the best condition/ protocols for yeast cell lysis within deep-well plates had to be established that would allow proper cooling, avoids cross contamination, and would extract lysate most efficiently. For the enrichment step, a custom-made solution for anti-GFP nanobody coated plates was initiated. In this context, I tested the most optimal plate material and coating with the aim of achieving best pull-down results and highest mass spectrometry compatibility. For the mass spectrometry sample preparation, several protocols were established to find a solution to keep the digestion and alkylation within the microtiter plate and in order to preserve Evotip compatibility. Initially, I explored different options in terms of LC or mass spectrometer and tried data independent acquisition modes before deciding on the use of the timsTOF Pro.

A major hurdle was the processing of the very large number of raw files, that initially took much more storage space than the Orbitrap output files. Particularly in early stages, limitations of the available software (initially only MaxQuant) was a major reason for delay and required many workarounds and tweaks.

These efforts have successfully enabled me to present in this doctoral thesis the most comprehensive and highly structured network of the yeast. Similar to human networks (on social media), the yeast interactome as described in the next chapter is highly connected with an average of 15 interactors, many of which are not reported. The rigorous workflow established here should allow similar interactome studies in other organisms (as demonstrated in Chapter

2.2). In addition, this work also provides a free web-portal to explore our datasets and thus serves as an important resource for other scientists.

2 Results

2.1 Article 1: The social architecture of an in-depth cellular protein interactome

André C. Michaelis, Andreas-David Brunner, Maximilian Zwiebel, Florian Meier, Maximilian T. Strauss, Isabell Bludau, Matthias Mann (2021). The social architecture of a near-complete cellular protein interactome. *Biorxiv, doi:10.1101/2021.10.24.465633.*

This publication contains the results of a near-complete protein-protein interactome in *S. cerevisiae.* Using affinity-purification coupled to mass spectrometry (AP-MS), I provide a map with high-quality interaction data, that triples and doubles the number of interactions and proteins, respectively, compared to the latest state-of-the-art reference data set *(*35*)*. Using AP-MS, this is the single-study derived interactome with the highest protein coverage in any organism yet. The majority of the reported interactions are new, based on a comparison with the broadly used BioGRID interaction database *(*92*)*. Building on previous studies from our group *(*43–45*)*, I developed a cell sample preparation and a mass spectrometry pipeline that would allow handling all of the about 4,200 GFP-tagged strains known to be expressed in yeast under standard growth conditions *(*19*)*. Using quantitative proteomics for the first time in a very large interaction screen, it was crucial to have very consistent handling in order to generate reproducible enrichment and background binders across all samples. The combination of an efficient lysis protocol and the latest generation of mass spectrometer allowed me to use 96-well plates throughout all steps. This “reduced” the sample number to 44 of these well plates.

The cell wall of yeast is tough and requires a special lysis protocol compared to other eukaryotic cells. While several options like cryogenic grinding as used previously *(*29, 32*)*, or proteolytic lysis *(*93*)* are available, they needed to be compatible with the high-throughput plate format and not interfere with the mass spectrometric workflow (as proteases would). I found that mechanical disruption fulfills those requirements best. It turned out to be important to use the correct low-protein binding equipment, a specific ratio of the right lysis buffer and glass beads, proper sealing of the plates while still allowing access to the samples, all while using optimized bead-beating conditions. Only a few devices allow parallel, high-frequency deep-well plate bead-beating. The cycles described in the methods part of the paper bring maximum lysis efficiency while keeping the temperature increase (which is crucial for maintaining protein interactions) with each cycle to a minimum.

For the pull-down, anti-GFP coated nanobody 96-well microtiter plates were available, but they turned out to have a coating that had unacceptably high contaminations for MS analysis. This made them incompatible with a single step protocol that would allow pull-down, washing, reduction, alkylation, and digestion within the same plate without transferring them. In cooperation with the company Chromotek, I tested several new production settings with a variety of plate materials to find the optimal setting for a MS compatible single step "in-well" digest. Those plates are now commercially available allowing other scientist to reproduce. For the denaturation and digestion protocol, some methods did not result in the efficient unfolding of stable proteins (as reflected in the absence of the GFP-tag) and some were not compatible with large-scale screens, nor the use of C18-material based purifications as it is required for the Evosep One. This included the commonly used SDC (sodium dodecyl sulfate) protocol *(*67*)* which requires heating to high temperature, which is impractical for large-scale analysis and an SDP-RPS (styrenedivinylbenzene- reverse phase sulfonate) based purification. Instead, I decided to use a classical high molar urea LysC digest followed by a low molar urea- one. This allowed the identification of e.g. GFP which I used in a tag-based abundance calculation later on. The LysC only digest improved results, likely due to in general better performance of TOF devices with slightly higher m/z peptides and the reduced missed-cleavage rate which I observed to be worse in a urea based LysC and trypsin digest. This is likely caused by the efficient digestion of LysC by trypsin in those denaturation conditions.

All optimized steps allowed the samples to be processed in a streamlined manner with only two major transfers: from the deep-well plate to the microtiter plate and then to the Evotips, resulting in high reproducibility.

The above detailed description should aid others to appreciate the steps and the underlying effort of the developed protocol that are only briefly described in the results-oriented paper. The motivation to optimize the workflow and to reduce and simplify all possible steps, was not only to generate the best feasible data in this study, but also to provide an easy protocol that will allow other groups to do similar experiments. Our dataset allows to select those baits that efficiently cover a part of the network of special interest. By doing this and by using the provided workflow, one can easily conduct new studies that for example investigate effects of specific perturbations. I also see this platform as a starting point for many global interaction screens to come, that will map differences between conditions and help to unravel new mechanists of the cell.

With this large coverage of expressed proteins, my results show that correlation analysis becomes a very powerful tool. This is because almost all yeast proteins are present in this dataset, and therefore correlations can be established for almost all of them. This is why we have very significant interactions that are only based on correlations. Examples include proteins that are not taggable such as the chaperonin containing t-complex. Overall, I find many very promising new interactions that are covered by several high confidence interactions.

Because large-scale data are sometimes hard to understand or even to access, I have put much effort in generating an easily accessible and visually appealing web application Maximillian Zwiebel was invaluable in this endeavor as he manifested most of our analysis pipeline into a corresponding code and transformed all my visions of how to browse the final data into an aesthetic, concise and easy to handle webpage (*www.yeast-interactome.org*).

The results described below belong to the manuscript which is published on *BioRxiv*.

The social architecture of an in-depth cellular protein interactome

André C. Michaelis[1], Andreas-David Brunner[1], Maximilian Zwiebel[1], Florian Meier[1,2], Maximilian T. Strauss[3], Isabell Bludau[1], Matthias Mann[1,3,#]

[1]Max-Planck Institute of Biochemistry, Martinsried, Germany; [2]Functional Proteomics, Jena University Hospital, Jena, Germany; [3]NNF Center for Protein Research, University of Copenhagen, Denmark

[#]Correspondence: *mmann@biochem.mpg.de*

Nearly all cellular functions are mediated by protein-protein interactions and mapping the interactome provides fundamental insights into the regulation and structure of biological systems. In principle, affinity purification coupled to mass spectrometry (AP-MS) is an ideal and scalable tool, however, it has been difficult to identify low copy number complexes, membrane complexes and those disturbed by protein-tagging. As a result, our current knowledge of the interactome is far from complete, and assessing the reliability of reported interactions is challenging. Here we develop a sensitive, high-throughput, and highly reproducible AP-MS technology combined with a quantitative two-dimensional analysis strategy for comprehensive interactome mapping of *Saccharomyces cerevisiae*. We reduced required cell culture volumes thousand-fold and employed 96-well formats throughout, allowing replicate analysis of the endogenous green fluorescent protein (GFP) tagged library covering the entire expressed yeast proteome. The 4159 pull-downs generated a highly structured network of 3,909 proteins connected by 29,710 interactions. Compared to previous large-scale studies, we double the number of proteins (nodes in the network) and triple the number of reliable interactions (edges), including very low abundant epigenetic complexes, organellar membrane complexes and non-taggable complexes interfered by abundance correlation. This nearly saturated interactome reveals that the vast majority of yeast proteins are highly connected, with an average of 15 interactors, the majority of them unreported so far. Similar to social networks between humans, the average shortest distance is 4.2 interactions. A web portal (*www.yeast-interactome.org*) enables exploration of our dataset by the network and biological communities and variations of our AP-MS technology can be employed in any organism or dynamic conditions.

The large-scale study of cellular interactomes by MS-based proteomics dates back almost 20 years *(*1*,* 2*)*, culminating in two studies in which nearly half the expressed yeast proteome was successfully purified with identified interactors *(*3*,* 4*)*. These datasets have been mined extensively, leading to a network-based view of the cellular proteome. Given the importance of the interactome for functional understanding and the dramatic improvements in MS-technology during the last decade *(*5*,* 6*)*, we set out to generate a substantially complete interactome of all proteins present in an organism in a given state. We made use of an endogenously GFP-tagged yeast library containing the 4159 proteins that were detectable by fluorescence under standard growth conditions *(*7*)*. Miniaturization and standardization of the workflow in combination with an ultra-robust liquid chromatography system with minimal overhead time coupled to a sensitive trapped ion mobility mass spectrometer employing the PASEF scan mode *(*8*,* 9*)*, resulted in very high data completeness across pull-downs. This workflow required only 1.5 mL instead of liters of yeast culture, provided a constant throughput of 60 pull-downs per day and allowed using the same conditions for soluble or membrane proteins of vastly different abundances (**Fig. 1A**).

Measurement of the yeast interactome

To test the quantitative reproducibility of our workflow, we performed 24 biological replicates of pull-downs of three nuclear complexes, which resulted in complete retrieval of these complexes from a single bait each, with 9% average coefficients of variation (CVs) of enriched complex members (**Fig. 1B**). This compares to a 69% repeatability of assigned interactions in the previous large-scale screens *(*10*)*.

Three layers of evidence help to establish an interaction between two proteins. The first two are statistically significant enrichment of the proteins in the forward and in the reverse pull-downs (where the prey pull-down significantly enriches the bait). Instead of employing only a t-test of bait pull-down against a pull-down of a strain only expressing GFP, we made use of our vast number of diverse GFP-tagged strains, to combine them into a single control group, thereby efficiently removing false positives not specifically binding to the bait (Methods: Enrichment analysis). Using this affinity enrichment (rather than affinity purification) concept *(*11*)*, we quantitatively compared all proteins across more than 8,000 pull-down measurements, making use of the profile similarities of interacting proteins in correlation analysis. This third evidence type turned out to be very informative due to the large quantitative accuracy combined with close to a complete set of “virtual controls” (Methods: Protein correlation, **Fig. 1C**).

We combined all three layers of each interaction into a single interaction score and retained those with a minimum score of 2, corresponding to (a) a single pull-down at 1% FDR or (b) a correlation z-score of at least five or (c) forward and reverse pull-downs at 5% FDR each, or (d) one at 5% FDR combined with a correlation z-score greater than four. To retrieve clusters and complexes from our interactome data, we used Markov clustering with the above-derived score as the edge weights, without any training or a priori knowledge (Methods: Network generation, **Fig. 1C**).

The replicate GFP pull-down measurement in the 4,147 yeast strains resulted in the enrichment of 82% of the baits (**Suppl. Fig. 1**). Our MS-data provided statistically significant evidence for a total of nearly 30,000 physical interactions, corresponding to an average of 15.2 interactions per protein. Most were supported by forward pull-down (38%), followed by forward pull-down and significant prey correlation (29%), whereas nearly all interactions with both forward and reverse evidence also had significant correlations (> 99%) **(Suppl. Fig. 2).**

Due to the limited overlap of the interactions reported by two previous large-scale studies (13% shared interactions), *Collins et al.* merged and reanalyzed these datasets to create a consensus network with 1,622 nodes *(*12*)*. Our data encompasses 95% of these, but places nearly the entire expressed yeast proteome in a network (3,909 nodes). Our dataset of 30,000 significant protein-protein interactions confirms 62% of the much smaller *Collins et al.* dataset (**Fig. 1E**). Based on a comparison with the BioGRID database *(*13*)*, over two-thirds of the interactions reported here are novel.

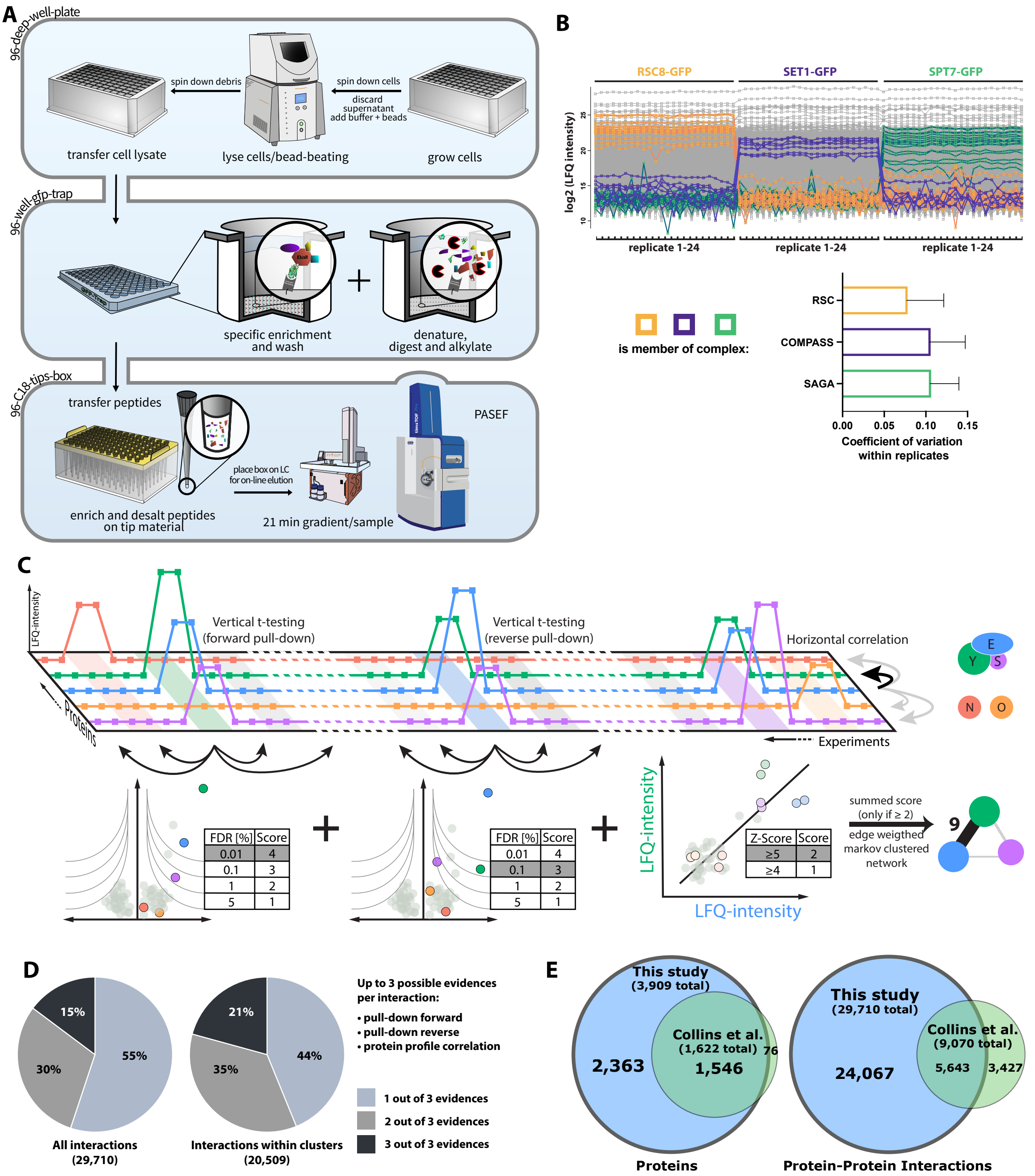


Figure 1. A comprehensive and scalable interactomics technology.

A) Sample preparation in 96-well format and mass spectrometric measurement: Each strain of the GFP-tagged library is lysed by mechanical disruption and transferred into anti-GFP nanobody coated microtiter plates, where weak interactions are preserved by gentle washing. After enzymatic "in-well" digestion, resulting peptides are transferred on standardized C_{18}-StageTips from which they are directly eluted into a standardized 60 samples/day gradient. Data is acquired in the PASEF scan-mode on a trapped ion mobility – Time of Flight mass spectrometer. **B)** Streamlined workflow and reduced transfer steps reduce the risk of manual errors and sample variation: Demonstration of workflow reproducibility and sensitivity on three nuclear complexes in biological replicates. Tagged members of each complex (baits) pull down the known preys in very similar amounts. Lower panel: bar plot of mean coefficient of variation with standard deviations. **C)** Two-dimensional interaction scoring: Columns represent pull-down experiments in replicates (light color). Squares depict intensities of detected proteins across the pull down-experiments. Three levels of evidence support each interaction: t-test of forward pull-down against complement experiments, t-test of reverse pull-down, and protein profile correlation – the correlated abundance profile against all other proteins across all experiments (z-scored, Methods: Protein correlation). **D)** Proportion of interactions backed by multiple layers of evidence. **E)** Overlap of proteins with at least one interactor and interactions detected in this study with the previous state-of-the-art network *(*12*)*.

Organization of protein-protein interactions in clusters

Markov clustering analysis - with our interaction scores as edge weights, condensed the network into 623 clusters, with about 20,000 interactions within them, most supported by at least two statistically significant levels of evidence (**Fig. 1D**). When we inspected known protein complexes from different cellular compartments, especially membrane complexes, we found them to recapitulate the literature to a large degree. Furthermore, we here retrieved 3628 interactions between membrane annotated proteins, compared to 853 in a dedicated membrane proteome *(*14*)*. This is shown exemplarily for the full retrieval of the endosomal retromer complex, the conserved oligomeric Golgi complex, and the plasma membrane exocyst complex (**Fig. 2A**). At the same time, our unbiased and high coverage analysis identified novel subunits with tight association to known complexes. For instance, three subunits of the essential endoplasmic reticulum (ER) membrane oligosaccharyl transferase (OST) complex - an integral component of the translocon - associated with α-1,2-mannosidase (Mns1; human homolog: MAN1B1), an enzyme that catalyzes the ER glycoprotein trimming reaction which is required for ER-associated protein degradation (ERAD). This indicates that the enzymatic activity of N-linked oligosaccharide chain addition is physically connected to the removal of a terminal sugar, at least in one isoform of the OST complex. The slow enzymatic activity of Mns1 acts as a timer *(*15*,* 16*)* and we speculate that it co-translationally primes stalled or erroneous proteins directly at its site of translocation for ERAD degradation. We also discovered a novel complex defined by three unreported interactions (all with the maximum interaction score of 10) between Tcd1, Tcd2 - mitochondrial proteins that are involved in tRNA base modification - and YGR012W, a protein of unknown function. A homolog of Tcd1 and Tcd2 in *E. coli* termed TcdA functions in a complex of three in the cyclization of an essential tRNA modification found in all three domains of life *(*17*)*.

Many biological complexes share members and these can be difficult to disentangle by clustering algorithms. We speculated that our highly quantitative data could nevertheless resolve these cases. Applying a network layout algorithm (Methods: Network generation) to members of the transcription factor TFIID and the SAGA complex, separately reconstructed these complexes, while correctly assigning shared members (**Fig. 2A**). At the global scale, we found that about two-thirds of all interactions connected members within clusters, whereas the remainder connected clusters to each other. For example, the cytoplasmatic signal recognition particle (SRP) is connected to another cluster containing the SRP-receptor (SRP101/102). The largest connected clusters were the small and large subunits of the ribosome, with 362 inter-complex connections.

Leveraging the common, endogenous GFP-tag on more than 3379 detected baits, we next investigated if the MS-signal of the GFP peptides could be used to quantify each bait. Indeed, these intensities correlated well (r = 0.77), with a recent compilation of yeast protein abundances *(*18*)* (**Fig. 2B**). This validates our interaction workflow and allows tag-based estimation of the relative abundances of proteins in a cluster, which is useful to determine their functional role *(*19*)*.

For some proteins, for example the members of the chaperonin containing t-complex (CCT), tagging is not possible because it interferes with protein stability or function *(*20*)*. Based on highly significant correlations between profiles of the subunits, CCT was nevertheless fully recovered (**Fig. 2C**). Besides the eight conserved, ring-forming members, we also detected a distinct set of 21 interacting proteins, about half of which had not been reported yet. Two of these were catalytic subunits of protein phosphatase 2A, suggesting regulatory functions, and others, such as tubulin and actin-related proteins (Tub1, Tub3, Arp1) major known folding substrates. CCT may have a restricted or broad set of folding substrates *(*21*)*, and our results quantitatively support the former possibility.

The above examples only scratch the surface of the interesting biological leads contained in the data. To allow ready exploration of interactions of interest, we created a web portal (*www.yeast-interactome.org*), which supplies statistical evidence for protein-protein associations, and summarizes the resulting clusters (**Fig. 2D**).

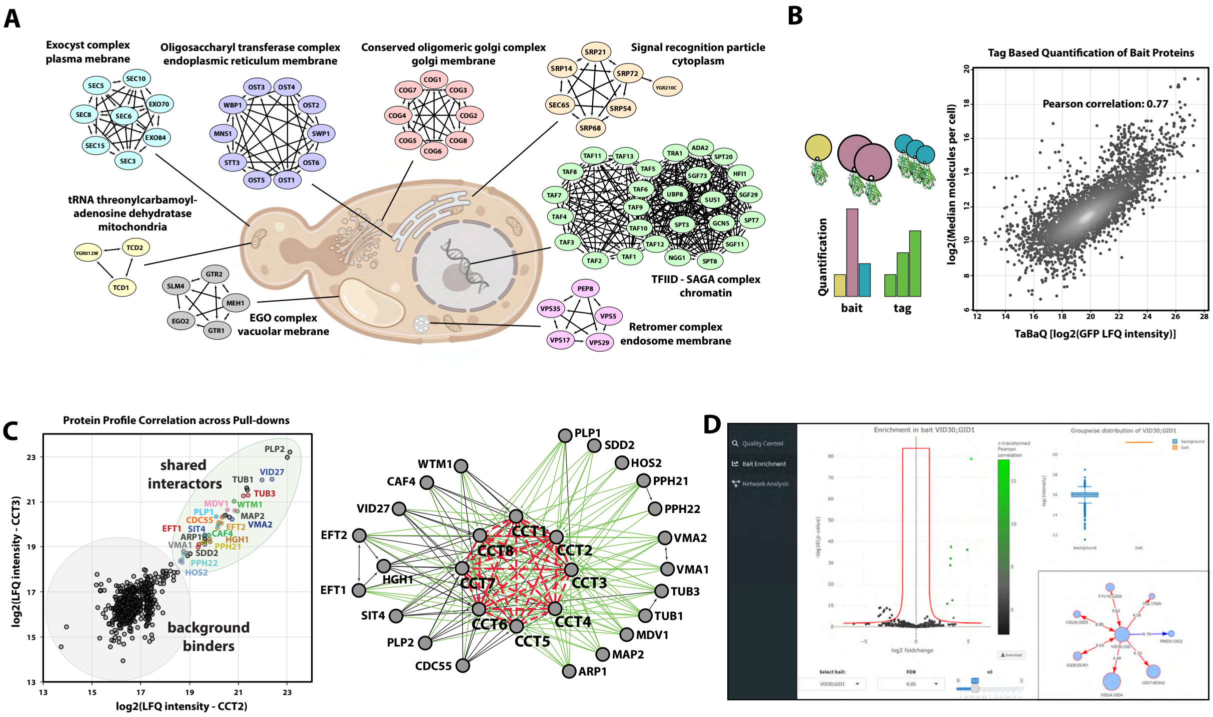

A
Exocyst complex plasma mebrane
SEC5 SEC10 EXO70 SEC8 SEC6 EXO84 SEC15 SEC3
Oligosaccharyl transferase complex endoplasmic reticulum membrane
OST3 OST4 WBP1 OST2 MNS1 SWP1 STT3 OST6 OST5 OST1
Conserved oligomeric golgi complex golgi membrane
COG1 COG7 COG3 COG4 COG2 COG5 COG8 COG6
Signal recognition particle cytoplasm
SRP21 SRP14 SRP72 YGR210C SEC65 SRP54 SRP68
tRNA threonylcarbamoyl-adenosine dehydratase mitochondria
YGR012W TCD2 TCD1
EGO complex vacuolar mebrane
GTR2 SLM4 MEH1 EGO2 GTR1
TFIID - SAGA complex chromatin
TAF11 TAF13 TRA1 ADA2 SPT20 TAF8 TAF5 SGF73 HFI1 TAF7 TAF6 UBP8 SUS1 SGF29 TAF4 TAF9 SPT3 GCN5 SPT7 TAF10 TAF3 TAF12 SGF11 TAF2 TAF1 NGG1 SPT8
Retromer complex endosome membrane
PEP8 VPS35 VPS5 VPS17 VPS29
B
Quantification
bait
tag
Tag Based Quantification of Bait Proteins
Pearson correlation: 0.77
log2(Median molecules per cell)
TaBaQ [log2(GFP LFQ intensity)]
12 14 16 18 20 22 24 26 28
6 8 10 12 14 16 18 20
C
Protein Profile Correlation across Pull-downs
shared interactors
background binders
PLP2 VID27 TUB1 TUB3 MDV1 WTM1 PLP1 MAP2 CDC55 VMA2 EFT1 SIT4 EFT2 ARP1 CAF4 HGH1 VMA1 PPH21 SDD2 PPH22 HOS2
log2(LFQ intensity - CCT3)
log2(LFQ intensity - CCT2)
13 15 17 19 21 23
PLP1 SDD2 WTM1 HOS2 CAF4 PPH21 VID27 PPH22 EFT2 CCT1 VMA2 CCT8 CCT2 VMA1 HGH1 EFT1 CCT7 CCT3 TUB3 SIT4 CCT6 CCT4 TUB1 CCT5 MDV1 PLP2 MAP2 CDC55 ARP1
D
Quality Control
Bait Enrichment
Network Analysis
Enrichment in bait VID30;GID1
-log10(p-value)
log2 foldchange
z-transformed Pearson correlation
Download
Groupwise distribution of VID30;GID1
background
bait
log(Intensity)
Select bait:
VID30;GID1
FDR
0.05
s0
0.5
FYV10;GID9 YDL176W VID28;GID5 VID30;GID1 RMD5;GID2 GID8;DCR1 GID7;MOH2 VID24;GID4
9.62 8.09 8.85 9.76 8.69 6.33 4.48

Figure 2. High-quality dataset for the exploration of the interactome.
A) Clusters derived from our interactome for a range of challenging complexes such as chromatin-associated, soluble and membrane-bound complex of various organelles. In each case, all known subunits were retrieved. **B)** Tag-based quantification allows retrieving abundance information for the baits in a generic manner (left panel). Correlation of tag peptide-based signals with a literature compilation of yeast protein abundances *(18)* (right panel). **C)** For the non-taggable chaperonin containing t-complex (CCT), profile correlation analysis nevertheless reveals its subunits and interactors. Interactions based on correlation only are shown in red (dashed) and unreported interactions with CCT in green. **D)** Web application that allows exploration of interaction data for interactions of interest. For all proteins, pull-downs are depicted as volcano plots together with a violin plot that shows the MS intensity of user-selected outliers. Subnetwork from pull-downs of the selected bait and reverse pull-downs or significant interactors.

Network architecture of the cellular interactome

The availability of data for large networks in systems ranging from power-grids, genetic networks to human social networks, has enabled the study of their underlying architecture, commonalities and differences *(22)*. This topic also has a long history in protein interaction networks. However, these analyses have been limited by the incompleteness of the data, especially in multicellular species *(23)*. With an in-depth protein-protein interaction map in hand, we compared its characteristics to networks in different domains. Yeast proteins are highly connected with an average of 15 and a median of 6 interactions per protein, significantly more than the human BioPlex interactome (average interactions: 8) *(24)* (**Fig. 3A**). Influential nodes – those with the highest number of normalized interactors (or degree centrality) – were more common than in the GitHub package dependency network, but less common than in a similarly-sized Facebook subnetwork (**Suppl. Fig. 4**). This high connectivity is reflected in a mean shortest path between yeast proteins of only 4.2, ranging from highly connected proteins with only three steps to less connected ones with an average of more than 7. (**Fig. 3B**). This is very similar to the 4.7 path-length for world-scale Facebook relationships *(*25*)*.

One of the key features for most real-life networks with complex topology in contrast to random networks is the scale-free power-law distribution of interactors *(*26*,* 27*)*. Scale-free network properties are thought to arise by preferential attachment over evolutionary time to already well-connected nodes and can be identified by a linear relation of the node degree or number of interactors with its frequency (number nodes with that degree) plotted in log-log space. While this has been hard to prove for biological networks, they rather appear to be exponential or have a truncated power-law degree distribution *(*28*)*, our yeast interactome clearly displays scale-free properties (**Fig. 3C**)**.** In accordance with previous protein-protein interaction networks *(*3*,* 29*)*, the exponent was below two, at the lower end of the two to four range of other scale-free networks.

The high connectivity of most proteins organizes almost all of them (3,827) into a single giant connected component, accompanied by 38 small components (82 proteins) (**Fig. 3D**). A total of 478 proteins were outside of the network because MS-analysis of their pull-downs only identified the bait itself. There was an significant enrichment for 87% of these baits (FDR<0.01%), indicating that there were no identifiable interactors under our standard conditions despite a successful pull-down (**Suppl. Fig. 3,** see volcano plots accessible via web-application).

We next investigated the large-scale organization of the yeast interactome using the Louvain community detection algorithm (Methods: Network comparisons). This revealed that yeast is organized in smaller communities than GitHub, ego-Facebook and also Bioplex (**Fig. 3E**). Important "bottleneck" proteins that are part of many shortest paths have a high "betweenness-centrality". The yeast interactome has comparably more of those central nodes and bioinformatic enrichment analysis highlighted proteins involved in "RNA polymerase II", "mitochondrial nucleoid", "gluconeogenesis" and "misfolded protein binding" (**Fig. 3F; Suppl. Table 1**).

Altogether, based on the total of 4,387 identified yeast proteins, only 10.9% had no discernable interaction partner, whereas 74.2% had at least two. Given that some of our baits will have context dependent interactions not captured here, our estimates are conservative and we conclude that almost all yeast proteins are "social".

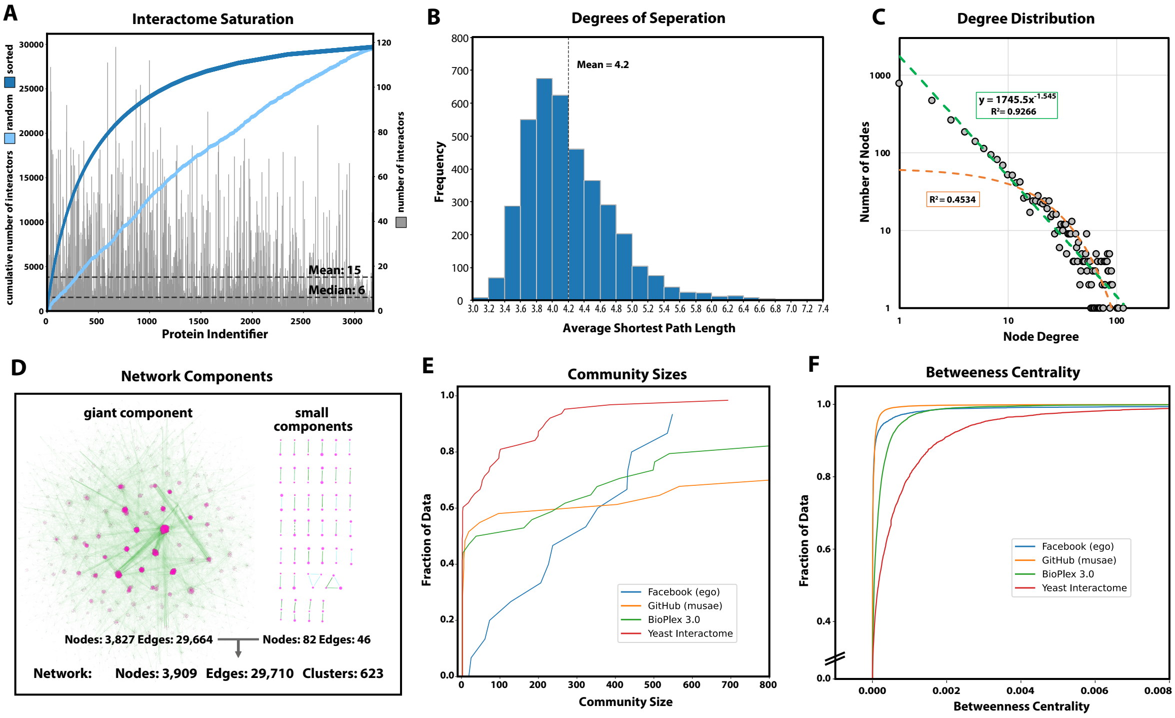


Figure 3. Properties of the protein interaction network.

A) Distribution of the number of interactors (grey). Sorted cumulative number of interactions reaches saturation at 30,000 interaction (blue) **B)** The distribution of average shortest path length between all possible pairs of nodes within the giant component shows a mean of 4.2 steps corresponding to 3.2 intermediaries ("degrees of separation") **C)** Power-law fit (green; equals a linear fit on a log-log scale) of the frequency of proteins with a given number of interactions highlights the scale-free properties of the network. Exponential fit depicted in orange **D)** Nearly all nodes of the network are connected with each other in the giant component. **E)** Cumulative distribution function of the community sizes (Louvain algorithm) detects more smaller communities for *S. cerevisiae*. **F)** Cumulative distribution function of betweenness centrality: The *S. cerevisiae* interactome has more nodes with a high betweenness-centrality than the comparison data sets.

Global organization in clusters highlights novel interactions

Intensive research over the last decades has made *S. cerevisiae* arguably the best understood single-cell eukaryotic organism, leading to the discovery of crucial conserved cellular functions, such as metabolic pathways, mechanisms of DNA replication and transcription, protein quality control and modifications that were later confirmed in human and other organisms. Nevertheless, our interactome still contained uncharacterized proteins or interactions not reported in the BioGRID database and thus providing novel biological insights (extended selection **Suppl. Fig. 6**). Furthermore, BioGRID has accumulated binding events from very disparate experiments without a common confidence score (133,900 physical interactions from about 10,000 publications). We reasoned that our homogeneous, high-quality data set would help biologists to highlight true positive interactors with biological relevance, several of whom we discuss below.

A total of eleven evidences connect the uncharacterized protein YDL176W with the conserved glucose-induced-degradation (GID) complex, only a few of which had been indicated by previous pull-down or genetic interaction data *(*3, 30*)* (**Fig. 4B**). These types of high-confidence associations assist in prioritizing interactions and form the basis for a detailed mechanism and structure discovery of a novel GID modulator. Similarly, our data ties the uncharacterized protein YJR011C to the conserved transcription and translation regulatory CCR4-Not complex *(*31, 32*)* via high-significant interactions to a majority of its subunits (**Fig. 4G**). Finally, YHR131C is linked to three and YLR407W to the fourth subunit of the kinase CK2 (**Fig. 4N**). We discovered an interaction of Cue4 – a protein of unknown function containing a ubiquitin-binding domain – with the ER membrane complex EMC, potential membrane protein chaperone (**Fig. 4L**). As Cue4 is a paralogue of Cue1 (coupling of ubiquitin conjugation to ER degradation), a component of ERAD *(*33*)*, this physical link and the known aggravating genetic interactions of Δ*cue1* with EMC knock-outs *(*34*)* suggests an ERAD related quality control mechanism for EMC.

The transcriptional regulator SWI/SNF unexpectedly interacts with the phosphate transporters Pho87 and Pho90 (**Fig. 4D**). Out of four plasma membrane phosphate transporters only Pho87 and Pho90 comprise a cytoplasmatic accessible SPX domain. While an SPX dependent phosphate sensing mechanism has been discovered in plants *(*35*)*, it remains elusive in *S. cerevisiae*. In *Arabidopsis* inositol pyrophosphate $InsP_8$ concentration increases under phosphate rich conditions and promotes the interaction between SPX domains and a four-stranded coiled-coil motif of phosphate starvation response transcription factors *(*36*)*. Strikingly the recently solved structure of SWI/SNF reveals such a coiled-coil four-helix-bundle at its spine region *(*37*)* providing a potential SPX interaction site. This raises the possibility of a novel cytoplasmatic sensing and retention mechanisms of this key transcriptional regulator which is known to be necessary for a phosphate starvation response *(*38, 39*)*. Interestingly, not only the SWI/SNF complex but also an SPX domain-containing phosphate transporter named XPR1 - which has recently been shown to be controlled by $InsP_8$ *(*40*)* - is present in humans.

Illustrating translational relevance, we expand the known interaction of the GTPase-activating protein Ira1/Ira2 (NF1/neurofibromin in humans) and Gpb1/Gpb2 (ETEA in humans) *(*41*)* by Trx2 a thioredoxin isoenzyme (human homolog: TXN) and Gpx1 (human homologs: GPX3-6), an antioxidant enzyme whose glutathione peroxidase activity is neuroprotective in models of Huntington's disease *(*42*)* (**Fig. 4C**).

Additionally, we find a new physical interaction between the two uncharacterized proteins YPR063C and YNR021W (**Suppl. Fig. 6**) whose dimerization and structure has just been predicted in a deep-learning approach *(*43*)*.

Apart from known and novel protein complexes, the yeast interactome depicted in **Fig. 4**, clearly shows evidence of high order connections. These often map to different compartments of the cell, such as the prominent connections between ribosomes in the cytoplasm and the nucleolus, its site of maturation or connect large and small ribosomal subunits that despite its “stickiness” are organized in individual clusters.

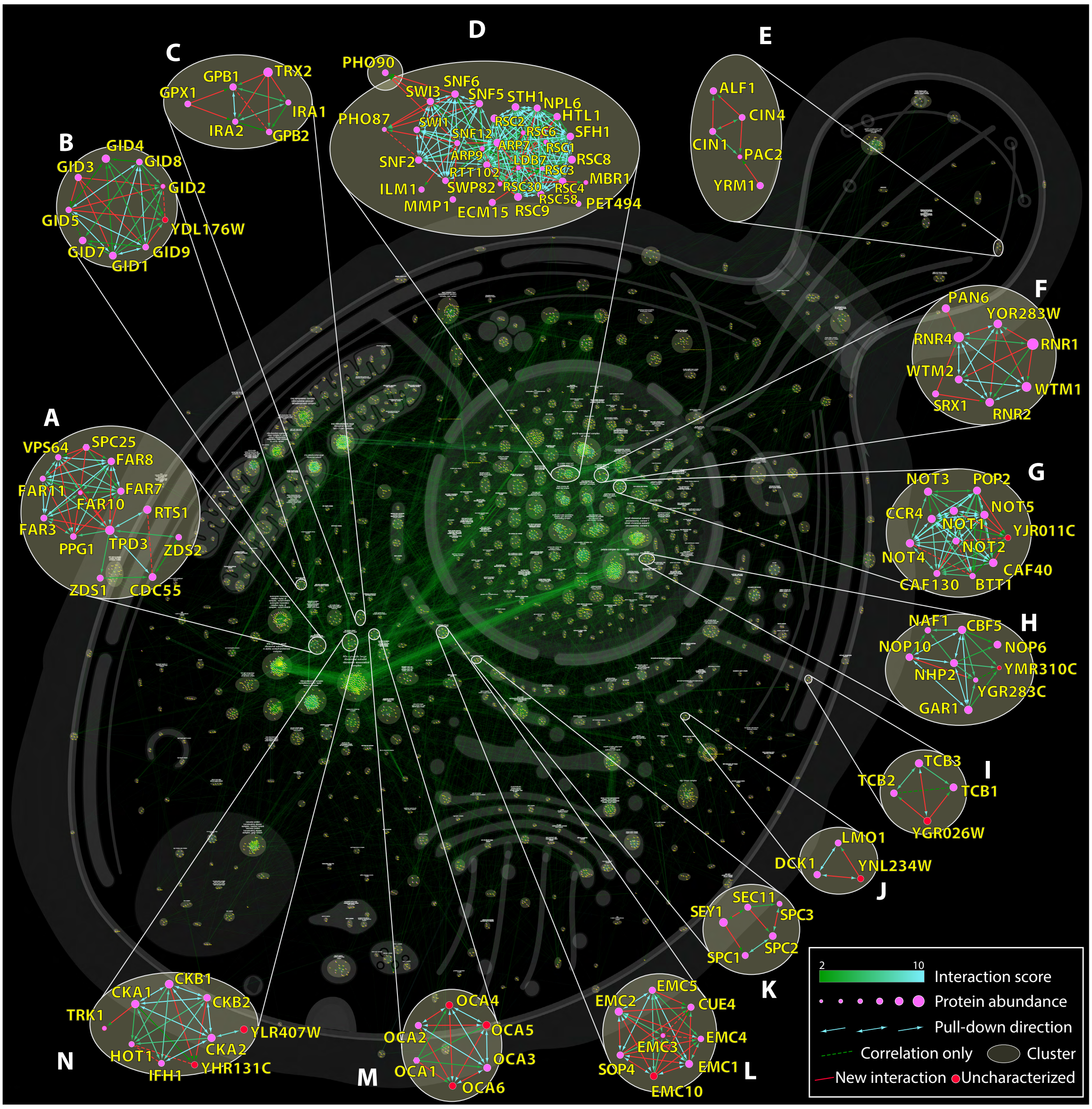


Figure 4. Network of an in-depth interactome highlighting novel interactions.
Cellular interaction map of all significant interactions. Clusters are highlighted by circles and cellular localization is indicated by most frequent GO term within a cluster. Enlargements show examples of either

novel interactions (based on BioGRID) or those that have not been described further as potential high significant interactor and interactions involving uncharacterized proteins. A full browsable and interactive version of this network can be found at our web application (*www.yeast-interactome.org*).

Outlook

Here we have developed and applied a novel and highly scalable interactome technology, enabling replicate measurement of the yeast network in a fraction of the measurement time and starting materials needed previously. Our screen reached near saturation and contained nearly all complexes expected under our experimental conditions (**Fig. 3A, Fig. 4**). Given its streamlined nature, our workflow can now readily be used in other endogenously tagged model organisms (44) or to study remodeling of the interactome in the presence of dynamic biological processes or perturbations. Similarly, we envision its use with other interaction technologies like BioID or APEX using tagged libraries that nowadays can be easily generated using the SWAp-Tag platform (45). The comprehensive yeast interactome data can further be used as prior knowledge for hypothesis-driven analysis of protein complexes, for example for native protein complex co-fractionation coupled to MS (46, 47). Additionally, we imagine that such interactome data could also be combined with MS-crosslinking studies and recent advances in computational prediction of protein structures from their sequences (48, 49) to yield complete structural models in many cases.

Acknowledgments. We thank the former members of the Stefan Jentsch lab for their support in yeast cell culture related issues, especially Jochen Rech for providing and handling the library. We greatly appreciate discussions with Marco Y. Hein and Fabian Hosp who were sharing their experience with large-scale pull-down screens and Brenda Schulman for project guidance and discussions. We thank all members of the Department of Proteomics and Signal Transduction for help and discussions and in particular Igor Paron and Christian Deiml for mass spectrometry assistance, Mario Oroshi for technological assistance and Medini Steger for manuscript proofreading. We thank our core facility for their great services and Bruker, Chromotek and Evosep for their continuous technical support.

Funding. This work was supported by the Max Planck Society for the Advancement of Science. I.B. acknowledges funding support from her Postdoc. Mobility fellowship granted by the Swiss National Science Foundation (P400PB_191046).

References

1. A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415, 141–147 (2002).

2. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, M. Tyers, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. 415, 180–183 (2002).

3. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, J. F. Greenblatt, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. 440, 637–643 (2006).

4. A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 440, 631–636 (2006).

5. R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function. *Nature*. 537, 347–355 (2016).

6. P. Lössl, M. van de Waterbeemd, A. J. Heck, The diverse and expanding role of mass spectrometry in structural and molecular biology. *Embo J*. 35, 2634–2657 (2016).

7. W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O'Shea, Global analysis of protein localization in budding yeast. *Nature*. 425, 686–691 (2003).

8. F. Meier, S. Beck, N. Grassl, M. Lubeck, M. A. Park, O. Raether, M. Mann, Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J Proteome Res*. 14, 5378–5387 (2015).

9. N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Müller, F. Meier, J. V. Olsen, O. Vorm, M. Mann, A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol Cell Proteom Mcp*. 17, 2284–2296 (2018).

10. J. Goll, P. Uetz, The elusive yeast interactome. *Genome Biol*. 7, 223 (2006).

11. E. C. Keilhauer, M. Y. Hein, M. Mann, Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Mol Cell Proteom Mcp*. 14, 120–35 (2015).

12. S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, N. J. Krogan, Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*. 6, 439–450 (2007).

13. R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, M. Tyers, The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 47, gky1079- (2018).

14. M. Babu, J. Vlasblom, S. Pu, X. Guo, C. Graham, B. D. M. Bean, H. E. Burston, F. J. Vizeacoumar, J. Snider, S. Phanse, V. Fong, Y. Y. C. Tam, M. Davey, O. Hnatshak, N. Bajaj, S. Chandran, T. Punna, C. Christopolous, V. Wong, A. Yu, G. Zhong, J. Li, I. Stagljar, E. Conibear, S. J. Wodak, A. Emili, J. F. Greenblatt, Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*. 489, 585–589 (2012).

15. G. Z. Lederkremer, M. H. Glickman, A window of opportunity: timing protein degradation by trimming of sugars and ubiquitins. *Trends Biochem Sci*. 30, 297–303 (2005).

16. C. Xu, D. T. W. Ng, Glycosylation-directed quality control of protein folding. *Nat Rev Mol Cell Bio*. 16, 742–752 (2015).

17. K. Miyauchi, S. Kimura, T. Suzuki, A cyclic form of N6-threonylcarbamoyladenosine as a widely distributed tRNA hypermodification. *Nat Chem Biol*. 9, 105–111 (2013).

18. B. Ho, A. Baryshnikova, G. W. Brown, Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst*. 6, 192-205.e3 (2018).

19. M. Y. Hein, N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman, M. Mann, A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*. 163, 712–723 (2015).

20. G. Pappenberger, E. A. McCormack, K. R. Willison, Quantitative Actin Folding Reactions using Yeast CCT Purified via an Internal Tag in the CCT3/γ Subunit. *J Mol Biol*. 360, 484–496 (2006).

21. J. Vallin, J. Grantham, The role of the molecular chaperone CCT in protein folding and mediation of cytoskeleton-associated processes: implications for cancer cell biology. *Cell Stress Chaperones*. 24, 17–27 (2019).

22. M. E. J. Newman, The structure and function of complex networks. *Arxiv* (2003), doi:10.1137/s003614450342480.

23. I. Bludau, R. Aebersold, Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat Rev Mol Cell Bio*, 1–14 (2020).

24. E. L. Huttlin, R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, J. Szpyt, B. M. Gassaway, A. Panov, H. Parzen, S. Fu, A. Golbazi, E. Maenpaa, K. Stricker, S. G. Thakurta, T. Zhang, R. Rad, J. Pan, D. P. Nusinow, J. A. Paulo, D. K. Schweppe, L. P. Vaites, J. W. Harper, S. P. Gygi, Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*. 184, 3022-3040.e28 (2021).

25. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, Four Degrees of Separation. *Arxiv* (2011).

26. A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks. *Science*. 286, 509–512 (1999).

27. M. Vidal, M. E. Cusick, A.-L. Barabási, Interactome Networks and Human Disease. *Cell*. 144, 986–998 (2011).

28. G. Lima-Mendez, J. van Helden, The powerful law of the power law and other myths in network biology. *Mol Biosyst*. 5, 1482–93 (2009).

29. E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wühr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. D. Camilli, J. A. Paulo, J. W. Harper, S. P. Gygi, The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. 162, 425–40 (2015).

30. I. Ulitsky, T. Shlomi, M. Kupiec, R. Shamir, From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. *Mol Syst Biol*. 4, 209 (2008).

31. R. Buschauer, Y. Matsuo, T. Sugiyama, Y.-H. Chen, N. Alhusaini, T. Sweet, K. Ikeuchi, J. Cheng, Y. Matsuki, R. Nobuta, A. Gilmozzi, O. Berninghausen, P. Tesina, T. Becker, J. Coller, T. Inada, R. Beckmann, The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science*. 368, eaay6912 (2020).

32. H. Liu, V. Badarinarayana, D. C. Audino, J. Rappsilber, M. Mann, C. L. Denis, The NOT proteins are part of the CCR4 transcriptional complex and affect gene expression both positively and negatively. *Embo J*. 17, 1096–1106 (1998).

33. T. Biederer, C. Volkwein, T. Sommer, Role of Cue1p in Ubiquitination and Degradation at the ER Surface. *Science*. 278, 1806–1809 (1997).

34. M. C. Jonikas, S. R. Collins, V. Denic, E. Oh, E. M. Quan, V. Schmid, J. Weibezahn, B. Schwappach, P. Walter, J. S. Weissman, M. Schuldiner, Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum. *Science*. 323, 1693–1697 (2009).

35. R. Wild, R. Gerasimaite, J.-Y. Jung, V. Truffault, I. Pavlovic, A. Schmidt, A. Saiardi, H. J. Jessen, Y. Poirier, M. Hothorn, A. Mayer, Control of eukaryotic phosphate homeostasis by inositol polyphosphate sensor domains. *Science*. 352, 986–990 (2016).

36. M. K. Ried, R. Wild, J. Zhu, J. Pipercevic, K. Sturm, L. Broger, R. K. Harmel, L. A. Abriata, L. A. Hothorn, D. Fiedler, S. Hiller, M. Hothorn, Inositol pyrophosphates promote the interaction of SPX domains with the coiled-coil motif of PHR transcription factors to regulate plant phosphate homeostasis. *Nat Commun*. 12, 384 (2021).

