Inferring protein from transcript abundances using convolutional neural networks

Dissertation von Patrick Maximilian Schwehn

München 2025

Inferring protein from transcript abundances using convolutional neural networks

Dissertation der Fakultät für Biologie
der Ludwig-Maximilian-Universität München

Patrick Maximilian Schwehn
München, 2025

Diese Dissertation wurde angefertigt
unter der Leitung von Prof. Dr. Pascal Falter-Braun
am Institut für Netzwerkbiologie
des Helmholtz Zentrums München

Erstgutachter:	Prof. Dr. Pascal Falter-Braun
Zweitgutachter:	Prof. Dr. Korbinian Schneeberger
Tag der Abgabe:	27.06.2025
Tag der mündlichen Prüfung:	23.02.2026

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass meine Dissertation selbständig und ohne unerlaubte Hilfsmittel angefertigt worden ist. Die vorliegende Dissertation wurde weder ganz noch teilweise bei einer anderen Prüfungskommission vorgelegt. Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den 27.06.2025

Patrick Maximilian Schwehn

Abstract

Progress in personalized medicine and crop optimization hinges on quantitative, systems-level insights into how genotype and environment shape cellular phenotypes. Because proteins and their interactions underlie nearly all cellular functions, fluctuations in their concentrations strongly influence phenotypic outcomes. Yet protein measurements remain far more expensive than high-throughput mRNA assays, which are therefore often used as surrogates even though transcript and protein levels correlate only imperfectly.

Recent advances in artificial intelligence offer a promising approach to reduce this inaccuracy. I therefore developed species-specific convolutional neural networks (CNNs) for *Homo sapiens* and *Arabidopsis thaliana* that predict protein abundances directly from paired transcript abundances and raw sequence data. Trained on matched transcriptome-proteome datasets, the models achieve coefficients of determination of 0.30 and 0.32, respectively. They improve accuracy for *H. sapiens* by about 40% over the best sequence-based approach and provide the first published model for *A. thaliana*. Analysis of the learned filter weights shows that the networks autonomously rediscover known regulatory motifs governing mRNA decay and translation while identifying several novel elements that merit experimental validation. An extended architecture that incorporates expression profiles of putative interaction partners does not yet surpass the sequence-only model, underscoring the need for larger, condition-rich training sets.

I also developed an automated image-analysis pipeline for high-throughput yeast two-hybrid systems. Classical computer-vision algorithms locate 96-well plate grids, and a second CNN scores yeast colony growth, generating quantitative protein-protein interaction networks from thousands of colonies. Integrating these curated interaction data with graph-convolutional modules alongside sequence features processed by conventional convolutions offers a promising route to further improve the accuracy of protein abundance prediction.

Collectively, these contributions demonstrate how CNNs can both extract and predict reliable biological information from raw experimental data, advancing systems-level modeling of gene expression and enabling scalable applications in personalized medicine and crop improvement.

Kurzzusammenfassung

Fortschritte in der personalisierten Medizin und der Optimierung von Nutzpflanzen erfordern quantitative, systemische Einblicke, wie Genotyp und Umwelt zelluläre Phänotypen prägen. Da Proteine und ihre Interaktionen nahezu alle zellulären Funktionen bestimmen, führen Schwankungen ihrer Konzentrationen zu erheblichen Veränderungen der Phänotypen. Proteinmessungen sind jedoch deutlich kostspieliger als Hochdurchsatz-mRNA-Analysen, die daher häufig als Annäherung dienen, obwohl Transkript- und Proteinkonzentrationen nur teilweise korrelieren.

Aktuelle Entwicklungen in der künstlichen Intelligenz bieten einen vielversprechenden Ansatz, diese Ungenauigkeit zu verringern. Daher habe ich für *Homo sapiens* und *Arabidopsis thaliana* speziesspezifische Convolutional Neural Networks (CNNs) entwickelt, die Protein- aus den zugehörigen Transkriptkonzentrationen und Rohsequenzdaten vorhersagen. Die auf gepaarten Transkriptom-Proteom-Datensätzen trainierten Modelle erreichen Bestimmtheitsmaße von 0,30 beziehungsweise 0,32. Damit steigern sie die Vorhersagegenauigkeit für *H. sapiens* um etwa 40 % gegenüber dem bislang besten sequenzbasierten Ansatz und stellen zugleich das erste veröffentlichte Modell für *A. thaliana* dar. Eine Analyse der gelernten Parameter zeigt, dass die Modelle eigenständig bekannte regulatorische Motive wiederentdecken, die den mRNA-Abbau und die Translation steuern, und mehrere bislang unbekannte Elemente identifizieren, die als Grundlage experimenteller Untersuchungen dienen können. Eine erweiterte Architektur, die Expressionsprofile potenzieller Interaktionspartner berücksichtigt, übertrifft das rein sequenzbasierte Modell bislang noch nicht und unterstreicht den Bedarf an größeren, konditionsreichen Trainingsdatensätzen.