37. Y. Han, A. A. Reyes, S. Malik, Y. He, Cryo-EM structure of SWI/SNF complex bound to a nucleosome. *Nature*. 579, 452–455 (2020).

38. P. Korber, S. Barbaric, The yeast PHO5 promoter: from single locus to systems biology of a paradigm for gene regulation through chromatin. *Nucleic Acids Res*. 42, 10888–10902 (2014).

39. P. D. Gregory, A. Schmid, M. Zavari, M. Münsterkötter, W. Hörz, Chromatin remodelling at the PHO8 promoter requires SWI–SNF and SAGA at a step subsequent to activator binding. *Embo J*. 18, 6407–6414 (1999).

40. X. Li, C. Gu, S. Hostachy, S. Sahu, C. Wittwer, H. J. Jessen, D. Fiedler, H. Wang, S. B. Shears, Control of XPR1-dependent cellular phosphate efflux by InsP8 is an exemplar for functionally-exclusive inositol pyrophosphate signaling. *Proc National Acad Sci*. 117, 3568–3574 (2020).

41. V. T. Phan, V. W. Ding, F. Li, R. J. Chalkley, A. Burlingame, F. McCormick, The RasGAP Proteins Ira2 and Neurofibromin Are Negatively Regulated by Gpb1 in Yeast and ETEA in Humans. *Mol Cell Biol*. 30, 2264–2279 (2010).

42. R. P. Mason, M. Casu, N. Butler, C. Breda, S. Campesan, J. Clapp, E. W. Green, D. Dhulkhed, C. P. Kyriacou, F. Giorgini, Glutathione peroxidase activity is neuroprotective in models of Huntington's disease. *Nat Genet*. 45, 1249–1254 (2013).

43. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. R. Ovchinnikov, J. Zheng, T. Ness, S. Banjade, S. R. Bagde, V. Stancheva, X. Li, K. Liu, Z. Zheng, D. Barerro, U. Roy, I. S. Fernandez, B. Szakal, D. Branzei, E. C. Greene, S. Biggins, S. Keeney,

E. A. Miller, J. C. Fromme, T. Hendrickson, Q. Cong, D. Baker, Structures of core eukaryotic protein complexes (2021), doi:10.1101/2021.09.30.462231.

44. N. H. Cho, K. C. Cheveralls, A.-D. Brunner, K. Kim, A. C. Michaelis, P. Raghavan, H. Kobayashi, L. Savy, J. Y. Li, H. Canaj, J. Y. S. Kim, E. M. Stewart, C. Gnann, F. McCarthy, J. P. Cabrera, R. M. Brunetti, B. B. Chhun, G. Dingle, M. Y. Hein, B. Huang, S. B. Mehta, J. S. Weissman, R. Gómez-Sjöberg, D. N. Itzhak, L. A. Royer, M. Mann, M. D. Leonetti, *Biorxiv*, in press, doi:10.1101/2021.03.29.437450.

45. I. Yofe, U. Weill, M. Meurer, S. Chuartzman, E. Zalckvar, O. Goldman, S. Ben-Dor, C. Schütze, N. Wiedemann, M. Knop, A. Khmelinskii, M. Schuldiner, One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat Methods*. 13, 371–8 (2016).

46. M. Heusel, I. Bludau, G. Rosenberger, R. Hafen, M. Frank, A. Banaei-Esfahani, A. Drogen, B. C. Collins, M. Gstaiger, R. Aebersold, Complex-centric proteome profiling by SEC-SWATH-MS. *Mol Syst Biol*. 15, e8438 (2019).

47. I. Bludau, Discovery–Versus Hypothesis–Driven Detection of Protein–Protein Interactions and Complexes. *Int J Mol Sci*. 22, 4450 (2021).

48. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 373, 871–876 (2021).

49. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*, 1–11 (2021).

50. F. Meier, A.-D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, N. Bache, O. Hoerning, J. Cox, O. Räther, M. Mann, Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics*. 17, 2534–2545 (2018).

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

52. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 26, 1367–1372 (2008).

53. N. Prianichnikov, H. Koch, S. Koch, M. Lubeck, R. Heilig, S. Brehmer, R. Fischer, J. Cox, *Mol Cell Proteom Mcp*, in press, doi:10.1074/mcp.tir119.001720.

54. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*. 32, 223–226 (2014).

55. Q. G. Gianetto, Y. Couté, C. Bruley, T. Burger, Uses and misuses of the fudge factor in quantitative discovery proteomics. *Proteomics*. 16, 1955–1960 (2016).

56. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 13, 2498–2504 (2003).

57. D. Pratt, J. Chen, D. Welker, R. Rivas, R. Pillich, V. Rynkov, K. Ono, C. Miello, L. Hicks, S. Szalma, A. Stojmirovic, R. Dobrin, M. Braxenthaler, J. Kuentzer, B. Demchak, T. Ideker, NDEx, the Network Data Exchange. *Cell Syst*. 1, 302–305 (2015).

58. M. Kucera, R. Isserlin, A. Arkhangorodsky, G. D. Bader, AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000research*. 5, 1717 (2016).

59. L. Oesper, D. Merico, R. Isserlin, G. D. Bader, WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code Biology Medicine*. 6, 7–7 (2011).

Experimental Methods

Cell growth. To achieve samples with similar cell numbers, pre-cultures of the *S. cerevisiae* GFP-tagged library were grown in YPD media (1% yeast extract, 2% bacto^{TM} peptone, 2% glucose) for two days in 2 mL, u-bottom shaped 96-deep-well plates. This allowed cell concentration convergence of different strains during the slow growing post-exponential phase. Cells were resuspended and 50 µl of each pre-culture was used to inoculate 1.5 mL of fresh YPD media (corresponding to an optical density of 0.5 at 600 nm) in 96-deep-well plates (LoBind®, 2 mL, cat no. 0030504305, Eppendorf AG, Hamburg, Germany). Plates were covered with an air permeable membrane and incubated while shaking at 300 rpm and 30 °C for 6 hours. This allowed the progression through the lag phase and three cell cycles followed by harvesting under standard growth conditions. Cells were pelleted in the 96-deep-well plates by centrifugation at 3500 rpm (= 2451 g) for 5 min. The supernatant was discarded by fast decanting and quick dabbing on paper towels. Plates with pellets were sealed with plastic covers and stored at -80 °C until cell lysis.

Cell lysis. Dee-well plates with cell pellets were thawed on ice for 5 min. 100 µl of glass beads (0.5 mm, acid-washed, cat no. G8772, Merck KGaA, Darmstadt, Germany) were added to each well using a 96-well bead dispenser (LabTIE International, Veenendaal, Netherlands). After 5 min 250 µl of 4 °C cold lysis buffer (50 mM Tris HCl pH 7.5, 150 mM NaCl, 5% glycerol, 0.05% IGEPAL CA-630, protease inhibitor EDTA-free (cOmplete^{TM}, 1 tablet per 50 mL, cat no. 11873580001, Merck KGaA, Darmstadt, Germany), 1 mM $MgCl_2$, 0.75 U/µL in-house *Serratia marcescens* endonuclease/SmDNase) were added. Plates were sealed using a heat sealer (S200, cat no. 5392000005, Eppendorf AG, Hamburg, Germany), the low profile plate adapter (cat no. 5392070020, Eppendorf AG, Hamburg, Germany) and transparent heat sealing films (cat no. 0030127838, Eppendorf AG, Hamburg, Germany) for 2 sec at 180 °C and immediately put back on ice. Cell lysis was performed within the 96-deep-well plates at 4 °C via bead-beating (2010 Geno/Grinder®, SPEX SamplePrep, Metuchen, NJ) for 4 cycles of 1.5 min each at 1750 rpm. Plates were cooled in ice water and covered with ice for 7 min in-between cycles and for 10 min after the last cycle. 4 plates were processed in parallel during bead-beating and top and bottom positions were switched at each cycle. Cell debris was spun down at max speed (4300 rpm = 4347 g) for 10 min at 4 °C. Plates were carefully put back on ice and immediately used for the pull-down protocol (Fig. 1A).

Interactor enrichment: Pull-downs and all sample handling steps were performed at 4 °C. Anti-GFP nanobody coated 96-well microtiter plates were custom made and optimized for this protocol allowing efficient and high reproducible "in-well" digestion, and mass spectrometry compatibility (plates are now commercially available as: GFP-Trap® Multiwell Plate, cat no. gtp-96, Chromotek GmbH, Martinsried, Germany). Plates were prepared with 200 µL wash buffer 1 (50 mM Tris HCl pH 7.5, 150 mM NaCl, 5% glycerol, 0.05% IGEPAL CA-630) per well on a shaker for 1 min at 800 rpm followed by removal of the buffer. The cell lysates were carefully transferred from the 96-deep-well plates by slow uptake of 175 µL supernatant without dislodging glass beads nor the cell debris pellet to the GFP-Trap plate. The GFP-Trap plate was incubated for 1 h at 800 rpm on a small stroke (3 mm) shaker (TiMix 5 control, Edmund Bühler GmbH, Tübingen, Germany) to enrich for GFP-tagged proteins and their interactors. Cell lysates were discarded and plate wells were washed twice with 200 µL wash buffer 1 and twice with wash buffer 2 (50 mM Tris HCl pH 7.5, 150 mM NaCl, 5% glycerol). To allow stable binding of unspecific background proteins – an important factor for label-free quantification – wash buffer was added slowly, and plates were not shaken during wash steps. Emptied, protein-enriched plates were covered and stored at -80 °C until mass spectrometry sample preparation (Fig. 1A).

Sample preparation for mass spectrometry. Protein-enriched GFP-Trap plates were brought to room temperature and 50 µL of digestion mix 1 (4.5 M urea, 1.5 M thiourea, 10 mM Tris HCl pH 8.5, 3 mM dithiothreitol, 2 ng/µL LysC) were added per well. Plates were incubated at 30 °C and 1000 rpm on a small stroke (3 mm) shaker. After 3 h, 100 µL of digestion mix 2 (10 mM Tris HCl pH 8.5, 7.5 mM chloroacetamide, 2 ng/µL LysC) were added and microtiter plates and lids were sealed with parafilm®. The plates were incubated overnight at 30 °C/800 rpm. The reaction was stopped and the sample was acidified with 15 µL of 10% TFA per well. Plates with peptides were stored at -80 °C till sample loading on EvoTips (Evosep, Odense, Denmark) (Fig. 1A).

Loading of peptide samples on Evotips. Evotips (Evosep, Odense, Denmark) were activated for 5 min in a 1-propanol Evotips-box reservoir at room temperature (RT), followed by a wash step with 50 µl buffer B (acetonitrile (ACN) with 0.1 % formic acid (FA)) and centrifugation at 500 g for 1 min at RT. The flow-through was discarded and Evotips were placed back into 1-Propanol. Evotips were conditioned with 50 µL of buffer A (ddH_2O with 0.1 % FA) and centrifugation at 500 g for 1.5 min at RT and were placed in a container with buffer A. 40 µL of thawed peptide sample were loaded and Evotips were centrifuged at 500 g for 1.5 min at RT and placed back in a container with buffer A. 200 µL of buffer A were added and partially washed through the Evotips by centrifugation at 500 g for 50 s. Evotips boxes with buffer A at the container bottom were placed on the Evosep One liquid chromatography (LC) platform (Evosep, Odense, Denmark) for LC-MS analysis. Pull-downs were acquired in technical duplicates and the injection order was reversed after the first measurement (Fig. 1A).

Liquid-chromatography. For separating peptides by hydrophobicity and eluting them into the mass spectrometer, we used the EvoSep One LC system and analyzed the yeast interactome pull-down proteomes with the standardized 21 min (60 samples per day) gradient. We employed a 15 cm × 150 µm inner diameter column with 1.9 µm C18 beads (PepSep, Marslev, Denmark) coupled to a 20 µm ID electrospray emitter (Bruker Daltonik GmbH, Bremen, Germany). The column was replaced between replicate measurements. Mobile phases A and B were 0.1 % FA in water and 0.1 % FA in ACN, respectively. The EvoSep system was coupled online to a trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer *(*50*)* (timsTOF Pro, Bruker Daltonik GmbH, Bremen, Germany) via a nano-electrospray ion source (Captive spray, Bruker Daltonik GmbH, Bremen, Germany). A 24-fraction library of wild-type *S. cerevisiae* was generated using the high-pH reversed-phase “spider-fractionator” *(*51*)* and data were acquired using the same sample set-up.

Mass spectrometry. Mass spectrometric analysis was performed in a data-dependent (dda) PASEF mode. For ddaPASEF, 1 MS1 survey TIMS-MS and 4 PASEF MS/MS scans were acquired per acquisition cycle. The cycle overlap for precursor scheduling was set to 2. Ion accumulation and ramp time in the dual TIMS analyzer was set to 50 ms each and we analyzed the ion mobility range from $1/K_0$ = 1.3 Vs cm^{-2} to 0.8 Vs cm^{-2}. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at $1/K_0$ = 1.6 VS cm^{-2} to 20 eV at $1/K_0$ = 0.6 Vs cm^{-2}. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker

Daltonik GmbH, Bremen, Germany). Precursors for MS/MS were picked at an intensity threshold of 2,000 arbitrary units (a.u.) and re-sequenced until reaching a “target value” of 24,000 a.u. considering a dynamic exclusion of 40 s elution. The capillary voltage was set to 1,750 V and dry gas temperature to 180 °C.

Raw data processing. MS raw files were processed using MaxQuant (v1.6.17.0) (52, 53), which extracts features from four-dimensional isotope patterns and associated MS/MS spectra, on a computing cluster (SUSE Linux Enterprise Server 15 SP2) utilizing UltraQuant (github.com/kentsisresearchgroup/UltraQuant). To allow processing in an acceptable time frame, RAW files were handled in 5 parallel batches of approximately 1700 files each containing plates equally distributed across the measurement period. Files were searched against the *S. cerevisiae* Uniprot databases (UP000002311_559292; canonical and isoform, reviewed-sp and unreviewed-tr from 02/2020). For high significance identification the false-discovery rates were reduced and controlled at 0.1% both on peptide spectral match (PSM) and protein levels. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification, while limiting the maximum peptide mass to 4,800 Da. Enzyme specificity was set to LysC cleaving C-terminal to lysine. A maximum of two missed cleavages were allowed. The parameter “type” was set to “TIMS-DDA” with “TIMS half width” at 4. The instrument was set to “Bruker TIMS” and main search peptide tolerance reduced to 8 ppm, the max. charge set to 5 and min. peak length to 3. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (4D-MBR) using a narrow elution match time window of 12 s and a reduced ion mobility window of 0.01 $1/K_0$. Protein quantification was performed by label-free quantification using a minimum ratio count of 2. The 24-fraction library was added as an additional parameter group with the same group-specific settings, but LFQ disabled and “separate LFQ in parameter groups” under global parameters enabled. The writing of additional tables was disabled for performance reasons.

Raw data availability. All mass spectrometry raw data and MaxQuant output tables have been deposited to the ProteomeXchange Consortium (54) via the PRIDEpartner repository with the dataset identifier available upon publication.

Data processing and normalization. Twelve outdated samples of the GFP library were eliminated. These included wrongly annotated ORFs that were merged with others: YAR044W, YPR090W, YDR474C, YFR024C, YJL021C, YJL017W, YGL046W, YFL006W, YGR272C, YBR100W, YJL018W, YJL012C-A. After the removal of potential contaminants, reverse and "only identified by site" hits, MaxQuant proteinGroups.txt output files from the 5 batches were merged using the majority protein IDs column. Values were filtered for two valid values within at least one replicate group. To adjust for potential differences between the 5 MaxQuant batches caused by the parallel applied label free normalization algorithm and for potential handling batch effects between 96-well plates, values were median normalized if there were more than 5% of valid values in each of the corresponding groups.

Missing value imputation. Missing values were imputed in a two-tiered approach. For proteins with measured values in more than 5% of all samples (or minimally 400 samples), a protein-specific missing value imputation approach was used. Here, a random value was sampled from a normal distribution with following properties: mean = median of all measured intensity values for the given protein, standard deviation = standard deviation of all measured intensity values for the given protein. Lower and upper bounds for the normal distribution were set to three standard deviations from the mean and minimally to zero. The function "rtruncnorm" from the R library "truncnorm" was employed. For proteins with less than 5% valid values (or in less than 400 samples), global metrics were employed for missing value imputation. Here, missing values were sampled from a normal distribution with the following parameters: mean = mean of all quantified values across all proteins and samples minus 1.8 times the standard deviation, standard deviation = the standard deviation of all quantified values across all proteins and samples multiplied by 0.3.

Protein correlation. Due to the large sample number that would negatively influence correlation, we chose a subsampling approach: For each protein pair across the sample profile, the top 2% of samples with the highest intensities for both proteins were selected (resulting in 2-4% depending on their overlap) and complemented by twice the number of randomly selected samples as background. The selected subset of samples was used to calculate the Pearson correlation coefficients of the protein pair (Fig. 1C). The effect of weighted correlation can be visualized by enabling "subsample values" under protein correlation in our web application (*yeast-interactome.org*). Since the distributions of correlation coefficients varies between proteins and in order to define a universal cut-off for significant correlations, correlation coefficients were normalized via row wise z-scoring. A z-scored Pearson correlation coefficient above 4 and 5 therefore corresponds to a chance probability of below $3.2*10^{-5}$ and $2.9*10^{-7}$, respectively.

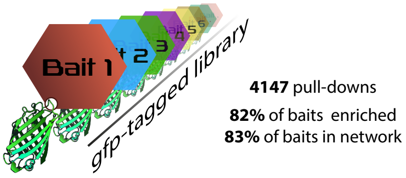
Enrichment analysis. A two-tailed Welch's t-test was performed on each replicate-grouped pull-down sample using all corresponding complement samples as a combined control *(*11*)*. Within the combined control group, samples with the highest bait correlation (top 5%) were excluded in order to provide a bait-unrelated control. FDR cutoff-lines were calculated using an analytical approach using an S0-parameter of 0.5 *(*55*)*.

Network generation. Interactions for the first two layers of evidence (forward and reverse pull-down) were defined between bait proteins and significantly enriched prey proteins from the t-tests. They were scored based on their FDR of 5%, 1%, 0.1% and 0.01% at 1, 2, 3 and 4, respectively ("score_FDR"). For the third layer of evidence, an interaction for z-scored Pearson correlation coefficients above 4 and 5 was scored at 1 and 2, respectively ("score_cor"). All three layers of evidence were combined into a single interaction score ranging from 1-10 ("score_FDR+cor"), thereby weighting interactions based on their experimental significance (Fig. 1C). Networks were created and exported into Cytoscape *(*56*)* for further analysis and visualization strategies. The network was filtered for interactions with a combined score equal to or above 2, thereby excluding interactions based only on a single t-test with an FDR of above 1% or a z-scored Pearson correlation coefficient of below 5. The Markov clustering algorithm was applied using the interaction score as edge weight and a granularity parameter of 2.5 while retaining inter-cluster edges. The "CompoundSpringEmbedder" (CoSE) layout algorithm was applied to single clusters. The network including edges (interactions) and nodes (proteins), annotations, and layouts can be downloaded as Cytoscape session at (*www.yeast-interactome.org*) or at the NDEX network database *(*57*)* via the UID available upon publication.

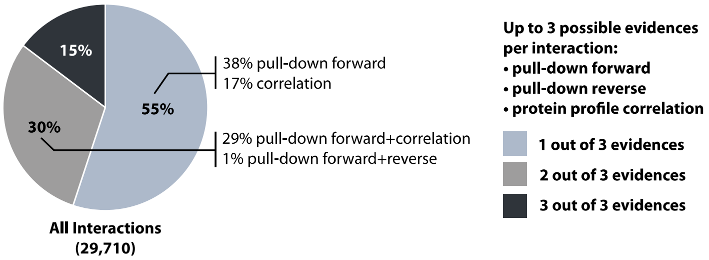
Organelle based mapping of clusters. Within the Cytoscape group preferences the attribute aggregation was enabled and "visualization for group" were set to "none". The WordCloud "minimum word occurrence" and the "max. words per label" was set to 1, and normalization to 0. To generate outcome with location specific words only, the excluded words list was extended by following terms: apparatus, matrix, membrane, intermembrane, chromosome, ii, protein, anchor, coated, cytoplasmic, iv, lipid, pass, peripheral, secreted, pit, side, single, centromere, type, endomembrane, tip, reticulum, body, localizes, kinetochore, gpi, note, neck, prospore, granule, replication. The "AutoAnnotate" plugin *(*58*)* was used to generate localization-based name for each markov cluster utilizing WordCloud *(*59*)* (most abundant word within "Subcellular localization [CC]"). Collapsed localization (collapse singleton clusters enabled) based labeled groups were organized using the "Boundary Layout" using self-defined areas. Node repulsion was increased to 1,000,000. For cluster annotation the standard complex name from EMBL Complexportal was used. For each cluster the two most frequent names were used, (minimum word occurrence: 2). The image of the background cell in Figure 4, the Cytoscape session and the web application is an adopted version from SwissBioPics by the Swiss-Prot group of the SIB Swiss Institute of Bioinformatics. Cell image in Figure 2A was created with BioRender.com.

Network comparisons. Network comparison analysis was performed in Python 3.8.1. Tabular data was loaded via the pandas package (1.3.1) and converted to a network via NetworkX (2.6.2). To calculate "Betweeness" and "Degree Centrality", the respective NetworkX functions were used. To perform community analysis, a Python implementation of the Louvain algorithm was used (https://github.com/taynaud/python-louvain, version 0.15). Cumulative distribution functions were plotted using the matplotlib-library (3.4.2) and NumPy (1.20.3). Reference datasets were downloaded from the Stanford Large Network Dataset Collection (http://snap.stanford.edu/data/) and the BioPlex Interactome homepage (https://bioplex.hms.harvard.edu/interactions.php). The accompanying notebook is available as Supplementary File "Yeast_Network_comparisons.ipynb". Gene annotation enrichment was performed using the 1D tool in Perseus (v.1.6.7.0). Annotation terms were filtered for 5% FDR (Benjamini–Hochberg correction) and a score above 0.

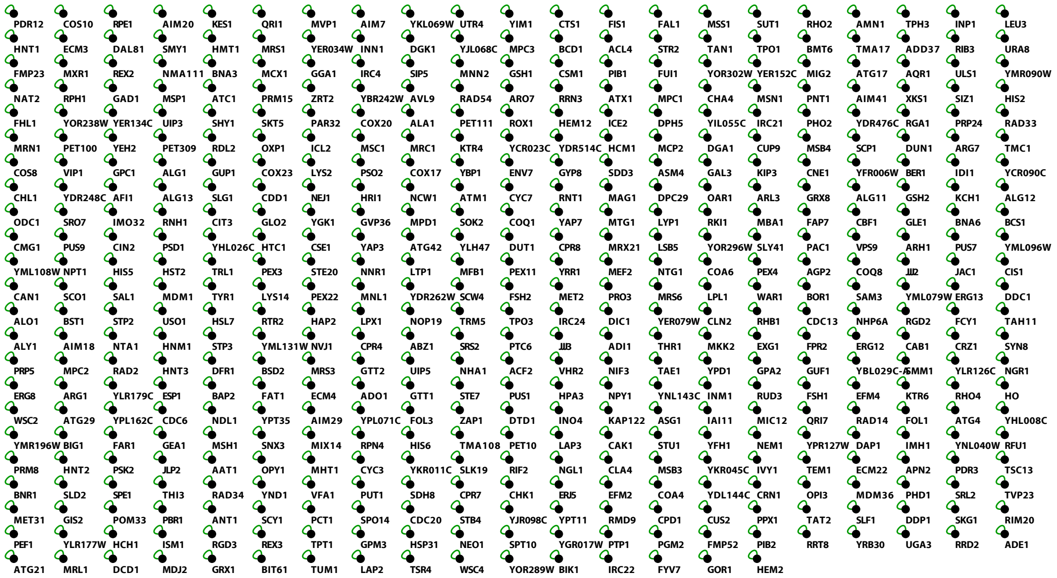
Supplementary Figures



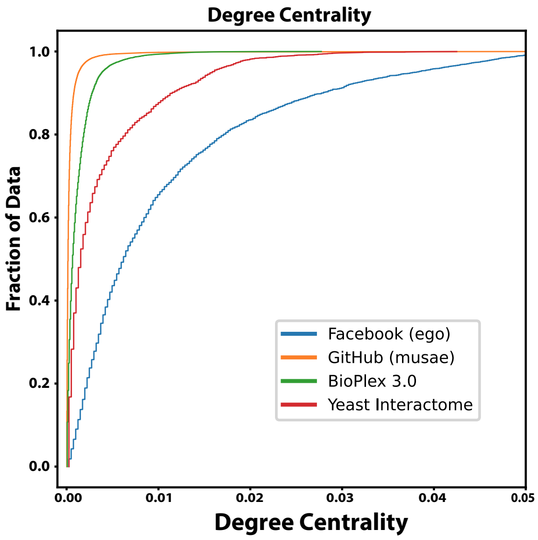
Supplementary Figure 1. Schematic of the GFP-tagged library. 4,147 different endogenous c-terminally tagged yeast strains *(7)* were used for 4,147 independent pull-down experiments. Each strain therefore allows the purification of the individually tagged protein (bait) and its specific interactors. The original library of 4159 strains was reduced by twelve strains to 4147, due to updates in ORF annotations (see methods: Data processing and normalization).



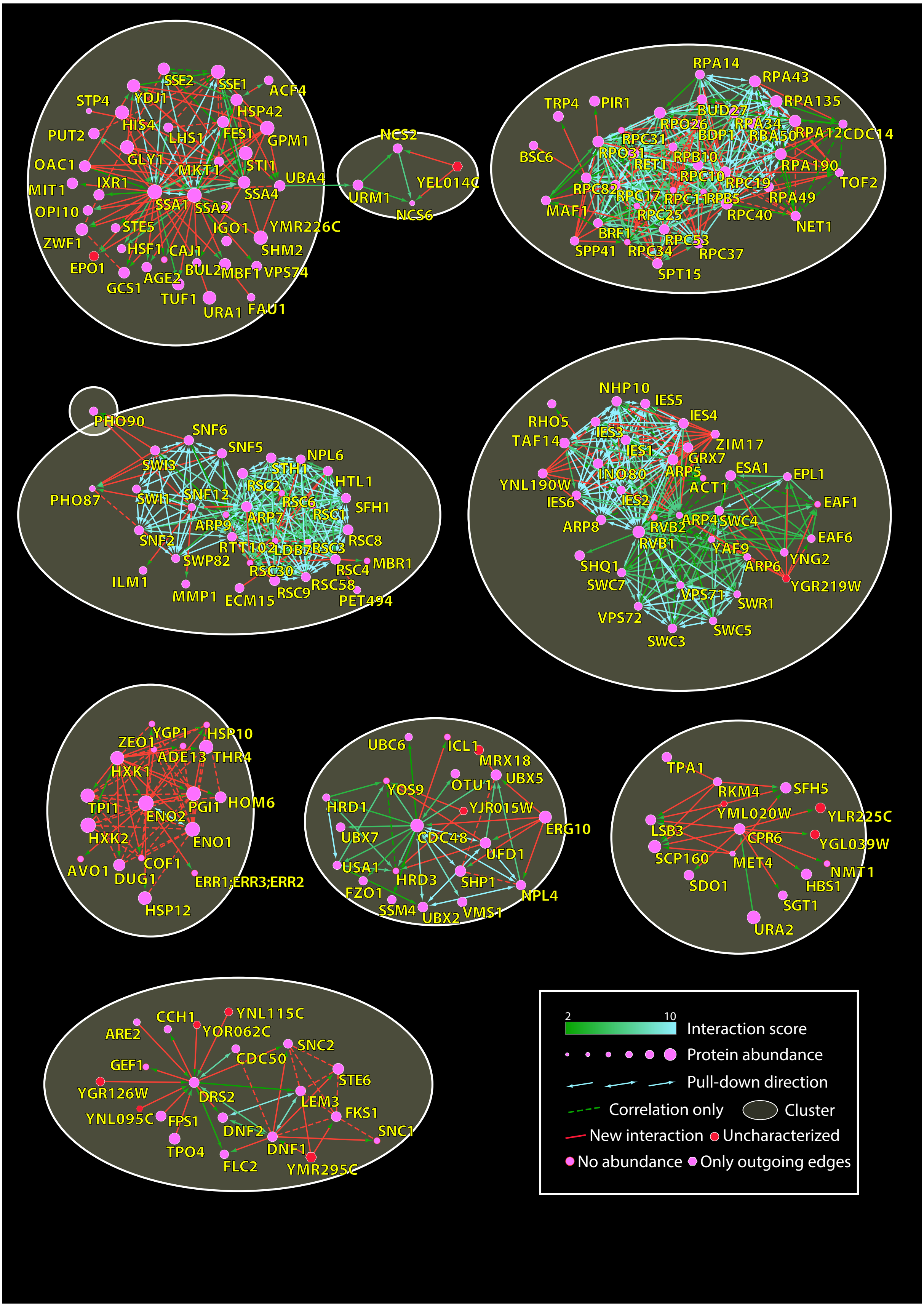
Supplementary Figure 2. Detailed proportion of interactions backed by multiple layers of evidence


PDR12 COS10 RPE1 AIM20 KES1 QRI1 MVP1 AIM7 YKL069W UTR4 YIM1 CTS1 FIS1 FAL1 MSS1 SUT1 RHO2 AMN1 TPH3 INP1 LEU3
HNT1 ECM3 DAL81 SMY1 HMT1 MRS1 YER034W INN1 DGK1 YJL068C MPC3 BCD1 ACL4 STR2 TAN1 TPO1 BMT6 TMA17 ADD37 RIB3 URA8
FMP23 MXR1 REX2 NMA111 BNA3 MCX1 GGA1 IRC4 SIP5 MNN2 GSH1 CSM1 PIB1 FUI1 YOR302W YER152C MIG2 ATG17 AQR1 ULS1 YMR090W
NAT2 RPH1 GAD1 MSP1 ATC1 PRM15 ZRT2 YBR242W AVL9 RAD54 ARO7 RRN3 ATX1 MPC1 CHA4 MSN1 PNT1 AIM41 XKS1 SIZ1 HIS2
FHL1 YOR238W YER134C UIP3 SHY1 SKT5 PAR32 COX20 ALA1 PET111 ROX1 HEM12 ICE2 DPH5 YIL055C IRC21 PHO2 YDR476C RGA1 PRP24 RAD33
MRN1 PET100 YEH2 PET309 RDL2 OXP1 ICL2 MSC1 MRC1 KTR4 YCR023C YDR514C HCM1 MCP2 DGA1 CUP9 MSB4 SCP1 DUN1 ARG7 TMC1
COS8 VIP1 GPC1 ALG1 GUP1 COX23 LYS2 PSO2 COX17 YBP1 ENV7 GYP8 SDD3 ASM4 GAL3 KIP3 CNE1 YFR006W BER1 IDI1 YCR090C
CHL1 YDR248C AFI1 ALG13 SLG1 CDD1 NEJ1 HRI1 NCW1 ATM1 CYC7 RNT1 MAG1 DPC29 OAR1 ARL3 GRX8 ALG11 GSH2 KCH1 ALG12
ODC1 SRO7 IMO32 RNH1 CIT3 GLO2 YGK1 GVP36 MPD1 SOK2 COQ1 YAP7 MTG1 LYP1 RKI1 MBA1 FAP7 CBF1 GLE1 BNA6 BCS1
CMG1 PUS9 CIN2 PSD1 YHL026C HTC1 CSE1 YAP3 ATG42 YLH47 DUT1 CPR8 MRX21 LSB5 YOR296W SLY41 PAC1 VPS9 ARH1 PUS7 YML096W
YML108W NPT1 HIS5 HST2 TRL1 PEX3 STE20 NNR1 LTP1 MFB1 PEX11 YRR1 MEF2 NTG1 COA6 PEX4 AGP2 COQ8 JJJ2 JAC1 CIS1
CAN1 SCO1 SAL1 MDM1 TYR1 LYS14 PEX22 MNL1 YDR262W SCW4 FSH2 MET2 PRO3 MRS6 LPL1 WAR1 BOR1 SAM3 YML079W ERG13 DDC1
ALO1 BST1 STP2 USO1 HSL7 RTR2 HAP2 LPX1 NOP19 TRM5 TPO3 IRC24 DIC1 YER079W CLN2 RHB1 CDC13 NHP6A RGD2 FCY1 TAH11
ALY1 AIM18 NTA1 HNM1 STP3 YML131W NVJ1 CPR4 ABZ1 SRS2 PTC6 JJJ3 ADI1 THR1 MKK2 EXG1 FPR2 ERG12 CAB1 CRZ1 SYN8
PRP5 MPC2 RAD2 HNT3 DFR1 BSD2 MRS3 GTT2 UIP5 NHA1 ACF2 VHR2 NIF3 TAE1 YPD1 GPA2 GUF1 YBL029C-A SMM1 YLR126C NGR1
ERG8 ARG1 YLR179C ESP1 BAP2 FAT1 ECM4 ADO1 GTT1 STE7 PUS1 HPA3 NPY1 YNL143C INM1 RUD3 FSH1 EFM4 KTR6 RHO4 HO
WSC2 ATG29 YPL162C CDC6 NDL1 YPT35 AIM29 YPL071C FOL3 ZAP1 DTD1 INO4 KAP122 ASG1 IAI11 MIC12 QRI7 RAD14 FOL1 ATG4 YHL008C
YMR196W BIG1 FAR1 GEA1 MSH1 SNX3 MIX14 RPN4 HIS6 TMA108 PET10 LAP3 CAK1 STU1 YFH1 NEM1 YPR127W DAP1 IMH1 YNL040W RFU1
PRM8 HNT2 PSK2 JLP2 AAT1 OPY1 MHT1 CYC3 YKR011C SLK19 RIF2 NGL1 CLA4 MSB3 YKR045C IVY1 TEM1 ECM22 APN2 PDR3 TSC13
BNR1 SLD2 SPE1 THI3 RAD34 YND1 VFA1 PUT1 SDH8 CPR7 CHK1 ERJ5 EFM2 COA4 YDL144C CRN1 OPI3 MDM36 PHD1 SRL2 TVP23
MET31 GIS2 POM33 PBR1 ANT1 SCY1 PCT1 SPO14 CDC20 STB4 YJR098C YPT11 RMD9 CPD1 CUS2 PPX1 TAT2 SLF1 DDP1 SKG1 RIM20
PEF1 YLR177W HCH1 ISM1 RGD3 REX3 TPT1 GPM3 HSP31 NEO1 SPT10 YGR017W PTP1 PGM2 FMP52 PIB2 RRT8 YRB30 UGA3 RRD2 ADE1
ATG21 MRL1 DCD1 MDJ2 GRX1 BIT61 TUM1 LAP2 TSR4 WSC4 YOR289W BIK1 IRC22 FYV7 GOR1 HEM2

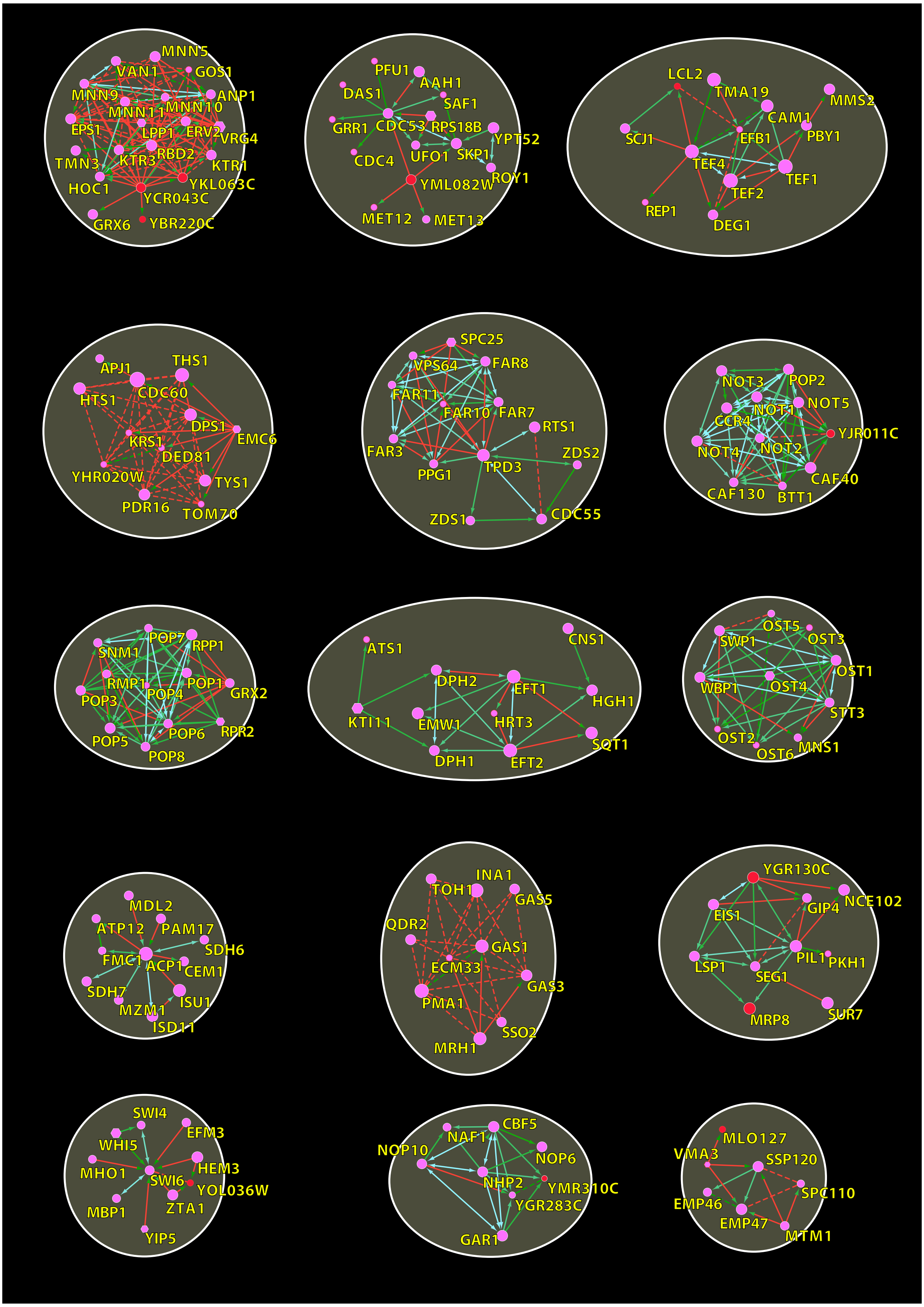
Supplementary Figure 3. “Asocial” proteins. Representation of 478 significantly enriched and detected bait proteins that lack any significant interactor under given conditions in this study. Green edges depict self-edges.



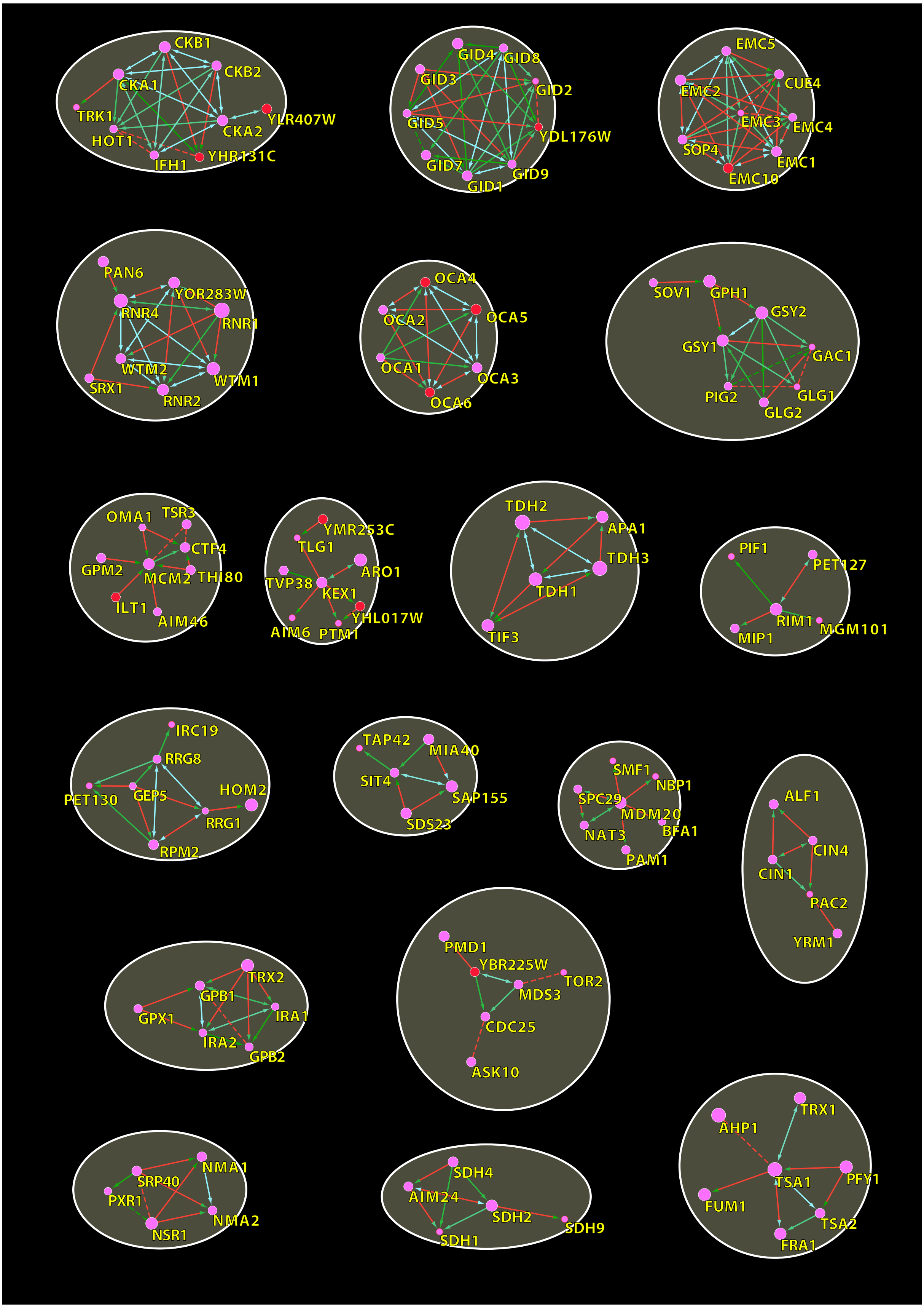
Supplementary Figure 4. Cumulative distribution function of the degree centrality. Comparison of different complex networks: *S. cerevisiae* has more influential (high degree centrality) nodes than BioPlex and GitHub, and less than Facebook.



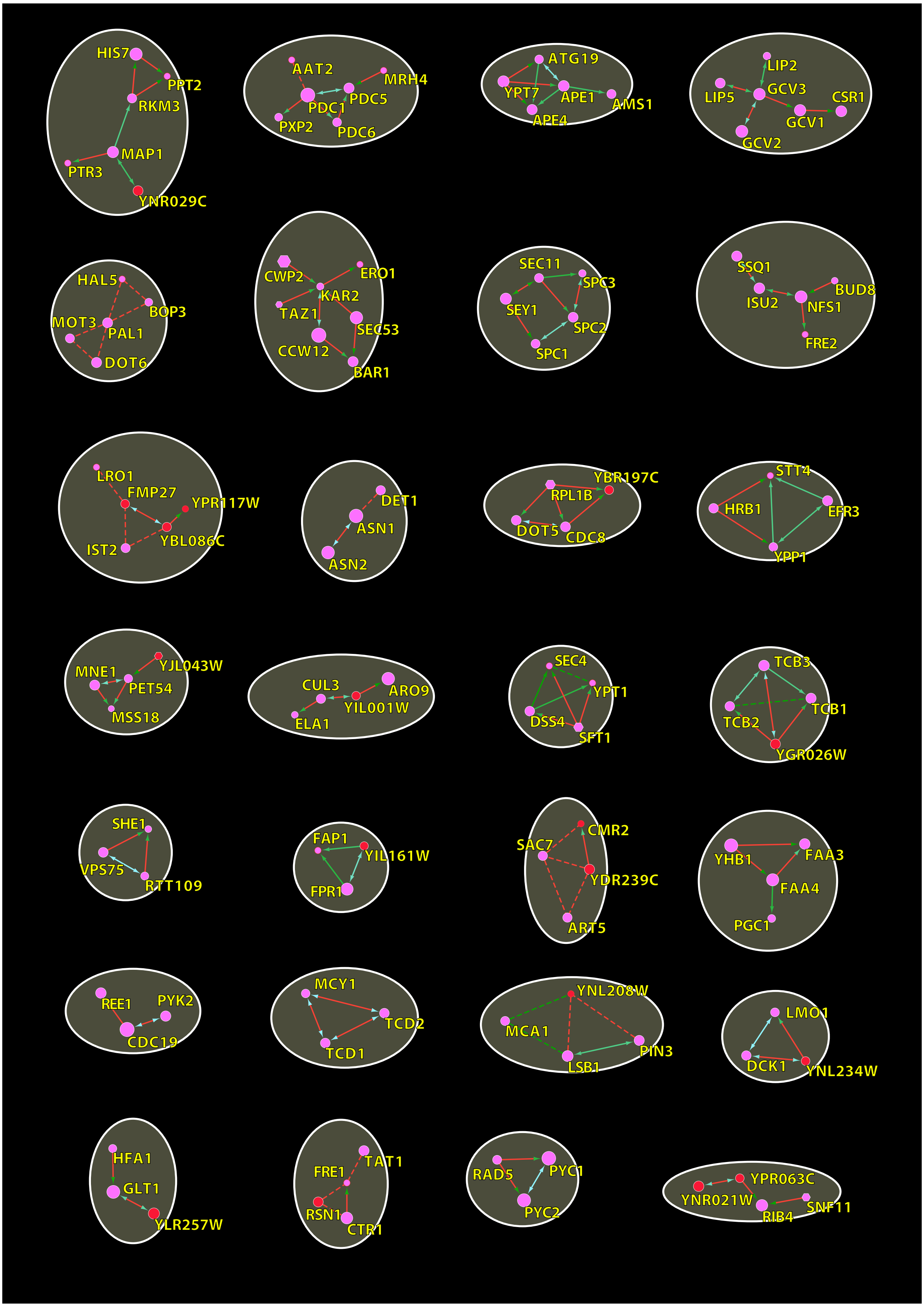
Supplementary Figure 6 (part 1/5). Extended selection of clusters involving proteins with novel interactions and/or uncharacterized proteins supported by multiple layers of evidence.



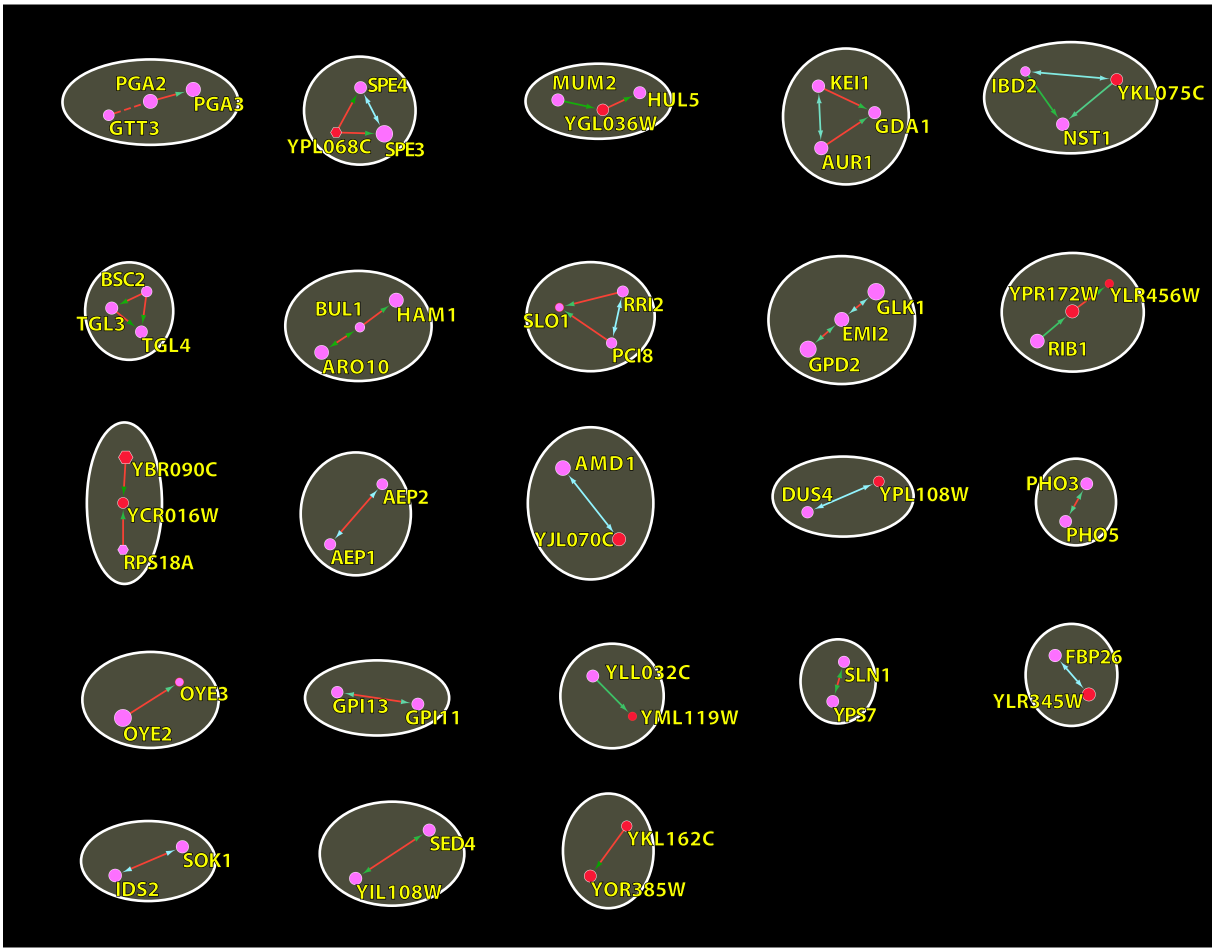
Supplementary Figure 6 (part 2/5). Continuation of Figure 6 part 2.



Supplementary Figure 6 (part 3/5). Continuation of Figure 6 part 3.



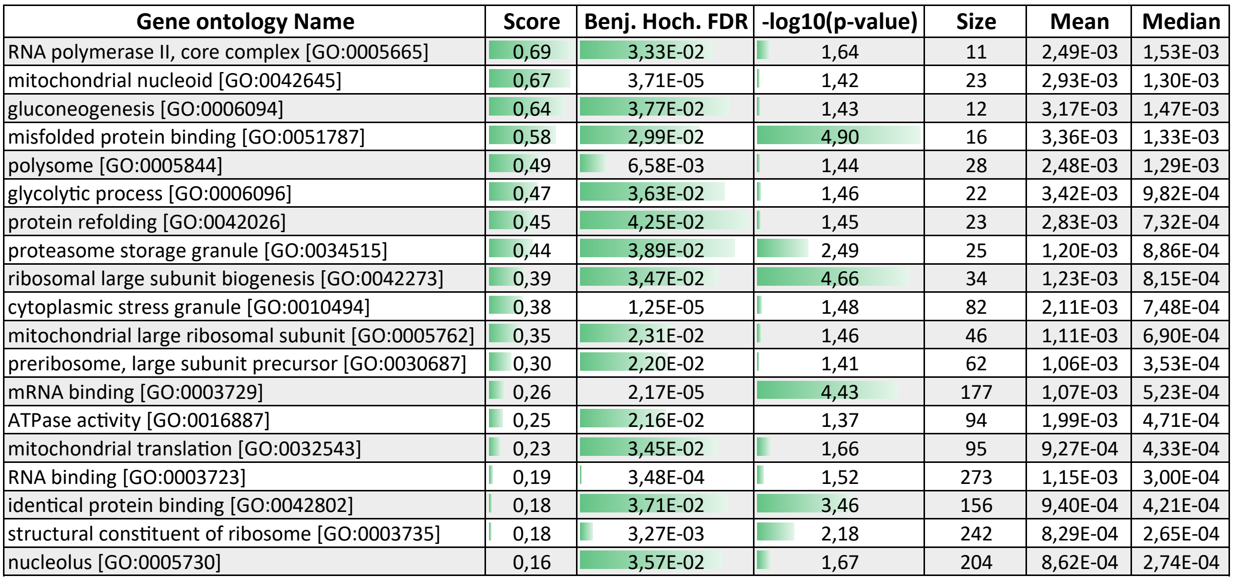
Supplementary Figure 6 (part 4/5). Continuation of Figure 6 part 4.



Supplementary Figure 6 (part 5/5). Continuation of Figure 6 part 5.

Supplementary Tables

Table 1. Gene ontology term enrichment



Gene ontology Name	Score	Benj. Hoch. FDR	-log10(p-value)	Size	Mean	Median
RNA polymerase II, core complex [GO:0005665]	0,69	3,33E-02	1,64	11	2,49E-03	1,53E-03
mitochondrial nucleoid [GO:0042645]	0,67	3,71E-05	1,42	23	2,93E-03	1,30E-03
gluconeogenesis [GO:0006094]	0,64	3,77E-02	1,43	12	3,17E-03	1,47E-03
misfolded protein binding [GO:0051787]	0,58	2,99E-02	4,90	16	3,36E-03	1,33E-03
polysome [GO:0005844]	0,49	6,58E-03	1,44	28	2,48E-03	1,29E-03
glycolytic process [GO:0006096]	0,47	3,63E-02	1,46	22	3,42E-03	9,82E-04
protein refolding [GO:0042026]	0,45	4,25E-02	1,45	23	2,83E-03	7,32E-04
proteasome storage granule [GO:0034515]	0,44	3,89E-02	2,49	25	1,20E-03	8,86E-04
ribosomal large subunit biogenesis [GO:0042273]	0,39	3,47E-02	4,66	34	1,23E-03	8,15E-04
cytoplasmic stress granule [GO:0010494]	0,38	1,25E-05	1,48	82	2,11E-03	7,48E-04
mitochondrial large ribosomal subunit [GO:0005762]	0,35	2,31E-02	1,46	46	1,11E-03	6,90E-04
preribosome, large subunit precursor [GO:0030687]	0,30	2,20E-02	1,41	62	1,06E-03	3,53E-04
mRNA binding [GO:0003729]	0,26	2,17E-05	4,43	177	1,07E-03	5,23E-04
ATPase activity [GO:0016887]	0,25	2,16E-02	1,37	94	1,99E-03	4,71E-04
mitochondrial translation [GO:0032543]	0,23	3,45E-02	1,66	95	9,27E-04	4,33E-04
RNA binding [GO:0003723]	0,19	3,48E-04	1,52	273	1,15E-03	3,00E-04
identical protein binding [GO:0042802]	0,18	3,71E-02	3,46	156	9,40E-04	4,21E-04
structural constituent of ribosome [GO:0003735]	0,18	3,27E-03	2,18	242	8,29E-04	2,65E-04
nucleolus [GO:0005730]	0,16	3,57E-02	1,67	204	8,62E-04	2,74E-04

Gene ontology term enrichment on betweenness-centrality of nodes (proteins) in the network (1-dimensional annotation enrichment, FDR < 5%, score > 0).

2.2 Article 2: OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization

N. H. Cho[1,]†, K. C. Cheveralls[1,]†, A.-D. Brunner[2,]†, K. Kim[1,]†, **A. C. Michaelis[2,]†**, P. Raghavan[1,]†, H. Kobayashi[1], L. Savy[1], J. Y. Li[1], H. Canaj[1], J. Y. S. Kim[1], E. Stewart[1], C. Gnann[1,3], F. McCarthy[1], J. P. Cabrera[1], R. M. Brunetti[4], B. B. Chhun[1], G. Dingle[5], M. Y. Hein[1], B. Huang[1,4,5], S. B. Mehta[1], J. S. Weissman[6,7], R. Gómez-Sjöberg[1], D. N. Itzhak[1], L. A. Royer[1], M. Mann[2,8], M. D. Leonetti[1,]*, (2021). **OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization**. *Biorxiv doi:10.1101/2021.03.29.437450.*

† equal contribution; * correspondence: *manuel.leonetti@czbiohub.org*

The work in this paper – in relation to the one presented before - expands the interactome from yeast to human cells. This large and fruitful collaboration between the Chan Zuckerberg Biohub in San Francisco and the Max-Planck Institute of Biochemistry in Munich under the lead of Manuel Leonetti und Matthias Mann, combined their expertise in fluorescence microscopy with ours in proteomics to generate an unprecedented protein localization and interaction map. By introducing a GFP-tag into 1,300 human HEK293T cells using CRISPR technology, we were able to use the florescent tag for confocal microscopy 3D-image rendering and at the same time for affinity-purification coupled to mass spectrometry for protein interaction detection. Key for this large-scale compatibility is the use of a split-GFP system described by the Weissmann lab *(*20*)*.

The split-GFP only uses a small part of GFP for endogenous tagging namely the last β-strand, strand 11. This is done in cells co-expressing the complement part of GFP (β-strand 1-10) forming a full functional version. Using only this small tag makes it possible to employ a small synthetic ssDNA oligos for CRIPR editing.

This project was performed by Andreas Brunner and myself on the side of the Mann lab for several years throughout almost all of our PhD times. We were responsible for all mass spectrometry-related tasks on the project running many hundreds of test samples and the final dataset of almost 4,000 runs. Due to the large overlap with the yeast interactome project many of my experience and developments gained there could be used to advance the human interactome as well.

The paper is published on *Biorxiv* and an updated version which has currently been resubmitted to *Science* after revision is included in this thesis.

Title:

OpenCell: endogenous tagging for the cartography of human cellular organization

Authors: Nathan H. Cho[1,†], Keith C. Cheveralls[1,†], Andreas-David Brunner[2,†], Kibeom Kim[1,†], André C. Michaelis[2,†], Preethi Raghavan[1,†], Hirofumi Kobayashi[1], Laura Savy[1], Jason Y. Li[1], Hera Canaj[1], James Y.S. Kim[1], Edna Stewart[1], Christian Gnann[1,3], Frank McCarthy[1], Joana P. Cabrera[1], Rachel M. Brunetti[4], Bryant B. Chhun[1], Greg Dingle[5], Marco Y. Hein[1], Bo Huang[1,4,5], Shalin B. Mehta[1], Jonathan S. Weissman[6,7], Rafael Gómez-Sjöberg[1], Daniel N. Itzhak[1], Loïc A. Royer[1], Matthias Mann[2,8], Manuel D. Leonetti[1,*]

Affiliations: [1] Chan Zuckerberg Biohub, San Francisco, USA; [2] Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany; [3] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH – Royal Institute of Technology, Stockholm, Sweden; [4] Department of Biochemistry and Biophysics, University of California, San Francisco, USA; [5] Department of Pharmaceutical Chemistry, University of California, San Francisco USA; [5] Chan Zuckerberg Initiative, Redwood City, USA; [6] Whitehead Institute, Koch Institute and Department of Biology, Massachusetts Institute of Technology, and Howard Hughes Medical Institute, Cambridge, USA; [7] Department of Cellular and Molecular Pharmacology, , University of California, San Francisco, USA; [8] NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

† equal contribution; * correspondence: manuel.leonetti@czbiohub.org

Abstract: **Elucidating the wiring diagram of the human cell is a central goal of the post-genomic era. We combined genome engineering, confocal live-cell imaging, mass spectrometry and data science to systematically map the localization and interactions of human proteins. Our approach provides a data-driven description of the molecular and spatial networks that organize the proteome. Unsupervised clustering of these networks delineates functional communities that facilitate biological discovery, and uncovers that RNA-binding proteins form a specific sub-group defined by unique interaction and localization properties. Furthermore, we discover that remarkably precise functional information can be derived from protein localization patterns, which often contain enough information to identify molecular interactions. Paired with a fully interactive website (opencell.czbiohub.org), we provide a resource for the quantitative cartography of human cellular organization.**

One Sentence Summary: CRISPR-based fluorescent tagging enables a systematic map of localization and interactions for human proteins.

Main Text:

Sequencing the human genome has transformed cell biology by defining the protein parts list that forms the canvas of cellular operation (*1*, *2*). This paves the way for elucidating how the ~20,000 proteins encoded in the genome organize in space and time to define the cell's functional architecture (*3*, *4*). Where does each protein localize within the cell? Can we comprehensively map how proteins assemble into larger functional communities? A main challenge to answering these fundamental questions is that cellular architecture is organized along multiple scales. Therefore, several approaches need to be combined for its elucidation (*5*). In a series of pioneering studies, human protein-protein interactions have been mapped using ectopic expression strategies with yeast two-hybrid (Y2H) (*6*) or epitope tagging coupled to immunoprecipitation-mass spectrometry (IP-MS) (*7*, *8*), while protein localization has been charted using immuno-fluorescence in fixed samples (*9*). A complementary approach is to directly modify genes in a genome by appending sequences that illuminate specific aspects of the corresponding proteins' function (commonly referred to as "endogenous tagging" (*10*)). For example, endogenously tagging a gene with a fluorescent reporter enables to image protein sub-cellular localization in live cells, and supports functional characterization in a native cellular environment (*10*, *11*). The use of endogenous tagging to study the organization of a eukaryotic cell is illustrated by seminal work in the budding yeast *S. cerevisiae*. There, libraries of tagged strains have enabled the comprehensive mapping of protein localization and molecular interactions across the yeast proteome (*12*–*14*). These libraries were made possible by the relative simplicity of homologous recombination and genome engineering in yeast (*15*). In human cells, earlier work has leveraged alternative strategies including expression from bacterial artificial chromosomes (*16*) or central-dogma tagging (*17*) because of the difficulty of site-specific gene editing. CRISPR-mediated genome engineering now allows for homologous recombination-based endogenous tagging to be applied for the interrogation of the human cell (*10*, *11*, *18*).

Here, we combine experimental and analytical strategies to create OpenCell, a proteomic map of human cellular architecture. We generated a library of 1,310 CRISPR-edited HEK293T cell lines harboring fluorescent tags on individual proteins, which we characterized by pairing confocal microscopy and mass spectrometry. Our dataset constitutes the most comprehensive live-cell image collection of human protein localization to date. In addition, integration of IP-MS using the fluorescent tags for affinity capture enables measurement of localization and interactions from the same samples. For a quantitative description of cellular architecture, we introduce a data-driven framework to represent protein interactions and localization features, supported by a new machine learning algorithm for image encoding. This approach allows us to delineate communities of functionally related proteins by unsupervised clustering and facilitates the generation of mechanistic hypotheses, including for proteins that had so far remained uncharacterized. We further demonstrate that the localization pattern of each protein is defined by unique and specific features that can be used for functional interpretation, to the point that spatial relationships often contain enough information to predict interactions at the molecular scale. Finally, our analysis enables an unsupervised description of the human proteome's organization, and highlights in particular that RNA-binding proteins exhibit unique functional signatures that shape the proteome's network.

Engineered cell library

Fluorescent protein (FP) fusions are versatile tools that can measure both protein localization by microscopy and protein-protein interactions by acting as affinity handles for IP-MS (*18*, *19*) (Fig. S1A). Here, we constructed a library of fluorescently tagged HEK293T cell lines by targeting human genes with the split-mNeonGreen2 system (*20*) (Fig. 1A). Split-FPs greatly simplify CRISPR-based genome engineering by circumventing the need for molecular cloning (*18*), and allowed us to generate endogenous genomic fusions (Fig. 1B) that preserve native expression regulation. A full description of our pipeline is available in the Methods section ((*21*) ; summarized in Fig. 1C through E). In brief, FP insertion sites (N- or C-terminus) were chosen on the basis of information from the literature or structural analysis (Fig. S1B; Table S1). For each tagged target we isolated a polyclonal pool of CRISPR-edited cells, which was then characterized by live-cell 3D confocal microscopy, IP-MS, and genotyping of tagged alleles by next-generation sequencing. Open-source software development and advances in instrumentation supported scalability (Fig. 1C). In particular, we developed *crispycrunch*, a CRISPR design software that enables guide RNA selection and homology donor sequence design (github.com/czbiohub/crispycrunch). We also fully automated the acquisition of data microscopy data in Python for on-the-fly computer vision and selection of desirable fields of view imaged in 96-well plates (github.com/czbiohub/2021-opencell-microscopy-automation). Our mass-spectrometry protocols use the high sensitivity of timsTOF instruments (*22*) which allowed miniaturization of IP-MS down to 0.8x10^6 cells of starting material (Fig. S1C; about a tenth of the material required in previous approaches (*7*, *8*)).

In total, we targeted 1757 genes, of which 1310 (75%) could be detected by fluorescence imaging and form our current dataset (full library details in Table S1). From these, we obtained paired

IP-MS measurements for 1260 targets (96%, Fig. 1D). The 1310-protein collection includes a balanced representation of the pathways, compartments and functions of the human proteome (Fig. S1D), with the exception of processes specific to mitochondria, organellar lumen or extracellular matrix. Indeed, the split-FP system tags a gene of interest with a short sequence (mNG11) while a larger FP fragment (mNG_21-10) is expressed separately (Fig. 1A). In the version used here, the mNG_21-10 fragment is expressed in the nucleo-cytoplasm and prevents access to proteins inside organellar compartments. Membrane proteins can be tagged as long as one terminus extends in the nucleo-cytoplasm. In future iterations, other split systems that contain compartment-specific signal sequences could be used to target organellar lumen (*23*).

Fluorescent tagging was readily successful for essential genes, suggesting that FP fusions are well tolerated (Fig. S2A). To evaluate other factors contributing to successful fluorescent detection, we measured RNA and protein concentration in HEK293T cells (Fig. S2B; using a 24-fraction scheme for deep proteome quantification; see fully annotated proteome in Table S2). This revealed that protein abundance is the main limitation to detection (Fig. 1D, S2C; see details for unsuccessful targets in Table S3); most successful targets are among the top 50% most abundant (Fig. S2D). Gene-editing efficiency was another important factor: among well-expressed targets, failure was correlated with significantly lower rates of homologous recombination (Fig. S2E), which would impair the selection of edited cells by fluorescence-activated cell sorting (FACS). Training a regression model revealed that the combination of protein abundance and editing efficiency could predict successful detection with 82% accuracy.

To maximize throughput, we used a polyclonal strategy to select genome-edited cells by FACS. Polyclonal pools contain cells with distinct genotypes. HEK293T are pseudo-triploid (*24*) and a single edited allele is sufficient to confer fluorescence. Moreover, various DNA repair mechanisms compete with homologous recombination for the resolution of CRISPR-induced genomic breaks (*25*) so that alleles containing non-functional mutations can be present in addition to the desired fusion alleles. However, such alleles do not support fluorescence and are therefore unlikely to impact other measurements, especially in the context of a polyclonal pool. We developed a stringent selection scheme to significantly enrich for fluorescent fusion alleles (Fig. S3A). Our final cell library has a median 61% of mNeonGreen-integrated alleles, 5% wild-type and 26% other non-functional alleles (Fig. S3B, full genotype information in Table S1).

Finally, we verified that our engineering approach maintained the endogenous abundance of the tagged target proteins. For this, we quantified protein expression by Western blotting using antibodies specific to proteins targeted in 12 different cell pools (Fig. S3C), and by single-shot mass spectrometry in 63 tagged lines (Fig. S3D). Both approaches revealed a median abundance of tagged targets in engineered lines at about 80% of untagged HEK293T control, with 5 outliers (8% of total) identified by proteomics (Fig. S3D, all within 3.5-fold of control). Importantly, the overall proteome composition was unchanged in all tagged lines (Fig. S3E-F). Overall, our gene-editing strategy preserves near-endogenous abundances and circumvents the limitations of ectopic overexpression (*11*, *26*, *27*), which include aberrant localization, changes in organellar morphology, and masking effects (see the examples of SPTLC1, TOMM20 and MAP1LC3B in Fig. S3G). Therefore, OpenCell supports the functional profiling of tagged proteins in their native cellular context.

Interactome analysis and stoichiometry-driven clustering

Affinity enrichment coupled to mass spectrometry is an efficient and sensitive method for the systematic mapping of protein interaction networks (*28*). We isolated tagged proteins ("baits") from cell lysates solubilized in digitonin, a mild non-ionic detergent that preserves the native structure and properties of membrane proteins (*29*). Specific protein interactors ("preys") were identified by proteomics from biological triplicate experiments (see Figure S4A-B and (*21*) for a detailed description of our statistical analysis, which builds upon established methods (*7*)). In total, the full interactome from our 1260 OpenCell baits includes 29,922 interactions between 5292 proteins (baits and preys, Fig. 2A, full interactome data in Table S4).

To assess the quality of our interactome, we estimated its precision (the fraction of true positive interactions over all interactions) and recall (the fraction of interactions identified compared to a ground truth set) using reference data (Fig. S4B). For recall analysis, we quantified the coverage in our data of interactions included in CORUM (*30*), a compendium of protein interactions manually curated from the literature. To estimate precision, we quantified how many of our interactions involved protein pairs expected to localize to the same broad cellular compartment (*31*) (Fig. S4B). To benchmark OpenCell against other large-scale interactomes, we compared its precision and recall to Bioplex (overexpression of HA-tagged baits (*8*, *32*)), the yeast-two-hybrid human reference interactome (HuRI (*6*)) and our own previous data (GFP fusions expressed from bacterial artificial chromosomes (*7*)) (Fig. S4C-E). We also calculated compression rates for each dataset as a measure of the overall richness in network patterns and motifs distinguishable from noise, which correlates with overall network quality: real-world networks contain redundant information which can be compressed, while pure noise is not compressible (see (*33*)) (Fig. S4F). Across all metrics, OpenCell outperformed previous approaches. OpenCell also includes many interactions not reported in previous datasets (Fig. S4E,G). Our interactome may better reflect biological interactions because it preserves near-endogenous protein expression.

A powerful way to interpret interactomes is to identify communities of interactors (*8*, *13*). To this end, we applied unsupervised Markov clustering (MCL) (*34*) to the graph of interactions defined by our data (5292 baits and preys). We first measured the stoichiometry of each interaction, using a quantitative approach we previously established (*7*). Interaction stoichiometry measures the abundance of a protein interactor relative to the abundance of the bait in a given immuno-precipitation sample. We have shown that stoichiometry can be interpreted as a proxy for interaction strength, and that interactions can be classified between core (i.e. high) and low stoichiometries (*7*). In our current data, both high- and low-stoichiometry interactions were significantly enriched for proteins pairs sharing gene ontology annotations (Fig. S4H). Using stoichiometry to assign weights to the edges in the interaction graph (Fig. 2B), a first round of MCL delineated inter-connected protein communities and led to better clustering performance than clustering based on connectivity alone (Fig. S4I). To better delineate stable complexes, we further refined each individual MCL community by additional clustering while removing low-stoichiometry interactions. The resulting sub-clusters outline core interactions within existing communities (Fig. 2B). Figure 2C illustrates how this unsupervised approach enables to delineate functionally related proteins: all subunits of the machinery responsible for the translocation of newly translated proteins at the ER membrane (SEC61/62/63) and of the EMC (ER Membrane Complex) are grouped within respective core interaction clusters, but both are part of the same larger MCL community. This mirrors the recently appreciated co-translational role of EMC for insertion of transmembrane domains at the ER (*35*). Additional proteins that have only recently been shown to act co-translationally are found clustering with translocon or EMC subunits, including ERN1 (IRE1) (*36*) and CCDC47 (*37*, *38*). Thus, clustering can facilitate mechanistic exploration by grouping proteins involved in related pathways. Overall, we identified 300 communities including a total of 2096 baits and preys (full details in Table S4). Ontology analysis revealed that these communities are significantly enriched for specific cellular functions, supporting their biological relevance (82% of all communities are significantly enriched for specific biological process or molecular function GO ontology terms; see Table S5 for complete analysis). A graph of interactions between communities reveals a richly inter-connected network (Fig. 2D), the structure of which outlines the global architecture of the human interactome (discussed further below).

A direct application of interactome clustering is to help elucidate the cellular roles of the many human proteins that remain poorly characterized (*39*). We identified poorly characterized proteins by quantifying their occurrence in article titles and abstracts from PubMed (Fig. 2E). Empirically, we determined that proteins in the bottom 10th percentile of publication count (corresponding to less than 10 publications) are very poorly annotated (Fig. 2E). This set encompasses a total of 251 proteins found in interaction communities for which our dataset offers potential mechanistic insights. For example, the proteins NHSL1, NHSL2 and KIAA1522 are all found as part of a community centered around SCAR/WAVE, a large multi-subunit complex nucleating actin polymerization (Fig. 2F). All three proteins share sequence homology and are homologous to NHS (Fig. S5A), a protein mutated in patients with Nance-Horan syndrome. NHS interacts with SCAR/WAVE components to coordinate actin remodeling (*40*). Thus, NHSL1, NHSL2 and KIAA1522 also act to regulate actin assembly. A recent mechanistic study supports this hypothesis: NHSL1 localizes at the cell's leading edge and directly binds SCAR/WAVE to negatively regulate its activity, reducing F-actin content in lamellipodia and inhibiting cell migration (*41*). The authors identified NHSL1's SCAR/WAVE binding sites, and we find these sequences to be conserved in NSHL2 and KIA1522 (Fig. 2F). Therefore, our data suggests that both NHSL2 and KIAA1522 are also direct SCAR/WAVE binders and possible modulators of the actin cytoskeleton.

Our data also sheds light on the function of ROGDI, whose variants cause Kohlschuetter-Toenz syndrome (a recessive developmental disease characterized by epilepsy and psychomotor regression (*42*)). ROGDI appears in the literature because of its association with disease, but no study, to our knowledge, specifically determines its molecular function. We first observed that ROGDI's interaction pattern closely matched that of three other proteins in our dataset: DMXL1, DMXL2 and WDR7 (Fig. 2G). This set exhibited a specific interaction signature with the v-ATPase lysosomal proton pump. All four proteins interact with soluble v-ATPase subunits (ATP6-V1), but not its intra-membrane machinery (ATP6-V0). DMXL1 and WDR7 interact with V1 v-ATPase, and their depletion in cells compromises lysosomal re-acidification (*43*). Sequence analysis showed that DMXL1 or 2, WDR7 and ROGDI are homologous to proteins from yeast and Drosophila involved in the regulation of assembly of the soluble V1 subunits onto the V0 transmembrane ATPase core (*44*, *45*) (Fig. S5B). In yeast, Rav1 and Rav2 (homologous to DMXL1/2 and ROGDI, respectively) form the stoichiometric RAVE complex, a soluble chaperone that regulates v-ATPase assembly (*45*). To assess the existence of a human RAVE-like complex, we generated new tagged cell lines for DMXL1 and 2, WDR7, and ROGDI. Because of the low abundance of these proteins, the localization of DMXL2 and ROGDI were not detectable but pull-downs of DMXL1 and WDR7 confirmed a stoichiometric interaction between DMXL1 and 2, WDR7 and ROGDI (Fig. 2G, right panels). No direct interaction between DXML1 and DMXL2 was detected, suggesting that they might nucleate two separate subcomplexes. Therefore, our data reveals a human RAVE-like complex comprising DMXL1 or 2, WDR7 and ROGDI, which we propose acts as a chaperone for v-ATPase assembly based on its yeast homolog. Altogether, these results illustrate how our data can facilitate the generation of new mechanistic hypotheses by combining quantitative analysis and literature curation.

Image dataset: localization annotation and self-supervised machine learning

A key advantage of our cell engineering approach is to enable the characterization of each tagged protein in live, unperturbed cells. To profile localization, we performed spinning-disk confocal fluorescence microscopy (63x 1.47NA objective) under environmental control (37°C, 5% CO_2), and imaged the 3D distribution of proteins in consecutive z-slices. Microscopy acquisition was fully automated in Python to enable scalability (Fig. S6A-B). In particular, we trained a computer vision model to identify fields of view (FOVs) with homogeneous cell density on-the-fly, which reduced experimental variation between images. Our dataset contains a collection of 6375 3D stacks (5 different FOVs for each target) and includes paired imaging of nuclei with live-cell Hoechst 33342 staining.

We manually annotated localization patterns by assigning each protein to one or more of 15 separate cellular compartments such as the nucleolus, centrosome or Golgi apparatus (Fig. 3A). Because proteins often populate multiple compartments at steady-state (*9*), we graded annotations using a three-tier system: grade 3 identifies prominent localization compartment(s), grade 2 represents less pronounced localizations, and grade 1 annotates weak localization patterns nearing our limit of detection (see Fig. S7A for two representative examples, full annotations in Table S6). Ignoring grade 1 annotations which are inherently less precise, 55% of proteins in our library were detected in multiple locations consistent with known functional relationships. for example, clear connections were observed between secretory compartments (ER, Golgi, vesicles, plasma membrane), or between cytoskeleton and plasma membrane (Fig. S7B, Table S6)). Many proteins are found in both nucleus and cytoplasm (21% of our library), highlighting the importance of the nucleo-cytoplasmic import and export machinery in shaping global cellular function (*46*, *47*). Importantly, because our split-FP system does not enable the detection of proteins in the lumen of organelles, multi-localization involving translocation across an organellar membrane (which is rare but does happen for mitochondrial or peroxisomal proteins) cannot be detected in our data.

To benchmark our dataset, we compared our localization annotations against the Human Protein Atlas (HPA), the reference antibody-based compendium of human protein localization (*9*). This revealed significant agreement between datasets: 75% of proteins share at least one localization annotation in common (Fig. 3B; this includes 25% of all proteins that share the exact same set of annotations, see full description in Table S7A). Because HPA mostly reports on cell lines other than HEK293T, a perfect overlap is not expected as proteins might differentially localize between related compartments in different cell types. However, the annotations for 147 proteins (11% of our data) were fully inconsistent between the two datasets (Fig. S7C). An extensive curation of the literature on the localization of those proteins allowed us to resolve discrepancies for 115 proteins (i.e., 78% of that set; full curation in Table S8). Of these, existing literature evidence supported the OpenCell results for 113 (98.3%) of the 115 cases (Fig. S7D). This validates that endogenous tagging can help refine the curation of localization in the human proteome. Finally, our dataset includes 350 targets that have orthologs in *S. cerevisiae*. Comparison between OpenCell and yeast localization annotations (*48*) revealed a high degree of concordance (Fig. S7E; Table S7B; 81% of proteins share at least one annotation in common, including 36% perfect matches).

While expert annotation remains the best performing strategy to curate protein localization (*49*, *50*), the low-dimensional description it allows is not well suited for quantitative comparisons. Recent developments in image analysis and machine learning offer new opportunities to extract high-dimensional features from microscopy images (*50*, *51*). Therefore, we developed a deep learning model to quantitatively represent the localization pattern of each protein in our dataset (*52*). Briefly, our model is a variant of an autoencoder (Fig. 3C): a form of neural network that learns to vectorize an image through paired tasks of encoding (from an input image to a vector in a latent space) and decoding (from the latent space vector to a new output image). After training, a consensus representation for a given protein can be obtained from the average of the encodings from all its associated images. This generates a high-dimensional "localization encoding" (Fig. 3C) that captures the complex set of features that define the spatial distribution of a protein at steady state and across many individual cells. One of the main advantages of this approach is that it is self-supervised. Therefore, as opposed to supervised machine learning strategies that are trained to recognize pre-annotated patterns (for example, manual annotations of protein localization (*50*)), our method extracts localization signatures from raw images without any *a priori* assumptions or manually assigned labels. To visualize the relationships between these high-dimensional encodings, we embedded the encodings for all 1,310 OpenCell targets in two dimensions using UMAP, an algorithm that reduces high-dimensional datasets to two dimensions (UMAP 1 and UMAP 2) while attempting to preserve the global and local structures of the original data (*53*). The resulting map is organized in distinct territories that closely match manual annotations (Fig. 3D, highlighting mono-localizing proteins). This validates that the encoding approach yields a quantitative representation of the biologically relevant information in our microscopy data. The separation of different protein clusters in the UMAP embedding (further discussed below) mirrors the fascinating diversity of localization patterns across the full proteome. Images from nuclear proteins offer compelling illustrative examples of this diversity and reveal how fine-scale details can define the localization of proteins within the same organelle (Fig. 3E).

Functional specificity of protein localization in the human cell

Extracting functional insights directly from cellular images is a major goal of modern cell biology and data science (*54*). In this context, our image library and associated machine learning encodings enable us to explore what degree of functional relationship can be inferred between proteins solely based on their localization. For this, we first employed an unsupervised Leiden clustering strategy commonly used to identify cell types in single-cell RNA sequencing datasets (*55*). Clusters group proteins that share similar localization properties (every protein in the dataset is included in a cluster); these groups can then be analyzed for how well they match different sets of ground-truth annotations (Fig. 4A). The average size of clusters is controlled by varying a hyper-parameter called resolution (Fig. S8A). Systematically varying clustering resolution in our dataset revealed that not only did low-resolution clusters delineate proteins belonging to the same organelles (Fig. 4A-B), clustering at higher resolution also enabled to delineate functional pathways and even molecular complexes of interacting proteins (Fig. 4A-C). This demonstrates that the spatial distribution of each protein in the cell is highly specific, to the point that proteins sharing closely related functions can be identified on the sole basis of the similarity between their spatial distributions. This is further illustrated by how finely high-resolution clusters encapsulate proteins specialized in defined cellular functions (Fig. 4C). For example, our analysis not only separated P-body proteins (cluster #83) from other forms of punctated cytoplasmic structures, but also unambiguously differentiated vesicular trafficking pathways despite their very similar localization patterns: the endosomal machinery (#40), plasma membrane endocytic pits (#117) or COP-II vesicles (#143) were all delineated with high precision (Fig. 4C). Among ER proteins, the translocon clusters with the SRP receptor, EMC subunits and the OST glycosylation complex, all responsible for co-translational operations (#9). This performance extends to cytoplasmic (Fig. S8A) and nuclear clusters (Fig. S8B), revealing that spatial patterning is not limited to membrane-bound organelles and that sub-compartments exist also in the nucleo-cytoplasm. An illustrative example is a cytoplasmic cluster (#17) formed by a group of RNA-binding proteins (including ATXN2L, NUFIP2 or FXR1, Fig. 4C) that separate into granules upon stress conditions (*56–59*). Stress granules are not formed under the standard growth conditions used in our experiments, but the ability of our analysis to cluster these proteins together reveals an underlying specificity to their cytoplasmic localization (i.e., "texture") even in the absence of stress.

A direct comparison between imaging and interactome data allows us to further examine the extent to which molecular-level relationships (that is, protein interactions) can be derived from a comparison of localization patterns. For OpenCell targets that directly interact, we compared the correlation between their localization encodings derived from machine learning (defining a "localization similarity") and the stoichiometry of their interaction. This "localization similarity" measures the similarity between the global steady-state distributions of two proteins, as opposed to a direct measure of co-localization. We find that most proteins interact with low stoichiometry (as we previously described (*7*)) and without strong similarities in their spatial distribution (Fig 4D, solid oval). This means that while low-stoichiometry interactors co-localize at least partially to interact, their global distribution within the cell is different at steady state. On the other hand, high stoichiometry interactors share very similar localization signatures (Fig 4D, dashed oval). Indeed, proteins interacting within stable complexes annotated in CORUM fall into this category (Fig 4E), and the localization signatures of different subunits from large complexes are positioned very closely in UMAP embedding (Fig. 4F). In an important correlate, we found that a high similarity of spatial distribution is a strong predictor of molecular interaction. Across the entire set of target pairs (predicted to interact or not), proteins that share high localization similarities are also very likely to interact (Fig. 4G). For example, target pairs with a localization similarity greater than 0.85 have a 58% chance of being direct interactors, and a 68% chance of being second-neighbors (i.e., sharing a direct interactor in common). This suggests that protein-protein interactions could be identified from a quantitative comparison of spatial distribution alone. To test this, we focused on FAM241A (C4orf32), a protein of unknown function that was not part of our original library and asked whether we could predict its interactions using imaging data alone, compared to the classical de-orphaning approach that uses interaction proteomics. We thus generated a FAM241A endogenous fusion that was analyzed with live imaging and IP-MS separately. Encoding its localization pattern using a "naïve" machine learning model that was never trained with images of this new target revealed a very high localization similarity with two subunits of the ER oligo-saccharyl transferase OST (>0.85 similarity to STT3B and OSTC), and high-resolution Leiden clustering placed FAM241A in an image cluster containing only OST subunits (Fig

4H, top). This analysis suggested that FAM241A is a high-stoichiometry interactor of OST. IP-MS identified that FAM241A was indeed a stoichiometric subunit of the OST complex (Fig. 4H, bottom). While the specific function of FAM241A in protein glycosylation remains to be fully elucidated, this proof-of-concept example establishes that live-cell imaging can be used as a specific readout to predict molecular interactions.

Collectively, our analyses establish that the spatial distribution of a given protein contains highly specific information from which precise functional attributes can be extracted by modern machine learning algorithms. In addition, we show that while high-stoichiometry interactors share very similar localization patterns, most proteins interact with low stoichiometry and share different localization signatures. This reinforces the importance of low-stoichiometry interactions for defining the overall structure of the cellular network, not only providing the "glue" that holds the interactome network together (*7*) but also connecting different cellular compartments.

RNA-binding proteins form a unique group in both interactome and spatial networks

To gain insight into global signatures that organize the proteome, we further examined the structures of our imaging and interactome datasets. First, we reduced the dimensionality of each dataset by grouping proteins into their respective spatial clusters (as defined by the high-resolution localization-based clusters in Figs. 4A, 4C) or interaction communities (as defined in Fig. 2B). We then separately clustered these spatial groups (Fig. S9A) and interaction communities (Fig. S9B) to formalize paired hierarchical descriptions of the human proteome organization. These hierarchies are highly structured and delineate clear groups of proteins (see comparison to hierarchies expected by chance, Fig. S9C). In both hierarchies, groups isolated at an intermediate hierarchical layer outline "modules" which are enriched for specific cellular functions or compartments (Fig. S9A-B; full ontology analysis in Suppl. Tables 5 & 9). At a higher layer, each dataset is partitioned into three "branches", which represent core signatures that shape the proteome's architecture from a molecular or spatial perspective (Fig. S9A-B). The structure of the localization-based hierarchy (Fig. S9A) recapitulates the human cell's architecture across its three key compartments (nucleus, cytoplasm, membrane-bound organelles, Fig. S10A-B), which validates the relevance of our unsupervised hierarchical analysis. This motivated a deeper examination of the hierarchical architecture of the interactome (Fig. S9B, ontology analysis in Table S5). We found that intermediate-layer modules of the interactome delineate specific cellular functions such as transcription or vesicular transport (Fig.

S9B), reflecting as expected that functional pathways are formed by groups of proteins that physically interact (*60*, *61*). More strikingly, the highest-layer structure showed that two of the three interactome branches were defined by clear functional signatures (Fig. S10C-E): branch B is significantly enriched in proteins that reside in or interact with lipid membranes, while branch C is significantly enriched in RNA-binding proteins (RNA-BPs) (Fig. 5B). This indicates that both membrane-related proteins and RNA-BPs interact more preferentially with each other than with other kinds of proteins in the cell.

That membrane-related proteins form a specific interaction group is perhaps not surprising as the membrane surfaces that sequester them within the three-dimensional cell will be partially maintained upon detergent solubilization. On the other hand, the fact that RNA-BPs also form a specific interaction group is unexpected, since our protein interactions were measured in nuclease-treated samples (*21*) in which most RNAs are degraded. This suggests that protein features beyond binding to RNAs themselves might drive the preferential interactions of RNA-BPs with each other. Therefore, we reasoned that the biophysical properties of proteins within each interactome branch might underly their segregation. Indeed, an analysis of protein sequence features revealed a separation of different biophysical properties in each branch (Fig. S10F-G). Branch B was enriched for hydrophobic sequences (Fig. 5C), consistent with its enrichment for membrane-related proteins, while branch C was enriched for intrinsic disorder (Fig. 5C). This is consistent with the fact that RNA-BPs are significantly more disordered than other proteins in the proteome (Fig. S11A, (*62*)). RNA-BPs are also among the most abundant in the cell (Fig. S11B), and form a higher number of interactions than other proteins (Fig. S11C-D).

IP-MS measures protein interactions *in vitro* after lysis and therefore does not directly address the spatial relationship between interacting proteins. Thus, we sought to further examine how RNA-BPs distribute in our live-cell imaging data. If RNA-BPs segregate into interacting groups *in vivo*, this should also manifest at the level of their intracellular localization: they should enrich in the same spatial clusters derived from our unsupervised machine learning analysis. Indeed, the distribution of RNA-BP content within spatial clusters revealed a significant over-representation of clusters that are either strongly enriched or depleted for RNA-BPs (Fig. 5D). Since spatial clusters can be interpreted as defining "micro-compartments" within the cell, both enrichment and depletion have functional implications: not only are RNA-BPs enriched within the same micro-compartments, they tend to also be excluded from others. 16 out of the 26 spatial clusters (62%) that are highly enriched in RNA-BPs include at least one protein involved in biomolecular condensation (as curated in PhaSepDB (*63*)), which might reflect a prevalent role for biomolecular condensation in shaping the RNA-BP proteome. Collectively, both interactome and imaging data underscore that RNA-BPs (a prevalent group of proteins that represents 13% of proteins expressed in HEK293T cells, see Table S2) form a distinct sub-group within the proteome characterized by unique properties.

These results motivated a broader analysis of the contribution of intrinsic disorder to the spatial organization of the proteome in our dataset. Plotting the distribution of mean intrinsic disorder within spatial clusters revealed a significant over-representation of clusters both enriched and depleted in disordered proteins (Fig. 5E). 26 out of 182 total spatial clusters were enriched for disordered proteins, covering 13% of the proteins in our imaging dataset. Overall, the extent to which disordered proteins segregate spatially is similar to the degree of segregation found for hydrophobic proteins: an analogous analysis revealed that 10% of proteins in our dataset are found within clusters significantly enriched for high hydrophobicity (Fig. S12E), which map to membrane-bound organelles (Fig. S12F). This supports the hypothesis that intrinsic disorder is as important a feature as hydrophobicity in organizing the spatial distribution of the human proteome. Consistent with our previous analysis, high-disorder clusters were enriched for RNA-BPs (Fig. 5F), with 15 out of these 26 clusters containing over 50% of RNA-BPs. High-disorder clusters were also enriched for proteins annotated to participate in biomolecular condensation (Fig. 5G), and were predominantly found in the nucleus (19 clusters, 73% of total, Fig. 5H). 5 out of 7 high-disorders clusters found in the cytosol delineate compartments for which biomolecular condensation has been proposed to play an important role (Fig. 5G), namely P-bodies (*64*), stress granules (*59*), centrosome (*65*), cell junctions (*66*) and the interface between cell surface and actin cytoskeleton (*67*).

Interactive data sharing at opencell.czbiohub.org

To enable widespread access to the OpenCell datasets, we built an interactive web application that provides side-by-side visualizations of the 3D confocal images and of the interaction network for each tagged protein, together with RNA and protein abundances for the whole proteome (Fig. 6). Our web interface is fully described in Suppl. Fig S12.

Discussion

OpenCell combines three strategies to augment the description of human cellular architecture. First, we present an integrated experimental pipeline for high-throughput cell biology, fueled by scalable methods for genome engineering, live-cell microscopy and IP-MS. Second, we provide an open-source resource of well-curated localization and interactome measurements, easily accessible through an interactive web interface at opencell.czbiohub.org. And third, we developed an analytical framework for the representation and comparison of interaction or localization signatures (including a self-supervised machine learning approach for image encoding). Finally, we demonstrate how our dataset can be used both for fine-grained mechanistic exploration (to explore the function of multiple proteins that were previously uncharacterized), as well as for investigating the core organizational principles of the proteome.

Our current strategy that combines split-FPs and HEK293T – a cell line that is heavily transformed but easily manipulatable – is mostly constrained by scalability considerations. Excitingly, technological advances are quickly broadening the set of cellular systems that can be engineered and profiled at scale. Advances in stem cell technologies enable the generation of libraries that can be differentiated in multiple cell types (*11*), while innovations in genome engineering (for example, by modulating DNA repair (*68*)) pave the way for the scalable insertion of gene-sized payload, for the combination of multiple edits in the same cell, or for increased homozygosity in polyclonal pools. In addition, recent developments in high-throughput light-sheet microscopy (*69*) might soon enable the systematic description of 4D intracellular dynamics (*70*).

A central feature of our approach is to use endogenous fluorescent tags to study protein function. Genome-edited cells enable to examine protein function at near-native expression levels (which can circumvent some limitations of over-expression (*71*)), and to measure protein localization in live cells (which can avoid artefacts caused by fixation or antibody labeling (*72*)). Comparing our data to the current reference datasets of protein-proteins interactions (Fig. S4C-F) or localization (Fig. S7C-D) highlights the performance of our strategy. In addition, our high success rate tagging essential genes (Fig. S2A; see also (*73*) in yeast) and the successful tagging of the near-complete yeast proteome (*14*, *73*) support that fluorescent tagging generally preserves normal protein physiology. However, limitations exist for specific protein targets. FPs are as big as an average human protein and their insertion can impair function or localization, for example by occluding important interaction interfaces or impairing sub-cellular targeting sequences. In other cases, tags can affect expression or degradation rates, which might explain why we find tagged proteins being expressed at 80% of their endogenous abundance, and 8% of targets in our dataset having outlier abundances at steady-state (Fig. S3D). Further, tagging often cannot discriminate between different isoforms of a protein (such as splicing or post-translationally modified variants). Finally, relying on endogenous expression can be an obstacle given the low concentration of most proteins in the human cell: even using a very bright FP like mNeonGreen (*74*), detecting proteins in the bottom 50% percentile of abundance is difficult (Fig.

S2D). Solutions to this obstacle include using FP repeats to increase signal (*18*, *23*) or using tags that bind chemical fluorophores (e.g., HaloTag (*75*)), which can be brighter than FPs or operate at wavelengths where cellular auto-fluorescence is decreased (*76*). Overall, the full description of human cellular architecture remains a formidable challenge which will require complementary methods being applied in parallel. The diversity of large-scale cell biology approaches is a solution to this problem (*6*, *8*, *9*, *11*, *31*, *70*, *77–80*). Mirroring the advances in genomics following the human genome sequence (*2*), open-source systematic datasets will likely play an important role in how the growth of cell biology measurements can be transformed into fundamental discoveries by an entire community (*81*).

In addition to presenting a resource of measurements and protocols, we also demonstrate how our data can be used to study the global signatures that pattern the proteome. Our analysis reveals that RNA-binding proteins, which form one of the biggest functional family in the cell, are characterized by a unique set of properties and segregate from other proteins in term of both interactions and spatial distribution. It would be fascinating to explore to which extent RNA itself might act as a structural organizer of the cellular proteome (*62*, *82*). This is for example the case for some non-coding RNAs whose main function is to template protein interactions to form nuclear bodies (*83*). High intrinsic disorder is one of the distinguishing features of RNA-BPs, which likely contributes to their unique properties. Beyond RNA-BPs, our data supports a general role for intrinsic disorder in shaping the spatial distribution of human proteins. For example, 13% of proteins in our dataset are found in spatial clusters that are significantly enriched for disordered proteins. This adds to the growing appreciation that intrinsic disorder, which is much more prevalent in eukaryotic vs. prokaryotic proteomes (*84*, *85*), plays a key role in the functional sub-compartmentalization of the eukaryotic nucleo- and cytoplasm in the context of biomolecular condensation (*86*).

Lastly, we show that the spatial distribution of each human protein is very specific, to the point that remarkably detailed functional relationships can be inferred on the sole basis of similarities between localization patterns – including the prediction of molecular interactions (which complements other studies (*87*)). This highlights that intracellular organization is defined by fine-grained features that go beyond membership to a given organelle. Our demonstration that self-supervised deep learning models can identify complex but deterministic signatures from light microscopy images opens exciting avenues for the use of imaging as an information-rich method for deep phenotyping and functional genomics (*51*). Because light microscopy is easily scalable, can be performed live and enables measurements at the single-cell level, this should offer rich opportunities for the full quantitative description of cellular diversity in normal physiology and disease.

Material and Methods

A complete description of our Material and Methods is found in the Supplementary Material online (*21*). This include methods for cell culture and CRISPR engineering, immuno-precipitation and mass spectrometry, live-cell imaging, and data analysis of both interactome and imaging datasets.

References and Notes

1. I. H. G. S. Consortium, Finishing the euchromatic sequence of the human genome. Nature. 431, 931–945 (2004).

2. L. Hood, L. Rowen, The Human Genome Project: big science transforms biology and medicine. Genome Med. 5, 79 (2013).

3. P. Nurse, J. Hayles, The Cell in an Era of Systems Biology. Cell. 144, 850–854 (2011).

4. F. D. Mast, A. V. Ratushny, J. D. Aitchison, Systems cell biology. The Journal of Cell Biology. 206, 695–706 (2014).

5. E. Lundberg, G. H. H. Borner, Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology. 20, 285–302 (2019).

6. K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Coté, M. Daley, S. Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovács, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S.-F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. D. Ridder, S. D. Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Taşan, A. Tejeda, V. Tropepe, J.-C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. D. L. Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, M. A. Calderwood, A reference map of the human binary protein interactome. Nature. 580, 1–7 (2020).

7. M. Y. Hein, N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman, M. Mann, A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell. 163, 712–723 (2015).

8. E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P.

Gygi, J. W. Harper, Architecture of the human interactome defines protein communities and disease networks. Nature. 545 (2017), doi:10.1038/nature22366.

9. P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. A. Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, E. Lundberg, A subcellular map of the human proteome. Science. 356, eaal3321 (2017).

10. H. Bukhari, T. Müller, Endogenous Fluorescence Tagging by CRISPR. Trends Cell Biol. 29, 912–928 (2019).

11. B. Roberts, A. Haupt, A. Tucker, T. Grancharova, J. Arakaki, M. A. Fuqua, A. Nelson, C. Hookway, S. A. Ludmann, I. A. Mueller, R. Yang, A. R. Horwitz, S. M. Rafelski, R. N. Gunawardane, Molecular biology of the cell, in press, doi:10.1091/mbc.e17-03-0209.

12. S. Ghaemmaghami, S. Ghaemmaghami, W.-K. Huh, K. Bower, K. Bower, R. W. Howson, A. Belle, A. Belle, N. Dephoure, N. Dephoure, E. K. O'Shea, J. S. Weissman, Global analysis of protein expression in yeast. Nature. 425, 737–741 (2003).

13. S. R. Collins, P. Kemmeren, X.-C. Zhao, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. Spencer, F. C. P. Holstege, F. C. P. Holstege, J. S. Weissman, N. J. Krogan, Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & cellular proteomics : MCP. 6, 439–450 (2007).

14. U. Weill, I. Yofe, E. Sass, B. Stynen, D. Davidi, J. Natarajan, R. Ben-Menachem, Z. Avihou, O. Goldman, N. Harpaz, S. Chuartzman, K. Kniazev, B. Knoblach, J. Laborenz, F. Boos, J. Kowarzyk, S. Ben-Dor, E. Zalckvar, J. M. Herrmann, R. A. Rachubinski, O. Pines, D. Rapaport, S. W. Michnick, E. D. Levy, M. Schuldiner, Genome-wide SWAp-Tag yeast libraries for proteome exploration. Nature Methods. 15 (2018), doi:10.1038/s41592-018-0044-9.

15. A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute, C. Cullin, A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. Nucleic acids research. 21, 3329–3330 (1993).

16. I. Poser, M. Sarov, J. R. A. Hutchins, J.-K. Hériché, Y. Toyoda, A. Pozniakovsky, D. Weigl, A. Nitzsche, B. Hegemann, A. W. Bird, L. Pelletier, R. Kittler, S. Hua, R. Naumann, M. Augsburg, M. M. Sykora, H. Hofemeister, Y. Zhang, K. Nasmyth, K. P. White, S. Dietzel, K. Mechtler, R. Durbin, A. F. Stewart, J.-M. Peters, F. Buchholz, A. A. Hyman, BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nature methods. 5, 409–415 (2008).