Darüber hinaus habe ich eine automatisierte Bildanalyse-Pipeline für Hochdurchsatz-Yeast-Two-Hybrid-Systeme entwickelt. Klassische Algorithmen lokalisieren das Raster der 96-Well-Platten und ein weiteres CNN bewertet das Wachstum der Hefekolonien, sodass aus Tausenden von Kolonien quantitative Protein-Protein-Interaktionsnetzwerke entstehen. Die Integration dieser kuratierten Interaktionsdaten mit Graph-Convolutional-Modulen und Sequenzmerkmalen, die durch konventionelle Convolutions verarbeitet werden, bietet einen vielversprechenden Ansatz, die Genauigkeit der Proteinkonzentrationsvorhersage weiter zu steigern.

Insgesamt zeigen diese Arbeiten, wie CNNs die Extraktion und Vorhersage zuverlässiger biologischer Informationen aus experimentellen Rohdaten ermöglichen. Dies fördert die systemische Modellierung der Genexpression und unterstützt das Hochskalieren von Anwendungen in der personalisierten Medizin und der Nutzpflanzenoptimierung.

Content

Abbreviations

cDNA ..... complementary DNA

CDS ..... coding sequence

CNN ..... convolutional neural network

DNA ..... deoxyribonucleic acid

DT ..... decision tree

EF ..... elongation factor

GCN ..... graph convolutional network

GWAS ..... genome-wide association study

IF ..... initiation factor

miRNA ..... microRNA

ML ..... machine learning

mRNA ..... messenger RNA

MS ..... mass spectrometry

poly(A) ..... polyadenylic acid

PPI ..... protein-protein interaction

PTR ..... protein-to-mRNA ratio

r ..... Pearson correlation coefficient

r^2 ..... coefficient of determination

RBP ..... RNA-binding protein

ReLU ..... rectified linear unit

RF ..... release factor

RISC ..... RNA-induced silencing complex

RNA ..... ribonucleic acid

RNA-Seq ..... RNA sequencing

rRNA ..... ribosomal RNA

siRNA ..... small interfering RNA

SNP.....single-nucleotide polymorphism

SVM .....support vector machine

tRNA ..... transfer RNA

UTR.....untranslated region

Y2H ..... yeast two-hybrid

List of Publications

1. Schwehn, P.M. and P. Falter-Braun, *Inferring protein from transcript abundances using convolutional neural networks.* BioData Min, 2025. **18**(1): p. 18.

2. Kim, D.K., et al., *A proteome-scale map of the SARS-CoV-2-human contactome.* Nat Biotechnol, 2023. **41**(1): p. 140-149.

3. Young, V., et al., *A gut meta-interactome map reveals modulation of human immunity by microbiome effectors.* bioRxiv, 2023: p. 2023.09.25.559292.

4. Kuhl-Nagel, T., et al., *Novel Pseudomonas sp. SCA7 Promotes Plant Growth in Two Plant Families and Induces Systemic Resistance in Arabidopsis thaliana.* Front Microbiol, 2022. **13**: p. 923515.

5. Sirazitdinova, E., et al., *Sewer Discharge Estimation by Stereoscopic Imaging and Synchronized Frame Processing.* Computer-Aided Civil and Infrastructure Engineering, 2018. **33**(7): p. 602-613.

Declaration of contribution as a co-author

Inferring protein from transcript abundances using convolutional neural networks

Contribution of co-authors: Pascal Falter-Braun critically edited the manuscript and approved its final version.

A proteome-scale map of the SARS-CoV-2-human contactome

My contribution: I evaluated the image data generated in our laboratory, generated and corrected the final interaction datasets, and performed network analyses.

A gut meta-interactome map reveals modulation of human immunity by microbiome effectors

My contribution: I evaluated the image data generated in our laboratory, generated and corrected the final interaction datasets, and performed network analyses.

München, den 27.06.2025

Patrick Maximilian Schwehn

München, den 27.06.2025

Prof. Dr. Pascal Falter-Braun

1. Introduction

Medical and agricultural advancements in recent decades have led to significant increases in life expectancy (The World Bank, 2024) and reductions in global hunger (Deutsche Welthungerhilfe et al., 2024). However, recent data indicate that these trends, which had progressed steadily for decades, have now plateaued. To sustain progress despite growing technical challenges, targeted approaches such as personalized medicine (Hamburg & Collins, 2010; Nicholson, 2006) and bioengineered, locally optimized crops are becoming even more essential.

To achieve these goals, the broader field of systems biology seeks to develop a comprehensive understanding of complex cellular processes (Ideker et al., 2001; Kitano, 2002b), with the long-term objective of fully deciphering and simulating them (Kitano, 2002a). This endeavor requires identifying and analyzing all cellular components, including deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, lipids, carbohydrates, metabolites, and micronutrients, and mapping their interactions (Joyce & Palsson, 2006). A dataset of this scope is necessary to describe a cell's complete initial state, which forms the basis for *in silico* predictions.