17. A. Sigal, T. Danon, A. Cohen, R. Milo, N. Geva-Zatorsky, G. Lustig, Y. Liron, U. Alon, N. Perzov, Generation of a fluorescently labeled endogenous protein library in living human cells. Nature protocols. 2, 1515–1527 (2007).

18. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, A scalable strategy for high-throughput GFP tagging of endogenous human proteins. Proceedings of the National Academy of Sciences of the United States of America. 113, E3501-8 (2016).

19. N. C. Hubner, A. W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman, M. Mann, Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. The Journal of cell biology. 189, 739–754 (2010).

20. S. Feng, S. Sekine, V. Pessino, H. Li, M. D. Leonetti, B. Huang, Improved split fluorescent proteins for endogenous protein labeling. Nature communications. 8, 370 (2017).

21. See supplementary Materials and Methods online.

22. F. Meier, S. Beck, N. Grassl, M. Lubeck, M. A. Park, O. Raether, M. Mann, Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. J Proteome Res. 14, 5378–5387 (2015).

23. D. Kamiyama, S. Sekine, B. Barsi-Rhyne, J. Hu, B. Chen, L. A. Gilbert, H. Ishikawa, M. D. Leonetti, W. F. Marshall, J. S. Weissman, B. Huang, Versatile protein tagging in cells with split fluorescent protein. Nature communications. 7, 11046 (2016).

24. Y.-C. Lin, M. Boone, L. Meuris, I. Lemmens, N. V. Roy, A. Soete, J. Reumers, M. Moisse, S. Plaisance, R. Drmanac, J. Chen, F. Speleman, D. Lambrechts, Y. V. de Peer, J. Tavernier, N. Callewaert, Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nature communications. 5, 4767 (2014).

25. S. Lin, B. Staahl, R. K. Alla, J. A. Doudna, Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife. 3 (2014), doi:10.7554/elife.04766.

26. J. B. Doyon, B. Zeitler, J. Cheng, A. T. Cheng, J. M. Cherone, Y. Santiago, A. H. Lee, T. D. Vo, Y. Doyon, J. C. Miller, D. E. Paschon, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov, D. G. Drubin, Rapid and efficient clathrin-mediated endocytosis revealed in genome-edited mammalian cells. Nature cell biology. 13, 331–337 (2011).

27. T. J. Gibson, M. Seiler, R. A. Veitia, The transience of transient overexpression. Nat Methods. 10, 715–721 (2013).

28. E. C. Keilhauer, M. Y. Hein, M. Mann, Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). Mol Cell Proteom Mcp. 14, 120–35 (2014).

29. J. A. Thomas, C. G. Tate, Quality Control in Eukaryotic Membrane Protein Overproduction. J Mol Biol. 426, 4139–4154 (2014).

30. M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, A. Ruepp, CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucleic Acids Res. 47, gky973- (2018).

31. D. N. Itzhak, S. Tyanova, J. Cox, G. H. Borner, Global, quantitative and dynamic mapping of protein subcellular localization. eLife. 5, 570 (2016).

32. E. L. Huttlin, R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, J. Szpyt, A. Panov, H. Parzen, S. Fu, A. Golbazi, E. Maenpaa, K. Stricker, S. G. Thakurta, R. Rad, J. Pan, D. P. Nusinow, J. A. Paulo, D. K. Schweppe, L. P. Vaites, J. W. Harper, S. P. Gygi, Biorxiv, in press, doi:10.1101/2020.01.19.905109.

33. L. Royer, M. Reimann, A. F. Stewart, M. Schroeder, Network Compression as a Quality Measure for Protein Interaction Networks. PLoS ONE. 7, e35729 (2012).

34. A. J. Enright, S. V. Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 30, 1575–1584 (2002).

35. M. J. Shurtleff, D. N. Itzhak, J. A. Hussmann, N. T. S. Oakdale, E. A. Costa, M. Jonikas, J. Weibezahn, K. D. Popova, C. H. Jan, P. Sinitcyn, S. S. Vembar, H. Hernandez, J. Cox, A. L. Burlingame, J. L. Brodsky, A. Frost, G. H. Borner, J. S. Weissman, The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. Elife. 7, e37018 (2018).

36. D. Acosta-Alvear, G. E. Karagöz, F. Fröhlich, H. Li, T. C. Walther, P. Walter, The unfolded protein response and endoplasmic reticulum protein targeting machineries converge on the stress sensor IRE1. Elife. 7, e43036 (2018).

37. P. T. McGilvray, S. A. Anghel, A. Sundaram, F. Zhong, M. J. Trnka, J. R. Fuller, H. Hu, A. L. Burlingame, R. J. Keenan, An ER translocon for multi-pass membrane protein biogenesis. Elife. 9, e56889 (2020).

38. P. J. Chitwood, R. S. Hegde, An intramembrane chaperone complex facilitates membrane protein biogenesis. Nature. 584, 630–634 (2020).

39. T. Stoeger, M. Gerlach, R. I. Morimoto, L. A. N. Amaral, Large-scale investigation of the reasons why potentially important genes are ignored. PLoS biology. 16, e2006643 (2018).

40. S. P. Brooks, M. Coccia, H. R. Tang, N. Kanuga, L. M. Machesky, M. Bailly, M. E. Cheetham, A. J. Hardcastle, The Nance–Horan syndrome protein encodes a functional WAVE homology domain (WHD) and is important for co-ordinating actin remodelling and maintaining cell morphology. Hum Mol Genet. 19, 2421–2432 (2010).

41. A.-L. Law, S. Jalal, F. Mosis, T. Pallett, A. Guni, S. Brayford, L. Yolland, S. Marcotti, J. A. Levitt, S. P. Poland, M. Rowe-Sampson, A. Jandke, R. Köchl, G. Pula, S. M. Ameer-Beg, B. M. Stramer, M. Krause, Biorxiv, in press, doi:10.1101/2020.05.11.083030.

42. A. Schossig, N. I. Wolf, C. Fischer, M. Fischer, G. Stocker, S. Pabinger, A. Dander, B. Steiner, O. Tönz, D. Kotzot, E. Haberlandt, A. Amberger, B. Burwinkel, K. Wimmer, C. Fauth, C. Grond-Ginsbach, M. J. Koch, A. Deichmann, C. von Kalle, C. R. Bartram, A. Kohlschütter, Z. Trajanoski, J. Zschocke, Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. Am J Hum Genetics. 90, 701–707 (2012).

43. M. Merkulova, T. G. Păunescu, A. Azroyan, V. Marshansky, S. Breton, D. Brown, Mapping the H+ (V)-ATPase interactome: identification of proteins involved in trafficking, folding, assembly and phosphorylation. Scientific Reports. 5 (2015), doi:10.1038/srep14827.

44. Y. Yan, N. Denef, T. Schüpbach, The Vacuolar Proton Pump, V-ATPase, Is Required for Notch Signaling and Endosomal Trafficking in Drosophila. Dev Cell. 17, 387–402 (2009).

45. T. Vasanthakumar, J. L. Rubinstein, Structure and Roles of V-type ATPases. Trends Biochem Sci. 45, 295–307 (2020).

46. D. Görlich, U. Kutay, TRANSPORT BETWEEN THE CELL NUCLEUS AND THE CYTOPLASM. Annu Rev Cell Dev Bi. 15, 607–660 (1999).

47. C. P. Lusk, M. C. King, The nucleus: keeping it together by keeping it apart. Curr Opin Cell Biol. 44, 44–50 (2017).

48. M. Breker, M. Gymrek, M. Schuldiner, A novel single-cell screening platform reveals proteome plasticity during yeast stress responsesYeast proteome plasticity. J Cell Biology. 200, 839–850 (2013).

49. D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, E. Lundberg, Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nat Biotechnol. 36, 820–828 (2018).

50. W. Ouyang, C. F. Winsnes, M. Hjelmare, A. J. Cesnik, L. Åkesson, H. Xu, D. P. Sullivan, S. Dai, J. Lan, P. Jinmo, S. M. Galib, C. Henkel, K. Hwang, D. Poplavskiy, B. Tunguz, R. D. Wolfinger, Y. Gu, C. Li, J. Xie, D. Buslov, S. Fironov, A. Kiselev, D. Panchenko, X. Cao, R. Wei, Y. Wu, X. Zhu, K.-L. Tseng, Z. Gao, C. Ju, X. Yi, H. Zheng, C. Kappel, E. Lundberg, Analysis of the Human Protein Atlas Image Classification competition. Nat Methods. 16, 1254–1261 (2019).

51. S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, A. E. Carpenter, Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. 20, 145–159 (2021).

52. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, Self-Supervised Deep-Learning Reveals High-Resolution Functional Features from Protein Localization Microscopy. in preparation.

53. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Arxiv (2018).

54. E. Meijering, A. E. Carpenter, H. Peng, F. A. Hamprecht, J.-C. Olivo-Marin, Imagining the future of bioimage analysis. Nat Biotechnol. 34, 1250–1255 (2016).

55. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep-uk. 9, 5233 (2019).

56. S. Markmiller, S. Soltanieh, K. L. Server, R. Mak, W. Jin, M. Y. Fang, E.-C. Luo, F. Krach, D. Yang, A. Sen, A. Fulzele, J. M. Wozniak, D. J. Gonzalez, M. W. Kankel, F.-B. Gao, E. J. Bennett, E.

Lécuyer, G. W. Yeo, Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. Cell. 172, 590-604.e13 (2018).

57. J.-Y. Youn, W. H. Dunham, S. J. Hong, J. D. R. Knight, M. Bashkurov, G. I. Chen, H. Bagci, B. Rathod, G. MacLeod, S. W. M. Eng, S. Angers, Q. Morris, M. Fabian, J.-F. Côté, A.-C. Gingras, High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. Mol Cell. 69, 517-532.e11 (2018).

58. H. Marmor-Kollet, A. Siany, N. Kedersha, N. Knafo, N. Rivkin, Y. M. Danino, T. G. Moens, T. Olender, D. Sheban, N. Cohen, T. Dadosh, Y. Addadi, R. Ravid, C. Eitan, B. T. Cohen, S. Hofmann, C. L. Riggs, V. M. Advani, A. Higginbottom, J. Cooper-Knock, J. H. Hanna, Y. Merbl, L. V. D. Bosch, P. Anderson, P. Ivanov, T. Geiger, E. Hornstein, Spatiotemporal Proteomic Analysis of Stress Granule Disassembly Using APEX Reveals Regulation by SUMOylation and Links to ALS Pathogenesis. Mol Cell. 80, 876-891.e6 (2020).

59. P. Yang, C. Mathieu, R.-M. Kolaitis, P. Zhang, J. Messing, U. Yurtsever, Z. Yang, J. Wu, Y. Li, Q. Pan, J. Yu, E. W. Martin, T. Mittag, H. J. Kim, J. P. Taylor, G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. Cell. 181, 325-345.e28 (2020).

60. M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, V. Pelechano, E. B. Styles, M. Billmann, J. van Leeuwen, N. van Dyk, Z.-Y. Lin, E. Kuzmin, J. Nelson, J. S. Piotrowski, T. Srikumar, S. Bahr, Y. Chen, R. Deshpande, C. F. Kurat, S. C. Li, Z. Li, M. M. Usaj, H. Okada, N. Pascoe, B.-J. S. Luis, S. Sharifpoor, E. Shuteriqi, S. W. Simpkins, J. Snider, H. G. Suresh, Y. Tan, H. Zhu, N. Malod-Dognin, V. Janjic, N. Przulj, O. G. Troyanskaya, I. Stagljar, T. Xia, Y. Ohya, A.-C. Gingras, B. Raught, M. Boutros, L. M. Steinmetz, C. L. Moore, A. P. Rosebrock, A. A. Caudy, C. L. Myers, B. Andrews, C. Boone, A global genetic interaction network maps a wiring diagram of cellular function. Science. 353, aaf1420 (2016).

61. M. A. Horlbeck, A. Xu, M. Wang, N. K. Bennett, C. Y. Park, D. Bogdanoff, B. Adamson, E. D. Chow, M. Kampmann, T. R. Peterson, K. Nakamura, M. A. Fischbach, J. S. Weissman, L. A. Gilbert, Mapping the Genetic Landscape of Human Cells. Cell. 174, 953-967.e22 (2018).

62. A. Balcerak, A. Trebinska-Stryjewska, R. Konopinski, M. Wakula, E. A. Grzybowska, RNA–protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. Open Biol. 9, 190096 (2019).

63. K. You, Q. Huang, C. Yu, B. Shen, C. Sevilla, M. Shi, H. Hermjakob, Y. Chen, T. Li, PhaSepDB: a database of liquid–liquid phase separation related proteins. Nucleic Acids Res. 48, D354–D359 (2019).

64. Y. Luo, Z. Na, S. A. Slavoff, P Bodies: Composition, Properties, and Functions. Biochemistry-us. 57, 2424–2431 (2018).

65. J. B. Woodruff, B. F. Gomes, P. O. Widlund, J. Mahamid, A. Honigmann, A. A. Hyman, The Centrosome Is a Selective Condensate that Nucleates Microtubules by Concentrating Tubulin. Cell. 169, 1066-1077.e10 (2017).

66. O. Beutel, R. Maraspini, K. Pombo-García, C. Martin-Lemaitre, A. Honigmann, Phase Separation of Zonula Occludens Proteins Drives Formation of Tight Junctions. Cell. 179, 923-936.e11 (2019).

67. S. Banjade, Q. Wu, A. Mittal, W. B. Peeples, R. V. Pappu, M. K. Rosen, Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. Proc National Acad Sci. 112, E6426–E6435 (2015).

68. S. Riesenberg, M. Chintalapati, D. Macak, P. Kanis, T. Maricic, S. Pääbo, Simultaneous precise editing of multiple genes in human cells. Nucleic acids research. 2, 163 (2019).

69. B. Yang, X. Chen, Y. Wang, S. Feng, V. Pessino, N. Stuurman, N. H. Cho, K. W. Cheng, S. J. Lord, L. Xu, D. Xie, R. D. Mullins, M. D. Leonetti, B. Huang, Epi-illumination SPIM for volumetric imaging with high spatial-temporal resolution. Nature methods. 16, 501–504 (2019).

70. Y. Cai, M. J. Hossain, J.-K. Hériché, A. Z. Politi, N. Walther, B. Koch, M. Wachsmuth, B. Nijmeijer, M. Kueblbeck, M. Martinic-Kavur, R. Ladurner, S. Alexander, J.-M. Peters, J. Ellenberg, Experimental and computational framework for a dynamic protein atlas of human cell division. Nature. 561, 411–415 (2018).

71. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 417, 399–403 (2002).

72. U. Schnell, F. Dijk, K. A. Sjollema, B. N. G. Giepmans, Immunolabeling artifacts and the need for live-cell imaging. Nature methods. 9, 152–158 (2012).

73. W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O'Shea, Global analysis of protein localization in budding yeast. Nature. 425, 686–691 (2003).

74. N. C. Shaner, G. G. Lambert, A. Chammas, Y. Ni, P. J. Cranfill, M. A. Baird, B. R. Sell, J. R. Allen, R. N. Day, M. Israelsson, M. W. Davidson, J. Wang, A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. Nature methods. 10, 407–409 (2013).

75. G. V. Los, L. P. Encell, M. G. McDougall, D. D. Hartzell, N. Karassina, C. Zimprich, M. G. Wood, R. Learish, R. F. Ohana, M. Urh, D. Simpson, J. Mendez, K. Zimmerman, P. Otto, G. Vidugiris, J. Zhu, A. Darzins, D. H. Klaubert, R. F. Bulleit, K. V. Wood, HaloTag: a novel protein labeling technology for cell imaging and protein analysis. ACS chemical biology. 3, 373–382 (2008).

76. L. D. Lavis, Chemistry Is Dead. Long Live Chemistry! Biochemistry-us. 56, 5165–5170 (2017).

77. C. D. Go, J. D. R. Knight, A. Rajasekharan, B. Rathod, G. G. Hesketh, K. T. Abe, J.-Y. Youn, P. Samavarchi-Tehrani, H. Zhang, L. Y. Zhu, E. Popiel, J.-P. Lambert, É. Coyaud, S. W. T. Cheung, D. Rajendran, C. J. Wong, H. Antonicka, L. Pelletier, B. Raught, A. F. Palazzo, E. A. Shoubridge, A.-C. Gingras, A proximity biotinylation map of a human cell. Biorxiv, 796391 (2019).

78. G. Gut, M. D. Herrmann, L. Pelkmans, Multiplexed protein maps link subcellular organization to cellular states. Science. 361, eaar7042 (2018).

79. J. R. A. Hutchins, Y. Toyoda, B. Hegemann, I. Poser, J.-K. Hériché, M. M. Sykora, M. Augsburg, O. Hudecz, B. A. Buschhorn, J. Bulkescher, C. Conrad, D. Comartin, A. Schleiffer, M. Sarov, A. Pozniakovsky, M. M. Slabicki, S. Schloissnig, I. Steinmacher, M. Leuschner, A. Ssykor, S. Lawo, L. Pelletier, H. Stark, K. Nasmyth, J. Ellenberg, R. Durbin, F. Buchholz, K. Mechtler, A. A. Hyman, J.-M. Peters, Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science (New York, N.Y.). 328, 593–599 (2010).

80. P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V.-N. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. M. Tillier, A. Paccanaro, E. M. Marcotte, A. Emili, A Census of Human Soluble Protein Complexes. Cell. 150, 1068–1081 (2012).

81. J. Ellenberg, J. R. Swedlow, M. Barlow, C. E. Cook, U. Sarkans, A. Patwardhan, A. Brazma, E. Birney, A call for public archives for biological image data. Nature Methods. 15, 849–854 (2018).

82. M. W. Hentze, A. Castello, T. Schwarzl, T. Preiss, A brave new world of RNA-binding proteins. Nat Rev Mol Cell Bio. 19, 327–341 (2018).

83. T. Chujo, T. Hirose, Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function. Mol Cells (2017), doi:10.14348/molcells.2017.0263.

84. Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky, L. Kurgan, Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci. 72, 137–151 (2015).

85. W. Basile, M. Salvatore, C. Bassot, A. Elofsson, Why do eukaryotic proteins contain more intrinsically disordered regions? Plos Comput Biol. 15, e1007186 (2019).

86. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. Science. 357, eaaf4382 (2017).

87. Y. Qin, C. F. Winsnes, E. L. Huttlin, F. Zheng, W. Ouyang, J. Park, A. Pitea, J. F. Kreisberg, S. P. Gygi, J. W. Harper, J. Ma, E. Lundberg, T. Ideker, bioRxiv, in press, doi:10.1101/2020.06.21.163709.

88. L. Hubert, P. Arabie, Comparing partitions. J Classif. 2, 193–218 (1985).

89. B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 46, gky384- (2018).

90. R. J. Emenecker, D. Griffith, A. S. Holehouse, metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Biophys J (2021), doi:10.1016/j.bpj.2021.08.039.

91. C. L. Young, Z. T. Britton, A. S. Robinson, Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. Biotechnol J. 7, 620–634 (2012).

92. G. Dingle, CrispyCrunch: High-throughput Design and Analysis of CRISPR+HDR Experiments, (available at https://blog.addgene.org/crispycrunch-high-throughput-design-and-analysis-of-crisprhdr-experiments).

93. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (New York, N.Y.). 337, 816–821 (2012).

94. N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Müller, F. Meier, J. V. Olsen, O. Vorm, M. Mann, A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics*. Mol Cell Proteomics. 17, 2284–2296 (2018).

95. F. Meier, A.-D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, N. Bache, O. Hoerning, J. Cox, O. Räther, M. Mann, Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer*. Mol Cell Proteomics. 17, i–2545 (2018).

96. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 26, 1367–1372 (2008).

97. N. Prianichnikov, H. Koch, S. Koch, M. Lubeck, R. Heilig, S. Brehmer, R. Fischer, J. Cox, MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics*. Mol Cell Proteomics. 19, 1058–1069 (2020).

98. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 32, 223–226 (2014).

99. P. Mertins, L. C. Tang, K. Krug, D. J. Clark, M. A. Gritsenko, L. Chen, K. R. Clauser, T. R. Clauss, P. Shah, M. A. Gillette, V. A. Petyuk, S. N. Thomas, D. R. Mani, F. Mundt, R. J. Moore, Y. Hu, R. Zhao, M. Schnaubelt, H. Keshishian, M. E. Monroe, Z. Zhang, N. D. Udeshi, D. Mani, S. R. Davies, R. R. Townsend, D. W. Chan, R. D. Smith, H. Zhang, T. Liu, S. A. Carr, Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography–mass spectrometry. Nat Protoc. 13, 1632–1661 (2018).

100. K. Drew, C. Lee, R. L. Huizar, F. Tu, B. Borgeson, C. D. McWhite, Y. Ma, J. B. Wallingford, E. M. Marcotte, Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology. 13, 932 (2017).

101. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25, 1422–1423 (2009).

102. A. Razavi, A. van den Oord, O. Vinyals, Generating Diverse High-Fidelity Images with VQ-VAE-2. Arxiv (2019).

103. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15 (2018).

104. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462 (2016).

105. T. Bonald, B. Charpentier, A. Galland, A. Hollocou, Hierarchical Graph Clustering using Node Pair Sampling. Arxiv (2018).

106. H. Mi, D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou, T. Mushayamaha, P. D. Thomas, PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49, gkaa1106- (2020).

Acknowledgements. We thank N. Neff, M. Tan, R. Sit and A. Detweiler for help with high-throughput sequencing; G. Margulis, A. Sellas, E. Ho and J. Mann for operational support; A. McGeever for help with web application architecture and deployment; and S. Schmid for critical feedback. M.D.L. thanks C. L. Tan for continuous discussions.

Funding. H.K. was supported by an International Research Fellowship from the Japan Society for the Promotion of Science. R.M.B. was supported by a NIH pre-doctoral fellowship (F31 HL143882). B.H. was supported by NIH (R01GM131641) and is a Chan Zuckerberg Biohub Investigator. J.S.W. was supported by NIH (1RM1HG009490) and is an investigator with the Howard Hughes Medical Institute.

Competing interests. J.S.W. declares outside interest in Chroma Therapeutics, KSQ Therapeutics, Maze Therapeutics, Amgen, Tessera Therapeutics and 5 AM Ventures. M. M. is an indirect shareholder in EvoSep Biosystems.

Author contributions:

Conceptualization: MDL, MM, JSW, BH, MYH
Methodology: MDL, MM, LAR, DNI, RGS, JSW, SBM, BH, MYH, KK, KCC, NHC
Investigation: NHC, KCC, ADB, KK, ACM, PR, HK, LS, JYL, HC, JYSK, EMS, CG, FM, JPC, RMB, BBC, GD, MYH, DNI, MDL
Visualization: MDL, KC, KK, MYH
Funding acquisition: MDL, MM, LAR, DIN, RSG, JSW, SBM, BH
Project administration: MDL
Supervision: MDL, MM, LAR, DIN, RSG, JSW, SBM, BH
Writing – original draft: MDL, KCC, JSW, MM, NHC, ADB, KK, ACM, PR, MYH
Writing – review & editing: MDL, KCC, KK, MYH

Data and materials availability:

Mass spectrometry raw interactome data and associated MaxQuant output tables are deposited to the ProteomeXchange Consortium via the PRIDEpartner repository (accession PXD024909). Raw microscopy images are hosted by AWS's Open Datasets Program at https://registry.opendata.aws/czb-opencell/.

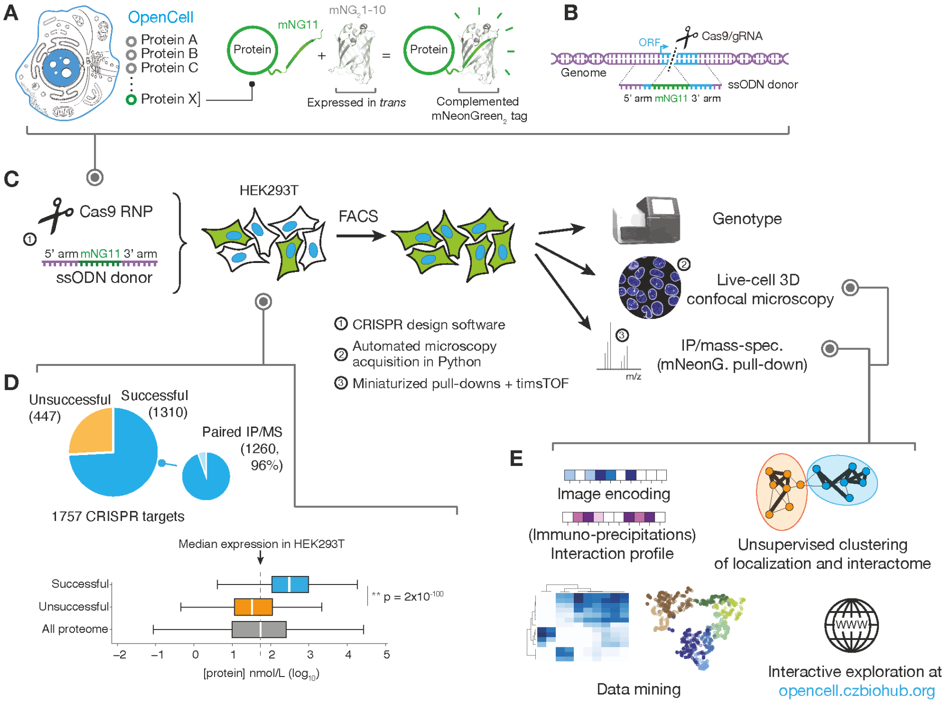


Figure 1: the OpenCell library. (A) Functional tagging with split-mNeonGreen$_2$. In this system, mNeonGreen$_2$ is separated into two fragments: a short mNG11 fragment, which is fused to a protein of interest, and a large mNG$_2$1-10 fragment, which is expressed separately in trans (that is, tagging is done in cells that have been engineered to constitutively express mNG$_2$1-10). **(B)** Endogenous tagging strategy: mNG11 fusion sequences are inserted directly within genomic open reading frames (ORFs) using CRISPR-Cas9 gene editing and homologous recombination with single-stranded oligonucleotides donors (ssODN). **(C)** The OpenCell experimental pipeline. See text for details. **(D)** Successful detection of fluorescence in the OpenCell library. Out of 1757 genes that were originally targeted, fluorescent signal was successfully detected for 1310 (top panel). Low protein abundance is the main obstacle to successful detection. Bottom left panel shows the full distribution of abundance for all proteins expressed in HEK293T vs. successfully or unsuccessfully detected OpenCell targets; boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x interquartile range. Median is indicated by a white line. P-value: Student's t-test. **(E)** The OpenCell data analysis pipeline, described in subsequent sections.

Figure 1

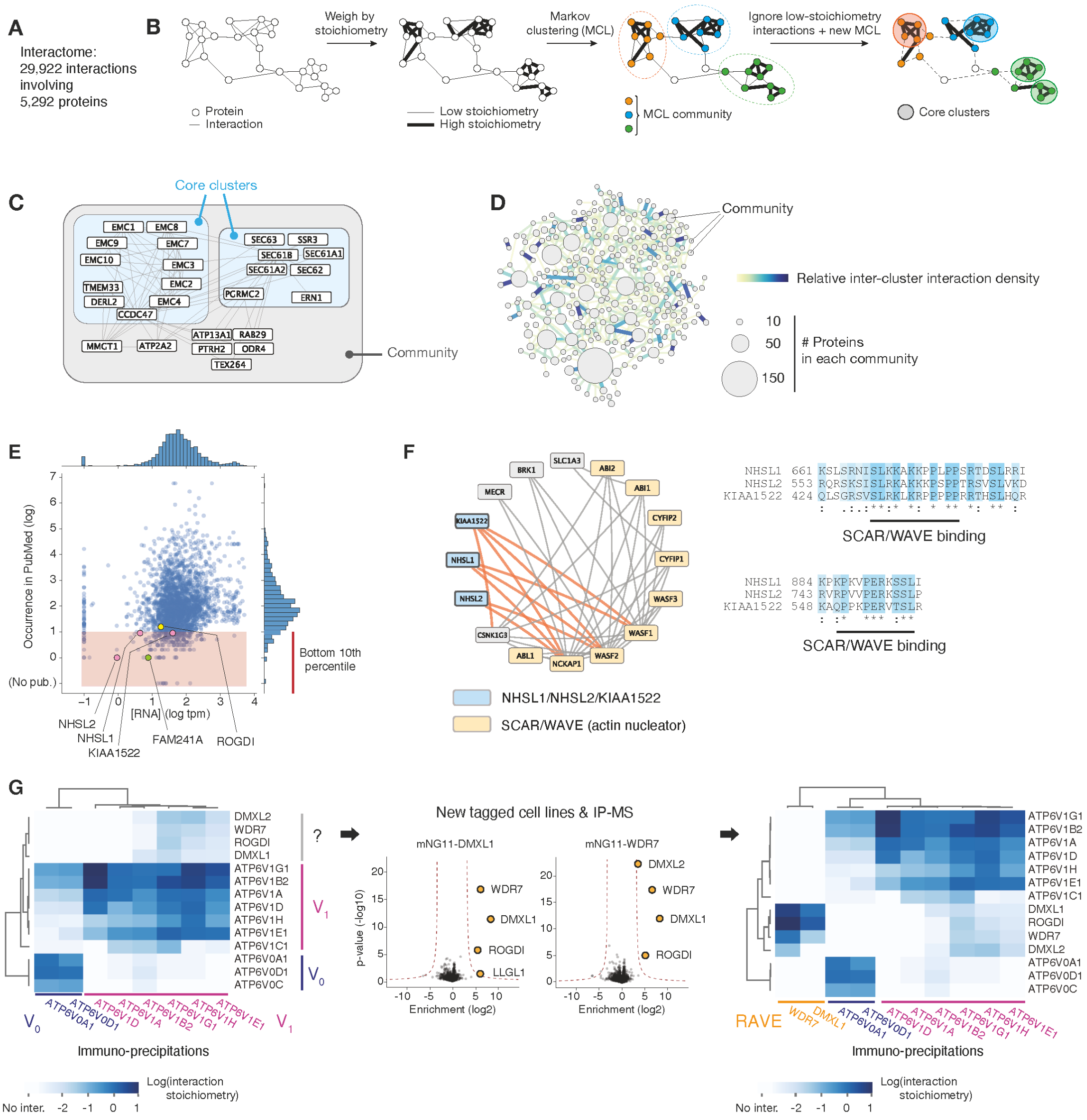


Figure 2: Protein interactome. (A) Overall description of the interactome. **(B)** Unsupervised Markov clustering of the interactome graph. **(C)** Example of community and core cluster definition for the translocon/EMC community. **(D)** The complete graph of connections between interactome communities. The density of protein-protein interactions between communities is represented by increased edge width. The numbers of targets included in each community is represented by circles of increasing diameters. **(E)** Distribution of occurrence in PubMed articles vs. RNA expression for all proteins found within interactome communities. The bottom 10th percentile of publication count (poorly characterized proteins) is highlighted. **(F)** NHSL1/NSHL2/KIAA1522 are part of the SCAR/WAVE community and share amino-acid sequence homology (right panel). **(G)** DMXL1/2, WDR7 and ROGDI form the human RAVE complex. Heatmaps represent the interaction stoichiometry of preys (lines) in the pull-downs of specific OpenCell targets (columns). See text for details.

Figure 2

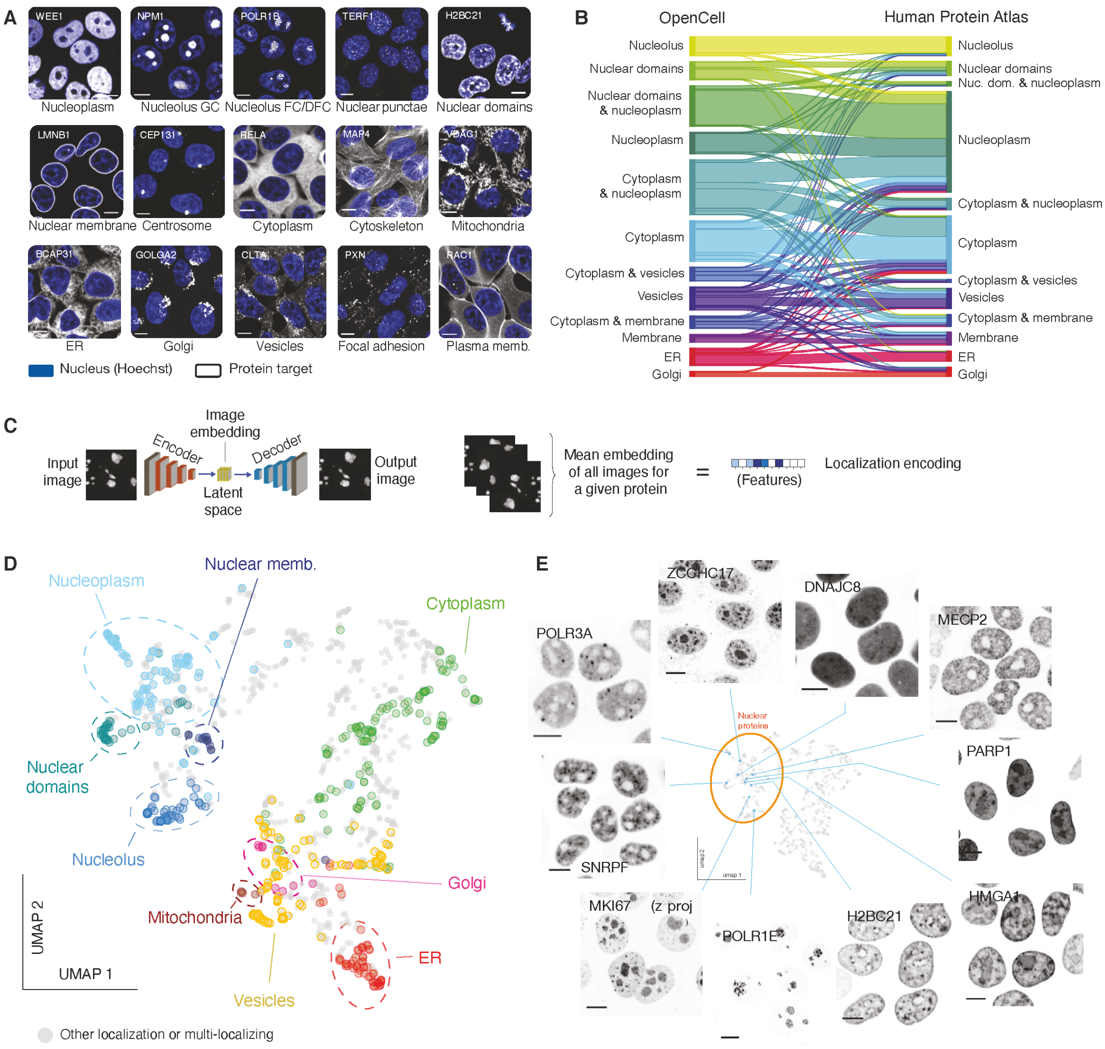


Figure 3: live-cell image collection. (A) The 15 cellular compartments segregated for annotating localization patterns. The localization of a representative protein belonging to each group is shown (greyscale, gene names in top left corners; scalebar: 10 μm). Nuclear stain (Hoechst) is shown in blue. "Nuclear domains" designate proteins with pronounced non-uniform nucleoplasmic localization, for example chromatin binding proteins. **(B)** Comparison of annotated localization for proteins included in both OpenCell and Human Protein Atlas datasets. In this flow diagram, colored bands represent groups of proteins that shared the same localization annotation in OpenCell, and the width of the band represents the number of proteins in each group. For readability, only the 12 most common localization groups are shown. Some multi-localization groups are included (e.g. "cytoplasm & nucleoplasm"). **(C)** Principle of localization encoding by self-supervised machine learning. See text for details. **(D)** UMAP representation of the OpenCell localization dataset, highlighting targets found to localize to a unique cellular compartment. **(E)** Representative images for 10 nuclear targets that exemplify the nuanced diversity of localization patterns across the proteome. Scale bars: 10 μm.

Figure 3

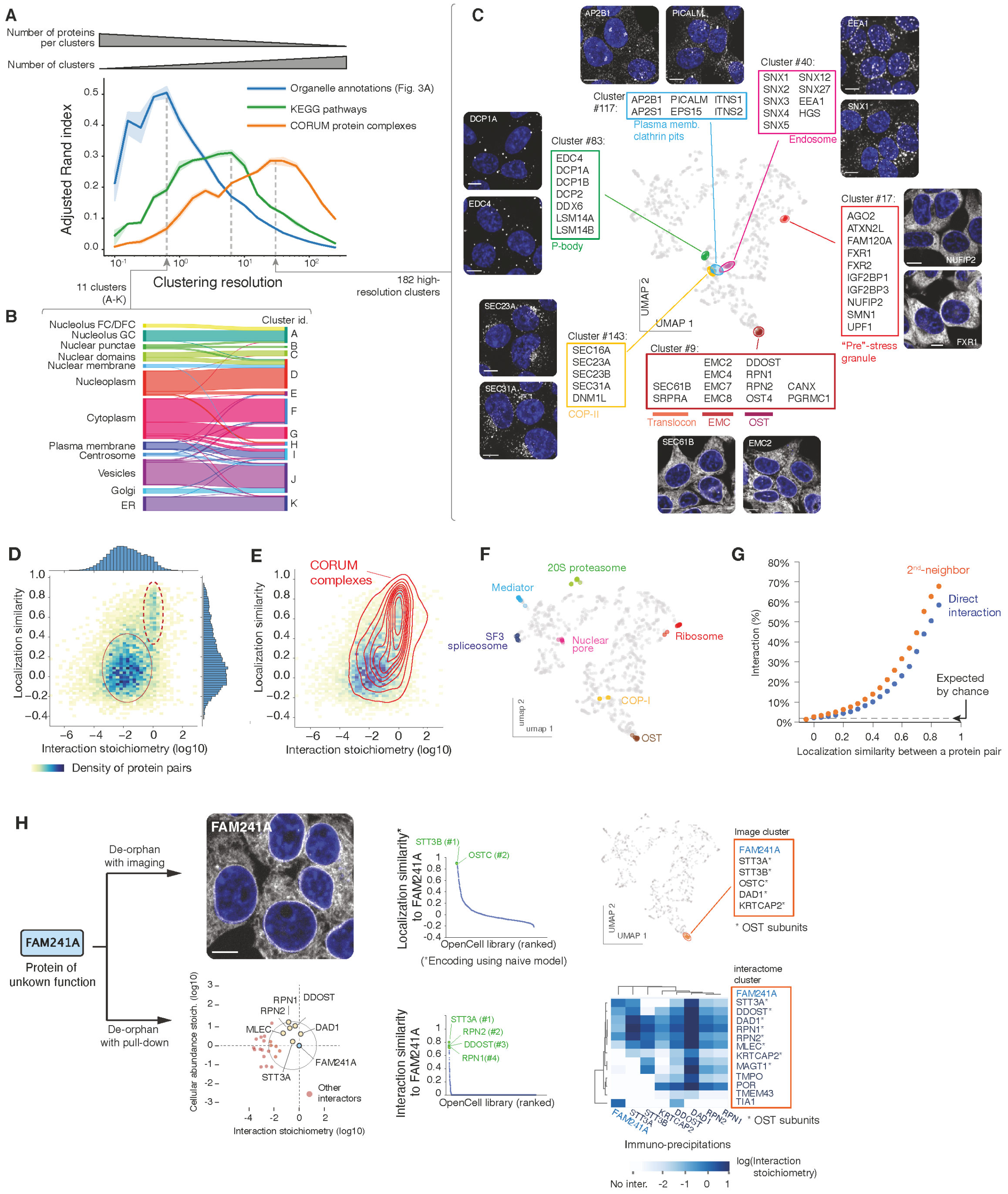


Figure 4
(legend on next page)

Figure 4: protein functional features derived from unsupervised image analysis. (A) Comparison of image-based Leiden clusters with ground-truth annotations. The Adjusted Rand Index (ARI, (86)) of clusters relative to three ground-truth datasets is plotted as a function of the Leiden clustering resolution. ARI (a metric between 0 and 1, see Materials and Methods) measures how well the groups from a given partition (in our case, the groups of proteins delineated at different clustering resolutions) match groups defined in a reference set. The amplitude of the ARI curves is approximately equal to the number of pairs of elements that partition similarly between sets; the resolution at which each curve reaches its maximum corresponds to the resolution that best captures the information in each ground-truth dataset. At a low resolution, Leiden clustering delineates groups that recapitulate about half of the organellar localization annotations, while at increasing resolutions, clustering recapitulates about a third of pathways annotated in KEGG, or molecular protein complexes annotated in CORUM. Shaded regions show standard deviations calculated from 9 separate repeat rounds of clustering, and average values are shown as a solid line. **(B)** High correspondence between low-resolution image clusters and cellular organelles. **(C)** Examples of functional groups delineated by high-resolution image clusters, highlighted on the localization UMAP. **(D)** Heatmap distribution of localization similarity (defined as the Pearson correlation between two deep learning-derived encoding vectors) vs. interaction stoichiometry between all interacting pairs of OpenCell targets. Two discrete sub-groups are outlined: low stoichiometry/low localization similarity pairs (solid line) and high stoichiometry/high localization similarity pairs (dashed line). **(E)** Probability density distribution of CORUM interactions mapped on the graph from (D). Contours correspond to iso-proportions of density thresholds for each 10th percentile. **(F)** Localization patterns of different subunits from example stable protein complexes, represented on the localization UMAP. **(G)** Frequency of direct (1st-neighbor) or once-removed (2nd neighbor, having a direct interactor in common) protein-protein interactions between any two pairs of OpenCell targets sharing localization similarities above a given threshold (x-axis). **(H)** Parallel identification of FAM241A as a new OST subunit by imaging or mass-spectrometry. See text for details.

Figure 4 (legend)

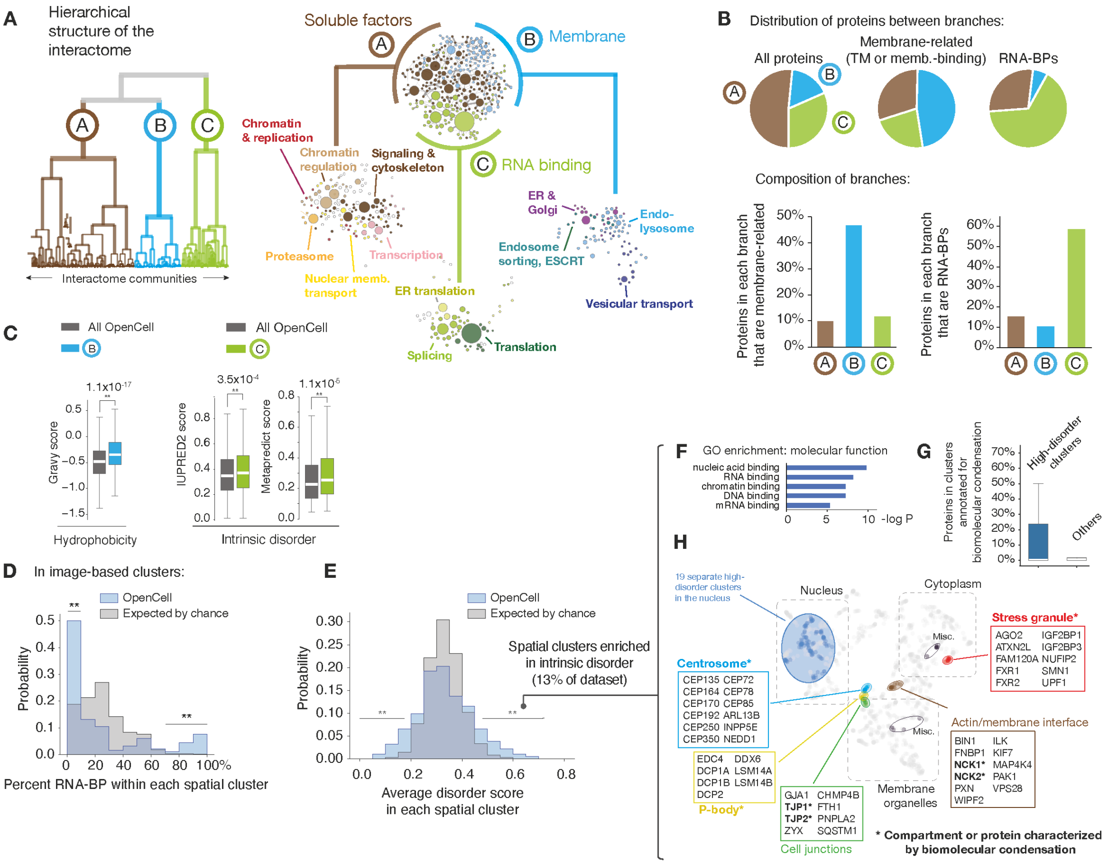
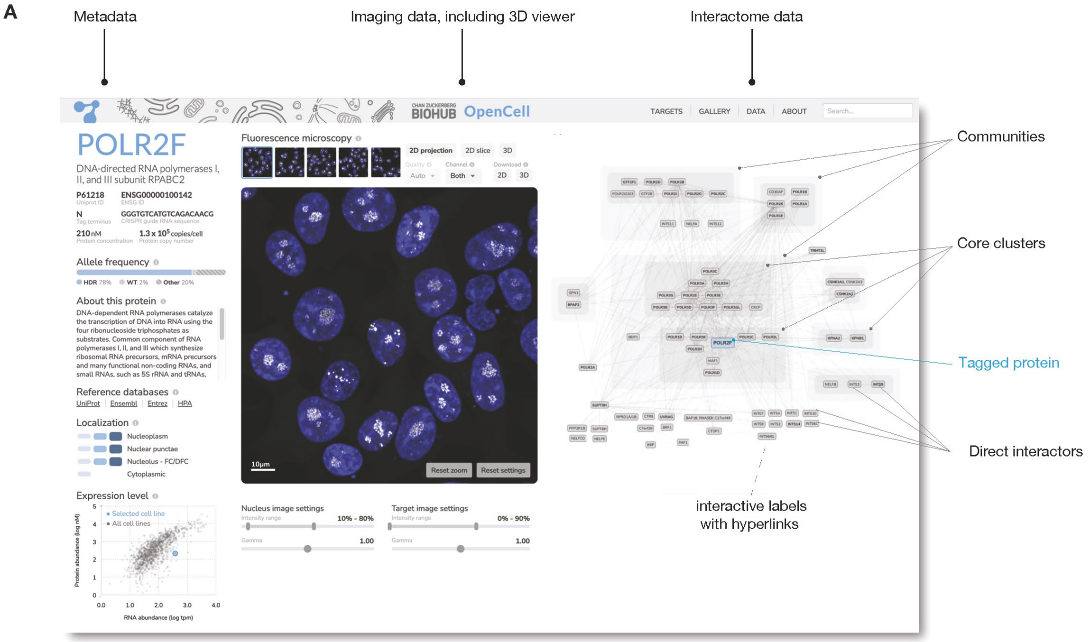


Figure 5: segregation of RNA-BPs in both interactome and imaging datasets. (A) Hierarchical structure of the interactome dataset, see full description in Figure S9B. **(B)** Distribution of membrane-related (transmembrane or membrane-binding) and RNA-BPs within the three interactome branches. **(C)** Distribution of intrinsic disorder in the RNA-BP branch of the interactome hierarchy (related to Figure S10). Two separate scores are shown for completeness: IUPRED2 (87), and metapredict (88), a new aggregative disorder scoring algorithm. Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x inter-quartile range. Median is represented by a white line. ** p < 10-4 (Student's t-test), exact p-values are shown. **(D)** Distribution of RNA-BP percentage across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1,000 times. Lines indicate parts of the distribution over-represented in our data vs control (**: p < $2x10^{-3}$, Fisher's exact t-test). **(E)** Distribution of disorder score (IUPRED2) across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1,000 times. Lines indicate parts of the distribution over-represented in our data vs control (**: p < $2x10^{-3}$, Fisher's exact t-test). **(F)** Ontology enrichment analysis of proteins contained in high-disorder spatial clusters (average disorder score > 0.45). Enrichment compares to the whole set of OpenCell targets (p-value: Fisher's exact test). **(G)** Prevalence of proteins annotated to be involved in biomolecular condensation in high-disorder vs. other spatial clusters. Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x inter-quartile range. Median is represented by a white line. Note that for both distributions, the median is zero. **(H)** Distribution of high-disorder spatial clusters in the UMAP embedding from Fig. 3D. Individual nuclear clusters are not outlined for readability. Multiple high-disorder spatial clusters include compartments or proteins known to be characterized by biomolecular condensation behaviors, which are marked by an asterisk.

Figure 5



POLR2F

DNA-directed RNA polymerases I, II, and III subunit RPABC2

P61218 Uniprot ID | ENSG00000100142 ENSG ID

N Tag terminus | GGGTGTCATGTCAGACAACG CRISPR guide RNA sequence

210 nM Protein concentration | 1.3 x 10^5 copies/cell Protein copy number

Allele frequency

HDR 78% | WT 2% | Other 20%

About this protein

DNA-dependent RNA polymerases catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. Common component of RNA polymerases I, II, and III which synthesize ribosomal RNA precursors, mRNA precursors and many functional non-coding RNAs, and small RNAs, such as 5S rRNA and tRNAs,

Reference databases

UniProt | Ensembl | Entrez | HPA

Localization

Nucleoplasm
Nuclear punctae
Nucleolus - FC/DFC
Cytoplasmic

Nucleus image settings
Intensity range 10% - 80%
Gamma 1.00

Target image settings
Intensity range 0% - 90%
Gamma 1.00

Figure 6: the OpenCell website. Shown is an annotated screenshot from our web-app at http://opencell.czbiohub.org, which is described in more details in Suppl Fig. S12.

Figure 6

Science
AAAS

Supplementary Materials for

OpenCell: endogenous tagging for the cartography of human cellular organization

Nathan H. Cho†, Keith C. Cheveralls†, Andreas-David Brunner†, Kibeom Kim†, André C. Michaelis†, Preethi Raghavan†, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y.S. Kim, Edna Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, Manuel D. Leonetti

Correspondence to: manuel.leonetti@czbiohub.org

This PDF file includes:

Materials and Methods
Figs. S1 to S12

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S9

Materials and Methods

Cell culture and CRISPR engineering

Cell culture. HEK-293T cells (ATCC CRL-3216) were cultured in DMEM high-glucose medium (Gibco, cat. #11965118) with 10% fetal bovine serum (Omega Scientific, cat. #FB-11), supplemented with 2mM glutamine (Gibco, cat. #25030081), penicillin and streptomycin (Gibco, cat. #15140163). All cell lines were maintained at 37°C and 5% CO2 and routinely tested for the absence of mycoplasma.

Fluorescent library design. mNeonGreen is monomeric green fluorescent protein ~2x brighter than GFP. We used the split-mNeonGreen$_2$ system for functional tagging, which separates last mNeonGreen$_2$ beta-strand (mNG11) from the rest of the fluorescent protein (mNG1-10)(*20*). Upon co-expression in the same cell, mNG1–10 and mNG11 stably assemble and reconstitute a functional FP. A parental cell line constitutively expressing mNG1-10 was first generated by lentiviral transduction (from pSFFV_mNG$_2$1-10, Addgene #82610). All successive cell lines were generated from this parental HEK293T$^{mNG1-10}$ cell line by incorporation of the mNG11 fragment at either N- or C- terminus of the genomic sequence of a target protein via CRISPR/Cas9 based genome editing. Our mNG11 fusion constructs include a HRV 3C cleavable linker(*91*) that can be used optionally for elution from an affinity capture matrix (16 a.a. tag + 14 a.a. linker, full sequences in Table S1). To minimize the risk of functional perturbation, we stringently selected integration sites (N- or C-terminus) by systematically curating the literature for data supporting the functional integrity of fusion proteins (or by requesting advice from cell biology experts for specific proteins). We also used 3D PDB structures whenever available to identify sites that avoid protein-protein interaction interfaces. See Table S1 for details. Because our split-FP system does not enable detection in the lumen of organelles (this requires split constructs harboring appropriate signal sequences(*23*)), fusions with membrane proteins were restricted to cytoplasmic termini, ensuring first that no annotated regulatory sequences (e.g., signal sequences) were compromised. In total, we used available supporting data to inform 62 % of insertion sites, and 3 % were constrained by membrane protein topology. In the absence of prior information, insertion choice was based on avoiding annotated regulatory sites. In the case of splice variants involving the terminus of choice, the main transcript expressed in HEK293T was used.

Overall genome engineering pipeline. To enable the expression of fluorescent fusion from endogenous genomic loci, we used an established high-throughput CRISPR/Cas9 method for gene editing by homologous recombination (*18*). In brief, *S. pyogenes* Cas9/guide RNA complexes were pre-assembled in vitro, mixed with short single-stranded oligo-nucleotide homology donors and delivered into HEK293T$^{mNG1-10}$ cells by electroporation in 96-well plates (see below). For each genomic insertion, the choice of guide RNA and associated homology donor sequence (which contains the mNG11 payload flanked by short sequences of genomic homology to the targeted insertion site) was automated using *crispycrunch*(*92*), an open-source CRISPR design software available at github.com/czbiohub/crispycrunch and as a web-app at crispycrunch.czbiohub.org/. *crispycrunch* selects a guide RNA closest to a desired genomic insertion site while also minimizing any off-target guide RNA activity, and if needed introduces silent mutations to inactivate guide

RNA biding and re-cutting after successful homologous recombination (*92*). gRNA and homology donor sequences for all targets are found in Table S1.

Cell engineering and selection *S. pyogenes* Cas9 protein (pMJ915 construct, containing two nuclear localization sequences) was expressed in E. coli and purified by the UC Berkeley Macrolab following protocols described by Jinek et al(*93*). Cells were synchronized by nocodazole treatment (200ng/mL for 15-18h_ to enhance homologous recombination(*25*). RNP complexes were freshly assembled with 50 pmol Cas9 protein and 65 pmol gRNA prior to electroporation, and combined with HDR template in a final volume of 10 µL. First, 0.5 µL gRNA (130 µM stock) was added to 2.35 µL high-salt RNP buffer {580 mM KCl, 40 mM Tris-HCl pH 7.5, 20% v/v glycerol, 2 mM TCEP-HCl pH 7.5, 2 mM MgCl2, RNAse-free} and incubated at 70°C for 5 min. 1.25 µL of Cas9 protein (40 µM stock in Cas9 buffer, ie. 50 pmol) was then added and RNP assembly carried out at 37°C for 10 min. Finally, HDR templates and sterile RNAse-free H2O were added to 10 µL final volume. Electroporation was carried out in Amaxa 96-well shuttle Nucleofector device (Lonza) using SF solution (Lonza) following the manufacturer's instructions. Cells were washed with PBS and resuspended to 10,000 cells/µL in SF solution (+ supplement) immediately prior to electroporation. For each sample, 20 µL of cells (ie. 200,000 cells) were added to the 10 µL RNP/template mixture. Cells were immediately electroporated using the CM130 program, after which 100µL of pre-warmed media was added to each well of the electroporation plate to facilitate the transfer of 25,000 cells to a new 96-well culture plate containing 150µL of pre-warmed media. Electroporated cells were cultured for >5 days and transferred to 12-well plates prior to selection by fluorescence-activated cell sorting (FACS). For each target, 1,200 cells from the top 1% fluorescent cell pool were isolated on a SH800 instrument (Sony biotechnology) and collected in 96-well plates.

Genotype analysis. For each polyclonal pool of engineered cells, the genotype of CRISPR-edited alleles was characterized by amplicon sequencing. Gene-specific primers were designed using Primer3, with a target amplicon length of 270bp and a maximum at 500bp. gDNA was first extracted by cell lysis using QuickExtract DNA Extraction Solution (Lucigen). From a confluent culture in 96-well plate, media was removed, cells were washed 1x in DPBS and resuspended in 50 µL QuickExtract. The cell layer was detached by repeated pipetting and transferred to a PCR plate for incubation. The lysate was incubated as follows {65°C for 20 min, 98°C for 5min, 4°C final}. gDNA was used directly from this preparation. Amplicon Libraries were created using a two-step PCR protocol: the first PCR amplifies the target genomic locus and adds universal amplification handle sequences, while the second PCR introduces index barcodes using the universal handles. PCR1: this PCR uses a "reverse touchdown" method designed to accommodate a number of different annealing temperatures for a number of different targets. 50-µL PCR reactions were set using 2x KAPA HiFi Hotstart reagents (Roche) with 2µL extracted gDNA, 80pmol each primer and betaine to 0.8M final concentration. PCR conditions: 95°C 3min; 3 cycles of {98°C for 20s, 63°C for 15s, 72°C for 20s}, 3 cycles of {98°C for 20s, 65°C for 15s, 72°C for 20s}, 3 cycles of {98°C for 20s, 67°C for 15s, 72°C for 20s}; 17 cycles {98°C for 20 s, 69°C for 15 s, 72°C for 20s} then 72°C for 1min; 4°C final. PCR2: amplicons were diluted 1:100 and 1 µL was used into a 40-µL barcoding reaction using 20 µL 2x KAPA HiFi Hotstart reagents (Roche) and 80pmol each barcoded primer. PCR conditions: 95°C 3min and 12 cycles of {98°C for 20s, 68°C for 15s, 72°C for 12s} then 72°C for 1min; 4°C final. Barcoded amplicons were analyzed using capillary electrophoresis (Fragment Analyzer, Agilent), pooled and purified using magnetic beads. Sequencing was performed on an Illumina Miseq V3 platform (paired-end 2x300bp) using standard P5/P7 primers. Genotype analysis was performed using CRISPRESSO2, which allowed to quantify three classes of alleles for each targeted locus: un-modified (wild-type), alleles integrated with mNG11 by homologous recombination, and alleles containing non-functional mutations as a result of competing DNA repair mechanisms. Primer sequences and genotype analysis for all targets are found in Table S1. Despite multiple attempts, genotyping PCR could not be successfully performed for 70 targets (5% of the total set), most often involving genes with extreme GC content or highly repetitive sequences.

Immuno-precipitation / mass-spectrometry

Overall strategy. mNG11-tagged proteins were isolated from digitonin-solubilized lysates using anti-mNeonGreen nanobody capture. Biological triplicate protein samples were digested "on-bead" for bottom-up proteomics analysis(*28*), and peptides were quantified using label-free mass spectrometry on a timsTOF Pro instrument (Bruker Daltonics).

Sample preparation. Confluent 12-well cultures (0.8x10^6 cells/sample) were washed twice with 1 ml of D-PBS (no divalent). 200 µl ice-cold lysis buffer A {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 2 mM MgOAc, 1 mM $CaCl_2$, 15% Glycerol, 1.5 % Digitonin (high purity, Calbiochem), Protease- and Phosphatase inhibitor (Halt, Pierce), 0.1% benzonase (Millipore Sigma)} were added to each well, cells were lysed by strong pipetting and the solution was transferred into a pre-chilled 96-well PCR plate. Per 96-well plate, 330 µl magnetic mNG-Trap slurry (magnetic agarose, Chromotek) was washed three times with buffer B {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 2 mM MgOAc, 1 mM $CaCl_2$, 15% Glycerol, 0.1 % Digitonin} and resuspended in 2,150 µl Buffer A. The cell lysate was incubated for 1h at 4°C, rotating. The insoluble cell fraction was pelleted for 30 min at 1800xg in a table-top centrifuge at 4°C, followed by supernatant transfer into a new plate pre-loaded with 20 µl of the washed bead slurry per well. Tagged proteins were captured by incubation for 2h at 4°C, rotating. Following capture and using a 96-well magnet, beads were washed (per well) with 200 µl buffer B (incubation for 5 min at 4°C, rotating), 2x 200 µl buffer B (no incubation) and a final 1x 200 µl buffer C to remove digitonin {50 mM HEPES pH 7.5, 150 mM KOAc, 5 mM NaCl, 15% Glycerol, 0.01% glyco-diosgenin (Avanti)}. Supernatant was removed and 50 µl of digestion buffer 1 {6 M Urea, 50 mM Tris-HCl, pH 8.5, 1 mM DTT, 2 ng/µl LysC protease (Wako Chemicals)} was added to each well, followed by overnight digestion at 30°C on a thermomixer, gently shaking. The next day, 100 µl digestion buffer 2 {50 mM Tris-HCl, pH 8.5, 8.25 mM iodoacetamide, 2 ng/µl LysC} was added to each well and incubated for ~6 hours at 30°C on a thermomixer in the dark, gently shaking. The digestion was finally quenched with 15 µl of 10 % TFA. Quenched samples were vortexed, flash-frozen and stored at -80 °C until further use for LC-MS analysis preparation.

EvoSep chromatography. We used the EvoSep liquid chromatography system for sample processing(*94*). EvoTips (EvoSep Gmbh) were activated for 5 min with 1-Propanol at RT, followed by a wash step with 50 µl Buffer A (99.9 % ddH2O, 0.1 % Formic Acid) and centrifugation at 600 xg for 1 min at RT. The flow-through was discarded and activated EvoTips were placed in an EvoTip-box reservoir filled with Buffer A. After on-bead digestion, captured protein samples were thawed for 5 min at 600 rpm and 25°C on a thermal shaker and placed on a 96-well magnet holder to remove magnetic beads. The whole sample (~150 µl) was transferred to activated EvoTips, followed by two consecutive centrifugation steps at 600xg for 1 min and RT, discarding flow-through after the first spin. Peptide-loaded EvoTips were washed once with 50 µl Buffer A and centrifuged at 600xg for 1 min at RT. The flow-through was discarded and 150 µl of Buffer A was added to each EvoTip followed by a centrifugation step for 20 sec at 600xg RT. Loaded EvoTips were then transferred into the 96-well EvoTip-box reservoir filled with Buffer A and transferred onto the EvoSep autosampler for LC-MS analysis. Pulldowns were acquired in triplicates and injected to the mass spectrometer while spacing replicates to prevent any bias.

Liquid-chromatography. For separating peptides by hydrophobicity and eluting them into the mass spectrometer, we used an EvoSep One1 liquid chromatography system (EvoSep, Gmbh) and analyzed purified petides with a standard 21 min method (60 samples per day). We used a 15 cm × 150 µm ID column with 1.9 µm C18 beads (PepSep) coupled to a 20 µm ID electrospray emitter (Bruker Daltonics). Mobile phases A and B were 0.1 % FA in water and 0.1 % FA in ACN, respectively. The EvoSep system was coupled online to a trapped ion mobility spectrometry quadrupole time-of-flight mass spectrometer(*95*) (timsTOF Pro, Bruker Daltonics) equipped with via a Captive nano-electrospray ion source.

Mass spectrometry. Mass spectrometric analysis was performed in a data-dependent (dda) PASEF mode. For ddaPASEF, 1 MS1 survey TIMS-MS and 4 PASEF MS/MS scans were acquired per acquisition cycle. The cycle overlap for precursor scheduling was set to 2. Ion accumulation and ramp time in the dual TIMS analyzer was set to 50 ms each and we analyzed the ion mobility range from 1/K0 = 1.3 Vs cm-2 to 0.8 Vs cm-2. Precursor ions for MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 in a total m/z range of 100-1,700 by synchronizing quadrupole switching events with the precursor elution profile from the TIMS device. The collision energy was lowered linearly as a function of increasing mobility starting from 59 eV at 1/K0 = 1.6 VS cm-2 to 20 eV at 1/K0 = 0.6 Vs cm-2. Singly charged precursor ions were excluded with a polygon filter (otof control, Bruker Daltonics). Precursors for MS/MS were picked at an intensity threshold of 2,000 arbitrary units (a.u.) and re-sequenced until reaching a 'target value' of 24,000 a.u. considering a dynamic exclusion of 40 s elution. Capillary voltage was set to 1,750 V and dry gas temperature to 180°C.

Raw Data Processing. MS raw files were processed using MaxQuant (v1.6.10.43)(*96*, *97*), which extracts features from four-dimensional isotope patterns and associated MS/MS spectra, on a computing cluster (SUSE Linux Enterprise Server 15 SP2) utilizing UltraQuant. Files were processed in several batches of appriximately 1000 files each and searched against the human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa). False-discovery rates were controlled at 1% both on peptide spectral match (PSM) and protein levels. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamido-methylation as fixed modification, while limiting the maximum peptide mass to 4,600 Da. Enzyme specificity was set to LysC cleaving c-terminal to lysine. A maximum of two missed cleavages were allowed. Maximum precursor and fragment ion mass tolerance were searched as default for TIMS-DDA data and the main search tolerance was reduced to 20 ppm. Peptide identifications by MS/MS were transferred by matching four-dimensional isotope patterns between the runs (MBR) with a 0.7-min retention-time match window and a 0.05 1/K0 ion mobility window. Protein quantification was performed by label-free quantification using a minimum ratio count of 1.

Data availability. All mass spectrometry raw data and MaxQuant output tables are deposited to the ProteomeXchange Consortium(*98*) via the PRIDEpartner repository and will be publicly available upon final publication (accession PXD024909).

Whole-cell abundance measurement by mass-spectrometry

Peptide preparation. HEK293T cells were grown in biological triplicate 15cm-plates, washed 2x in ice-cold PBS and lysed in { 2.5% SDS sodium dodecyl-sulfate; 50 mM Tris pH 8.1 }. Lysis was performed at 95°C for 5 min, followed by probe sonication. Lysates were cleared by centrifugation, protein amount was measured by BCA assay, and lysates were precipitated with 5 volumes of acetone. Pellets were resuspended in 50 mM Tris pH 8.1 containing 8 M urea, reduced with 1 mM DTT and alkylated with 5 mM IAA before initiation of digestion overnight with LysC at an enzyme-to-protein ratio of 1:100. The digest mixture was diluted four-fold, and trypsin was added at an enzyme-to-protein ratio of 1:100 for 6 h, followed by an additional aliquot of trypsin overnight. The digestion reaction was stopped by acidifying the sample adding TFA to 1%, placed on ice for 10min and centrifuged at 4 degree C, 21000g for 20min. The resulting peptide supernatant was then desalted using mixed mode Strata-XC SPE cartridge. Briefly, the cartridge was prepared by activating with methanol, conditioning with 80% acetonitrile/0.1% TFA and equilibrated with 0.2% TFA. The acidified peptides were then added, washed with 99% isopropanol/0.1% TFA, 2 x 0.2% TFA washes, 1x 0.1% formic acid and eluted with 60% acetonitrile/0.5% ammonium hydroxide. The eluted peptides were flash frozen and then dried down.

Fractionation. To obtain achieve measurement depth, peptides from the triplicate experiment were further separated in 24 fractions using C18 chromatography. Peptides were resuspended in buffer A (10 mM ammonium bicarbonate) and injected onto a 4.6 × 250-mm 3.5-µm Zorbax 300 Extend-C18 column. Peptides were separated on a non-linear gradient as described in (*99*), using the following composition of buffer B (10 mM ammonium bicarbonate, 90% acetonitrile). Peptide fractions were frozen at −80 °C before centrifugal evaporation.

Mass spectrometry. Peptides were resuspended in 2% ACN with 0.1% TFA before loading onto a 25 cm x 75 µm ID, 1.6 µm C18 column (IonOpticks) maintained at 40°C. Peptides were separated with an EASY-nLC 1200 system (Thermo Fisher Scientific, San Jose, CA) at a flow rate of 300 nl min-1 using a binary buffer system of 0.1% FA (buffer A) and 80% acetonitrile with 0.1% FA (buffer B) in a two-step gradient, from 3% to 27% B in 105 min and from 27% to 40% B in 15min. All fractions were analyzed on a Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with a nanoFlex ESI source operated at 1550 volts, RF lens set to 30%, operated in data dependent acquisition mode with a duty cycle time of 1 sec. Full MS scans were acquired with a m/z scan range of 375-1500 m/z in the Orbitrap mass analyzer (FTMS) with a resolution of 240k. Selected precursor ions were subjected to fragmentation using higher-energy collisional dissociation (HCD) with a quadrupole isolation window of 0.7 m/z, and normalized collision energy of 31%. HCD fragments were analyzed in the Ion Trap mass analyzer (ITMS) set to Turbo scan rate. Fragmented ions were dynamically excluded from further selection for a period of 45 sec. The AGC target was set to 1,000,000 and 10,000 for full FTMS and ITMS scans, respectively. The maximum injection time was set to Auto for both full FTMS and ITMS scans.

Live-cell imaging

Sample preparation. Live-cell imaging was performed on 96-well glass-bottom plates (Greiner Bio One, cat. #655891) coated with 50µg/ml fibronectin (Corning, cat. #356008). Cells were seeded on an imaging plate 28-32 hours before imaging at 15,000 cells per. Before imaging, cells were counter-stained with the live-cell DNA dye Hoechst 33342 (Invitrogen, cat. #H3570) by incubation for 30 minutes at 37°C in 150 µl of Hoechst diluted to 1µg/mL in culture media. Media was then replaced with phenol-free DMEM (Gibco, cat. #21063029) supplemented with 10% FBS. Hoechst staining was performed three to four hours prior to imaging to provide the cells time to recover from any mechanical stress due to medium changes.

Live-cell fluorescence microscopy. Cells were imaged on a DMI-8 inverted microscope (Leica) equipped with a Dragonfly spinning-disk confocal system (Andor), a 63x 1.47NA oil objective (Leica), and a 16-bit iXon Ultra 888 EMCCD camera (Andor, pixel size: 13x13 µm^2). A pinhole size of 40µm was used with an EM gain of 400. Cells were maintained at 37°C and 5% CO2 during image acquisition by a stage-top incubator (Okolab, H101-K-Frame). The microscope was controlled using the open-source microscope-control software MicroManager (version 1.4.22).

Automated confocal acquisition. We automated the imaging of 96-well plates using a custom acquisition script, written in Python, combined with a custom MicroManager plugin (mm2python; github.com/czbiohub/mm2python) to expose the MicroManager APIs in a Python environment. This script selected optimal fields of view (FOVs) at which to acquire confocal z-stacks by using a pre-trained machine-learning model to assign a quality score to the FOVs at a set of different positions in each well. Briefly, at each position, the script acquired a single 2D snapshot of the Hoechst staining, segmented the nuclei in the snapshot, and calculated an array of features associated with the distribution of nuclei within the FOV. The script then used a pre-trained random-forest regression model (see below) to predict a quality score for the FOV from this set of features. This process was repeated at each of 25 different positions in each well, and then the script selected the positions with the highest-scoring FOVs to revisit for confocal z-stack acquisition. At each of these selected positions, the focal plane was centered on the cell layer using a laser-based Adaptive Focus Control system (Leica) and confocal z-stacks, consisting of 110 z-slices at a spacing of 0.2µm, were acquired. The exposure settings for the mNeonGreen channel were determined dynamically for each target using a custom auto-exposure algorithm that iteratively adjusted the exposure time and laser power until the maximum pixel intensity was just below or just above an intensity of 2^{15} (half of the full dynamic range of the camera). For dim targets for which this condition could not be met, the script fell back to a hard-coded absolute maximum exposure time and laser power to minimize both acquisition time and photobleaching. The exposure settings for the Hoechst stain were manually selected and held constant for all targets. The random-forest regression model used by the script to predict the FOV quality scores was trained prior to acquisition using a set of 3800 FOV snapshots that were manually assigned to one of three grades: “poor,” “mediocre,” or “good.” These grades were mapped to a continuous response variable by assigning the values of -1, 0, and 1, respectively, and a random forest regression model (scikit-learn) was trained to predict this value. The out-of-bag estimated R^2 was 0.86 and scores predicted for a withheld set of test snapshots were also evaluated by manual inspection. The trained model was cached and imported at acquisition time by the acquisition script. The acquisition script, trained FOV-scoring model, autoexposure algorithm, and other associated microscope-control methods are available online at github.com/czbiohub/2021-opencell-microscopy-automation.

Data analysis – proteomics

Statistical detection of protein interactions. Statistical analysis was performed according to methods described in Hein et al. (*7*), with modifications. Protein identifications were filtered, removing common contaminants, hits to the reverse decoy database as well as proteins only identified by modified peptides. We required that each protein be quantified in all replicates from the IP-MS samples of at least one cell line and used log2 MaxQuant LFQ intensities for all analyses. Rather than imputing missing values, robust null control sets were generated for statistical enrichment analysis of each protein group by pooling triplicate data from an average of 349 unrelated samples. In this approach, rather than using a single control we measure enrichment in a specific sample against an entire cohort of ~349 unrelated tagged cell lines. We have previously described (*7*, *28*) how this enables a better estimation of the null distribution and leads to more robust identification of interactions. The null control sets might contain triplicate samples that are outliers and would be considered significant interactions. The presence of these samples lead to underestimation of enrichment and could mask some significant interactions. We systematically removed these outliers from the negative control sets using a Student's t-test and excluding any sample of triplicates that had a p-value < 0.001. From the filtered pool, we approximated the true mean and the true standard deviation of the null set by bootstrapping via sampling with replacement. The approximated mean and standard deviation of the null set was then used for the final Student's t-test to calculate the statistical significance of the triplicate means. Any missing values in the triplicate sample set were then replaced with the mean of the null set. Enrichment was calculated by subtracting the mean of the triplicates from the mean of the null set, and was normalized to account for variability within each protein through division by the standard deviation of the null control set. Our statistical strategy to define significant interactors is described in Figure S4A-B and supported by a quantitative estimation of precision and recall.

Precision / recall analysis of the interactome. For a quantitative evaluation of our statistical approach, and to compare the quality of OpenCell against reference interactome datasets, we created a framework the precision and recall in interaction data. In the absence of established ground truth for human protein interactions, we indirectly derived measurements of precision and recall. For recall, we calculated the coverage in a given dataset of interactions curated in the human CORUM database(*30*), as a percentage of all possible CORUM interactions given the set of baits in that dataset. For calculating precision, we used the assumption that two interactors should have localization patterns that at least partially overlap. As an independent ground truth set for protein localization, we used the quantitative analysis of the HeLa proteome from Itzhak et al. (*31*). Using these annotations, we categorized localization into four broad classes: exclusively nuclear, exclusively cytoplasmic, exclusively organellar, and multi-localizing (i.e., any non-exclusive localization). To calculate precision, we consider any two interactors that overlap in exclusive localization to be true positives, and those that do not overlap localization annotations at all to be false positives, with multi-localizing proteins allowed to interact agnostically (Fig. S4B).

Protein stoichiometry measurements. Calculation of interaction stoichiometries was performed as in Hein et al by dividing LFQ intensities by the number of theoretically observable peptides for each protein. We defined the "interaction stoichiometry" as the stoichiometry of the abundance of a given interactor, relative to the abundance of the corresponding bait, in a given pull-down. We also defined a "cellular abundance stoichiometry" as the stoichiometry of the abundance of a given interactor, relative to the abundance of the corresponding bait, in a whole cell lysate. For proteins that were not detected in whole cell lysates (due to lack of measurable peptides, for example in the absence of lysine residues), protein abundances were imputed from RNA-Seq data by interpolating from a linear regression of RNA-Seq tpm vs. protein abundance measured by mass spectrometry in our dataset.

Network Analysis. For graph-based clustering of the entire interactome network, we weighted edges using the interaction stoichiometry between each pair of interacting proteins. We utilized Markov clustering(*34*) at various inflation parameters and evaluated clustering performance using the k-clique method described in Drew et al(*100*) using CORUM complexes as the ground truth. To eliminate complexes with many shared proteins, the Jaccard distance was calculated between all pairs of complexes, and pairs of complexes were merged if the distance was below 0.6. Our final clustering analysis used an inflation parameter of 3.0 (Fig. S4I). The clusters were pruned to remove any node included in a cluster on the basis of a single edge. The resulting clusters correspond to the protein "communities" described in the text. We then utilized another round of MCL clustering to identify core-clusters within each community by considering only highly stoichiometric interactions (interaction stoichiometries between 0.05 and 10, and cellular abundance stoichiometry between 0.1 and 10). The resulting core-clusters represented highly stable core clusters within the original communities.

Measurement of biophysical properties of proteins. Biophysical properties were calculated using the *ProteinAnalysis* package from BioPython(*101*). Hydrophobicity scores were calculated using the *gravy* method of that module to compute the Gravy index. Calculation of disorder in protein regions was performed using the IUPred2A algorithm(*89*) or metapredict, a recent agglomerative algorithm (*90*). Scores were averaged across the sequences of each protein. Scores computed across the whole proteome are included in Table S2.

Data analysis – imaging

Consensus localization encodings. Protein localization patterns were encoded from the raw confocal images using a customized variant of the vector-quantized autoencoder architecture VQ-VAE-2(*102*). The image preprocessing, autoencoder architecture, and model training are described in detail in an accompanying manuscript(*52*). Briefly, confocal z-stacks were reduced to two dimensions by a maximum-intensity z-projection and normalized to control for variation in intensity. Regions of interest 200x200 pixels in size were centered on individual nuclei and cropped from each z-projection to generate a set of 50-200 cropped images for each tagged protein. These images were randomly partitioned into a training set and a test set. After training the model on the images in the training set, the images in the test set were encoded, and the resulting latent-space vectors from the VQ2 layer of the network were flattened to obtain a localization encoding for each image in the test set as a 9216-dimensional vector. The encodings of all images for each tagged protein were then averaged to obtain a single consensus encoding for each tagged protein. The matrix of consensus encodings for all OpenCell targets are available on Figshare at:

https://figshare.com/articles/dataset/Consensus_protein_localization_encodings_for_all_OpenCell_targets/16754965

Comparison of OpenCell and Human Protein Atlas (HPA) localization annotations. The v20 dataset of HPA localization annotations was first obtained from the HPA website (https://v20.proteinatlas.org/download/subcellular_location.tsv.zip). To compare OpenCell localization annotations to HPA annotations, it was necessary to reconcile the OpenCell and HPA localization categories as the ontologies used to annotate the two datasets vary slightly. To do so, a set of 'consensus' annotation categories were defined for the most common localization categories, as described in Table S7 (see sheet: "annotation-definitions").

Because wide-spread multi-localization of proteins complicates direct comparisons, we focused first on comparing the "main" localization annotations provided by each dataset. After mapping to these consensus categories, grade-2 and grade-3 OpenCell annotations were compared to their corresponding HPA 'main location' annotations and categorized as either exact matches, partial matches, or entirely discrepant. Exact matches were targets whose sets of consensus annotations were identical in the OpenCell and HPA datasets; partial matches were targets with at least one of the same consensus annotations in the OpenCell and HPA datasets. The list of exact and partial matches, and the sets of consensus OpenCell and HPA annotations, are provided in Table S7.

To refine the list of proteins that did not share any matching annotation across the datasets, minor localization annotations were considered (grade 1 in OpenCell, "additional" localization in HPA), as well localization between closely related organelles (for example, ER and Golgi), which could explain differences between the datasets as they probe localization in different cell lines. As a result, a final list of 147 proteins for which the two dataset were fully discrepant was obtained. The full analysis of proteins from that list using literature curation is presented in Table S8.

Analysis of image localization encodings. The matrix of consensus localization encodings for all OpenCell targets was analyzed using the *scanpy* package(*103*). Briefly, the dimensionality of the consensus encodings was reduced using PCA and the first 200 PCs, which captured 96% of the variance, were retained for downstream analysis. The UMAP algorithm (*53*) was used to embed the encodings in two-dimensional space using 10 nearest neighbors, the Euclidean distance metric, and a minimum embedding distance of zero. The encodings were clustered using the Leiden graph-based clustering algorithm(*55*) with a resolution parameter of 30 and the weighted adjacency matrix calculated by the UMAP algorithm (again with 10 nearest neighbors). Finally, the Pearson correlation coefficient between the top 200 PCs of the localization encodings was used to quantify the localization similarity between OpenCell targets.

Image-based clustering. OpenCell targets were clustered on the basis of their consensus encodings using the Leiden graph-based clustering algorithm (*55*) and the weighted adjacency matrix calculated by the UMAP algorithm with 10 nearest neighbors. The Leiden algorithm depends upon a single ‘resolution’ hyperparameter that determines the number of clusters. To quantify clustering performance as a function of this hyperparameter, the Adjusted Rand Index (*88*) was used to compare the Leiden clusters to ground-truth datasets. The ARI is near zero for random clustering and is equal to one when clustering perfectly matches the ground-truth labels. Three different ground-truth datasets were used that capture biological relationships at three different scales: manual OpenCell localization annotations (organelle scale), KEGG pathways (https://www.genome.jp/kegg/ (*104*)), and CORUM complexes (http://mips.helmholtz-muenchen.de/corum/ (*30*). OpenCell targets that were in more than one ground-truth cluster were excluded from this analysis, as the ARI is defined only for hard clustering (that is, sample-cluster assignments that are one-to-one). The ARI was calculated with respect to each of the ground-truth datasets at a range of values of the Leiden resolution hyperparameter; the global maxima in the resulting ARI curves correspond to the clustering resolutions that best capture the information in each ground-truth dataset. To control for the stochasticity of the Leiden algorithm, the ARI curve was calculated from the average of the curves for nine random seeds.

Hierachical analysis of interactions and localization patterns.

Hierarchical clustering of interactome and image-localization clusters. To explore the relationships between the 182 localization clusters or the 300 interactome communities, we employed the Paris algorithm, an agglomerative graph-based hierarchical clustering algorithm(*105*). The algorithm was initialized with a network of nodes representing the initial clusters (either the localization clusters or the interactome communities) and edge weights between the initial clusters were calculated according to the definition of the cluster pair sampling ratio used in the Paris algorithm.

Gene Ontology enrichment analysis. To analyze enrichment of GO terms in a given hierarchical protein group, we utilized the PANTHER gene list analysis API (*106*) using Fisher exact test for significance testing. Enrichment of GO terms was tested against a reference set of either all OpenCell targets for the imaging dataset, or all proteins found in communities for the interactome dataset.

OpenCell web portal development

The OpenCell web portal is a full-stack web application. The frontend (that is, the web interface itself) is written with React, a modern JavaScript library for building modular user interfaces. The backend is a PostgreSQL database paired with a REST API written in Python using Flask and SQLAlchemy. Together, the database and API provide the metadata, protein interaction data, and the confocal image data required to populate the frontend. For efficiency, the 3D confocal stacks are transferred to the client as two-dimensional tiled arrays of confocal slices, saved as compressed JPEG images to enable fast download times. To maximize responsiveness, the web app makes API requests dynamically and asynchronously so that it loads, in parallel, only the data required to update the state of the app in response to a given user input. Both the backend and frontend are built using many open-source packages. In particular, the 3D rendering of confocal stacks relies on Three.js, the interactive scatterplots are built with d3.js, and the interaction networks are built with Cytoscape.js. The backend is built with SQLAlchemy and Flask and also leverages the Python data-science stack, including pandas, NumPy, SciPy, and scikit-image. All source code for the application is available on GitHub at github.com/czbiohub/opencell-portal-pub

Figure generation

Data analysis was performed in Python. Figures were generated in Python using matplotlib or seaborn, with the exception of the protein-protein interactions / network visualizations, which were generated using Cytoscape. The code and data used to generate the figures can be found on GitHub at github.com/czbiohub/2021-opencell-figures.

Supplementary references

91. C. L. Young, Z. T. Britton, A. S. Robinson, Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. Biotechnol J. 7, 620–634 (2012).

92. G. Dingle, CrispyCrunch: High-throughput Design and Analysis of CRISPR+HDR Experiments, (available at https://blog.addgene.org/crispycrunch-high-throughput-design-and-analysis-of-crisprhdr-experiments).

93. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (New York, N.Y.). 337, 816–821 (2012).

94. N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Müller, F. Meier, J. V. Olsen, O. Vorm, M. Mann, A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics*. Mol Cell Proteomics. 17, 2284–2296 (2018).

95. F. Meier, A.-D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, N. Bache, O. Hoerning, J. Cox, O. Räther, M. Mann, Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer*. Mol Cell Proteomics. 17, i–2545 (2018).

96. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 26, 1367–1372 (2008).

97. N. Prianichnikov, H. Koch, S. Koch, M. Lubeck, R. Heilig, S. Brehmer, R. Fischer, J. Cox, MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics*. Mol Cell Proteomics. 19, 1058–1069 (2020).

98. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 32, 223–226 (2014).

99. P. Mertins, L. C. Tang, K. Krug, D. J. Clark, M. A. Gritsenko, L. Chen, K. R. Clauser, T. R. Clauss, P. Shah, M. A. Gillette, V. A. Petyuk, S. N. Thomas, D. R. Mani, F. Mundt, R. J. Moore, Y. Hu, R. Zhao, M. Schnaubelt, H. Keshishian, M. E. Monroe, Z. Zhang, N. D. Udeshi, D. Mani, S. R. Davies, R. R. Townsend, D. W. Chan, R. D. Smith, H. Zhang, T. Liu, S. A. Carr, Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography–mass spectrometry. Nat Protoc. 13, 1632–1661 (2018).

100. K. Drew, C. Lee, R. L. Huizar, F. Tu, B. Borgeson, C. D. McWhite, Y. Ma, J. B. Wallingford, E. M. Marcotte, Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology. 13, 932 (2017).

101. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25, 1422–1423 (2009).

102. A. Razavi, A. van den Oord, O. Vinyals, Generating Diverse High-Fidelity Images with VQ-VAE-2. Arxiv (2019).

103. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15 (2018).

104. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462 (2016).

105. T. Bonald, B. Charpentier, A. Galland, A. Hollocou, Hierarchical Graph Clustering using Node Pair Sampling. Arxiv (2018).

106. H. Mi, D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou, T. Mushayamaha, P. D. Thomas, PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49, gkaa1106- (2020).

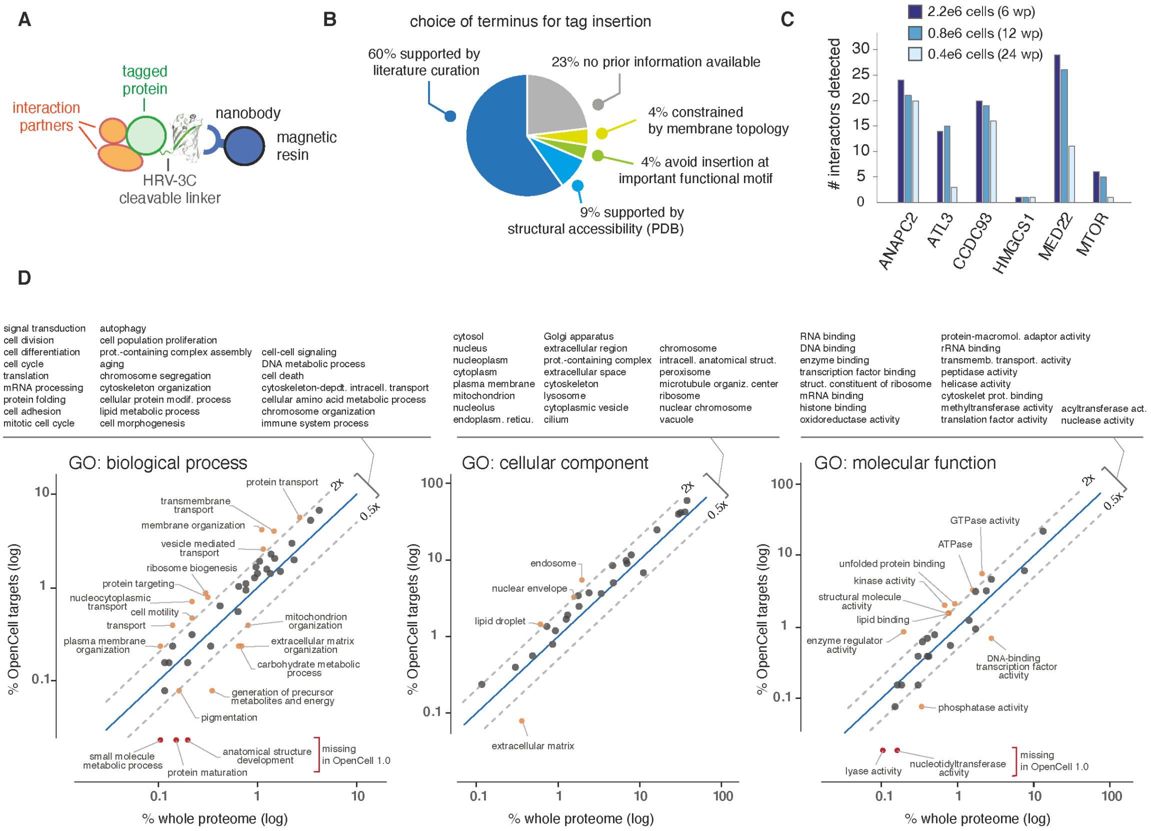

A
tagged protein
interaction partners
nanobody
magnetic resin
HRV-3C cleavable linker
B
choice of terminus for tag insertion
60% supported by literature curation
23% no prior information available
4% constrained by membrane topology
4% avoid insertion at important functional motif
9% supported by structural accessibility (PDB)
C
2.2e6 cells (6 wp)
0.8e6 cells (12 wp)
0.4e6 cells (24 wp)
# interactors detected
0
5
10
15
20
25
30
ANAPC2
ATL3
CCDC93
HMGCS1
MED22
MTOR
D
signal transduction
cell division
cell differentiation
cell cycle
translation
mRNA processing
protein folding
cell adhesion
mitotic cell cycle
autophagy
cell population proliferation
prot.-containing complex assembly
aging
chromosome segregation
cytoskeleton organization
cellular protein modif. process
lipid metabolic process
cell morphogenesis
cell-cell signaling
DNA metabolic process
cell death
cytoskeleton-depdt. intracell. transport
cellular amino acid metabolic process
chromosome organization
immune system process
GO: biological process
protein transport
transmembrane transport
membrane organization
vesicle mediated transport
ribosome biogenesis
protein targeting
nucleocytoplasmic transport
cell motility
transport
plasma membrane organization
mitochondrion organization
extracellular matrix organization
carbohydrate metabolic process
generation of precursor metabolites and energy
pigmentation
small molecule metabolic process
protein maturation
anatomical structure development
missing in OpenCell 1.0
2x
0.5x
10
1
0.1
% OpenCell targets (log)
0.1
1
10
% whole proteome (log)
cytosol
nucleus
nucleoplasm
cytoplasm
plasma membrane
mitochondrion
nucleolus
endoplasm. reticu.
Golgi apparatus
extracellular region
prot.-containing complex
extracellular space
cytoskeleton
lysosome
cytoplasmic vesicle
cilium
chromosome
intracell. anatomical struct.
peroxisome
microtubule organiz. center
ribosome
nuclear chromosome
vacuole
GO: cellular component
endosome
nuclear envelope
lipid droplet
extracellular matrix
2x
0.5x
100
10
1
0.1
% OpenCell targets (log)
0.1
1
10
100
% whole proteome (log)
RNA binding
DNA binding
enzyme binding
transcription factor binding
struct. constituent of ribosome
mRNA binding
histone binding
oxidoreductase activity
protein-macromol. adaptor activity
rRNA binding
transmemb. transport. activity
peptidase activity
helicase activity
cytoskelet prot. binding
methyltransferase activity
translation factor activity
acyltransferase act.
nuclease activity
GO: molecular function
GTPase activity
ATPase
unfolded protein binding
kinase activity
structural molecule activity
lipid binding
enzyme regulator activity
DNA-binding transcription factor activity
phosphatase activity
lyase activity
nucleotidyltransferase activity
missing in OpenCell 1.0
2x
0.5x
100
10
1
0.1
% OpenCell targets (log)
0.1
1
10
100
% whole proteome (log)

Figure S1

Fig. S1: Experimental pipeline (related to Fig. 1). (A) IP-MS using FP capture. All mNG11 tagging constructs also include an HRV-3C cleavable linker for optional release from the capture resin. **(B)** Justifying the choice of tag insertion in engineered cell lines. To inform tag insertion sites, we used a combination of existing data from the literature suggesting preservation of properties, 3D structures of protein complexes from the PDB and sequence analysis to avoid important functional motifs. 4% of insertion sites were constrained by the topology of transmembrane protein targets (fusion to cytosolic termini), and for 23% of targets no prior data was available. See details in Table S1. **(C)** Sensitivity of interaction proteomics detection on a timsTOF instrument. The number of interactors detected in pull-downs from 6 different targets is shown, varying the amount of input material. To balance sensitivity and scalability, 0.8e6 cells were used for high-throughput assays (12 well-plate, wp). **(D)** Distribution of gene ontology annotations in the OpenCell library (successful targets) compared the whole proteome. Over- and under-represented terms are outlined. Because organellar organization and transport between organelles are foundational to human cellular architecture, proteins in these groups are slightly enriched in our library. Under-represented groups are mostly comprised of proteins in compartments that are not accessible to our tagging strategy (mitochondrial functions, extracellular matrix) or proteins that are typically present at low copy numbers and therefore difficult to detect at endogenous levels (transcription factors).

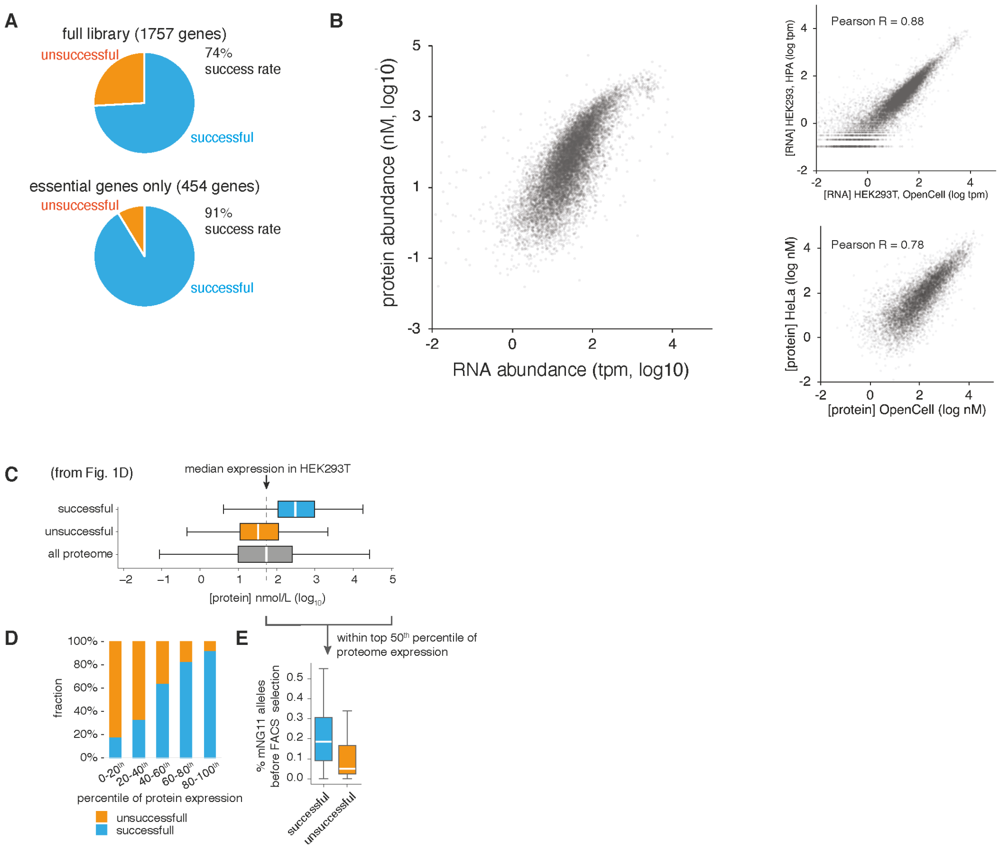

A
full library (1757 genes)
unsuccessful
74%
success rate
successful
essential genes only (454 genes)
unsuccessful
91%
success rate
successful
B
protein abundance (nM, log10)
5
3
1
-1
-3
-2
0
2
4
RNA abundance (tpm, log10)
Pearson R = 0.88
[RNA] HEK293, HPA (log tpm)
[RNA] HEK293T, OpenCell (log tpm)
Pearson R = 0.78
[protein] HeLa (log nM)
[protein] OpenCell (log nM)
C
(from Fig. 1D)
median expression in HEK293T
successful
unsuccessful
all proteome
–2
–1
0
1
2
3
4
5
[protein] nmol/L (log10)
D
100%
80%
60%
40%
20%
0%
fraction
0-20th
20-40th
40-60th
60-80th
80-100th
percentile of protein expression
unsuccessfull
successfull
E
within top 50th percentile of proteome expression
% mNG11 alleles before FACS selection
0.5
0.4
0.3
0.2
0.1
0.0
successful
unsuccessful

Figure S2

Fig. S2: Cell line generation (related to Fig. 1). (A) Success rate for the generation and detection by imaging of fluorescently tagged cell lines are compared for the whole set of targets we attempted, and the subset of these that are essential genes. **(B)** Correlation of protein and RNA abundance in HEK293T cells (OpenCell). For comparison purposes, RNA and protein abundances in our dataset are compared to two external references: HEK293 cell line RNASeq from the Human Protein Atlas, and the HeLa proteome published in (*7*). In both cases, our data correlates well with existing references. **(C)** Repeated from Fig. 1C. **(D)** Fluorescent detection success rates for proteins at different percentiles of abundance in the proteome. **(E)** To evaluate the influence of CRISPR editing efficiency on the ability to successfully select fluorescently tagged cells, we genotyped 432 cell lines from our library before FACS sorting (these lines were randomly selected). After FACS sorting (top 1% fluorescent cells, see Fig. S3A), all lines were imaged by fluorescence microscopy. Within this set, no fluorescence could be detected in 99 lines (23% of total). For well-expressed proteins (top 50th percentile of abundance in the whole proteome), un-successful detection is correlated with low rates of CRISPR-mediated homologous recombination before FACS selection. A low rate of homologous recombination likely prevents the successful selection of a fluorescent pool by FACS.