In addition to internal cellular factors, external environmental influences play a crucial role in shaping cellular behavior. Cells do not exist in isolation and are continually exposed to a broad spectrum of external stimuli, including physical conditions such as light (Terzaghi & Cashmore, 1995), temperature (Somero, 2020), and mechanical stress (Janmey & McCulloch, 2007), as well as biological interactions with other cells (Armingol et al., 2021), bacteria (Lebeer et al., 2010), and the extracellular matrix (Frantz et al., 2010). Accurately simulating a living cell under physiologically relevant conditions therefore requires models that integrate dynamic environmental inputs with internal molecular interaction networks.

Although significant progress has been made in measuring various *omics* layers, comprehensive datasets that simultaneously capture all *omics* modalities alongside physical conditions and biological interactions have yet to be published (Misra et al., 2018). Because most single-cell measurement techniques are destructive, generating fully integrated datasets depends on major advances in scalable, nondestructive live-cell imaging methods such as fluorescent probes (Zhang et al., 2002) and video mass spectrometry (MS) (Mizuno et al., 2008). In the meantime, many large-scale research initiatives focus on statistically integrating the individual contributions of specific *omics* layers to particular phenotypes and genetic disease mechanisms (Karczewski & Snyder, 2018).

A substantial proportion of genetic diseases is associated with altered gene expression levels rather than with complete loss of function or structural disruption of proteins (Cookson et al.,

2009). Such changes can arise from genetic variation in both the gene itself and the regulatory elements that influence its expression (Section 1.1), as well as from environmental factors. Since the advent of next-generation DNA sequencing, research has increasingly focused on elucidating the roles of genetic variants across the entire gene expression pathway, including transcription, translation, and post-translational modification (Ansorge, 2009). These advances have also introduced major data science challenges, as the vast amounts of data must be processed and interpreted effectively.

Genome-wide association studies (GWAS) address this complexity by identifying genetic variants, such as single-nucleotide polymorphisms (SNPs), nucleotide insertions or deletions, and nucleotide repeat expansions, that are significantly associated with specific phenotypes (Wang et al., 2005). By analyzing genomic data across large cohorts alongside detailed phenotypic measurements, GWAS narrow the genomic search space to a limited set of candidate loci, allowing researchers to focus experimental validation on a manageable number of genomic regions.

While GWAS have identified numerous genetic variants associated with specific phenotypes, many appear to have little or no observable effect (Manolio, 2010). In many cases, variants reside in noncoding regions of the genome or alter protein amino acid sequences in ways that do not significantly affect protein function. Nevertheless, some variants, even those involving subtle nucleotide changes, can have significant consequences by altering a protein's structure, function, or expression level (Wang & Moult, 2001). Because proteins are key effectors of most biological activities, such disruptions can propagate through cellular networks and lead to complex multifactorial diseases, including Huntington's disease (Bates et al., 2015), sickle cell disease (Rees et al., 2010), adenosine deaminase deficiency (Flinn & Gennery, 2018), and cystic fibrosis (O'Sullivan & Freedman, 2009).

In many drug-discovery efforts, characterizing protein expression levels is crucial because they provide insight into potential protein-protein interactions (PPIs) (Phizicky & Fields, 1995) and their functional significance (Scott et al., 2016). Understanding how nucleotide changes and alterations in regulatory factor expression affect protein expression is therefore essential for deciphering disease mechanisms and developing effective treatments. Quantifying expression levels is thus a prerequisite for linking genetic variants to gene expression.

Protein expression levels in cells are most commonly measured via MS-based analysis (Aebersold & Mann, 2003). Unfortunately, high-throughput MS is resource-intensive, requiring costly instrumentation and extensive sample preparation (Section 1.2). By contrast, RNA expression levels can be quantified more readily with RNA sequencing (RNA-Seq) (Wang et al.,

2009). Consequently, many large-scale gene expression datasets report messenger RNA (mRNA) levels instead of protein levels (Brawand et al., 2011; Morley et al., 2004; Oleksiak et al., 2002).

However, the relationship between mRNA and protein abundance is not uniform across genes (Mergner et al., 2020; Wang et al., 2019) and depends on factors such as translational efficiency (Gingold & Pilpel, 2011), mRNA and protein degradation rates (Hargrove & Schmidt, 1989), and post-translational modifications (Mann & Jensen, 2003). Using mRNA abundance as a proxy for protein abundances can therefore introduce significant inaccuracies, for example, in PPI applications, where protein levels directly influence functional outcomes (Rao et al., 2014). *In silico* models that estimate protein from mRNA abundance thus offer a promising strategy to improve the accuracy and interpretability of analyses that currently rely on mRNA-based approximations.

As noted earlier, SNPs associated with diseases illustrate how subtle changes in the initial genetic sequence can lead to substantial alterations in protein expression levels. A similar principle, where small changes in initial conditions produce disproportionately large effects, applies to nonlinear dynamic systems (Hilborn, 2000), often described as chaotic. In physics and engineering, simple examples include the double pendulum (Shinbrot et al., 1992) and the three-body problem (Valtonen & Karttunen, 2009), while more complex examples include phenomena such as weather dynamics (Tsonis & Elsner, 1989). Although such systems follow deterministic laws, their extreme sensitivity to initial conditions and the practical limits on measuring those conditions sharply constrain long-term predictive power.