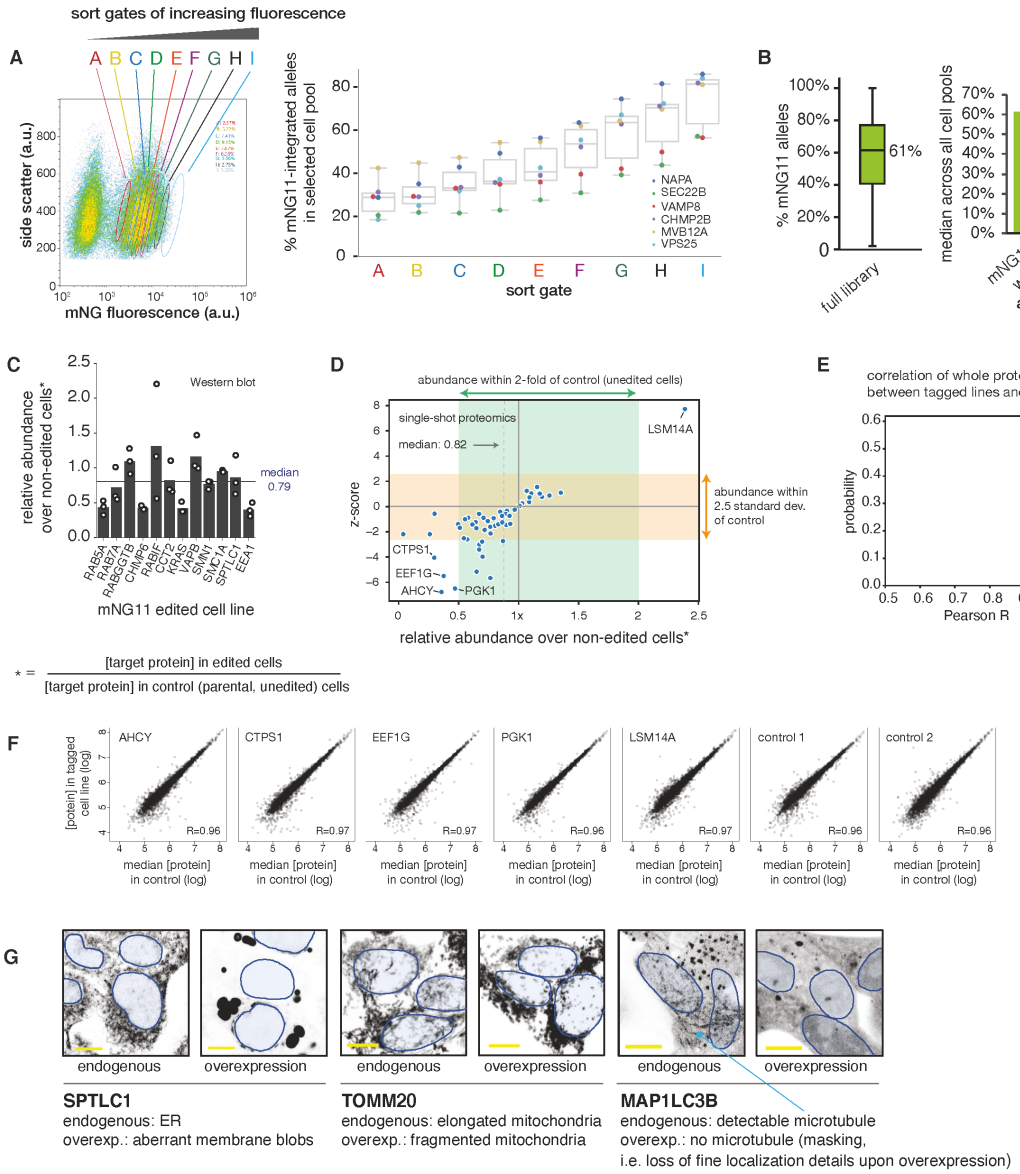
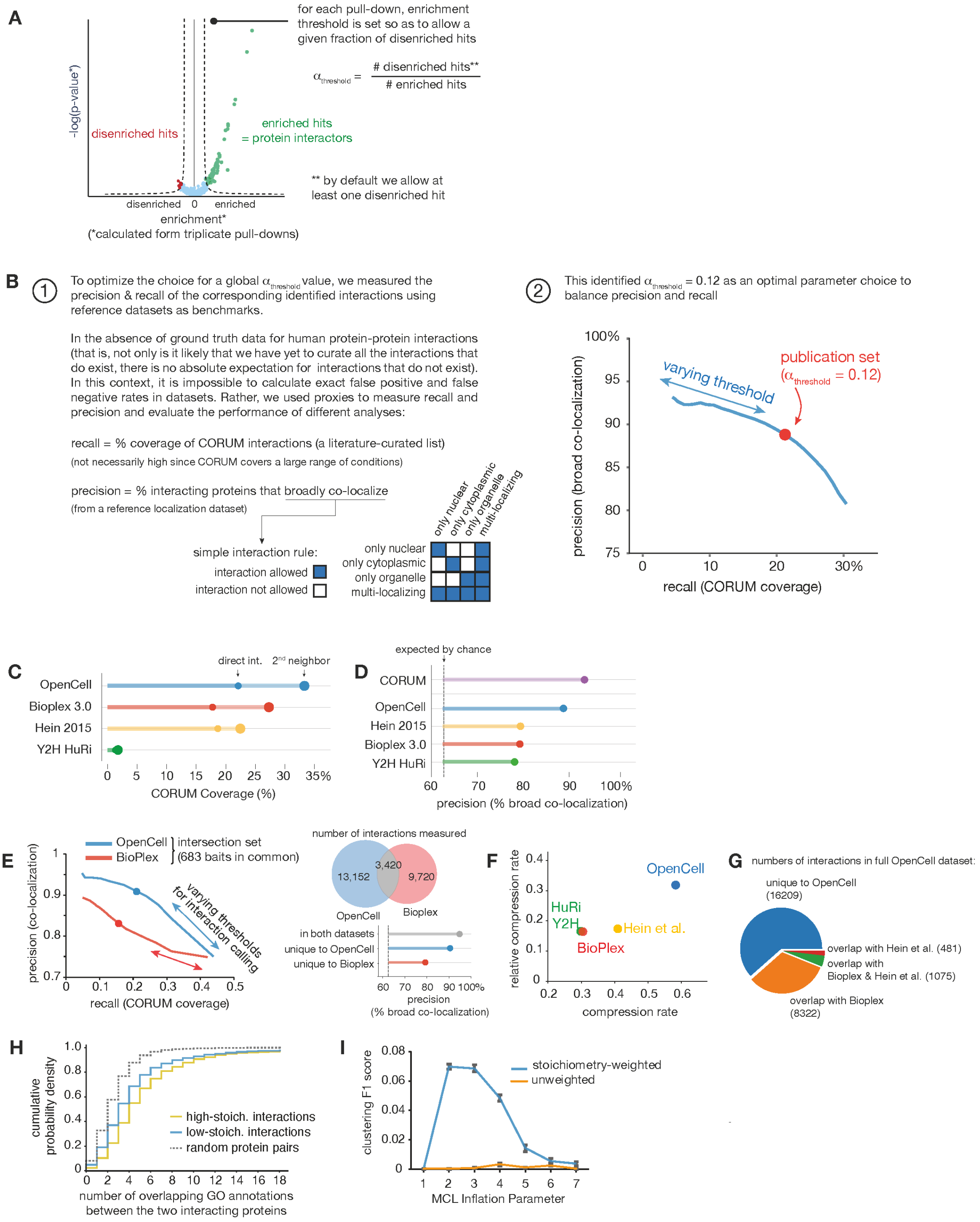

sort gates of increasing fluorescence
A
A B C D E F G H I
side scatter (a.u.)
mNG fluorescence (a.u.)
% mNG11-integrated alleles in selected cell pool
sort gate
NAPA
SEC22B
VAMP8
CHMP2B
MVB12A
VPS25
B
% mNG11 alleles
61%
full library
median across all cell pools
mNG11
wild-type
other
allele type
C
relative abundance over non-edited cells*
Western blot
median 0.79
RAB5A
RAB7A
RABGGTB
CHMP6
RABIF
CCT2
KRAS
VAPB
SMN1
SMC1A
SPTLC1
EEA1
mNG11 edited cell line
D
abundance within 2-fold of control (unedited cells)
single-shot proteomics
median: 0.82
LSM14A
z-score
abundance within 2.5 standard dev. of control
CTPS1
EEF1G
AHCY
PGK1
relative abundance over non-edited cells*
E
correlation of whole proteome content between tagged lines and control:
probability
Pearson R
* = [target protein] in edited cells / [target protein] in control (parental, unedited) cells
F
[potein] in tagged cell line (log)
AHCY
CTPS1
EEF1G
PGK1
LSM14A
control 1
control 2
R=0.96
R=0.97
median [protein] in control (log)
G
endogenous
overexpression
SPTLC1
endogenous: ER
overexp.: aberrant membrane blobs
TOMM20
endogenous: elongated mitochondria
overexp.: fragmented mitochondria
MAP1LC3B
endogenous: detectable microtubule
overexp.: no microtubule (masking,
i.e. loss of fine localization details upon overexpression)

Figure S3

Fig. S3: Cell library characterization and quality control (related to Fig. 1). (A) Optimization of sorting strategy. Polyclonal cell pools were sorted using gates of increasing fluorescence (left panel) and genotyped to quantify the enrichment for mNG11-inserted alleles (right panel, showing data for 6 different target genes). This informed our final sorting strategy in which the top 1% of fluorescent cells (gate I) were selected. **(B)** Genotype analysis of the polyclonal OpenCell library. A single allele is required for fluorescence, but our cell collection is enriched for homozygous insertions. In total, mNG11 insertions account for 61% (median) of alleles in a given cell pool across the full library (Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x interquartile range). The median values of mNG11 integrated alleles, *wt* alleles and other alleles are shown on the right. **(C)** Measurement of target protein abundance in final selected cell pools vs. parental cell line, by quantitative Western blotting. **(D)** Measurement of target protein abundance in final selected cell pools vs. parental cell line, by single-shot mass spectrometry. In these experiments, tagged lines are measured in a single replicate and compared to 6 replicates of non-edited control cell lines. Outliers targets are defined by an abundance that deviates by more than 2.5 standard deviations and by more than 2-fold of their abundance in the controls. The 5 outlier lines are outlined. **(E)** Distribution of Pearson correlation values measuring the overall correlation of abundances for all cellular proteins in each tagged cell line vs. median control. **(F)** For the outliers outlined in (D), correlation of abundances for all cellular proteins in the tagged cell line vs. median control. The abundance correlations for two individual control repeats are shown for reference. **(G)** Examples of overexpression artifacts. Single z-slice confocal images are shown (scale bar: 10 µm). Endogenously tagged lines and their equivalent overexpression constructs were not imaged using the same laser power, so that signal intensities are not directly comparable. Nuclei are shown as blue outlines (nuclei can be located in a different z-plane than the one shown). “Masking effects” are defined as the loss of fine localization details upon overexpression.



1. To optimize the choice for a global $\alpha_{threshold}$ value, we measured the precision & recall of the corresponding identified interactions using reference datasets as benchmarks.

In the absence of ground truth data for human protein-protein interactions (that is, not only is it likely that we have yet to curate all the interactions that do exist, there is no absolute expectation for interactions that do not exist). In this context, it is impossible to calculate exact false positive and false negative rates in datasets. Rather, we used proxies to measure recall and precision and evaluate the performance of different analyses:

recall = % coverage of CORUM interactions (a literature-curated list)

(not necessarily high since CORUM covers a large range of conditions)

precision = % interacting proteins that broadly co-localize

(from a reference localization dataset)

simple interaction rule:

interaction allowed ■

interaction not allowed □

	only nuclear	only cytoplasmic	only organelle	multi-localizing
only nuclear	■	□	□	■
only cytoplasmic	□	■	□	■
only organelle	□	□	■	■
multi-localizing	■	■	■	■

2. This identified $\alpha_{threshold}$ = 0.12 as an optimal parameter choice to balance precision and recall

Figure S4

Fig. S4: Interactome analysis (related to Fig. 2). (A) Strategy for defining enrichment threshold to define interactions. Our strategy builds upon methods described by Hein et al (*7*). Here we use a quantitative approach to define enrichment thresholds dynamically for each replicate set, globally constrained by the parameter $a_{\text{threshold}}$. **(B)** To optimize parameter choice, we measured how precision (% co-localization) and recall (% CORUM coverage) of the corresponding interaction network varied with $a_{\text{threshold}}$. This informed a final value of 0.12. **(C)** Comparing interaction recall (% CORUM coverage) of OpenCell vs. other large-scale interactomes, including direct or 2nd-neighbor interactions (i.e., sharing a direct interactor in common). **(D)** Comparing interaction precision (% co-localization) of OpenCell vs. other large-scale interactomes. CORUM interactions are shown as a reference. **(E)** Direct comparison of OpenCell vs. Bioplex 3.0 on identical bait set. Both datasets use the same HEK293T cell line and share a large number (683) of baits in common. Precision and recall analysis by varying threshold for interaction detection ($a_{\text{threshold}}$ in OpenCell and *pInt* in Bioplex) is shown for the intersection set of 683 baits (dots represent values using thresholds used for final publication sets in both studies). For these set of overlapping baits, OpenCell also includes many new measured interactions for that intersection set of baits (right panel, top). The interactions unique to OpenCell have high precision values (right panel, bottom). **(F)** Compressibility analysis (*32*) of OpenCell vs. other large-scale interactomes. **(G)** Number of interactions measured in OpenCell (in the full dataset) that were also measured in Hein et al. (*7*) or BioPlex 3.0. **(H)** Distribution of GO annotation overlap between protein pairs identified in low-stoichiometry and high-stoichiometry interactions. **(I)** MCL clustering performance (F1 score) using stoichiometry-weighted or unweighted interaction graphs, derived from CORUM interactions as described in Drew et al (*90*).

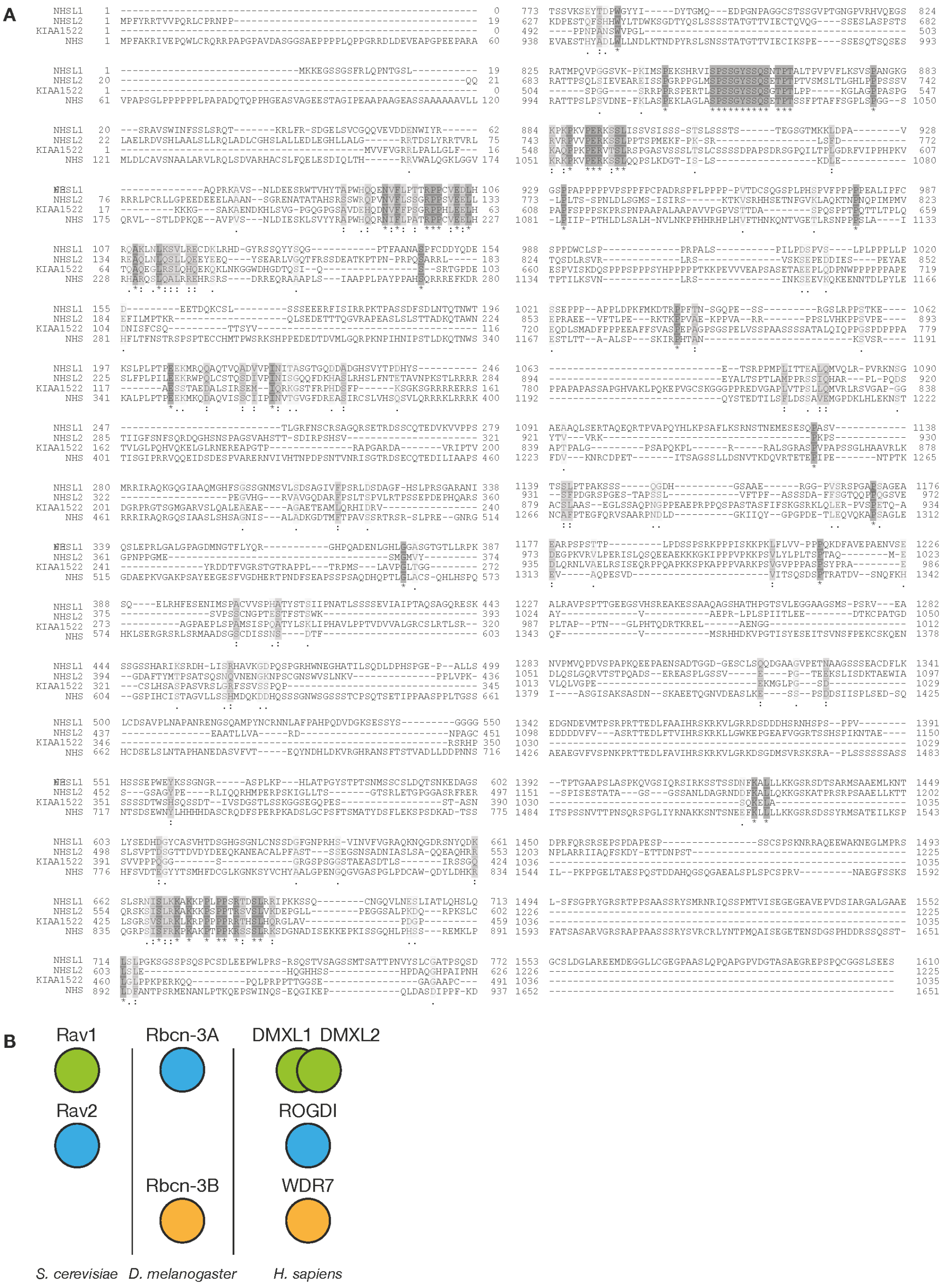

A
NHSL1
NHSL2
KIAA1522
NHS
B
Rav1
Rav2
Rbcn-3A
Rbcn-3B
DMXL1
DMXL2
ROGDI
WDR7
S. cerevisiae
D. melanogaster
H. sapiens

Figure S5

Fig. S5: Sequence analysis of orphan proteins (related to Fig. 2). (A) Amino-acid sequence alignment between human NHSL1, NSHL2, KIAA1522 and NHS. **(B)** Correspondence of RAVE complex members in *S. cerevisiae, D. melanogaster* and *H. sapiens.* Note that in *S. cerevisiae* RAVE also includes Skp1, not depicted here.

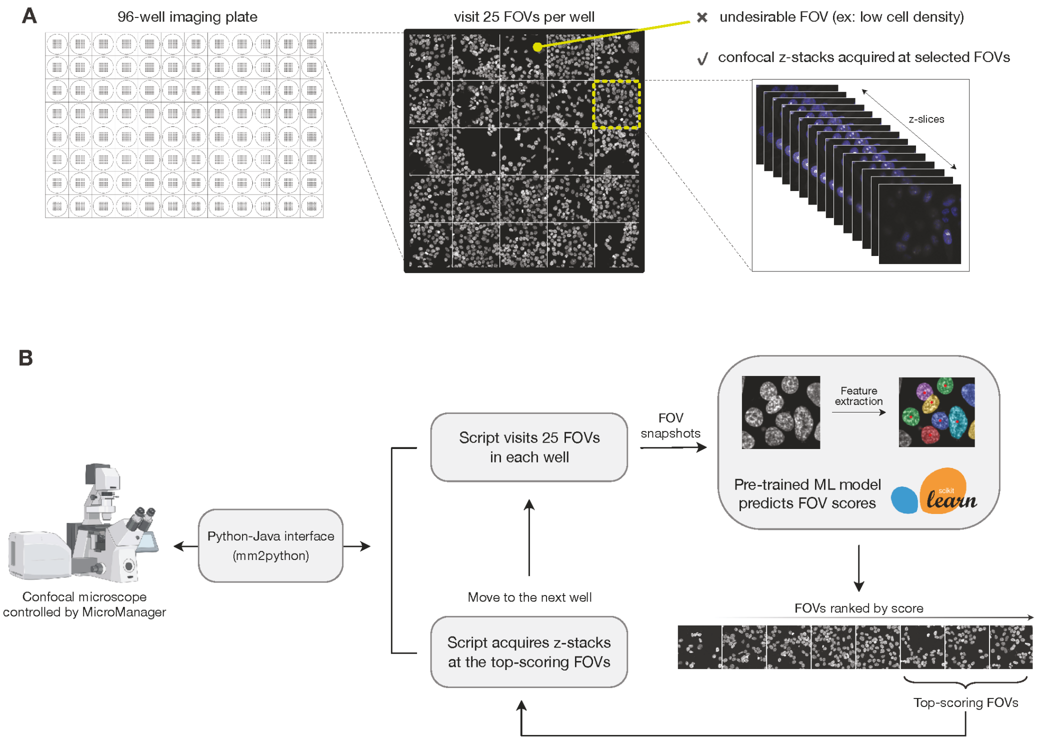

A
96-well imaging plate
visit 25 FOVs per well
× undesirable FOV (ex: low cell density)
√ confocal z-stacks acquired at selected FOVs
z-slices
B
Confocal microscope
controlled by MicroManager
Python-Java interface
(mm2python)
Script visits 25 FOVs
in each well
FOV
snapshots
Feature
extraction
Pre-trained ML model
predicts FOV scores
scikit
learn
Move to the next well
FOVs ranked by score
Script acquires z-stacks
at the top-scoring FOVs
Top-scoring FOVs

Figure S6

Fig. S6: Computer vision for automated microscopy acquisition (related to Fig. 3). (A) To automate microscopy acquisition on 96-well plates and to limit experimental variability between imaging sessions (e.g., to limit variations in cell density) we paired an acquisition script, written in Python, with a pre-trained machine learning model to select field of views (FOVs) on-the-fly during the acquisition. A total of 25 FOVs are sampled per well in a single z-plane, and desirable FOVs are selected for further 3D confocal acquisition on the basis of a score predicted by the pre-trained model. **(B)** Microscopy automation workflow. Microscope hardware is controlled by a Python-based acquisition script via an open-source MicroManager-Python bridge (mm2python; https://github.com/czbiohub/mm2python). This approach enables us to combine custom acquisition logic with the rich ecosystem of Python-based machine-learning packages. Here, we use the scikit-image package to extract features from each FOV snapshot, then use a pre-trained random-forest regression model (scikit-learn) to predict a quality score for the FOV. This process is not computationally expensive and requires less than a second; the FOV score can therefore be used immediately to determine whether the script should acquire a z-stack or else move on to the next position. To maximize the quality of our confocal z-stacks, however, we chose to visit and score all 25 FOVs in each well, then re-visit the top-scoring FOVs for confocal z-stack acquisition.

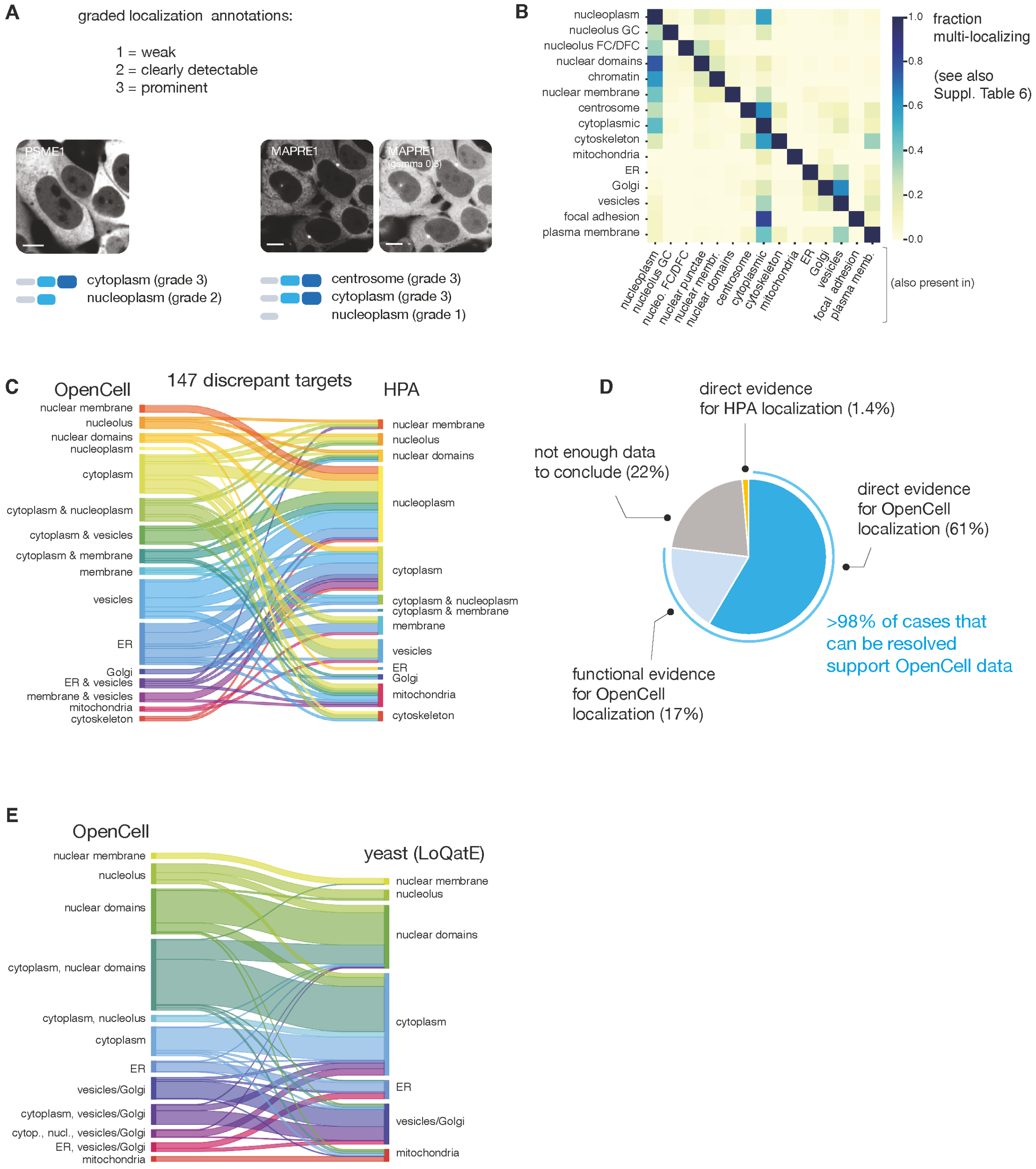

A
graded localization annotations:
1 = weak
2 = clearly detectable
3 = prominent
PSME1
MAPRE1
MAPRE1 (gamma 0.5)
cytoplasm (grade 3)
nucleoplasm (grade 2)
centrosome (grade 3)
cytoplasm (grade 3)
nucleoplasm (grade 1)
B
nucleoplasm
nucleolus GC
nucleolus FC/DFC
nuclear domains
chromatin
nuclear membrane
centrosome
cytoplasmic
cytoskeleton
mitochondria
ER
Golgi
vesicles
focal adhesion
plasma membrane
nucleoplasm
nucleolus GC
nucleo. FC/DFC
nuclear punctae
nuclear membr.
nuclear domains
centrosome
cytoplasmic
cytoskeleton
mitochondria
ER
Golgi
vesicles
focal adhesion
plasma memb.
(also present in)
1.0
0.8
0.6
0.4
0.2
0.0
fraction multi-localizing
(see also Suppl. Table 6)
C
OpenCell
147 discrepant targets
HPA
nuclear membrane
nucleolus
nuclear domains
nucleoplasm
cytoplasm
cytoplasm & nucleoplasm
cytoplasm & vesicles
cytoplasm & membrane
membrane
vesicles
ER
Golgi
ER & vesicles
membrane & vesicles
mitochondria
cytoskeleton
nuclear membrane
nucleolus
nuclear domains
nucleoplasm
cytoplasm
cytoplasm & nucleoplasm
cytoplasm & membrane
membrane
vesicles
ER
Golgi
mitochondria
cytoskeleton
D
direct evidence for HPA localization (1.4%)
not enough data to conclude (22%)
direct evidence for OpenCell localization (61%)
>98% of cases that can be resolved support OpenCell data
functional evidence for OpenCell localization (17%)
E
OpenCell
yeast (LoQatE)
nuclear membrane
nucleolus
nuclear domains
cytoplasm, nuclear domains
cytoplasm, nucleolus
cytoplasm
ER
vesicles/Golgi
cytoplasm, vesicles/Golgi
cytop., nucl., vesicles/Golgi
ER, vesicles/Golgi
mitochondria
nuclear membrane
nucleolus
nuclear domains
cytoplasm
ER
vesicles/Golgi
mitochondria

Figure S7

Fig. S7: The OpenCell image dataset (related to Fig. 3). (A) Principle of graded localization annotation (manual annotations). **(B)** Fraction of multi-localization between cellular compartments. Complete localization annotations can be found in Table S6. **(C)** Comparison of annotated localization for proteins in OpenCell and Human Protein Atlas (HPA, version v20) datasets for which annotations are inconsistent. **(D)** Extensive literature curation allows to resolve 77% of OpenCell/HPA discrepancies (full details in Table S8). Here "direct evidence" refers to proteins for which localization has been directly measured in published studies, while "functional evidence" refers to proteins for which localization might not have been directly measured, but for which literature establishes a function that is predictive of a specific localization. For example, SCFD1 is a protein whose main known function is to regulate transport between ER and Golgi. This qualifies as "functional evidence". It is annotated as localized in the ER and Golgi in OpenCell, and in the nucleoplasm (main) and cytosol (additional) in HPA. **(E)** Comparison of annotated localization for 350 orthologous proteins in OpenCell and *S. cerevisiae* yeast (from LoQaTe (*47*)). Note that in yeast Golgi and vesicles are difficult to distinguish.

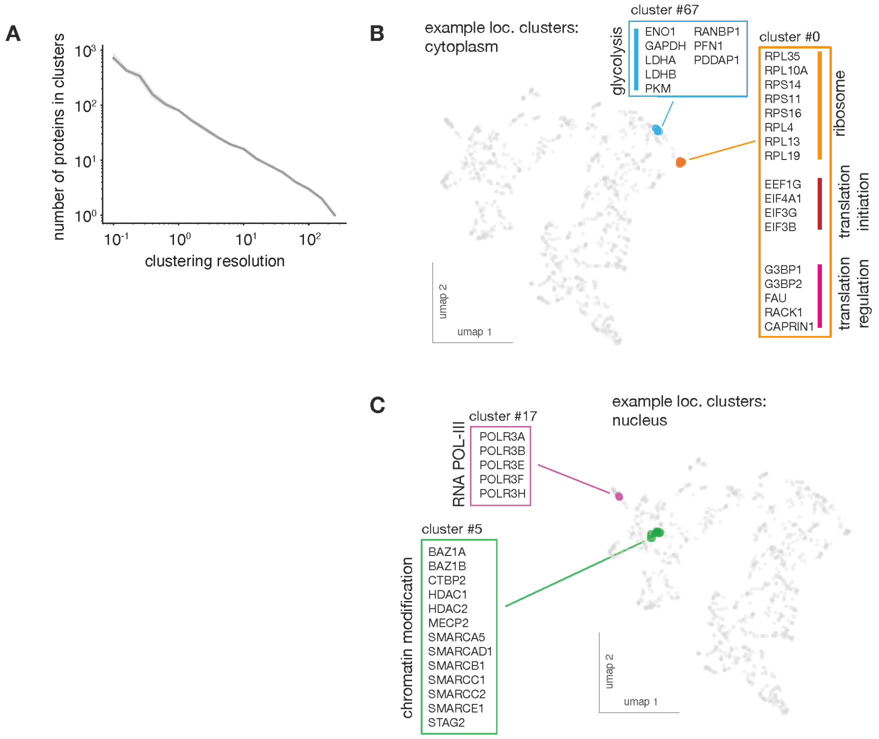

A
number of proteins in clusters
10^3
10^2
10^1
10^0
10^{-1}
10^0
10^1
10^2
clustering resolution
B
example loc. clusters:
cytoplasm
cluster #67
glycolysis
ENO1
GAPDH
LDHA
LDHB
PKM
RANBP1
PFN1
PDDAP1
cluster #0
RPL35
RPL10A
RPS14
RPS11
RPS16
RPL4
RPL13
RPL19
ribosome
EEF1G
EIF4A1
EIF3G
EIF3B
translation initiation
G3BP1
G3BP2
FAU
RACK1
CAPRIN1
translation regulation
umap 2
umap 1
C
cluster #17
RNA POL-III
POLR3A
POLR3B
POLR3E
POLR3F
POLR3H
example loc. clusters:
nucleus
cluster #5
chromatin modification
BAZ1A
BAZ1B
CTBP2
HDAC1
HDAC2
MECP2
SMARCA5
SMARCAD1
SMARCB1
SMARCC1
SMARCC2
SMARCE1
STAG2
umap 2
umap 1

Figure S8

Fig. S8: high-resolution image clusters (related to Fig. 4C). (A) Size of clusters **C** (number of proteins in each cluster) as a function of clustering resolution. Shaded regions show standard deviations calculated from 9 separate repeat rounds of clustering, and average values are shown as a solid line. **(B), (C)** Examples of clusters of cytoplasmic **(B)** and nuclear **(C)** proteins.

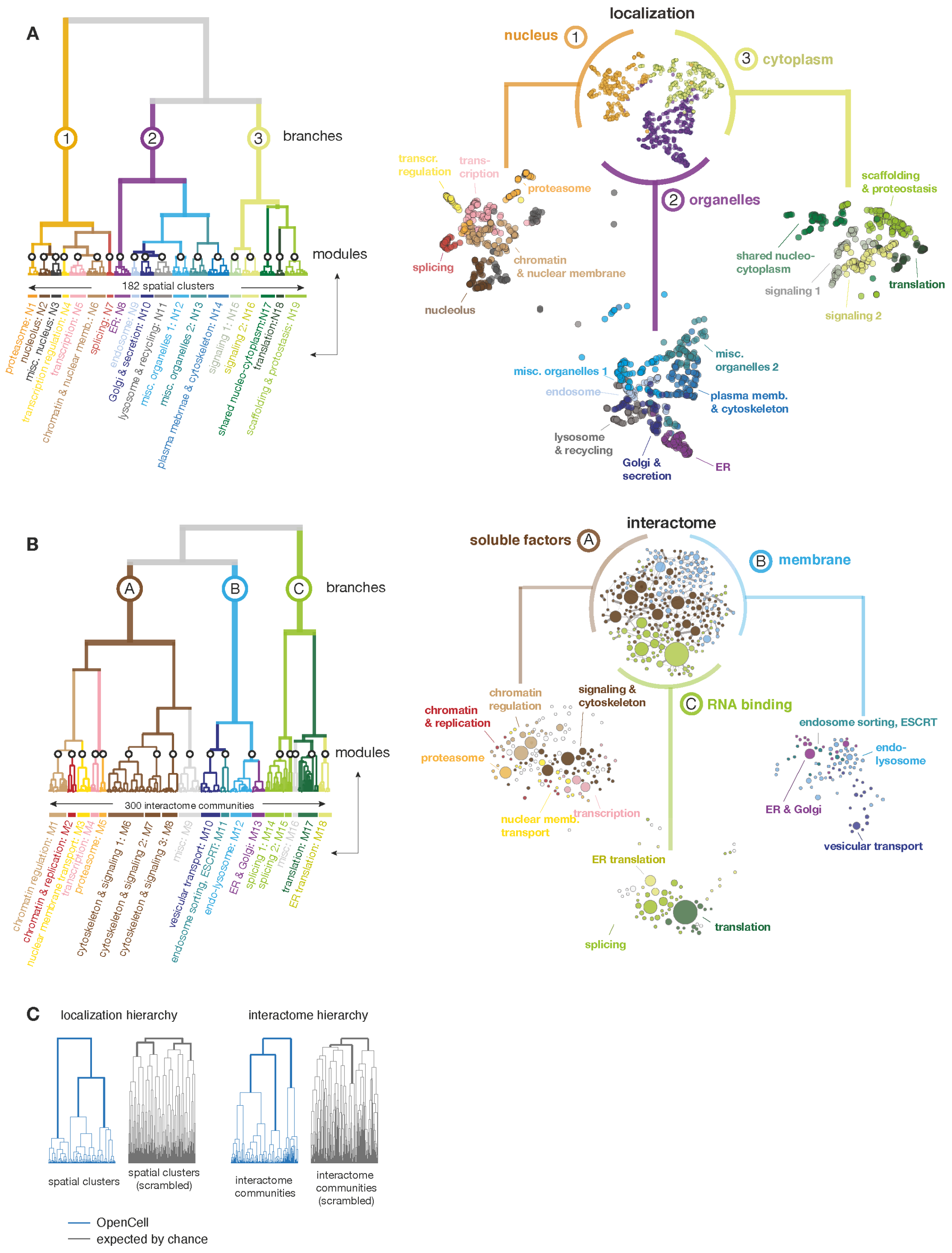

A
branches
1
2
3
modules
182 spatial clusters
proteasome: N1
nucleolus: N2
misc. nucleus: N3
transcription regulation: N4
transcription: N5
chromatin & nuclear memb.: N6
splicing: N7
ER: N8
endosome: N9
Golgi & secretion: N10
lysosome & recycling: N11
misc. organelles 1: N12
misc. organelles 2: N13
plasma mebrnae & cytoskeleton: N14
signaling 1: N15
signaling 2: N16
shared nucleo-cytoplasm:N17
translation:N18
scaffolding & protostasis: N19
localization
nucleus 1
3 cytoplasm
2 organelles
transcr. regulation
trans- cription
proteasome
splicing
chromatin & nuclear membrane
nucleolus
scaffolding & proteostasis
shared nucleo- cytoplasm
signaling 1
translation
signaling 2
misc. organelles 1
misc. organelles 2
endosome
plasma memb. & cytoskeleton
lysosome & recycling
Golgi & secretion
ER
B
branches
A
B
C
modules
300 interactome communities
chromatin regulation: M1
chromatin & replication: M2
nuclear membrane transport: M3
transcription: M4
proteasome: M5
cytoskeleton & signaling 1: M6
cytoskeleton & signaling 2: M7
cytoskeleton & signaling 3: M8
misc: M9
vesicular transport: M10
endosome sorting, ESCRT: M11
endo-lysosome: M12
ER & Golgi: M13
splicing 1: M14
splicing 2: M15
misc: M16
translation: M17
ER translation: M18
interactome
soluble factors A
B membrane
C RNA binding
chromatin regulation
signaling & cytoskeleton
chromatin & replication
proteasome
nuclear memb. transport
transcription
endosome sorting, ESCRT
endo- lysosome
ER & Golgi
vesicular transport
ER translation
translation
splicing
C
localization hierarchy
interactome hierarchy
spatial clusters
spatial clusters (scrambled)
interactome communities
interactome communities (scrambled)
OpenCell
expected by chance

Figure S9

Fig. S9: Full hierarchical structure of interactome and localization datasets (related to Fig. 5). Dendrograms represent the hierarchical relationships connecting **(A)** the full set of protein communities identified in the interactome (see Fig. 2) or **(B)** the full set of high-resolution clusters identified in the image collection (see Fig. 4C). For each dataset, an intermediate layer of hierarchy separates 18-19 modules, while an upper hierarchical layer delineates three separate branches. Modules and branches are annotated on the basis of gene ontology enrichment analysis (see Suppl. Tables 5 & 9). Right-hand panels present the topological arrangement of branches (top) and modules (bottoms) in each dataset, highlighted from the full graph of connections between interaction communities ("interactome", see Fig. 2D) or from the localization UMAP ("localization", see Fig. 4C). The color codes between interactome and localization datasets are not directly comparable (i.e. same colors are not meant to represent the same exact set of proteins). **(C)** The hierarchical structures derived from interactome (left) and localization (right) datasets are compared to the hierarchical structures derived from "scrambled" controls – that is, to the hierarchical structure that is expected by chance given the proteins present in our dataset. Controls are generated by randomly shuffling the membership of each protein between spatial clusters or interaction communities. The number of proteins in each cluster or community was preserved from the original data.

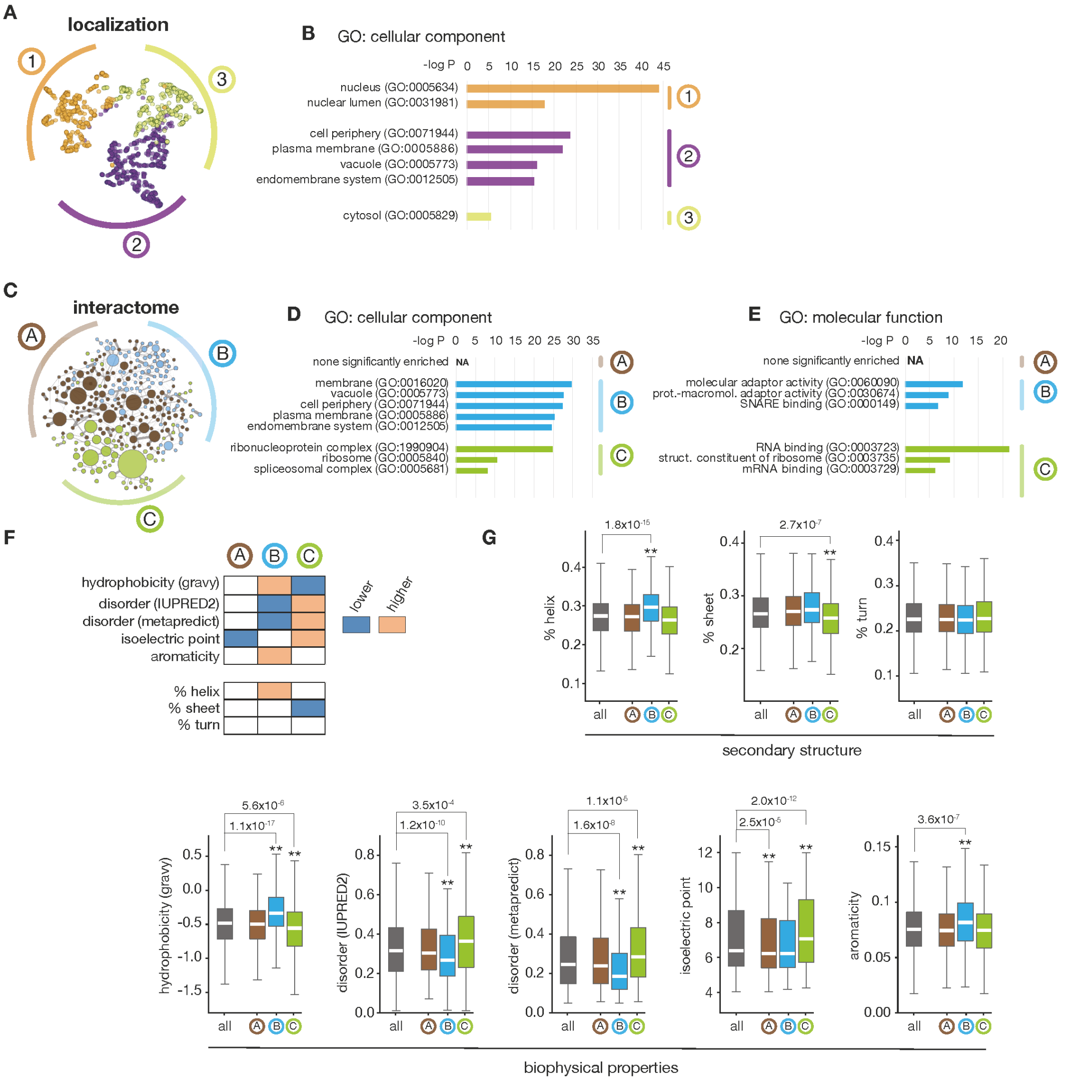

A
localization
1
2
3
B
GO: cellular component
-log P 0 5 10 15 20 25 30 35 40 45
nucleus (GO:0005634)
nuclear lumen (GO:0031981)
cell periphery (GO:0071944)
plasma membrane (GO:0005886)
vacuole (GO:0005773)
endomembrane system (GO:0012505)
cytosol (GO:0005829)
C
interactome
A
B
C
D
GO: cellular component
-log P 0 5 10 15 20 25 30 35
none significantly enriched NA
membrane (GO:0016020)
vacuole (GO:0005773)
cell periphery (GO:0071944)
plasma membrane (GO:0005886)
endomembrane system (GO:0012505)
ribonucleoprotein complex (GO:1990904)
ribosome (GO:0005840)
spliceosomal complex (GO:0005681)
E
GO: molecular function
-log P 0 5 10 15 20
none significantly enriched NA
molecular adaptor activity (GO:0060090)
prot.-macromol. adaptor activity (GO:0030674)
SNARE binding (GO:0000149)
RNA binding (GO:0003723)
struct. constituent of ribosome (GO:0003735)
mRNA binding (GO:0003729)
F
hydrophobicity (gravy)
disorder (IUPRED2)
disorder (metapredict)
isoelectric point
aromaticity
% helix
% sheet
% turn
lower
higher
G
1.8x10-15
2.7x10-7
**
% helix
% sheet
% turn
all
secondary structure
5.6x10-6
1.1x10-17
3.5x10-4
1.2x10-10
1.1x10-5
1.6x10-8
2.0x10-12
2.5x10-6
3.6x10-7
hydrophobicity (gravy)
disorder (IUPRED2)
disorder (metapredict)
isoelectric point
aromaticity
biophysical properties

Figure S10

Fig. S10: Biophysical & ontology analysis of the main branches from interactome and localization hierarchies (related to Fig.s 5 and S9). (A) The three branches derived from the image-based hierarchy (see Fig. S9A). **(B)** Enrichment analysis of GO annotations in the hierarchical branches, testing GO term enrichment of proteins in each branch against all proteins in the interactome (Fisher's exact test, showing annotations enriched at $p < 10^{-10}$ and excluding near-synonymous annotations). **(C)** The three branches derived from the interactome hierarchy (see Fig. S9B). **(D), (E)** Enrichment analysis of GO annotations in the hierarchical branches, testing GO term enrichment of proteins in each branch against all proteins in the interactome (Fisher's exact test, showing annotations enriched at $p < 10^{-10}$ and excluding near-synonymous annotations). **(F)** Heat-map representing significance testing of biophysical properties of protein sequences in the 3 branches. P-values were obtained using Student's t-test comparing proteins belonging to a specific hierarchical branch against all proteins in the three branches. **(G)** Box plots representing the significance testing of biophysical properties described in (F). Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x inter-quartile ranges. Median is represented by a white line. ** $p < 10^{-3}$ (Student's t-test), exact p-values are shown.

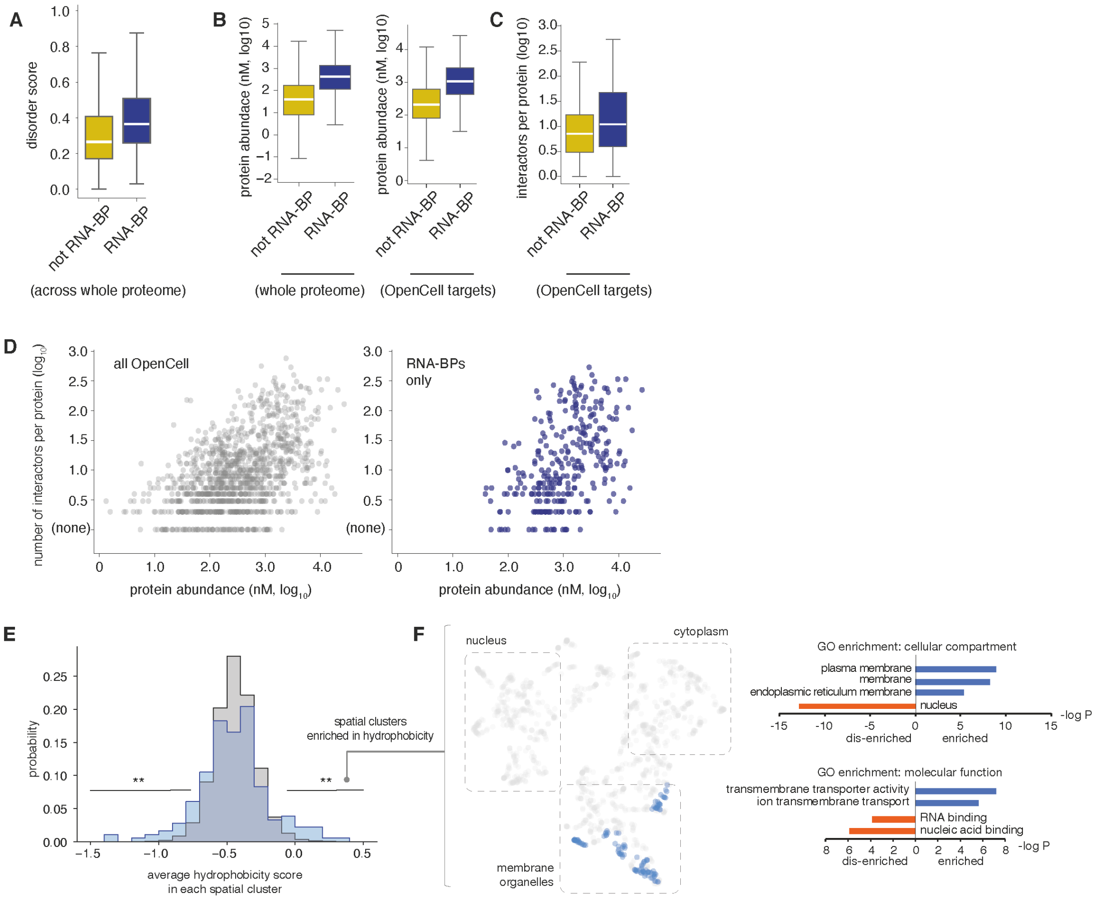
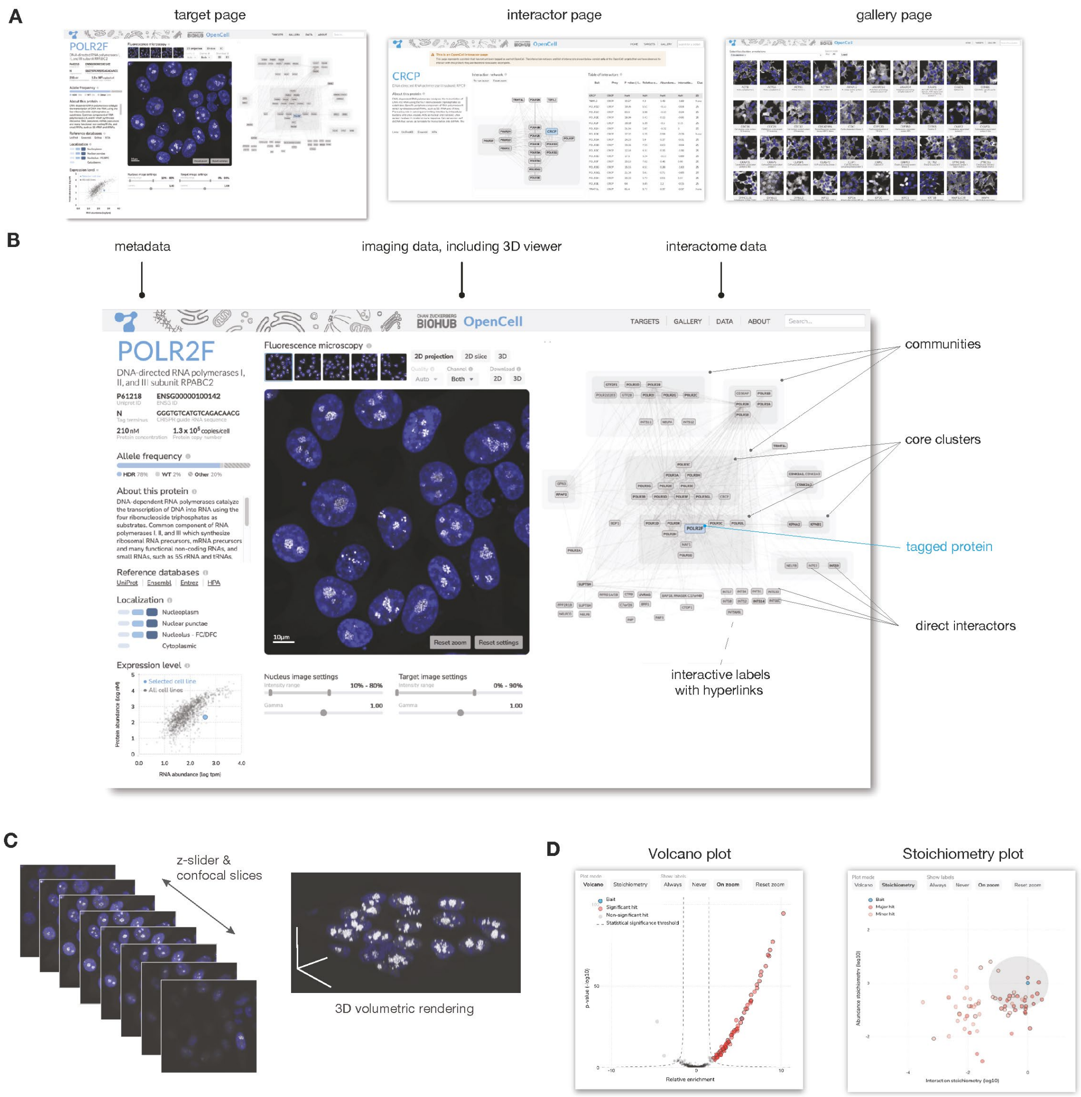

A
disorder score
1.0
0.8
0.6
0.4
0.2
0.0
not RNA-BP
RNA-BP
(across whole proteome)
B
protein abundace (nM, log10)
5
4
3
2
1
0
−1
−2
not RNA-BP
RNA-BP
(whole proteome)
protein abundace (nM, log10)
4
3
2
1
0
not RNA-BP
RNA-BP
(OpenCell targets)
C
interactors per protein (log10)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
not RNA-BP
RNA-BP
(OpenCell targets)
D
number of interactors per protein (log10)
all OpenCell
3.0
2.5
2.0
1.5
1.0
0.5
(none)
0
1.0
2.0
3.0
4.0
protein abundance (nM, log10)
RNA-BPs only
3.0
2.5
2.0
1.5
1.0
0.5
(none)
0
1.0
2.0
3.0
4.0
protein abundance (nM, log10)
E
probability
0.25
0.20
0.15
0.10
0.05
0.00
**
**
spatial clusters enriched in hydrophobicity
−1.5
−1.0
−0.5
0.0
0.5
average hydrophobicity score in each spatial cluster
F
nucleus
cytoplasm
membrane organelles
GO enrichment: cellular compartment
plasma membrane
membrane
endoplasmic reticulum membrane
nucleus
-15
-10
-5
0
5
10
15
-log P
dis-enriched
enriched
GO enrichment: molecular function
transmembrane transporter activity
ion transmembrane transport
RNA binding
nucleic acid binding
8
6
4
2
0
2
4
6
8
-log P
dis-enriched
enriched

Figure S11

Fig. S11: Unique properties of RNA-binding proteins (RNA-BPs, related to Fig. 5). (A) Distribution of disorder score (IUPRED2) for RNA-BPs vs. non-RNA-BPs across the whole proteome. **(B)** Distribution of protein abundance for RNA-BPs vs non-RNA-BPs across the whole proteome (left) and across OpenCell targets only (right). **(C)** Distribution of number of interactors for RNA-BPs vs non-RNA-BPs across OpenCell targets. **(D)** For each OpenCell target, the number of interactors is plotted as a function of protein abundance. The subset of targets that are RNA-BPs is highlighted on the right-hand panel. Note: for boxplots in (A), (B), (C) and (D), boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x interquartile range. Median is represented by a white line. **(E)** Distribution of hydrophobicity score (gravy) across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1,000 times. Lines indicate parts of the distribution over-represented in our data vs control (**: $p < 2x10^{-3}$, Fisher's exact t-test). **(F)** Distribution of high-hydrophobicity spatial clusters (average hydrophobicity score > -0.1) in the UMAP embedding from Fig. 3D (left), and ontology enrichment analysis of proteins contained in these clusters (right). Enrichment compares to the whole set of OpenCell targets (p-value: Fisher's exact test).



POLR2F

DNA-directed RNA polymerases I, II, and III subunit RPABC2

P61218 Uniprot ID	ENSG0000100142 ENSG ID
N Tag terminus	GGGTGTCATGTCAGACAACG CRISPR guide RNA sequence
210 nM Protein concentration	1.3 x 10^5 copies/cell Protein copy number

Allele frequency

HDR 78% WT 2% Other 20%

About this protein

DNA-dependent RNA polymerases catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. Common component of RNA polymerases I, II, and III which synthesize ribosomal RNA precursors, mRNA precursors and many functional non-coding RNAs, and small RNAs, such as 5S rRNA and tRNAs.

Reference databases

UniProt | Ensembl | Entrez | HPA

Localization

Nucleoplasm
Nuclear punctae
Nucleolus - FC/DFC
Cytoplasmic

Expression level

● Selected cell line
● All cell lines

Protein abundance (log nM)

RNA abundance (log tpm)

Figure S12

Fig. S12: Interactive data exploration at opencell.czbiohub.org. (A) The three principal pages of the OpenCell web app. From left to right: the target page, interactor page, and gallery page. **(B)** The target page consists of three columns. The leftmost column contains the functional annotation for the target from UniProt, links to other databases, our manually-assigned localization annotations, and measures of protein expression. The middle column contains the image viewer, and the rightmost column the interaction network. **(C)** The image viewer allows the user to scroll through the confocal z-slices using a slider or to visualize the z-stack in 3D as a volume rendering. In either mode, the user can pan and zoom by clicking, dragging, and scrolling. **(D)** The interaction network can be toggled with two alternative, complementary visualizations of the target's protein interactions: a volcano plot of relative enrichment vs. p-value and a scatterplot of interaction stoichiometry vs. cellular abundance stoichiometry. In both the network view and the scatterplots, the user can click on an interactor to open the target or the interactor page for the corresponding protein.

List of Supplementary Tables

Note: each Supplementary Table contains a specific "read_me" tab that describes its content in detail.

Table S1.
The OpenCell library (includes target information, library design and genotype data). Related to Fig. 1.

Table S2.

Annotated HEK293T proteome (includes RNA and protein abundance data, biophysical properties and ontologies relevant to the analyses presented in this paper). Related to Fig. 1.

Table S3.
Properties of successful vs. unsuccessful edited targets. Related to Fig. 1 & S2.

Table S4.
The OpenCell interactome (quantitative description of interactions). Related to Fig. 2.

Table S5.
Clustering analysis of the interactome (analysis of MCL clustering and subsequent hierarchical analyses). Related to Figs. 2 & S9.

Table S6.
The OpenCell localization dataset and annotations. Related to Fig. 3.

Table S7
Comparison of OpenCell to Human Protein Atlas (Table S7A) or yeast (Table S7B) localization annotations. Related to Fig. 3B & S7.

Table S8.
Resolving discrepancies between OpenCell and Human Protein Atlas annotations by literature curation. Related to Fig. S7C.

Table S9.
Clustering analysis of the imaging dataset (analysis of Leiden clustering and subsequent hierarchical analyses from high-resolution clusters). Related to Figs. 4 & S9.

2.3 Article 3: DIA-based systems biology approach unveils E3 ubiquitin ligase-dependent responses to a metabolic shift

Karayel, O., **Michaelis, A. C.**, Mann, M., Schulman, B. A. & Langlois, C. R. **DIA-based systems biology approach unveils E3 ubiquitin ligase-dependent responses to a metabolic shift.** Proc National Acad Sci 117, 32806–32815 (2020). doi: *10.1073/pnas.202019711*

This study connects to the previous ones via its systems biology approach to use optimized 96-well plate compatible sample preparation and shortest possible mass spectrometry measurement times to screen for hundreds of near-complete proteomes under perturbation conditions in *S. cerevisiae*. In order to allow proteome measurements in about 20 min we used a data-independent acquisition approach and we were able to reduce gap times drastically by optimizing HPLC settings. This allowed us to measure several hundreds of proteomes in a few days.

In the screen we applied several perturbation conditions that included heat shock, osmotic stress, growth on ethanol, and starvation conditions. While those distinct responses provide a comprehensive resource, they also unveiled a carbon source dependent GID E3 ligase dependent regulation, which is an important cellular regulator for metabolic switches.

This study shows that global approaches are necessary to observe and understand the complex dependencies that shape the cell and its responses to environmental changes. Systems biology approaches like this enable the fast screening of many samples to discover the most prominent cause of a response efficiently and unbiasedly. Here, in comparison to the previous described project, we offer a solution to use a nano-flow HPLC instead of a high-flow system for short gradients with reduced gap times on an Orbitrap platform.

This study was a great collaboration between the Mann lab, represented by Ozge Karayel and myself and the Schulman lab, represented by Christine Langlois and was published in *PNAS*.

DIA-based systems biology approach unveils E3 ubiquitin ligase-dependent responses to a metabolic shift

Ozge Karayel[a,1], André C. Michaelis[a], Matthias Mann[a,2], Brenda A. Schulman[b,2], and Christine R. Langlois[b,1,2]

[a]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany; and [b]Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

Contributed by Brenda A. Schulman, November 8, 2020 (sent for review September 28, 2020; reviewed by Angus I. Lamond and Alexander Varshavsky)

The yeast *Saccharomyces cerevisiae* is a powerful model system for systems-wide biology screens and large-scale proteomics methods. Nearly complete proteomics coverage has been achieved owing to advances in mass spectrometry. However, it remains challenging to scale this 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 workflow 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 and comparisons of hundreds of nearly complete yeast proteomes in only a few days. We evaluate its capability by characterizing changes in the yeast proteome in response to environmental perturbations, identifying distinct responses to each of them and providing a comprehensive resource of these responses. Apart from rapidly recapitulating previously observed responses, we characterized carbon source-dependent regulation of the GID E3 ligase, an important regulator of cellular metabolism during the switch between gluconeogenic and glycolytic growth conditions. This unveiled regulatory targets of the GID ligase during a metabolic switch. Our comprehensive yeast system readout pinpointed effects of a single deletion or point mutation in the GID complex on the global proteome, allowing the identification and validation of targets of the GID E3 ligase. Moreover, this approach allowed the identification of targets from multiple cellular pathways that display distinct patterns of regulation. Although developed in yeast, rapid whole-proteome–based readouts can serve as comprehensive systems-level assays in all cellular systems.

yeast systems biology | mass spectrometry | proteomics | stress | GID E3 ligase

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 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 (2–4). Despite the importance of cellular stress responses, our understanding of how cellular pathways interact during adaptation remains incomplete. Therefore, knowing 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 resources in yeast, including deletion libraries (5, 6), green fluorescent protein–tagged libraries (7, 8), overexpression libraries (9), and the recently developed SWAp-tag library (10–12), has made yeast a premier model system for conducting transcriptomics, proteomics, interactomics, or metabolomics screens (13–19). Indeed, systems-wide biology screens and large-scale proteomics were both pioneered in the yeast model. Furthermore, the cellular interaction networks and molecular mechanisms ascertained in yeast can be readily applied to other systems (20–22).

Early genome-wide studies showed that over 4,000 proteins are expressed during log-phase growth in yeast and this organism was the first whose entire proteome was mapped by mass spectrometry (MS)-based proteomics (23). Subsequently, yeast has served as a model of choice for the development of ever-more-sensitive and faster proteomics workflows (23–34). Remarkably, the optimized sample preparation coupled with MS analysis performed on the Orbitrap hybrid mass spectrometer allowed identification of around 4,000 yeast proteins over a 70-min liquid chromatography (LC)-MS/MS run (24, 30). However, the necessity of technological expertise and lengthy analysis times for 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) (35, 36), we hypothesized that it would now be possible to obtain accurate and high yeast proteome coverage by a straightforward and rapid single-run approach, enabling researchers to

Significance

We use a single-run, data-independent acquisition–based mass spectrometry approach to generate and compare dozens of yeast proteomes in less than a day, and provide a comprehensive resource detailing changes to the yeast proteome following commonly used stress treatments in yeast. Our systems biology approach identifies and validates regulatory targets of an E3 ubiquitin ligase during a metabolic switch, providing insights into the interplay of metabolic pathways. The speed, simplicity, and scalability of this workflow makes it particularly well-suited for screens in any cellular system to investigate specific effects of deletions or mutants or other perturbations to obtain the response of biological system on a global level.

Author contributions: O.K., M.M., B.A.S., and C.R.L. designed research; O.K. and C.R.L. performed research; O.K., A.C.M., and C.R.L. contributed new reagents/analytic tools; O.K. and C.R.L. analyzed data; and O.K., M.M., B.A.S., and C.R.L. wrote the paper.

Reviewers: A.I.L., University of Dundee; and A.V., California Institute of Technology.

The authors declare no competing interest.

See online for related content such as Commentaries.

[1]O.K. and C.R.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mmann@biochem.mpg.de, Schulman@biochem.mpg.de, or langlois@biochem.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2020197117/-/DCSupplemental.

First published December 7, 2020.

easily study biological processes on a global scale. Such a system could then serve as a template for more complex proteomes, including the human proteome.

One mechanism of 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 (37). 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 sparing the yeast from energetically costly metabolic pathways that are unnecessary in fermentable carbon sources (38–40). In addition, ubiquitin ligases also serve as crucial regulators in response to oxidative, heavy metal, and protein folding stresses (41–43). Despite the importance of ubiquitination during cellular adaptation, our knowledge of the E3-dependent responses to cellular perturbation remains incomplete.

Here, we describe a systems biology approach employing rapid, single-run, data-independent (DIA) mass spectrometric analysis, which we use 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. Our DIA-based approach is sufficiently sensitive and robust to detect quantitative proteome remodeling in response to all these stresses. We then apply this methodology to probe a specific biological question to identify novel regulation by the GID E3 ligase during a metabolic switch. We use a combination of a core subunit deletion and a structure-based catalytic mutant to identify all of the known substrates of the GID E3 ligase and discover two previously unknown targets which display distinct patterns of regulation.

Results

Streamlined and Scalable Yeast Proteome Analysis Employing DIA. In order to establish a fast and scalable single-run analysis approach for yeast proteome profiling, we explored a DIA strategy on an Orbitrap mass spectrometer. Unlike data-dependent acquisition (DDA), a DIA method isolates coeluting peptide ions together in predefined mass windows, fragmenting and analyzing all ions simultaneously (36). This strategy overcomes the limited sequencing speed of sequential DDA, enabling fast and scalable single-shot analysis workflows. On Orbitrap-based mass analyzers, it yields substantially higher number of identified proteins with unprecedented quantitative accuracy (44). To generate a yeast-specific and comprehensive spectral library that is generally used for this approach, we cultured yeast under various growth and stress conditions. After extraction and digestion of proteins, we separated peptides obtained from each condition by basic reversed-phase (RP) chromatography into eight fractions. The resulting 64 fractions (8 fractions × 8 conditions) were measured using a DDA method with a 23-min LC gradient and analyzed with the Spectronaut software (Fig. 1*A*). Together with LC overhead time this took about half an hour, allowing for 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 the expressed yeast proteome according to a previous report that computationally aggregated 21 different large-scale datasets (45). The median sequence coverage was 27% and on average 12 peptides were detected per protein.

Combined with our own comprehensive spectral library, the 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; Fig. 1 *B* and *C*). This implies that ∼73% of proteins in the deep yeast spectral library were matched into the single runs. Note that the single runs represent only yeast grown in rich media (yeast extract peptone dextrose [YPD]), whereas the library combines the proteomes of yeast grown under several growth conditions and therefore contains proteins which are not expressed during growth in YPD. Therefore, the degree of proteome completeness is likely much higher than 73%. Measurements were highly reproducible with Pearson coefficients greater than 0.92 between replicates (*SI Appendix*, Fig. S1*A*) and coefficients of variation <20% for 68% of all common proteins between the six replicates. In comparison, a single-run, data-dependent acquisition strategy with the same LC gradient quantified only 11,883 peptides and 2,289 distinct proteins on average (Fig. 1 *B* and *C*). To more directly compare the performance of the 23-min DIA method to the DDA method we analyzed the same sample with increasing gradient lengths. We could only reach the same depth using the DDA method with at least 180-min-long LC gradients (33,425 peptides and 3,435 proteins) (Fig. 1 *B* and *C*). Thus, the DIA method allows us to obtain coverage comparable to DDA in a high-throughput and in-depth fashion while taking considerably less MS time.

Large-Scale and Quantitative Analysis of Yeast Stress Response in Half a Day. Using this DIA-based systems biology approach, we next comprehensively and quantitatively analyzed 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 the rapid DIA method. We quantified 3,506 distinct proteins in total (Fig. 1*D* and Dataset S1). Reproducibility was high, with Pearson correlations >0.93 between the three biological replicates (*SI Appendix*, Fig. S1*B*). Strikingly, over 90% of all detected proteins were consistently quantified at varying levels across all conditions (Dataset S1). Principal component analysis (PCA) demonstrated that the first component accounted for 13% of the variability and segregated with the different conditions and growth media as the major effectors (Fig. 1*E*).

We first looked more closely at the differences in protein expression during growth in YPD (rich media) and SC (synthetic complete media), the two most common growth media used in yeast research. YPD and SC media differ in their nutrient composition as well as their pH. During growth in YPD, the three most significantly up-regulated proteins (Sit1, Ctr1, and Enb1) are regulators of copper and iron transport (Fig. 1 *F*, *Right* and *G*), consistent with the fact that copper and iron are limiting factors for the growth of yeast at more alkaline pH (46). Conversely, during growth in SC, many mitochondrial proteins were up-regulated compared to YPD (Fig. 1 *F*, *Left* and *G*), including the cytochrome *c* oxidase subunits Cox8, Cox2, and Cox5a, the mitochondrial adenosine 5′-triphosphate (ATP) synthase Atp20, and the mitochondrial aminopeptidase Icp55. Yeast mitochondria reproduce through fission and must be inherited by daughter cells during cell division (47). The up-regulation of many mitochondrial proteins is thus consistent with the faster growth rate of our yeast strains in SC compared to YPD. Because the choice of media is often considered crucial in experimental design, these data on differentially regulated proteins in pathways of interest provide an important resource for yeast biologists.

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. Each produced a discrete stress response, resulting in synthesis or

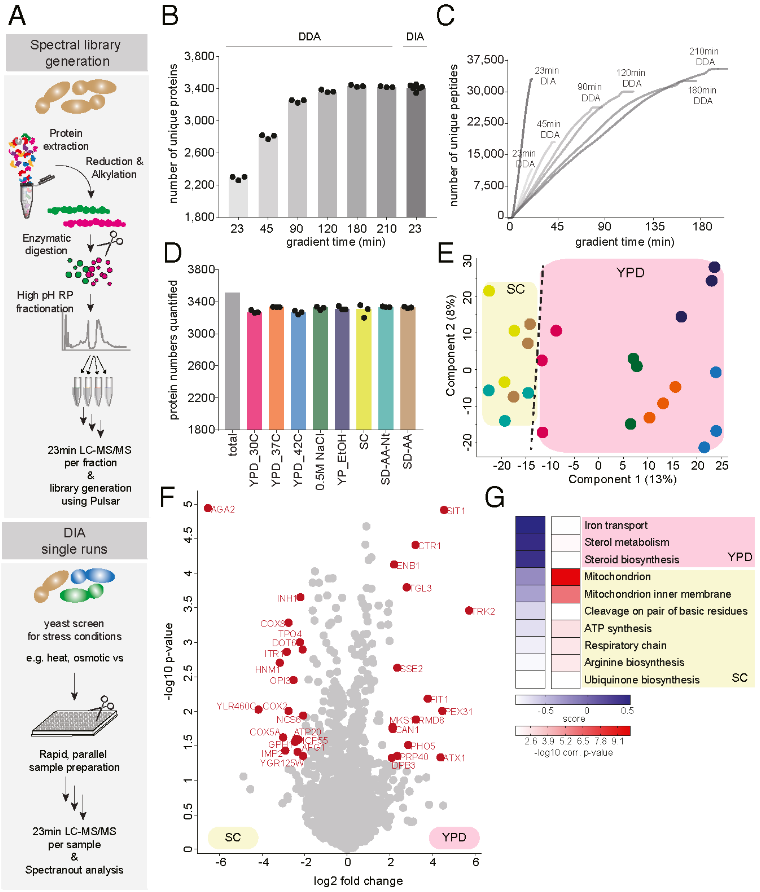


Fig. 1. Fast and scalable yeast proteome analysis using DIA. (*A*) Experimental workflow for yeast spectral library construction (*Top*) and fast, single-run DIA-based analysis of yeast proteomes (*Bottom*). (*B*) Number of identified proteins using DDA with varying LC gradient lengths compared to 23-min DIA. (*C*) Cumulative number of identified unique yeast peptides over time using DDA with varying LC gradient lengths and the 23-min DIA method. (*D*) Number of quantified proteins in growth and stress conditions. (*E*) PCA of 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 ($P < 0.05$ and at least fourfold change in both directions). (*G*) GO-term enrichment in the YPD vs. SC fold change dimension (one-dimensional annotation enrichment, FDR <5%). Terms with positive enrichment scores are enriched in YPD over SC and vice versa.

degradation of a distinct set of proteins (Fig. 2*A*). For example, 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 (48–50). Importantly, our data also revealed that the heat-shock response is dose-dependent, with higher induction of the stress response at 42 °C compared to 37 °C (Fig. 2 *A*, green cluster and *B*). Yeast experiencing osmotic shock, on the other hand, induced distinct proteome changes, with the most enriched Gene Ontology (GO) term under this condition being actin-cortical patch (*SI Appendix*, Fig. S2 *A* and *B*). This is consistent with the fact that yeast cells rapidly disassemble and remodel the actin cytoskeleton during osmotic stress and favor the formation of actin patches over filaments, a mechanism that lowers the turgor pressure and allows continued growth of yeast under high osmolarity (51, 52). In addition, one of the most up-regulated proteins during osmotic stress is Ena1 (*SI Appendix*, Fig. S2*A*), a sodium efflux pump that plays a crucial role in allowing salt tolerance (53). Growth during amino acid or nitrogen starvation primarily resulted in the induction of amino acid biosynthetic pathways, with arginine and cysteine synthesis being particularly up-regulated (*SI Appendix*, Fig. S2 *C*–*F*).

In addition to temperature and nutrient availability, carbon source is a crucial determinant of yeast growth. We compared the proteomes of yeast grown in the aerobic carbon source, glucose, with the nonfermentable carbon source, ethanol. Yeast will preferentially metabolize aerobic carbon sources, such as glucose, when they are present in the media. When only nonfermentable carbon sources, such as ethanol, are present, yeast cells will instead 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 (54, 55). Consistent with this, we observe a general up-regulation of mitochondrial proteins and those involved in the tricarboxylic acid cycle during growth in ethanol (Fig. 2 *A*, light blue cluster, *C*, and *D*). In addition, many proteins involved in carbon metabolism are differentially regulated in glucose and ethanol-containing media. For example, we see a greater than 16-fold up-regulation of the gluconeogenic enzymes Fbp1, Pck1, and Icl1 (Fig. 2*C*). In the absence of glucose, both Hxt7, a glucose transporter, and Hxk1,

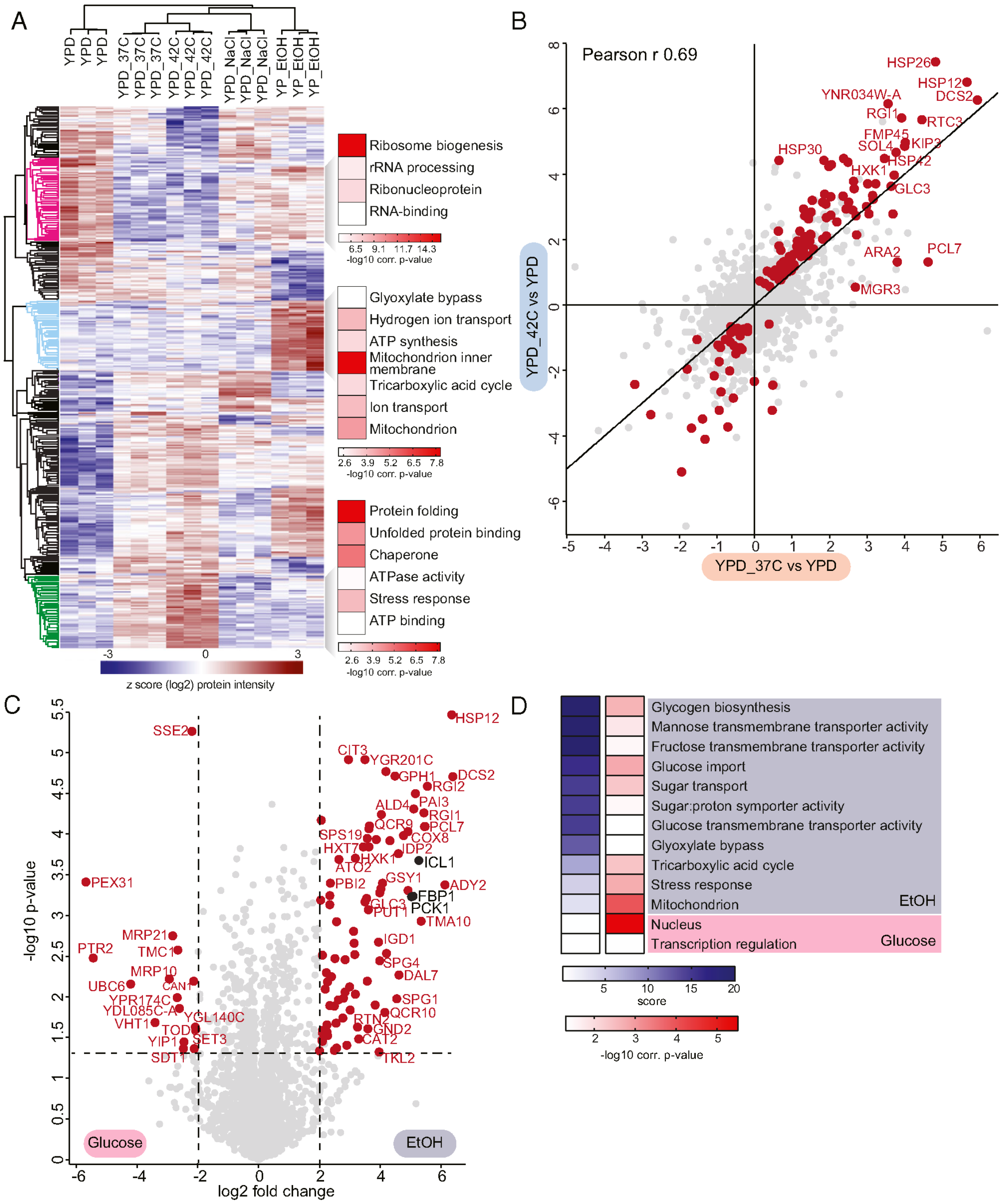


Fig. 2. Large-scale and quantitative analysis of yeast proteomes under different stresses. (*A*) Heat map of z-scored protein abundances (log2 DIA intensities) of the differentially expressed proteins (ANOVA, FDR <0.01) after hierarchical clustering of stress conditions performed in YPD and YPE. Fisher exact test was performed to identify significantly enriched GO terms in the most prominent profiles (FDR <5%). (*B*) Correlation of log2 fold-changes of all the quantified proteins during heat shock. The proteins that change significantly in either 37 °C or 42 °C compared to 30 °C YPD control are colored in red (*t* test, FDR <5%). (*C*) Volcano plot of the (−log10) *P* values vs. the log2 protein abundance differences between glucose starvation (ethanol) vs. YPD. Red dots indicate significantly different proteins, determined based on $P < 0.05$ and at least fourfold change in both directions. (*D*) GO-term enrichment in the ethanol vs. YPD fold change dimension (one-dimensional annotation enrichment, FDR <5%). Terms with positive enrichment scores are enriched in stress condition over glucose (YPD) control and vice versa.

a hexokinase, are up-regulated (Fig. 2*C*), 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 the current stressor but also facilitates a rapid response to shifts in environmental conditions (40, 56). Moreover, apart from identifying proteins that have altered levels in response to a shift in environmental conditions, we also accurately determined their fold changes, giving valuable insight into the protein content under different stress and growth conditions that is indispensable for systems-level modeling.

Taken together, our results indicate that the fast and robust DIA-based approach described here can reliably and quantitatively retrieve the known differences and even reveal new and biologically meaningful regulation of protein expression, thereby providing a near-comprehensive resource for yeast researchers and a valuable platform to support future studies in quantitative biology.

Global Regulation of the Yeast Proteome during Glucose Starvation and Recovery. To gain better insights into how yeast regulate metabolism in response to a change in carbon source, we next expanded our analysis to investigate glucose starvation and glucose recovery. Yeast cultures were first grown to logarithmic phase in glucose then switched to media containing ethanol as a nonfermentable carbon source. Following 19 hours of growth in ethanol, glucose was replenished and the yeast were allowed to recover for 30 minutes or 2 hours (Fig. 3*A*). In these growth conditions, we quantified 3,602 distinct proteins in total (Dataset S2). The first PCA component segregated the growth conditions, with glucose being largely separated from the ethanol and recovery conditions (Fig. 3*B* and *SI Appendix*, Fig. S3*A*). To further investigate the regulation of metabolism in alternate carbon sources, we compared the proteome changes with those of the transcriptome. PCA analysis of the transcriptome also showed that the first component separated the growth conditions. Interestingly, in this case cells grown in ethanol were largely separated from the glucose (never starved) and glucose recovery conditions (Fig. 3*C* and *SI Appendix*, Fig. S3*B*), suggesting that during this metabolic shift yeast cells remodel their gene expression first through rapid changes in transcription, which facilitates production of new proteins, and then remove proteins that are no longer required.

Several regulatory mechanisms contribute to carbohydrate metabolism, including allosteric regulation, reversible enzyme inactivation through covalent modifications, and irreversible loss of enzyme activity through proteolysis (reviewed in ref. 57). 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 E3 ubiquitin ligase. GID E3 ligase subunits are 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 a yet-inactive anticipatory complex: GID^{Ant}. Following glucose replenishment, the substrate receptor, Gid4, is rapidly induced and joins the complex, allowing the recognition and subsequent degradation of the gluconeogenic proteins Fbp1, Mdh2, Icl1, and Pck1 via the Pro/N-degron pathway (Fig. 3*D*) (38–40, 58–61). Indeed, our analysis confirmed that most components of the GID E3 ligase are up-regulated around fourfold during growth in ethanol, with the exception of Gid4, which is rapidly and transiently up-regulated within 30 min of glucose replenishment (Fig. 3*E*).

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 the major contributors to the segregation based on growth condition (Fig. 3*F*). While the GID E3 ligase is known to be an important contributor to the regulation of yeast metabolism during the switch from

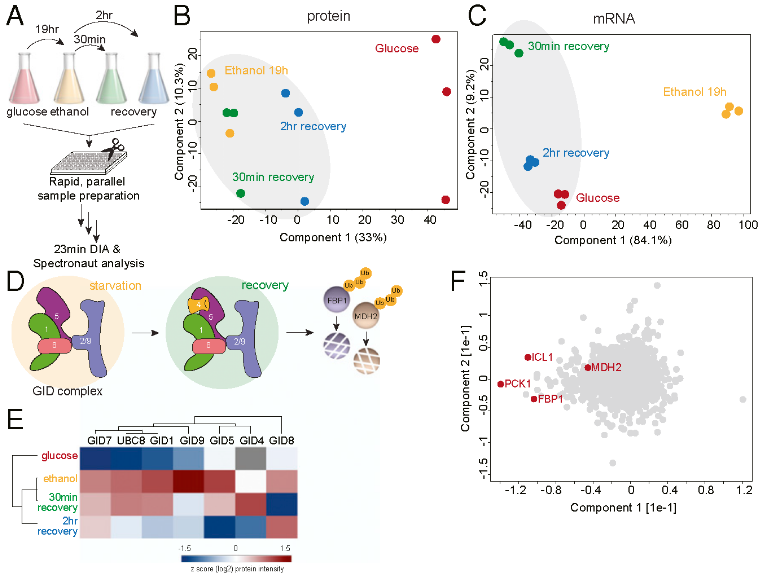


Fig. 3. Global proteome changes of yeast under glucose starvation and recovery. (*A*) Rapid yeast proteome profiling under glucose starvation and recovery. (*B* and *C*) PCA plot of growth conditions along with their biological replicates based on their protein expression (*B*) and mRNA abundance (*C*) 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, including Fbp1 and Mdh2. (*E*) Heat map of z-scored protein abundances (log2) of the GID complex subunits under glucose starvation and recovery in wild-type yeast cells. (*F*) PCA plot of proteins during glucose starvation and recovery. Proteins marked in red represent the known GID complex substrates.

gluconeogenic to glycolytic conditions, and is thought to have additional substrates, the lack of an obvious phenotype in GID mutants has made the identification of further substrates challenging. Thus, we applied a DIA-based workflow to search for novel regulatory targets of the GID E3 ligase.

Identifying GID Ligase-Dependent Regulation during Recovery from Carbon Starvation. The structure and molecular mechanism of the GID E3 ligase are known, but the pathways it regulates are only beginning to be elucidated (38, 40, 61). While the role of the GID ligase in the regulation of gluconeogenesis is well-characterized, the conservation of this multiprotein complex throughout eukaryotes suggests that it likely regulates additional pathways. For example, the GID/CTLH complex has a role in erythropoiesis and spermatogenesis in human cells and in embryogenesis in *Drosophila* (62–65). Thus, we set out to uncover additional pathways regulated by the GID E3 ligase in yeast by utilizing a combination of mutants. First, we used a deletion of the substrate receptor, Gid4, which targets proteins with either an N-terminal proline or a proline at position 2 via the Pro/N-degron pathway (38, 59, 60, 66). Deletion of Gid4 therefore should prevent substrate binding to the GID complex and thereby inhibit degradation. However, Gid4, while conserved in human cells, is not conserved throughout all eukaryotes. For example, the GID complex in *Drosophila* lacks an identifiable Gid4 homolog (65), suggesting an alternate mode of recognition. In addition, in yeast, the protein Gid10 has been identified as an alternate substrate receptor of the GID complex (40, 67), although no Gid10-dependent cellular substrates have been identified to date. To identify pathways regulated by the GID complex by an alternative recognition pathway, we used a structure-based point mutant in the RING-domain-containing subunit, Gid2^{K365A}, which eliminates catalytic activity without altering folding or complex assembly (40).

We compared the transcriptomes and proteomes of wild-type yeast to yeast containing either a Gid4 deletion or a Gid2 mutant (Gid2^{K365A}) grown under the glucose starvation and recovery conditions described previously. Each condition was measured in triplicate using the rapid DIA method (Fig. 4*A*). Importantly, there were no GID-dependent differences in messenger RNA (mRNA) levels following glucose replenishment (*SI Appendix*, Fig. S4*A*), demonstrating that the GID E3 ligase does not regulate protein synthesis but rather the fate of existing proteins. To confirm that a DIA-based approach would be able to recognize bona fide GID substrates, we first examined the expression patterns of the well-characterized substrates Fbp1 and Mdh2. Indeed, in wild-type cells, Fbp1 and Mdh2 protein levels are induced during growth in ethanol and then turned over within 2 hours of glucose recovery, with Fbp1 and Mdh2 protein levels reduced by around eightfold and 5.7-fold, respectively. As expected, both proteins are also stabilized in the GID4-deleted and gid2-mutant cells (Fig. 4 *B* and *C*), confirming that we can robustly identify changes in expression of known substrates.

To identify novel targets, 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) its levels should decrease during glucose replenishment, and 3) after 2 h of glucose replenishment it should have a higher expression level in the GID4-deleted and/or gid2-mutant cells, compared to wild type (*SI Appendix*, Fig. S4*B*). This provided a list of 31 proteins, including all four known GID substrates (Fbp1, Mdh2, Pck1, and Icl1) (*SI Appendix*, Fig. S4*C*). To further prioritize candidates, we limited our search to proteins with an N-terminal proline or a proline in the second position, a genetic and structural requirement of all known cellular substrates (38, 40, 60). The resulting list of seven proteins consisted of the four known substrates, the transcription factor Azf1, and the metabolic enzymes Aro10 and Acs1 (Fig. 4*D*). Interestingly, Azf1 has already been implicated in regulation of GID4 transcription (68), suggesting its up-regulation in the GID-deficient cells may be a cellular compensation mechanism. However, because we did not observe any GID-dependent mRNA expression changes (*SI Appendix*, Fig. S4*A*), 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 GID targets in vivo, we used the promoter reference technique (38, 69), a transcription-

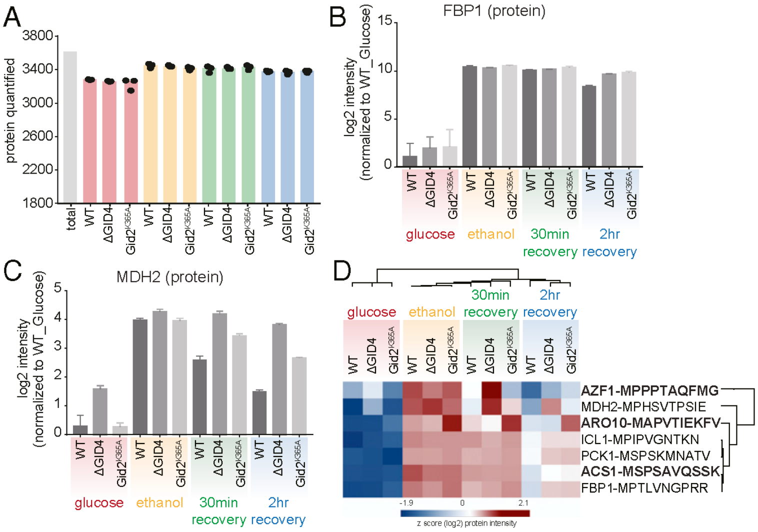


Fig. 4. Rapid and robust DIA-based approach identifies GID substrates during recovery after glucose starvation. (*A*) Number of quantified proteins in wild type (WT), ΔGID4, and Gid2^{K365A} yeast cells during glucose starvation and recovery. (*B* and *C*) Bar graphs showing abundances (log2) of Fbp1 (*B*) and Mdh2 (*C*) proteins that are normalized to WT glucose (never starved) condition in WT, ΔGID4, and Gid2^{K365A} yeast cells during glucose starvation and recovery. (*D*) Heat map of z-scored protein abundances (log2) of the proteins which have the criteria of GID substrates.

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 (Fig. 5*A*). 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, allowing the fate of the existing protein to be monitored. Importantly, this method selectively terminates synthesis of our test proteins, and 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 (Fig. 5*B*), while the Aro10 protein is stabilized in GID2-deleted but not GID4-deleted cells (Fig. 5*C*), indicating a potential Gid4-independent regulation. Thus, Acs1 and Aro10 are confirmed to be regulatory targets of the GID E3 ligase during the switch from gluconeogenic to glycolytic conditions.

Discussion

Here, we described a straightforward, streamlined, and reproducible systems biology approach for yeast proteome profiling using DIA to analyze biological pathways much faster and with greater depth. The 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 nonspecialized research groups. Despite its simplicity, it robustly and quantitatively profiles hundreds of largely covered yeast proteomes [80% of the expressed proteome at normal growth conditions (7)] within an unprecedented throughput (100 samples in ~2.2 days).

The ability of cells to adapt to stress or changes in environmental conditions relies on extensive proteome remodeling (70–75). Understanding these changes provides broad insight into the molecular mechanisms underlying many processes including heat stress, adaptation to nutrient availability, and regulation of cell division. Applying the DIA-based approach to profile protein levels during response to several stress and growth conditions demonstrated its systems-wide robustness and specificity. In addition, our work provides an in-depth resource on stress mediators regulated at the protein level, which will complement the widely available yeast transcriptome data and further allow yeast researchers to probe numerous biological pathways of interest, including stress response pathways, autophagy, and nutrient signaling pathways.

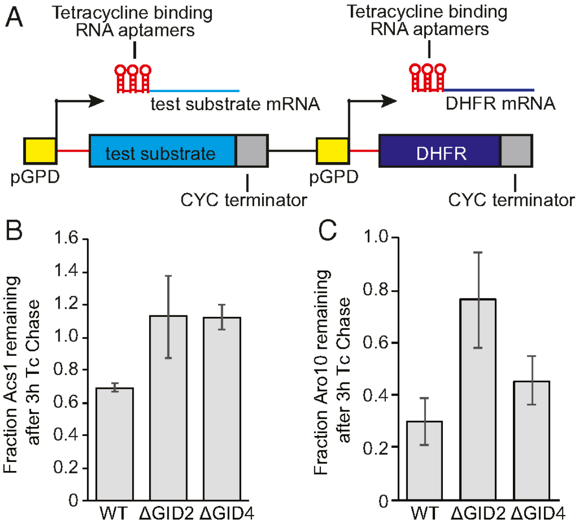


Fig. 5. In vivo validation of GID targets. (*A*) Schematic of constructs used in the promoter reference technique. (*B* and *C*) Quantification of Acs1 (*B*) and Aro10 (*C*) degradation, based on at least four independent replicates. Bars represent mean values, and error bars represent SD.

In addition to identifying proteome changes during stress, we used the DIA-based systems biology approach to identify proteins that are regulated by the GID E3 ubiquitin ligase, a key regulator in the switch from gluconeogenic to glycolytic conditions (54, 61, 76). Despite the importance of the GID complex in metabolic regulation, identification of additional substrates has been hindered by the lack of an obvious phenotype, variable kinetics of protein degradation, and the necessity for a sensitive readout. Our generic and unbiased approach, however, robustly identified two protein regulatory targets of the GID complex, Acs1 and Aro10, further highlighting the importance and need for quantitative proteome datasets to provide a 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 encodes 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 nonfermentable carbon sources (77). During glycolytic growth, however, the main energy flux does not require Acs1/2 function, Acs1 expression is suppressed, and existing Acs1 protein must be degraded. Aro10 encodes a phenylpyruvate decarboxylase that catalyzes an irreversible step in the Ehrlich pathway, which provides a more energetically favorable means of NADH (reduced nicotinamide-adenine dinucleotide) regeneration during anaerobic growth. Following glucose replenishment, NADH is regenerated through glycolysis, and thus Aro10 function is no longer required (78, 79).

Here, we show that both Acs1 and Aro10 turnover are dependent on the catalytic activity of the GID complex, via its RING-containing subunit, Gid2. Intriguingly, only Acs1 turnover is dependent on the well-characterized substrate receptor, Gid4, suggesting an alternate mode of recognition for Aro10. Indeed, an additional substrate receptor, Gid10, has recently been identified (40, 67), raising the possibility that Aro10 may be the first substrate identified in this recognition pathway. Alternatively, Aro10 recognition may be facilitated by a yet-to-be identified substrate receptor or an alternative mechanism. In either case, the regulation of Aro10 suggests that the GID E3 ligase may function with separable catalytic and substrate recognition elements, a mechanism previously described for SCF (Skp1-Cullin-Fbox) E3 ligases (80, 81) that provides a flexible means for linking a single E3 to a greater number of substrates. Intriguingly, expression of the GID substrate receptors is induced during several other cellular stresses, including osmotic shock, heat shock, and nitrogen starvation (40, 67, 71), suggesting that the GID complex may play an important role in rewiring metabolic pathways during adaptation to a wide variety of stress conditions.

Taken together, the GID-dependent regulation of Acs1 and Aro10, along with the previously known substrates, suggests that the GID complex is a multifunctional metabolic regulator that influences multiple cellular pathways simultaneously to allow for an efficient switch from gluconeogenic to glycolytic conditions. Moreover, our findings demonstrate that the DIA-based systems biology approach is capable of simultaneously identifying changes to multiple cellular pathways which are integrated to maintain cellular homeostasis. While we here identified specific targets of an E3 ligase, this workflow can be readily adopted by the community to probe numerous cellular pathways, including kinase signaling pathways or cell-cycle-dependent changes. Furthermore, its speed allows the analyzing of at least 15 conditions, in triplicate, per day, making it particularly well-suited for screens. For example, the effect of each of the ~80 yeast E3 ligases on the global proteome could be ascertained in just 5 days, or each of the ~117 yeast kinases in about 1 week. In addition, the DIA-based workflow can be easily adapted to identify changes in posttranslational modifications including phosphorylation, ubiquitination, and acetylation, when coupled with an enrichment step (44, 82–84).

Thus, the speed and reproducibility of the DIA-based approach presented here allows researchers to probe complex biological pathways and identify novel regulatory mechanisms. We are currently integrating an HPLC system into our approach as it eliminates the overhead time between sample pickup and start of MS measurement by using preformed gradients (85). Simplified workflows like the one described here could be extended to other organisms, generating high-quality quantitative proteome datasets which are required to explain biological processes on a system-wide level (86, 87). Furthermore, we believe that library-free approaches using prediction tools will further increase the speed of DIA-based proteome profiling workflows like the one presented here. Given that the expressed human proteome (around 15,479 proteins, https://www.proteomicsdb.org) is only around three times larger than the expressed yeast proteome [5,391 proteins, (45)], with only three fold increase in proteomic depth, we anticipate fast single run DIA approaches will also be suitable for rapid generation of human proteomes.

Materials and 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, and 2% glucose) or SC (0.67% yeast nitrogen base without amino acids, 2% glucose, containing 87.5 mg/L alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, leucine, lysine, methionine, myo-inositol, isoleucine, phenylalanine, proline, serine, threonine, tyrosine and valine, 43.7 mg/L histidine, tryptophan and uracil, 22.5 mg/L adenine, and 8.7 mg/L *para*-aminobenzoic acid) media. Unless otherwise specified, yeast cultures were grown at 30 °C. For heat-shock conditions, yeast cultures were grown in YPD to an optical density at 600 nm (OD_{600}) of 1.0 and then shifted to the indicated temperature for 1 h. For osmotic shock conditions, yeast cells were grown in YPD to and OD_{600} of 1.0, pelleted at 3,000 rpm for 3 min, and resuspended at an OD_{600} of 1.0 in prewarmed YPD + 0.5 M NaCl. For glucose starvation, yeast cells were grown in YPD to an OD_{600} of 1.0 to 2.0, pelleted at 3,000 rpm for 3 min, washed once with YPE (1% yeast extract, 2% peptone, and 2% ethanol), resuspended in prewarmed YPE at an OD_{600} of 1.0, and grown at 30 °C for 19 h. For glucose recovery, yeast cells were pelleted after 19 h of growth in YPE, resuspended to an OD_{600} of 1.0 in YPD, and allowed to grow at 30 °C for 30 min or 2 h. For amino acid starvation, yeast cells were grown in SC to an OD_{600} of 1.0 to 2.0, pelleted at 3,000 rpm for 3 min, washed once with SD-AA (0.67% yeast nitrogen base without amino acids, 2% glucose, and 20 mg/L uracil), resuspended in SD-AA to an OD_{600} of 1.0, and allowed to grow for 1 h. For nitrogen depletion, yeast cells were grown in SC to an OD_{600} of 1.0 to 2.0, pelleted at 3,000 rpm for 3 min, washed once with SD-N (0.17% yeast nitrogen base without amino acids or ammonium sulfate and 2% glucose), resuspended in SD-N to an OD_{600} of 1.0, and allowed to grow for 1 h. For proteomics analysis, 50 ODs of cells were pelleted at 3,000 rpm for 3 min, 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.

Protein Degradation Assays (Promoter Reference Technique). Protein degradation assays using the promoter reference technique were done as previously described (69). Plasmids used are listed in Table 2. 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 h, 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 0.8 mL of 0.2 M NaOH, followed by incubation on ice for 20 min, and then pelleted at 11,200 × *g* for 1 min. The supernatant was removed and the pellet resuspended in 50 μL HU buffer and incubated at 70 °C for 10 min. The lysate was precleared by centrifugation at 11,200 × g for 5 min and then loaded onto a 12% sodium dodecyl sulfate polyacrylamide gel. Protein samples were transferred to a nitrocellulose membrane and then visualized by Western blot using αFLAG (F1804; Sigma) and α-hemagglutinin (H6908; Sigma) primary antibodies and Dylight 633 goat anti-Mouse (35512; Invitrogen) and Dylight 488 goat anti-rabbit (35552; Invitrogen) secondary antibodies. Membranes were imaged on a typhoon scanner (Amersham). Bands were quantified with ImageStudio software (LI-COR).

Table 1. Yeast strains used in this study

Strain	Genotype	Source
BY4741	*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*	Euroscarf
CRLY12	*BY4741 GID4::KANMX*	This study
CRLY30	*BY4741 GID2::KANMX*	This study
CRLY131	*BY4741 gid2::3xFLAG-GID2K365A*	Ref. 40

mRNA Sequencing. Harvested and frozen cells were sent to Novogene Co., Ltd. (Hong Kong) for RNA extraction, library preparation, mapping, and bioinformatics analysis. Briefly, 3 μg of RNA was used for library generation using NEB Next Ultra RNA Library Prep Kit for Illumina (NEB). The library preparations were sequences on an Illumina Hi-Seq platform and 125-base pair (bp)/150-bp paired-end reads were generated. Reads were indexed using Bowtie v2.2.3 and paired-end clean reads were aligned to the reference genome using TopHat v2.0.12. HTSeq v0.6.1 was used to count the read numbers mapped to each gene, and then FPKM (expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced) of each gene was calculated based on the length of the gene and read counts mapped to the gene. The transcriptome data analysis was performed as explained in *Data Processing and Bioinformatics Analysis*.