Because genetic variation can give rise to deterministically unpredictable systems, heuristic methods, particularly machine learning (ML) approaches (Section 1.3), are promising candidates for modeling, predicting, and exploring gene expression regulation (Larranaga et al., 2006). Each gene is subject to unique regulatory mechanisms that influence its protein-to-mRNA ratio (PTR). These factors are partly encoded in the nucleotide and amino acid sequences and partly reflected in the interaction profiles of the corresponding mRNA and protein. By analyzing protein abundance alongside sequence information and the expression levels of potential interaction partners, ML can infer gene-specific patterns that govern the PTR.

Despite the proliferation of ML models, most studies report little more than performance metrics and continue to treat the trained models as black boxes. Consequently, the information encoded in their learned parameters remains largely unexplored. Yet ML models do not need to remain opaque but can be examined in detail (Xu et al., 2019) to verify whether known biological patterns have been captured and to uncover novel insights that can guide future experiments. Developing

and interpreting ML models for predicting protein abundance from transcript abundance therefore requires a thorough understanding of gene expression, its regulatory mechanisms, and mRNA and protein degradation (Section 1.1), as well as familiarity with common ML methods (Section 1.3).

1.1. Gene expression, regulation, and degradation

Gene expression encompasses the entire set of processes that produce functional proteins from DNA's genetic code. These processes include transcription of DNA into RNA, RNA splicing into mRNA, mRNA capping, mRNA polyadenylation, export of mRNA from the nucleus to the cytoplasm, translation into protein, and post-translational protein modification (Berg et al., 2013). Because each step depends on the availability and activity of specific proteins and RNAs, gene expression is continuously shaped by the expression of its own regulators. The following concise overview of relevant molecular partners and their mechanisms illustrates why fully deterministic modeling of gene expression and its regulation quickly becomes intractable.

During transcription, RNA polymerases bind to a gene's promoter region, unwind the double-stranded DNA, and synthesize a complementary single-stranded pre-mRNA. Capping enzymes add a guanine cap to the 5′ end of the pre-mRNA, whereas polyadenylic acid (poly(A)) polymerase adds a poly(A) tail to the 3′ end. Both structures serve as binding sites for proteins that govern export and degradation. The pre-mRNA contains exons interspersed with noncoding introns, from which the spliceosome excises the introns and ligates the exons, producing mature mRNA that includes untranslated regions (UTRs) and protein-coding sequences (CDSs). Some introns are further processed into microRNAs (miRNAs) or small interfering RNAs (siRNAs) with additional cellular functions. Exportins then transport the mature mRNA through nuclear pore complexes into the cytoplasm, where ribosomal binding and translation occur.

Translation comprises initiation, elongation, and termination. During initiation, complexes of ribosomal RNA (rRNA) and proteins, together with several initiation factors (IFs) (Pestova et al., 2001), bind to the start codon of the mRNA. An initiator transfer RNA (tRNA) delivers the first amino acid to this codon in conjunction with an IF. During elongation, the ribosome advances codon by codon along the mRNA. Elongation factors (EFs) deliver cognate tRNAs to the ribosome, each carrying the amino acid specified by the current codon (Voorhees & Ramakrishnan, 2013). This cycle repeats until the ribosome encounters a stop codon for which no corresponding tRNA exists. Instead, release factors (RFs) bind the stop codon and trigger the release of both the mRNA and the nascent polypeptide (Nakamura et al., 1996). Finally, the

ribosome dissociates into its subunits, which can be reused in subsequent rounds of translation. Multiple ribosomes often translate the same mRNA simultaneously, thereby increasing protein output.

Translational regulation operates at both gene-specific and global levels (Merchante et al., 2017). Gene-specific mechanisms include upstream alternative start codons (Barbosa et al., 2013) or complex 5′ UTR structures (Hinnebusch et al., 2016) that inhibit initiation. Optimized synonymous codon usage, which leaves the amino acid sequence unchanged, modulates elongation efficiency according to tRNA availability (Hanson & Coller, 2018). At the global scale, stress-induced hyperphosphorylation of eukaryotic IF 2α blocks delivery of the initiator tRNA in plants (Wek, 2018). Likewise, heat shock protein 70 undergoes conformational changes at elevated temperatures, reducing its affinity for several eukaryotic EFs and thereby stalling elongation in mice and humans (Shalgi et al., 2013). Conversely, Jumonji domain containing 4 hydroxylates eukaryotic RF 1 in mammals, accelerating termination (Feng et al., 2014).

Gene- and context-specific regulation is also mediated by sequence-specific RNA-binding proteins (RBPs). In plants, for example, far-red light activates the photoreceptor protein PENTA1, which binds protochlorophyllide reductase mRNA, prevents ribosome association, and thus represses translation (Paik et al., 2012). In addition, sequence-specific miRNAs induce mRNA silencing by associating with the RNA-induced silencing complex (RISC) (Bartel, 2004). The miRNA binds an Argonaute protein within RISC, guiding the complex to a complementary mRNA sequence. Once bound, RISC can block ribosome progression or initiate endonucleolytic cleavage of the target, leading to mRNA degradation.

In mammals, more than 1,500 of roughly 20,000 protein-coding genes encode RBPs, underscoring their central role in post-transcriptional gene expression regulation (Turner & Diaz-Munoz, 2018). Although some RBPs function directly in fundamental processes such as signaling, cytoskeletal organization, or enzymatic activity, most act primarily as regulators of gene expression. RBPs assemble into large protein complexes such as RNA polymerase, capping enzyme, and poly(A) polymerase, and they participate in RNA-containing ribonucleoprotein complexes including the spliceosome, ribosome, ribonuclease, and RISC. RBP activity is further modulated by noncoding RNAs (ncRNAs) and PPIs, adding yet another regulatory layer (Ramanathan et al., 2019).

Beyond transcription and translation rates, the PTR also depends on mRNA and protein degradation. mRNA decay is initiated either by poly(A)-specific ribonuclease, which removes the 3′ poly(A) tail (Gao et al., 2000), or by decapping complexes that excise the 5′ guanine cap (Mugridge et al., 2016). Both mechanisms expose the transcript to additional exonucleases,

including the exosome, promoting 5′ to 3′ or 3′ to 5′ degradation. Additional regulation is provided by poly(A)-binding proteins, which shield the poly(A) tail from deadenylases (Kuhn & Wahle, 2004), and by eukaryotic IF 4E, which binds the 5′ guanine cap and hinders decapping enzymes (Mugridge et al., 2016).

A major pathway for protein degradation is ubiquitination (Pickart & Eddins, 2004). During ubiquitination, ubiquitin is repeatedly conjugated to substrate proteins, marking them for destruction by the 26S proteasome (Hershko & Ciechanover, 1992). The cascade involves three enzyme classes. First, a ubiquitin-activating enzyme (E1 enzyme) hydrolyzes adenosine triphosphate and forms a bond with ubiquitin. Second, a ubiquitin-conjugating enzyme (E2 enzyme) accepts ubiquitin from the E1 enzyme and interacts with a ubiquitin ligase (E3 ligase). Third, the E3 ligase transfers ubiquitin to a lysine residue on the target protein. The 26S proteasome subsequently removes the ubiquitin chain via deubiquitinating enzymes and unfolds and degrades the substrate.

Many E3 ligases possess modular domains that confer substrate specificity (Komander & Rape, 2012). For example, within the Skp, Cullin, F-box containing complex (Deshaies, 1999), the F-box subunit recognizes specific sequence motifs on the target protein. Another large E3 ligase, the anaphase-promoting complex (Page & Hieter, 1999), detects short linear motifs rather than folded domains. Overall, the *Homo sapiens* (*H. sapiens)* genome encodes at least eight E1 enzymes (Schulman & Harper, 2009), about 40 E2 enzymes (Stewart et al., 2016), and between 500 and 1,000 E3 ligases (Nakayama & Nakayama, 2006). By contrast, the *Arabidopsis thaliana* (*A. thaliana)* genome possesses two known E1 enzymes (Bachmair et al., 2001), 37 known E2 enzymes (Kraft et al., 2005), and more than 1,500 E3 ligases (Kelley, 2018).

This overview highlights the complexity and diversity of gene expression, which emerge from the combined action of PPIs, RBPs, and ribonucleoprotein complexes that govern regulation as well as the degradation or silencing of mRNAs and proteins. Yet a comprehensive mechanistic understanding of the relative contribution of each mechanism is still missing. Even if every mechanism were cataloged, simulating the millions of RNA and protein molecules within a single cell would be computationally prohibitive, making heuristic approaches attractive. Whether dissecting these contributions experimentally or building heuristic models that capture their dynamics, quantitative information on both mRNA and protein abundance is indispensable. The next section introduces the principal experimental strategies for quantifying mRNA and protein levels.

1.2. Measuring mRNA and protein abundances

The quantification of mRNA and protein abundances shares many technical steps, particularly in sample preparation and data analysis. In RNA-Seq, the most widely used method for quantitative mRNA analysis, sample preparation begins with isolating RNA from cells, enriching for specific RNA species such as mRNA, and reverse transcribing the RNA into complementary DNA (cDNA) (Mortazavi et al., 2008). The cDNA is then enzymatically fragmented, fluorescently labeled nucleotides are incorporated, and the fragments are sequenced on a high-throughput DNA sequencer (Shendure & Ji, 2008). Each nucleotide emits light at a characteristic wavelength when excited by a laser, allowing the fragments to be read base by base.