Sample Preparation for MS Analysis. Sodium deoxycholate (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 to 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 (27) 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 40 mM and 10 mM, respectively. Samples were digested overnight at 37 °C using trypsin (1:100 wt/wt; Sigma-Aldrich) and LysC (1/100 wt/wt; Wako). The following day, peptide material was desalted using SDB-RPS StageTips (Empore) (27). Briefly, samples were diluted with 1% trifluoroacetic acid (TFA) in isopropanol to a final volume of 200 μL and loaded onto StageTips and subsequently washed with 200 μL of 1% TFA in isopropanol and 200 μL 0.2% TFA/2% ACN (acetonitrile). Peptides were eluted with 80 μl of 1.25% Ammonium hydroxide (NH4OH)/80% ACN and dried using a SpeedVac centrifuge (Concentrator Plus; Eppendorf). Samples were resuspended in buffer A* (0.2% TFA/2% ACN) prior to LC-MS/MS analysis. Peptide concentrations were measured optically at 280 nm (Nanodrop 2000; Thermo Scientific) and subsequently equalized using buffer A*. Three hundred nanograms of 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 StageTips (discussed above). Around 8 or 30 μg of peptides were fractionated into 8 or 24 fractions, respectively, by high-pH reversed-phase chromatography as described earlier (88). 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-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 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 min,

Table 2. Plasmids used in this study

Plasmid		Source
CRLP47	pRS313-P_{TDH3}(modified)-Aro10$_{3xFlag}$-CYC-p_{TDH3}(modified)-$_{flag}$DHFR$_{ha}$-CYC	This study
CRLP48	pRS313-P_{TDH3}(modified)-Acs1$_{3xFlag}$-CYC-p_{TDH3}(modified)-$_{flag}$DHFR$_{ha}$-CYC	This study

95% in a minute, and stayed at 95% for 3.5 min. The mass spectrometer was operated in Top10 data-dependent mode (DDA) with a full scan range of 300 to 1,650 *m/z* at 60,000 resolution with an automatic gain control (AGC) target of 3e6 and a maximum fill time of 20 ms. Precursor ions were isolated with a width of 1.4 *m/z* and fragmented by higher-energy collisional dissociation (HCD) (normalized collision energy [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 to 1,650 *m/z*, a maximum fill time of 60 ms, and an 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 HCD (NCE 25.5, 27, and 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 (UniProt, 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 (89). Protein intensities and mRNA abundances were log2-transformed for further analysis. The datasets 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 downshift = 1.8). PCA analysis of stress and growth conditions and biological replicates was performed as previously described in ref. 90. Multisample 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 datasets. For truncation, we used permutation-based false discovery rate (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 the 1D tool in Perseus. Annotation terms were filtered for 5% FDR after Benjamini–Hochberg correction.

Data Availability. All MS proteomics data have been deposited on ProteomeXchange via the PRIDE database with the dataset identifier PXD021559. All other data supporting findings of this study are available within this paper, *SI Appendix*, and Datasets S1 and S2.

ACKNOWLEDGMENTS. 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, Kirby Swatek, Laura A. Hehl, 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 assistance and helpful discussions. B.A.S. is funded from the Deutsche Forschungsgemeinschaft (SCHU 3196/1-1). B.A.S. and M.M. are supported by the Max Planck Society.

1. X. Sui *et al.*, Widespread remodeling of proteome solubility in response to different protein homeostasis stresses. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2422–2431 (2020).
2. N. Berner, K. R. Reutter, D. H. Wolf, Protein quality control of the endoplasmic reticulum and ubiquitin-proteasome-triggered degradation of aberrant proteins: Yeast pioneers the path. *Annu. Rev. Biochem.* **87**, 751–782 (2018).
3. J. Hanna, A. Guerra-Moreno, J. Ang, Y. Micoogullari, Protein degradation and the pathologic basis of disease. *Am. J. Pathol.* **189**, 94–103 (2019).
4. M. S. Hipp, S. H. Park, F. U. Hartl, Proteostasis impairment in protein-misfolding and -aggregation diseases. *Trends Cell Biol.* **24**, 506–514 (2014).
5. L. M. Steinmetz *et al.*, Systematic screen for human disease genes in yeast. *Nat. Genet.* **31**, 400–404 (2002).
6. G. Giaever *et al.*, Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).
7. S. Ghaemmaghami *et al.*, Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
8. W. K. Huh *et al.*, Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
9. G. M. Jones *et al.*, A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. *Nat. Methods* **5**, 239–241 (2008).
10. U. Weill *et al.*, Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods* **15**, 617–622 (2018).
11. I. Yofe *et al.*, One library to make them all: Streamlining the creation of yeast libraries via a SWAp-tag strategy. *Nat. Methods* **13**, 371–378 (2016).
12. M. Meurer *et al.*, Genome-wide C-SWAT library for high-throughput yeast genome tagging. *Nat. Methods* **15**, 598–600 (2018).
13. M. Costanzo *et al.*, A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
14. A. C. Gavin *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
15. Y. Ho *et al.*, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180–183 (2002).
16. J. L. DeRisi, V. R. Iyer, P. O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
17. A. Kumar *et al.*, Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
18. D. A. Lashkari *et al.*, Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13057–13062 (1997).
19. U. Nagalakshmi *et al.*, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
20. M. Costanzo *et al.*, Global genetic networks and the genotype-to-phenotype relationship. *Cell* **177**, 85–100 (2019).
21. G. E. Janssens *et al.*, Protein biogenesis machinery is a driver of replicative aging in yeast. *eLife* **4**, e08527 (2015).
22. D. Petranovic, J. Nielsen, Can yeast systems biology contribute to the understanding of human disease? *Trends Biotechnol.* **26**, 584–590 (2008).
23. L. M. de Godoy *et al.*, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
24. A. S. Hebert *et al.*, The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).
25. S. Marguerat *et al.*, Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151**, 671–683 (2012).
26. J. B. Müller *et al.*, The proteome landscape of the kingdoms of life. *Nature* **582**, 592–596 (2020).
27. N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
28. N. Nagaraj *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. *Mol. Cell. Proteomics* **11**, M111.013722 (2012).
29. P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, R. Aebersold, Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138**, 795–806 (2009).
30. A. L. Richards *et al.*, One-hour proteome analysis in yeast. *Nat. Protoc.* **10**, 701–714 (2015).
31. B. Soufi *et al.*, Global analysis of the yeast osmotic stress response by quantitative proteomics. *Mol. Biosyst.* **5**, 1337–1346 (2009).
32. S. S. Thakur *et al.*, Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10**, M110.003699 (2011).
33. K. J. Webb, T. Xu, S. K. Park, J. R. Yates 3rd, Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res.* **12**, 2177–2184 (2013).
34. R. Wu *et al.*, Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. *Mol. Cell. Proteomics* **10**, M111.009654 (2011).
35. L. C. Gillet *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 (2012).
36. C. Ludwig *et al.*, Data-independent acquisition-based SWATH-MS for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).

37. A. Varshavsky, The ubiquitin system, an immense realm. *Annu. Rev. Biochem.* **81**, 167–176 (2012).
38. S. J. Chen, X. Wu, B. Wadas, J. H. Oh, A. Varshavsky, An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes. *Science* **355**, eaal3655 (2017).
39. M. Hämmerle *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. *J. Biol. Chem.* **273**, 25000–25005 (1998).
40. S. Qiao *et al.*, Interconversion between anticipatory and active GID E3 Ubiquitin ligase conformations via metabolically driven substrate receptor assembly. *Mol. Cell* **77**, 150–163.e9 (2020).
41. P. Kaiser, N. Y. Su, J. L. Yen, I. Ouni, K. Flick, The yeast ubiquitin ligase SCFMet30: Connecting environmental and intracellular conditions to cell division. *Cell Div.* **1**, 16 (2006).
42. A. Kobayashi *et al.*, Oxidative and electrophilic stresses activate Nrf2 through inhibition of ubiquitination activity of Keap1. *Mol. Cell. Biol.* **26**, 221–229 (2006).
43. D. D. Zhang, M. Hannink, Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress. *Mol. Cell. Biol.* **23**, 8137–8151 (2003).
44. D. B. Bekker-Jensen *et al.*, Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* **11**, 787 (2020).
45. B. Ho, A. Baryshnikova, G. W. Brown, Unification of protein abundance datasets yields a quantitative Saccharomyces cerevisiae proteome. *Cell Syst.* **6**, 192–205.e3 (2018).
46. R. Serrano, D. Bernal, E. Simón, J. Ariño, Copper and iron are the limiting factors for growth of the yeast Saccharomyces cerevisiae in an alkaline environment. *J. Biol. Chem.* **279**, 19698–19704 (2004).
47. J. Bereiter-Hahn, Behavior of mitochondria in the living cell. *Int. Rev. Cytol.* **122**, 1–63 (1990).
48. R. Gomez-Pastor, E. T. Burchfiel, D. J. Thiele, Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 4–19 (2018).
49. F. U. Hartl, Molecular chaperones in cellular protein folding. *Nature* **381**, 571–579 (1996).
50. R. M. Vabulas, S. Raychaudhuri, M. Hayer-Hartl, F. U. Hartl, Protein folding in the cytoplasm and the heat shock response. *Cold Spring Harb. Perspect. Biol.* **2**, a004390 (2010).
51. A. Blomberg, L. Adler, Physiology of osmotolerance in fungi. *Adv. Microb. Physiol.* **33**, 145–212 (1992).
52. S. Chowdhury, K. W. Smith, M. C. Gustin, Osmotic stress and the yeast cytoskeleton: Phenotype-specific suppression of an actin mutation. *J. Cell Biol.* **118**, 561–571 (1992).
53. M. Platara *et al.*, The transcriptional response of the yeast Na(+)-ATPase ENA1 gene to alkaline stress involves three main signaling pathways. *J. Biol. Chem.* **281**, 36632–36642 (2006).
54. J. M. Gancedo, Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev.* **62**, 334–361 (1998).
55. H. J. Schüller, Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr. Genet.* **43**, 139–160 (2003).
56. I. Tagkopoulos, Y. C. Liu, S. Tavazoie, Predictive behavior within microbial genetic networks. *Science* **320**, 1313–1317 (2008).
57. J. A. Barnett, K. D. Entian, A history of research on yeasts 9: Regulation of sugar metabolism. *Yeast* **22**, 835–894 (2005).
58. S. J. Chen, A. Melnykov, A. Varshavsky, Evolution of substrates and components of the Pro/N-degron pathway. *Biochemistry* **59**, 582–593 (2020).
59. A. Varshavsky The N-end rule pathway and regulation by proteolysis. *Protein Sci.* **20**, 1298–1345 (2011).
60. C. Dong *et al.*, Molecular basis of GID4-mediated recognition of degrons for the Pro/N-end rule pathway. *Nat. Chem. Biol.* **14**, 466–473 (2018).
61. O. Santt *et al.*, The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism. *Mol. Biol. Cell* **19**, 3323–3333 (2008).
62. W. X. Cao *et al.*, Precise temporal regulation of post-transcriptional repressors is required for an orderly Drosophila maternal-to-zygotic transition. *Cell Rep.* **31**, 107783 (2020).
63. S. Puverel, C. Barrick, S. Dolci, V. Coppola, L. Tessarollo, RanBPM is essential for mouse spermatogenesis and oogenesis. *Development* **138**, 2511–2521 (2011).
64. S. Soni, S. Bala, M. Hanspal, Requirement for erythroblast-macrophage protein (Emp) in definitive erythropoiesis. *Blood Cells Mol. Dis.* **41**, 141–147 (2008).
65. M. Zavortink *et al.*, The E2 Marie Kondo and the CTLH E3 ligase clear deposited RNA binding proteins during the maternal-to-zygotic transition. *eLife* **9**, e53889 (2020).
66. T. Tasaki, S. M. Sriram, K. S. Park, Y. T. Kwon, The N-end rule pathway. *Annu. Rev. Biochem.* **81**, 261–289 (2012).
67. A. Melnykov, S. J. Chen, A. Varshavsky, Gid10 as an alternative N-recognin of the Pro/N-degron pathway. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15914–15923 (2019).
68. M. G. Slattery, D. Liko, W. Heideman, The function and properties of the Azf1 transcriptional regulator change with growth conditions in Saccharomyces cerevisiae. *Eukaryot. Cell* **5**, 313–320 (2006).
69. J. H. Oh, S. J. Chen, A. Varshavsky, A reference-based protein degradation assay without global translation inhibitors. *J. Biol. Chem.* **292**, 21457–21465 (2017).
70. D. B. Berry, A. P. Gasch, Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. *Mol. Biol. Cell* **19**, 4580–4587 (2008).
71. A. P. Gasch *et al.*, Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
72. J. S. Hahn, Z. Hu, D. J. Thiele, V. R. Iyer, Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol. Cell. Biol.* **24**, 5249–5256 (2004).
73. S. Lindquist, The heat-shock response. *Annu. Rev. Biochem.* **55**, 1151–1191 (1986).
74. K. A. Morano, C. M. Grant, W. S. Moye-Rowley, The response to heat shock and oxidative stress in Saccharomyces cerevisiae. *Genetics* **190**, 1157–1195 (2012).
75. M. Mühlhofer *et al.*, The heat shock response in yeast maintains protein homeostasis by chaperoning and replenishing proteins. *Cell Rep.* **29**, 4593–4607.e8 (2019).
76. J. Regelmann *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. *Mol. Biol. Cell* **14**, 1652–1663 (2003).
77. M. A. van den Berg *et al.*, The two acetyl-coenzyme A synthetases of Saccharomyces cerevisiae differ with respect to kinetic properties and transcriptional regulation. *J. Biol. Chem.* **271**, 28953–28959 (1996).
78. L. A. Hazelwood, J. M. Daran, A. J. van Maris, J. T. Pronk, J. R. Dickinson, The Ehrlich pathway for fusel alcohol production: A century of research on Saccharomyces cerevisiae metabolism. *Appl. Environ. Microbiol.* **74**, 2259–2266 (2008).
79. E. J. Pires, J. A. Teixeira, T. Brányik, A. A. Vicente, Yeast: The soul of beer's aroma–A review of flavour-active esters and higher alcohols produced by the brewing yeast. *Appl. Microbiol. Biotechnol.* **98**, 1937–1949 (2014).
80. C. Bai *et al.*, SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. *Cell* **86**, 263–274 (1996).
81. D. Skowyra, K. L. Craig, M. Tyers, S. J. Elledge, J. W. Harper, F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* **91**, 209–219 (1997).
82. F. M. Hansen *et al.*, Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. bioRxiv: 10.1101/2020.07.24.219055 (25 July 2020).
83. A. Stukalov *et al.*, Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV. biorxiv:10.1101/2020.06.17.156455 (17 June 2020).
84. J. Baeza *et al.*, Revealing dynamic protein acetylation across subcellular compartments. *J. Proteome Res.* **19**, 2404–2418 (2020).
85. N. Bache *et al.*, A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
86. Y. Ahmad, A. I. Lamond, A perspective on proteomics in cell biology. *Trends Cell Biol.* **24**, 257–264 (2014).
87. M. Larance, A. I. Lamond, Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* **16**, 269–280 (2015).
88. N. A. Kulak, P. E. Geyer, M. Mann, Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
89. S. Tyanova *et al.*, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
90. S. J. Deeb *et al.*, Machine learning-based classification of diffuse large B-cell lymphoma patients by their protein expression profiles. *Mol. Cell. Proteomics* **14**, 2947–2960 (2015).

PNAS
www.pnas.org

Supplementary Information for

DIA-based systems biology approach unveils novel E3-dependent responses to a metabolic shift

Ozge Karayel [1,a], André C. Michaelis [1], Matthias Mann [1,b], Brenda A. Schulman [2,b] and Christine R. Langlois [2,a,b]

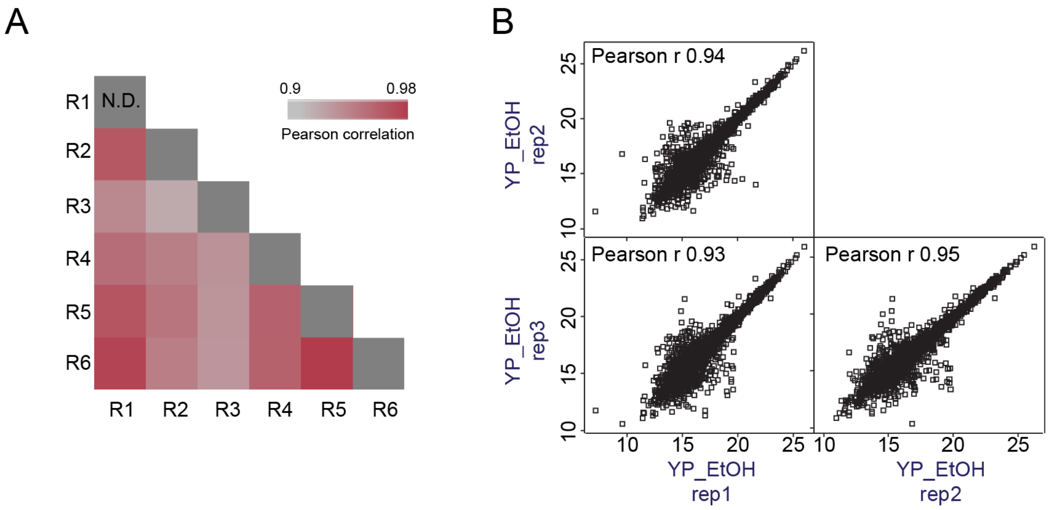
[1] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2] Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[a] These authors contributed equally
[b] Corresponding authors

E-mail: mmann@biochem.mpg.de, schulman@biochem.mpg.de, and langlois@biochem.mpg.de

This PDF file includes:

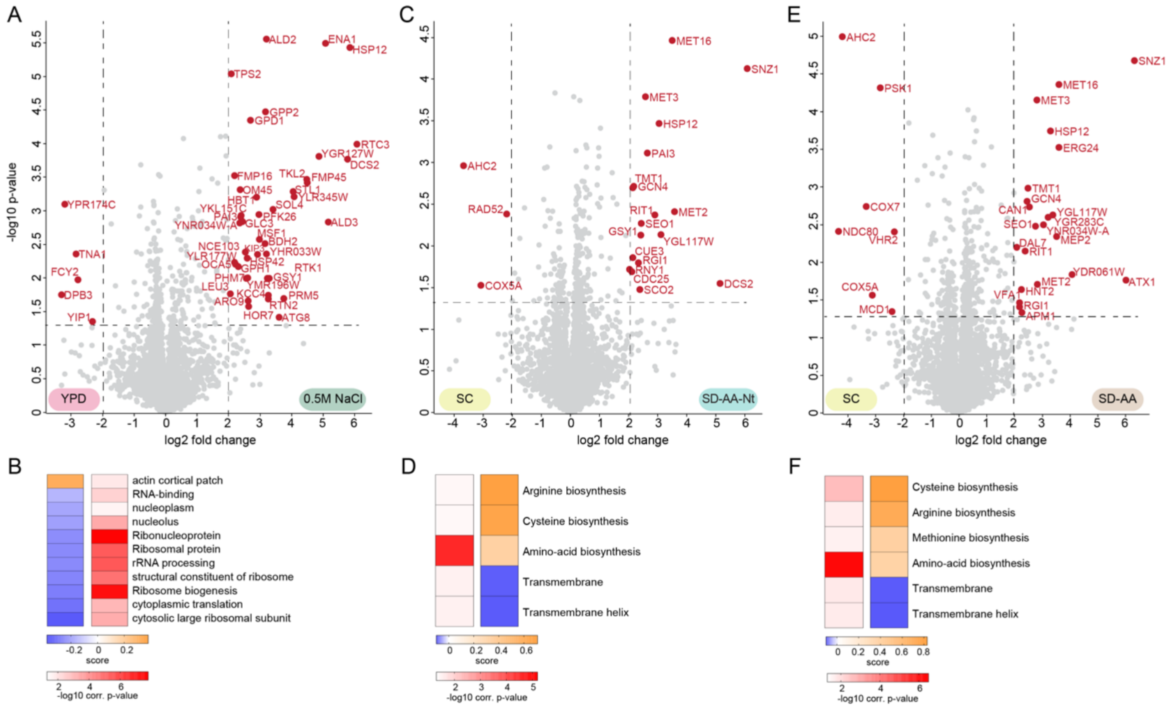
Figures S1 to S4

SUPPLEMENTARY FIGURES



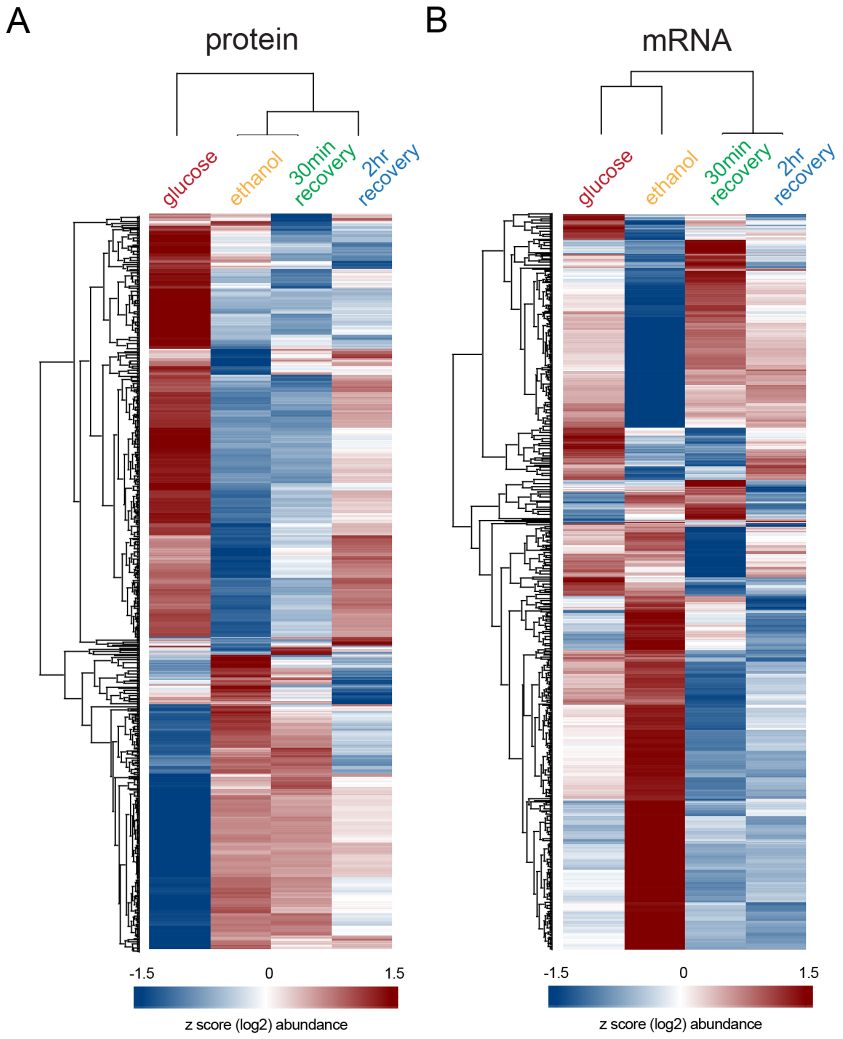
Supplementary Figure 1. Reproducibility of the DIA-based workflow

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.



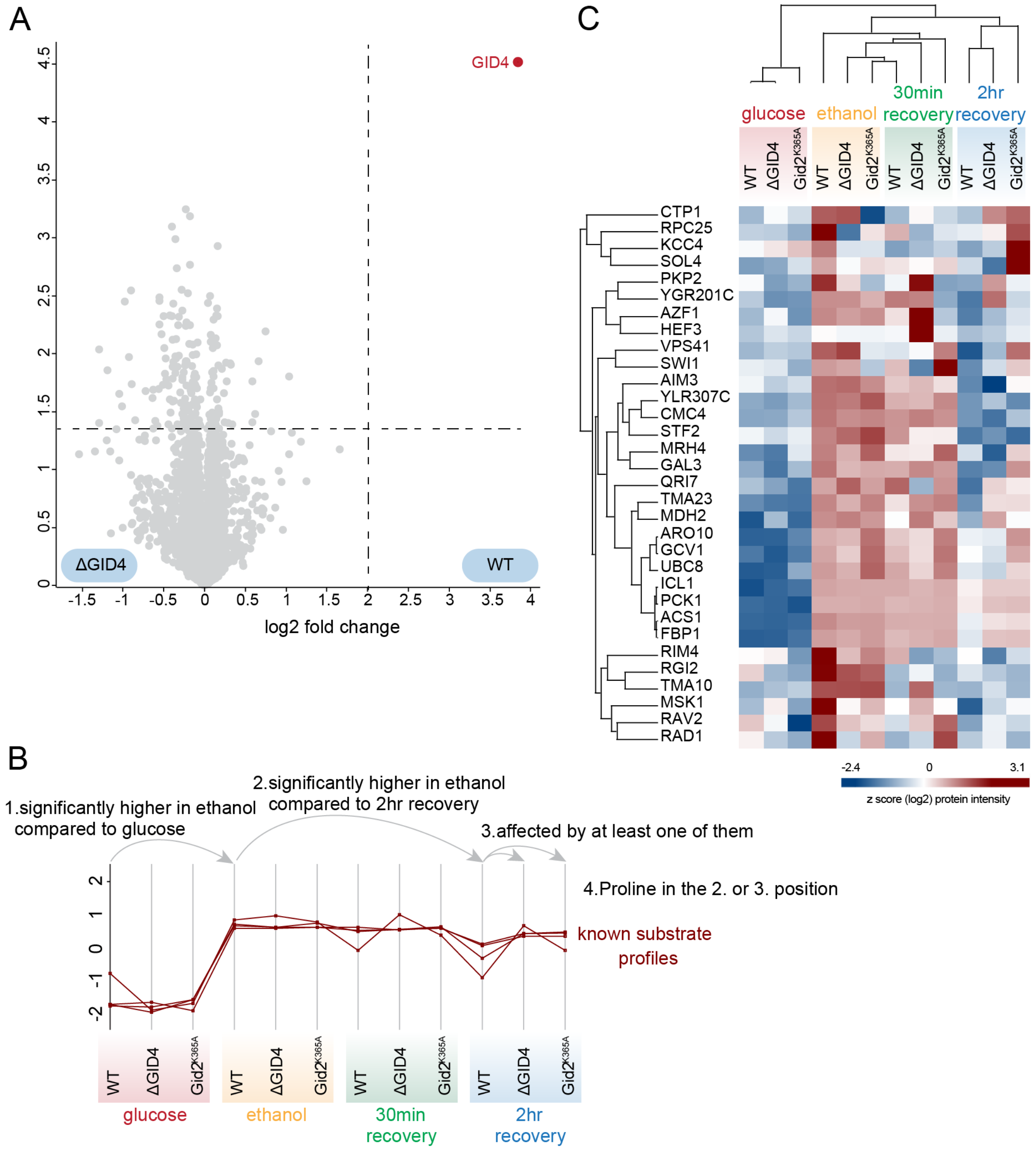
Supplementary Figure 2. Mapping changes to the yeast proteome in response to osmotic shock, and amino acid and nitrogen depletion

Volcano plot of the (-log10) p-values vs. the log2 protein abundance differences between 0.5M NaCl (osmotic shock) vs. YPD (A), SD-AA-Nt vs. SC (C), and SD-AA vs. SC (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 0.5M NaCl (osmotic shock) vs. YPD (B), SD-AA-Nt vs. SC (D), and SD-AA vs. SC (F). fold change dimension (1D enrichment, FDR < 5%). Terms with positive enrichment scores are enriched in stress condition over YPD or SC 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) (log2) 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 wildtype vs. Gid4 (substrate receptor) deletion. GID4 (shown in red) 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 targets which meet the first three conditions in panel B during glucose starvation and recovery.

3 Concluding Remarks and Outlook

3.1 Systems Biology: Unknot the Unknown

The thesis presented here provides systems biology approaches and applies them to the discovery of global protein-protein interactions in yeast and human cells, as well a study to quickly screen whole proteomes in yeast in response to different perturbation conditions. The resulting interactome for yeast is a resource of high-quality interactions that will help scientists around the world to gain novel functional insights of the cell.

We also provide a powerful way for scientists to validate every single interaction of interest. Our web application is readily accessible and reports how and why an interaction is included. In our bait-enrichment section one can validate the underlying p-values and t-test differences while the correlation section shows which samples cause a correlation to which degree. A separate quality control tab provides insights into the completeness of a sample in terms of quantified proteins.

From a network perspective, we have investigated and highlighted the “social” character of the interactome based on the observation that most proteins are involved in interactions and that there is an average shortest path of 4.2 interactions between any two proteins. This comes very close to the distance that separates people in the social network Facebook: 4.5 connections - the modern version of the more famous six-degrees-of-separation *(*94*,* 95*)*. While a scale-free attribute for protein-protein interaction has been claimed in most studies, they sometimes rather appear to be exponential or truncated *(*23*)*. In this study I show a clear power-law distribution that helps to clarify the higher order structure of protein interaction networks and secondly serves as an indicator for high data completeness.

The resulting network map of yeast protein-protein interactions clustered nicely into structures that represent known complexes and at the same time uncovered many novel associations and assigned potential functions to yet uncharacterized proteins. I have highlighted only a few of several new discoveries in the paper while many more are depicted in the supplementary information. The Markov clustering algorithm employed here is based on a random walk simulation in which I used the developed score as edge weights. The structured outcome and the quite complete reflection of protein complexes in clusters without *priori* knowledge supports the quality of the experimental results and validates our scoring approach, since the score is used as an edge weight and therefore directly influences the clustering.

While we have exhaustively explorated the yeast interactome under standard growth conditions, our platform opens up possibilities for large-scale screens under stress conditions. The complexity of human cells and their reduced susceptibility to genetic modifications are reason for the still not completed interactome. The human OpenCell interactome presented here tackles those difficulties by using a split protein system. This strategy allows large-scale endogenous tagging of cells while the sensitivity or our mass spectrometry pipeline of cells allows us to grow them in only 12-well plates - enabling efficient processing of samples. OpenCell provides the largest confocal microscopy library to date and enables interaction exploration of about 1,300 endogenous tagged proteins both derived from the same cell population.

3.2 A Hairy Situation

If you were reading diligently through this work or have followed the link to the yeast interactome webpage, you might remember the figure that depicts all the interactions detected in this study within a yeast cell in a structured manner. Why do we not remember any similar representations of protein-protein interaction maps from previous studies? Are they any available? In my opinion, the simple explanation is that they do exist, but they are not very memorable. Almost all large-scale interaction screens resulted in a network structure that is so tightly interconnected and impenetrable that they are often referred to as “hairball”. A representation of a network that is a “hairball” is not really useful. One possible reason why this is not the case for the yeast interactome in this study appears to be the unique clustering, that is in turn enabled by an excellent score and underlying data. Another thing that improves the random walk of the Markov clustering is redundancy. Redundancy is the key to good interactomes and builds on many different pull-downs that confirm the same interactions. A complex of 4 members for instance can consist of up to 12 interactions (counting reverse experiments separately) and when including correlations even up to 18 interactions. This redundancy helps the random walk of the Markov clustering to find what truly belongs together. If one weights those edges during the random walk this effect becomes even stronger, but only if the weight corresponds to a proper increase in likelihood that an interaction is true. One can assume for instance that a high FDR (false discovery rate) interaction is more likely to reflect an important interaction than one with a very low one. This differentiation must be made and if the enrichment works well then this can be reflected in our score that is used as a weight.

It is possible that human interactomes by themselves have a “hairier” nature, due their increased complexity. Hoverer, the number of protein coding genes is “only” about three times higher and the human cell is organized into protein complexes as well. A potential reason for the lack of organization in human interactomes might be that they are yet incomplete. The above-mentioned redundancy can only be achieved when the coverage of all expressed proteins is reached. In this case the use of correlation analysis becomes much more powerful. Still, I believe that to some degree the limitations of existing interactomes are also caused by many false positives and generally by suboptimal data quality.

3.3 From the Past to …

At the beginning of this millennium the first AP-MS studies in yeast were conducted *(*27*,* 28*)*. Combined, a team of 84 scientists established the first larger dataset of protein-protein interactions, a milestone in proteomics and cell biology. This was soon to be followed by two greatly extended versions as a result of combined forces of a similarly large group of scientists around the world *(*29*,* 30*)*. This massive effort was certainly enough reason not to contemplate a next-generation version of an interactome in yeast and the focus for good reasons shifted to other organism. Still, it is surprising that a vision of redoing the yeast interactome on an improved platform has not been put forward for so long. It is not only about redoing interactomics in yeast to gain improved data, but also about doing it much more effortlessly, on a much smaller scale, and in a higher throughput to smoothen the path for future screens. A cell even as simple as the yeast is rarely growing in the perfect conditions that we call standard growth condition, it is rather subjected to environmental challenges, like depletion of nutrients, change of temperature, or damaging agents. Observing the changes on a global interaction level that occur as a response to such events will help to understand the mechanism behind it. Some of the reactions to the plans of this project that can be summarized in “but, hasn’t this been done already?” reflect the mindset that might impede progress in some areas. From my own experience I can tell that pull-downs are the daily bread and butter of a cell biologist. The constant search for interaction partners of a protein of interest – being it under stress or standard condition – in order to confirm a hypothesis or to explain a mechanism, is a daily routine for so many scientists. This work is tedious, redundant, error prone and very small-scale. Scientists ask themselves why this has not already been done for everyone else. If we include every protein and every condition, it might be a long way, but we have to start.

Luckily the Mann lab, in which all this began also started in recent years to explore new ways of how to conduct, minimize and analyze interaction experiments by using non-quantitative proteomics *(*43–45*)*.

3.4 The Future of Interactomics

For the near future for yeast interactomics I see further miniaturization and reduction in measurement time ahead. The material derived from my currently developed protocols is already sufficient for 3-4 injections. Given the already drastically increased sensitivity of the next generation of timsTOF Pro, this already implies a potential reduction of input material of a factor of at least four. This alone would in theory allow a switch of the library format from 96-well to 384-well plates, reducing the handling load from the current 44 to 11 plates. In order to accomplish this, one would need to address how growth, lysis and enrichment steps perform in the smaller format.

I also see potential in reducing the library complexity by using a “smart-selection” of baits, based on my interaction data. One can exchange some of the large redundancy in the data for a reduced library and would still get a very similar network information. This could potentially further halve the library resulting in about 6 plates. During the 21 min gradient runs we detected on average about 1,500 proteins per run. This large number is helpful for label-free normalization and quantification but is more than sufficient. Here it will be worth to explore the quality of the data acquired under even shorter gradients. Assuming a 12 min or even a 6 min gradient corresponding to 100 and 200 samples per day for 6 x 384-well plates this would allow the measurement of a complete interactome (excluding replicates) in 24 or 12 days respectively.

A promising development was recently made by putting proteomics on a microfluidic platform *(*96*)*. This on-chip AP-MS is still on its way to provide large-scale application compatibility but it requires very low input material and makes hope for drastic improvements in the field of interactomics in the mid and long-term future.

4 References

1. M. Newman, *Networks* (OUP Oxford, 2018; https://books.google.de/books?id=YdZjDwAAQBAJ).

2. D. Piovesan, M. Giollo, C. Ferrari, S. C. E. Tosatto, Protein function prediction using guilty by association from interaction networks. *Amino Acids*. 47, 2583–2592 (2015).

3. F. Crick, Central Dogma of Molecular Biology. *Nature*. 227, 561–563 (1970).

4. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, I. H. G. S.

Consortium, Initial sequencing and analysis of the human genome. *Nature*. 409, 860–921 (2001).

5. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The Sequence of the Human Genome. *Science*. 291, 1304–1351 (2001).

6. I. Bludau, R. Aebersold, Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat Rev Mol Cell Bio*, 1–14 (2020).

7. C. L. Meyerkord, H. Fu, *Protein-Protein Interactions: Methods and Applications* (Springer New York, 2015; https://books.google.de/books?id=4wdVvgAACAAJ), *Methods in Molecular Biology*.

8. K. Iyer, L. Bürkle, D. Auerbach, S. Thaminy, M. Dinkel, K. Engels, I. Stagljar, Utilizing the Split-Ubiquitin Membrane Yeast Two-Hybrid System to Identify Protein-Protein Interactions of Integral Membrane Proteins. *Sci Signal*. 2005, pl3–pl3 (2005).

9. H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, M. Vidal, High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Sci New York N Y*. 322, 104–10 (2008).

10. X. Liu, W. Yang, Q. Gao, F. Regnier, Toward chromatographic analysis of interacting protein networks. *J Chromatogr A*. 1178, 24–32 (2008).

11. M. A. Skinnider, N. E. Scott, A. Prudova, C. H. Kerr, N. Stoynov, R. G. Stacey, Q. W. T. Chan, D. Rattray, J. Gsponer, L. J. Foster, An atlas of protein-protein interactions across mouse tissues. *Cell*. 184, 4073-4089.e17 (2021).

12. F. Liu, P. Lössl, R. Scheltema, R. Viner, A. J. R. Heck, Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat Commun*. 8, 15473 (2017).

13. J. D. Martell, T. J. Deerinck, Y. Sancak, T. L. Poulos, V. K. Mootha, G. E. Sosinsky, M. H. Ellisman, A. Y. Ting, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat Biotechnol*. 30, 1143–1148 (2012).

14. K. J. Roux, D. I. Kim, M. Raida, B. Burke, A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biology*. 196, 801–810 (2012).

15. T. C. Branon, J. A. Bosch, A. D. Sanchez, N. D. Udeshi, T. Svinkina, S. A. Carr, J. L. Feldman, N. Perrimon, A. Y. Ting, Efficient proximity labeling in living cells and organisms with TurboID. *Nat Biotechnol*. 36, 880–887 (2018).

16. A.-C. Gingras, K. T. Abe, B. Raught, Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Curr Opin Chem Biol*. 48, 44–54 (2019).

17. J. H. Morris, G. M. Knudsen, E. Verschueren, J. R. Johnson, P. Cimermancic, A. L. Greninger, A. R. Pico, Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nat Protoc*. 9, 2539–2554 (2014).

18. E. L. Rodriguez, S. Poddar, S. Iftekhar, K. Suh, A. G. Woolfork, S. Ovbude, A. Pekarek, M. Walters, S. Lott, D. S. Hage, Affinity chromatography: A review of trends and developments over the past 50 years. *J Chromatogr B*. 1157, 122332 (2020).

19. W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O’Shea, Global analysis of protein localization in budding yeast. *Nature*. 425, 686–691 (2003).

20. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc National Acad Sci*. 113, E3501–E3508 (2016).

21. A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks. *Science*. 286, 509–512 (1999).

22. R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks. *Nature*. 406, 378–382 (2000).

23. G. Lima-Mendez, J. van Helden, The powerful law of the power law and other myths in network biology. *Mol Biosyst*. 5, 1482–93 (2009).

24. S. Larochelle, Putting the pieces together. *Nat Methods*. 12, 21–21 (2015).

25. J. Fenn, M. Mann, C. Meng, S. Wong, C. Whitehouse, Electrospray ionization for mass spectrometry of large biomolecules. *Science*. 246, 64–71 (1989).

26. N. Methods, Nature, N. Biotechnology, N. C. B. and N. Protocols, Nature Milestones - Mass Spectrometry. *Nature Methods* (2015).

27. A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415, 141–147 (2002).

28. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, M. Tyers, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. 415, 180–183 (2002).

29. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, J. F. Greenblatt, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. 440, 637–643 (2006).

30. A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 440, 631–636 (2006).

31. G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Séraphin, A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*. 17, 1030–1032 (1999).

32. N. J. Krogan, M. Kim, S. H. Ahn, G. Zhong, M. S. Kobor, G. Cagney, A. Emili, A. Shilatifard, S. Buratowski, J. F. Greenblatt, RNA Polymerase II Elongation Factors of Saccharomyces cerevisiae: a Targeted Proteomics Approach. *Mol Cell Biol*. 22, 6979–6992 (2002).

33. J. Goll, P. Uetz, The elusive yeast interactome. *Genome Biol*. 7, 223 (2006).

34. D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardiu, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras, A. I. Nesvizhskii, The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. *Nat Methods*. 10, 730–736 (2013).

35. S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, N. J. Krogan, Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*. 6, 439–450 (2007).

36. B. Ho, A. Baryshnikova, G. W. Brown, Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst*. 6, 192-205.e3 (2018).

37. E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wühr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. D. Camilli, J. A. Paulo, J. W. Harper, S. P. Gygi, The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. 162, 425–40 (2015).

38. E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, J. W. Harper, Architecture of the human interactome defines protein communities and disease networks. *Nature*. 545, 505–509 (2017).

39. M. E. Sowa, E. J. Bennett, S. P. Gygi, J. W. Harper, Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell*. 138, 389–403 (2009).

40. J. A. Ranish, E. C. Yi, D. M. Leslie, S. O. Purvine, D. R. Goodlett, J. Eng, R. Aebersold, The study of macromolecular complexes by quantitative proteomics. *Nat Genet*. 33, 349–355 (2003).

41. W. X. Schulze, M. Mann, A Novel Proteomic Screen for Peptide-Protein Interactions. *J Biol Chem*. 279, 10756–10764 (2004).

42. J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol Cell Proteom Mcp*. 13, 2513–26 (2014).

43. E. C. Keilhauer, M. Y. Hein, M. Mann, Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Mol Cell Proteom Mcp*. 14, 120–35 (2015).

44. M. Y. Hein, N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman, M. Mann, A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*. 163, 712–723 (2015).

45. F. Hosp, R. A. Scheltema, H. C. Eberl, N. A. Kulak, E. C. Keilhauer, K. Mayr, M. Mann, A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. *Mol Cell Proteomics*. 14, 2030–2041 (2015).

46. W. Wien, Über positive Elektronen und die Existenz hoher Atomgewichte. *Ann Phys-berlin*. 318, 669–677 (1904).

47. J. J. Thomson, Further experiments on positive rays. *Lond Edinb Dublin Philosophical Mag J Sci*. 24, 209–253 (1912).

48. J. J. Thomson, Bakerian Lecture :—Rays of positive electricity. *Proc Royal Soc Lond Ser Contain Pap Math Phys Character*. 89, 1–20 (1913).

49. J. J. Thomson, Rays of Positive Electricity and their Application to Chemical Analysis. *Nature*. 92, 549–550 (1914).

50. J. H. Gross, *Mass Spectrometry* (Springer, 2017).

51. J. J. Thomson, A New Method of Chemical Analysis. *Sci Am*. 73, 41–42 (1912).

52. F. W. Aston, Neon. *Nature*. 104, 334–334 (1919).

53. F. W. Aston, The Constitution of the Elements. *Nature*. 105, 547–547 (1920).

54. F. W. Aston, XI. The mass spectra of chemical elements . (Part 3.). *Lond Edinb Dublin Philosophical Mag J Sci*. 42, 140–144 (1921).

55. F. W. ASTON, The Isotopes of Germanium. *Nature*. 111, 771–771 (1923).

56. F. W. ASTON, The Isotopes of Tin. *Nature*. 109, 813–813 (1922).

57. F. W. Aston, The Mass-spectrum of Iron. *Nature*. 110, 312–313 (1922).

58. F. W. ASTON, Isotopes and Atomic Weights. *Nature*. 105, 617–619 (1920).

59. N. M. A. 2020, The traps of Paul and Dehmelt (2020), (available at https://www.nobelprize.org/prizes/physics/1989/9747-the-traps-of-paul-and-dehmelt/).

60. N. M. A. 2020, The Nobel Prize in Chemistry 2002, (available at https://www.nobelprize.org/prizes/chemistry/2002/summary/).

61. M. S. Wilm, M. Mann, Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int J Mass Spectrom*. 136, 167–180 (1994).

62. M. Mann, M. Wilm, Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci*. 20, 219–224 (1995).

63. M. Wilm, M. Mann, Analytical Properties of the Nanoelectrospray Ion Source. *Anal Chem*. 68, 1–8 (1996).

64. K. R. Jennings, *A History of European Mass Spectrometry* (IM Publications, 2012).

65. D. B. Bekker-Jensen, A. Martínez-Val, S. Steigerwald, P. Rüther, K. L. Fort, T. N. Arrey, A. Harder, A. Makarov, J. V. Olsen, A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol Cell Proteomics*. 19, 716–729 (2020).

66. V. Marx, A dream of single-cell proteomics. *Nat Methods*. 16, 809–812 (2019).

67. N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods*. 11, 319–24 (2014).

68. J. B. Müller, P. E. Geyer, A. R. Colaço, P. V. Treit, M. T. Strauss, M. Oroshi, S. Doll, S. V. Winter, J. M. Bader, N. Köhler, F. Theis, A. Santos, M. Mann, The proteome landscape of the kingdoms of life. *Nature*. 582, 592–596 (2020).

69. N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Müller, F. Meier, J. V. Olsen, O. Vorm, M. Mann, A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol Cell Proteom Mcp*. 17, 2284–2296 (2018).

70. F. Meier, P. E. Geyer, S. V. Winter, J. Cox, M. Mann, BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*. 15, 440–448 (2018).

71. C. Wichmann, F. Meier, S. V. Winter, A.-D. Brunner, J. Cox, M. Mann, MaxQuant.Live enables global targeting of more than 25,000 peptides. *Mol Cell Proteom Mcp*. 18, 982–994 (2019).

72. F. Meier, A.-D. Brunner, M. Frank, A. Ha, I. Bludau, E. Voytik, S. Kaspar-Schoenefeld, M. Lubeck, O. Raether, R. Aebersold, B. C. Collins, H. L. Röst, M. Mann, Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): Bottom-up proteomics with near optimal ion usage. *Biorxiv*, 656207 (2020).

73. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods*. 17, 41–44 (2020).

74. F. Meier, A.-D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, N. Bache, O. Hoerning, J. Cox, O. Räther, M. Mann, Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics*. 17, 2534–2545 (2018).

75. M. Y. Hein, K. Sharma, J. Cox, M. Mann, *Handbook of Systems Biology* (Elsevier Inc., 2013), *Section I: Components of Biological Systems*.

76. J. R. Wiśniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis. *Nat Methods*. 6, 359–362 (2009).

77. F. Coscia, S. Doll, J. M. Bech, L. Schweizer, A. Mund, E. Lengyel, J. Lindebjerg, G. I. Madsen, J. M. Moreira, M. Mann, A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *J Pathology*. 251, 100–112 (2020).

78. M. L. Nielsen, M. Vermeulen, T. Bonaldi, J. Cox, L. Moroder, M. Mann, Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry. *Nat Methods*. 5, 459–60 (2008).

79. L. Tsiatsiani, A. J. R. Heck, Proteomics beyond trypsin. *Febs J*. 282, 2612–2626 (2015).

80. J. Rappsilber, M. Mann, Y. Ishihama, Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*. 2, 1896–1906 (2007).

81. J. C. Rogers, R. D. Bomgarden, Modern Proteomics – Sample Preparation, Analysis and Practical Applications. *Adv Exp Med Biol*, 43–62 (2016).

82. M. Gilar, P. Olivova, A. E. Daly, J. C. Gebler, Orthogonality of Separation in Two-Dimensional Liquid Chromatography. *Anal Chem*. 77, 6426–6434 (2005).

83. A. J. Alpert, Modern Proteomics – Sample Preparation, Analysis and Practical Applications. *Adv Exp Med Biol*, 23–41 (2016).

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

85. A. M. Haag, Modern Proteomics – Sample Preparation, Analysis and Practical Applications. *Adv Exp Med Biol*. 919, 157–169 (2016).

86. A. Makarov, Orbitrap journey: taming the ion rings. *Nat Commun*. 10, 3743 (2019).

87. E. S. Hecht, M. Scigelova, S. Eliuk, A. Makarov, Encyclopedia of Analytical Chemistry, 1–40 (2006).

88. F. Fernandez-Lima, D. A. Kaplan, J. Suetering, M. A. Park, Gas-phase separation using a trapped ion mobility spectrometer. *Int J Ion Mobil Spectrom Official Publ Int Soc Ion Mobil Spectrom*. 14, 93–98 (2011).

89. J. A. Silveira, M. E. Ridgeway, F. H. Laukien, M. Mann, M. A. Park, Parallel accumulation for 100% duty cycle trapped ion mobility-mass spectrometry. *Int J Mass Spectrom*. 413, 168–175 (2017).

90. F. Meier, S. Beck, N. Grassl, M. Lubeck, M. A. Park, O. Raether, M. Mann, Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J Proteome Res*. 14, 5378–5387 (2015).

91. N. Prianichnikov, H. Koch, S. Koch, M. Lubeck, R. Heilig, S. Brehmer, R. Fischer, J. Cox, *Mol Cell Proteom Mcp*, in press, doi:10.1074/mcp.tir119.001720.

92. R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, M. Tyers, The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 47, gky1079- (2018).

93. D. W. Burden, *Guide to the Disruption of Biological Samples* (Random Primers, 2012), vol. 12.

94. S. Edunov, S. Bhagat, M. Burke, C. Diuk, I. O. Filiz, Three and a half degrees of separation - Facebook Research, (available at https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/).

95. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, Four Degrees of Separation. *Arxiv* (2011).

96. C. Furlan, R. A. M. Dirks, P. C. Thomas, R. C. Jones, J. Wang, M. Lynch, H. Marks, M. Vermeulen, Miniaturised interaction proteomics on a microfluidic platform with ultra-low input requirements. *Nat Commun*. 10, 1525 (2019).

5 Acknowledgments

First and foremost, I want to express my deepest gratitude to Matthias Mann for your help and decision that enabled my start in this fantastic department and for your guidance, patience, and support throughout my projects. Providing this environment of excellent science and people is what makes being part of this group genuinely amazing.

I also would like to express my great appreciation to Stefan Jentsch for his trust in me. Remembered as a great person and scientist like no other.

Thanks to Brenda Schulman for your support at the right time and for being part of my committee.

Thanks for welcoming me into the office, withstanding all my questions, and showing me everything I needed to know in the world of proteomics: Oezge Karayel and Martin Steger.

To my other mates in the “dark office”: thank you for an always fun and balanced atmosphere: Fynn Hansen, Jakob Bader, and Matteo Pecoraro.

Daniel Hornburg, thank you for being such an open and good-minded friend and for introducing me into the world of Max P.

For answering all my endless questions on this tall guy, everyone keeps calling TIM without losing patience: thanks to Andreas Brunner and Florian Meier.

Thanks to McBugFix Max - veggie döner without - Zwiebel, please! And the clinical office.

Thanks to my interactomics predecessors Fabian Hosp and Marco Hein for discussions and sharing your experiences.

To the constants in the department: Igor Paron, Alison Dalfovo, Medini Steger, Mario Oroshi, Theresa Schneider, and Helen Delos Santos, thanks for keeping this group running.

To everyone in the Mann department for being so great and helpful!

For an amazing time in the MCB department, thanks to everyone! Especially, Markus Höpfler, Irina Heckmann, Fabian den Brave, Claudio Lademann, Benjamin Anstett, Florian Paasch and Alex Strasser.

Thanks to my mom, dad, and sister for your unconditional support and faith.

-

To my dear wife, Ellen.