In MS-based proteomics, the predominant technique for quantitative protein analysis, sample preparation starts with extracting proteins from cells, denaturing them, and enzymatically digesting them into smaller peptide fragments (Peng et al., 2003). A coupled liquid chromatography system separates the peptides by their hydrophobicity and transfers them into a mass spectrometer, which measures each peptide's characteristic mass-to-charge ratio.

Both techniques require extensive computational analysis to obtain absolute or relative abundance values. For RNA-Seq, raw reads are aligned to a reference genome (Li & Homer, 2010) and normalized to enable comparisons across samples, for example, as reads per kilobase of transcript per million mapped reads (Dillies et al., 2013). For MS data, raw spectra are matched against spectral databases to identify peptides (Lam et al., 2007) before peptide intensities are aggregated, aligned, and normalized, with additional corrections for missing peptides (Neilson et al., 2011). Because both workflows depend on alignment to reference sequences, they are inherently limited to previously annotated genes and peptides.

Practical considerations such as cost and throughput can be estimated from publicly available price lists of academic core facilities. State-of-the-art bulk RNA-Seq performed on instruments such as the Illumina NovaSeq 6000 or NovaSeq X typically costs about US $10,000 per run, yields up to 10 billion paired-end reads, and routinely detects roughly 20,000 genes (National Cancer Institute, 2025; Stanford Medicine, 2025). By adding unique nucleotide barcodes, up to 200 samples can be multiplexed in a single run and later separated computationally. In comparison, cutting-edge MS-based proteomics costs about US $5,000 per run (Harvard Medical School, 2025; University of Florida, 2025), produces approximately 600,000 scans, and identifies around 8,000 proteins (Li et al., 2020). With isotopic labeling, about 20 samples can be analyzed in a single MS run.

This discrepancy in feature depth means that many more MS runs are required to obtain statistically reliable protein abundance estimates, especially because a typical mammalian cell contains about two orders of magnitude more protein molecules than mRNA molecules. These practical limitations help explain why large-scale proteomics remains technically challenging and underscore the value of *in silico* approaches that infer protein levels from transcript abundances. The following sections introduce fundamental ML concepts and review prior applications to sequence-derived data in light of recent advances in the field.

1.3. Machine learning

Machine learning encompasses a broad range of techniques that are commonly divided into unsupervised methods, such as clustering and dimensionality reduction, and supervised methods, including regression and classification (Alpaydin, 2014). Unsupervised approaches are valuable for exploring underlying data structures without predefined labels, but they do not directly support prediction tasks that involve known input-output relationships. Therefore, for applications that aim to predict gene expression levels with defined output values, supervised learning is more appropriate.

Before optimizing a model, researchers must design a suitable architecture. Although manual design remains common, developing and fine-tuning these architectures is challenging and requires both a solid mathematical foundation and a deep understanding of the specific problem. Support vector machines (SVMs) (Cortes & Vapnik, 1995) and decision trees (DTs) (Quinlan, 1986) are among the simplest ML architectures for supervised tasks such as classification and regression. An SVM learns a weight vector that is applied to the input through a dot product, whereas a DT learns a hierarchy of threshold-based if-then rules.

Modern architectures increase expressive power by stacking many such units into layers, yielding deep neural networks. By progressively modifying these units, for example, replacing global dot products with local convolutions, substituting nonlinear functions for hard if-then splits, or making more radical changes such as adding recurrent connections (Williams & Zipser, 1989), memory cells (Schmidhuber, 1992), or attention mechanisms (Bahdanau et al., 2014), researchers have developed increasingly specialized architectures. Notable examples include convolutional neural networks (CNNs) (Section 1.3) (Lecun & Bengio, 1994), long short-term memories (Hochreiter & Schmidhuber, 1997), and transformers (Vaswani et al., 2017).

Once the architecture's hyperparameters, such as kernel size in SVMs or tree depth in DTs, have been selected, each trainable parameter is randomly initialized (Glorot & Bengio, 2010). Using annotated training data, the parameters are then optimized by gradient descent to minimize a user-defined loss function that measures the discrepancy between predicted and desired outputs (Amari, 1993). Selecting an appropriate loss function is critical to ensure that gradient-based optimization not only updates the model parameters but also aligns the model's predictions with the intended objective (Wang et al., 2020). Common loss functions include mean squared error for regression and cross-entropy for classification tasks. In many implementations, these loss functions are combined with activation functions applied to the model's final output to match the probability or frequency distribution of the target values. For regression, activation functions such as the logarithm, hyperbolic tangent, or sigmoid can be used, whereas for classification tasks the softmax function, an extension of the sigmoid, is predominantly applied.

Before optimization, the dataset is divided into multiple equal subsets, or folds (Browne, 2000). One fold is reserved as test data, whereas the remaining folds serve as training data. Repeatedly training the same architecture with different train-test splits is termed k-fold cross-validation. To exploit parallel computing, the training data are further divided into batches, each containing multiple data points (Alpaydin, 2014). During a single step, all data points in a batch are processed in parallel. Afterward, their gradients are averaged and used to update the model's parameters. Because these gradients are computed independently, the averaged gradient is scaled by a learning rate adjusted for the batch size to prevent divergence during training. After one epoch, all batches have been processed, the model is evaluated on the test fold, the order of the training data is reshuffled, new batches are generated, and training resumes. Typically, optimization continues until changes in the loss fall below a specified threshold or the loss begins to increase, indicating potential overfitting. This iterative process ensures that changes in both parameters and hyperparameters can be evaluated statistically for significance.

A key challenge for simple SVMs and DTs is that they require explicit definition and preprocessing of input features such as edges, corners, or peaks in image and speech data; keywords in text data; and nonlinear transformations such as histograms of intensities or word occurrences (Deserno, 2011). A common feature selection strategy is to evaluate features individually, retain those that improve validation performance, and iteratively combine them until additional features no longer yield gains. Alternatively, analyzing the learned weights can deterministically identify relevant input features. Once a model has been trained, it can be examined in two main ways. First, the learned parameters can be visualized directly, for example, by displaying kernels as heatmaps or clustering them to reveal recurring patterns. Second, the

outputs at each layer can be probed with saliency maps to understand how the network processes specific input features (Simonyan et al., 2014).

In recent sequence-based applications, supervised methods have been used to predict binding sites (Alipanahi et al., 2015; Munusamy et al., 2017; Zhuang et al., 2019), protein structure (Baker & Sali, 2001; Jumper et al., 2021; Senior et al., 2020), mRNA abundance (Agarwal & Shendure, 2020; Zrimec et al., 2020), and PTR (Buric et al., 2025; Cuperus et al., 2017; Eraslan et al., 2019; Hebditch et al., 2017). These studies often rely on manually defined sequence-derived features, including codon usage patterns, start- and stop-codon contexts, and short linear motifs, alongside secondary or tertiary structure characteristics of both mRNA and proteins, as well as physicochemical features such as hydrophobicity, isoelectric point, and folding energy.

Although manually defined features provide concrete starting points, this approach faces several limitations. First, it is practically infeasible to exhaustively consider all possible combinations of biologically relevant features given current dataset sizes. Second, manually defined features are typically simple and rigid, making it difficult to capture complex patterns such as combinatorial motifs. Third, there is a high risk of tailoring models so specifically to the training data that they lose generalizability and fail on unseen data. Consequently, the next step is to integrate modules capable of learning useful features automatically from raw input data. One widely adopted solution, particularly successful in image and speech recognition, is to stack convolutional layers followed by dense layers that act as the final classifier, thereby forming CNNs.

1.4. Convolutional Neural Networks

Convolutional operations have long been used in hand-crafted feature extraction tasks, such as edge detection (Canny, 1986), corner detection (Harris & Stephens, 1988), and blob detection (Huertas & Medioni, 1986), across image, speech, and time series analysis. Ongoing advances in computing power have made ML-based optimization of convolutional parameters feasible (Lecun et al., 1998), and user-friendly frameworks like TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019) have established CNNs as the default solution. With the growing availability of large-scale image and speech datasets online, pretrained deep CNNs containing numerous convolutional layers and millions of trainable parameters, such as VGG19 (Simonyan & Zisserman, 2015) and ResNet50 (He et al., 2016), have made sophisticated CNN-based models accessible even for small-scale projects.

Deep CNNs typically combine linear transformations, such as convolutions and dot products, with nonlinear transformations, such as rectified linear units (ReLUs) (Nair & Hinton, 2010), Leaky-ReLUs (Maas et al., 2013), and exponential linear units (Clevert et al., 2016), into functional units known as layers. These layers are reused throughout the network, each time with distinct learnable parameters. Nonlinear transformations greatly enhance model capacity by enabling the network to suppress irrelevant features in the input signal. Because early computer vision approaches relied on handcrafted operators, the kernel sizes adopted in CNNs were initially influenced by those traditional filters. The number of kernels within a layer and the overall depth of the network are limited primarily by the amount of available training data. A common practice is to use a small number of kernels in the initial layers and to increase this number after each downsampling operation. The final layers in CNNs are typically dense layers with activation functions analogous to the linear classifier used in an SVM.

Unlike fully connected classifiers such as SVMs or DTs, convolutional filters are initially connected only to small, local patches of the input signal. Consequently, CNNs must deliberately expand their receptive field, which describes the region of the input that contributes to an output value, so that the network can capture dependencies across distant regions of an image or speech signal (Luo et al., 2016). Because standard convolutions enlarge the receptive field only linearly, downsampling operations such as pooling or strided convolutions are often inserted between convolutional layers. Common pooling strategies, such as max or mean pooling over non-overlapping windows, allow the receptive field to grow exponentially with depth (Boureau et al., 2010). An alternative is dilated convolution, which inserts gaps filled with zeros between kernel elements (Yu & Koltun, 2015), enabling exponential growth of the receptive field without loss of resolution or increased computational cost.

Although CNNs can learn relevant features automatically, their hyperparameters still require extensive tuning (Bergstra et al., 2011). Hyperparameters such as kernel size, kernel count, activation type, and layer depth have a pronounced impact on generalizability. For emerging signal-processing tasks, such as analyzing biological sequences, prior domain knowledge can guide initial hyperparameter choices. In sequence analysis, convolutions can extract nucleotide motifs directly from raw DNA or RNA, reducing dependence on predefined motif libraries. Initial models can nonetheless leverage experimentally validated motifs (Dinkel et al., 2014; Kawaguchi & Bailey-Serres, 2005) as informative priors for hyperparameter selection.

Among the earliest sequence-based CNNs was DeepBind (Alipanahi et al., 2015), which predicts protein-DNA and protein-RNA binding affinities directly from raw sequences. The network comprised a single convolutional layer with 12 kernels of length 6, followed by a dense output layer. In transcriptomics, Xpresso (Agarwal & Shendure, 2020) predicts mRNA expression from

promoter sequences using two convolutional layers with 128 and 32 kernels of lengths 6 and 9, respectively, and two intermediate dense layers with 64 and 2 units. Zrimec et al. (2020) extended this concept to full gene sequences, employing three convolutional layers with 128, 32, and 64 kernels of lengths 40, 30, and 30, respectively, and two intermediate dense layers with 128 and 32 units. Collectively, these studies illustrate a steady progression toward deeper CNNs for sequence-based inference.

In proteomics, notable SVM-based models include Protein-Sol (Hebditch et al., 2017) for predicting protein solubility and the model by Eraslan et al. (2019) for inferring protein abundance from mRNA levels. DT-based models for inferring protein abundance from mRNA levels, such as that of Buric et al. (2025), have also been introduced. These models rely on manually engineered sequence features, such as nucleotide and amino acid composition, as well as brute-force-selected upstream and downstream motifs. A CNN-based study by Cuperus et al. (2017) assessed the impact of 5′ UTR variants on protein levels both *in vitro* and *in silico* using three convolutional layers with 128 kernels of length 13, followed by a dense layer with 6 units.

Across both transcriptomic and proteomic studies, the convolutional kernels and manually selected features frequently target short linear motifs of 6 to 13 nucleotides or amino acids and treat the CDS separately from the 5′ and 3′ UTRs. Histograms of nucleotide and amino acid composition are also common. The next logical step is to build CNNs that jointly process the CDS, 5′ UTR, 3′ UTR, and composition histograms to improve protein abundance prediction.

1.5. Aim of the thesis

The aim of this thesis is to employ CNNs for gene expression analysis and network biology. In publication A, I develop a custom CNN architecture to predict protein concentrations from mRNA concentrations. For this purpose, I use recently generated matched transcriptome-proteome datasets from the representative model organisms *H. sapiens* (Wang et al., 2019) and *A. thaliana* (Mergner et al., 2020) as training data. I also analyze the learned parameters of the trained CNNs to determine whether the automatically learned features are biologically meaningful. In publications B and C, I develop an automated pipeline to analyze image data from the yeast two-hybrid (Y2H) system and generate quantitative PPI networks. Classical image-processing methods are applied to localize the Petri dish and identify the position and orientation of the 96-well plate pattern, while a second CNN automatically scores yeast culture growth.

2. Discussion

In publication A, I developed a CNN that predicts gene-specific protein concentrations directly from corresponding mRNA abundances. The model incorporates sequence-derived features from both mRNA and protein sequences and, in an extended version, also includes the expression levels of correlated transcripts. With this approach, protein concentrations can be predicted for previously unseen transcripts and tissue types by using only their mRNA sequences and abundances. On independent test sets, the sequence-based model achieved coefficients of determination (r^2) of 0.30 for *H. sapiens* and 0.32 for *A. thaliana*. For *H. sapiens*, the model improves prediction accuracy by approximately 40% relative to the best sequence-based approach of Eraslan et al. (2019), which did not employ CNNs and reached an r^2 of 0.22. For *A. thaliana*, this work represents the first attempt to predict protein concentrations from matched transcript abundances.

Although previous sequence-based models have been developed for *H. sapiens*, they differ from the approach presented here. For example, the models by Hebditch et al. (2017) and Stefanini et al. (2023) rely primarily on sequence-derived features without incorporating measured mRNA abundances. As a result, their predictions reflect the intrinsic translational potential of each gene rather than the context-specific protein levels dictated by current transcriptional activity. These models effectively average protein predictions over multiple data points, leading to systematically lower accuracy ($r^2 < 0.20$). Eraslan et al. (2019) adopted a sequence- and expression-based approach, but instead of a single global predictor, they retrained an independent model for each tissue to capture unique expression characteristics.

While tissue-specific training can improve performance within any single known tissue, it considerably limits generalizability. When a novel tissue type or condition is encountered, the model must be retrained, which is impractical in large-scale applications where proteomics data are scarce. In contrast, the CNN-based model presented here is trained once per species, separately for *H. sapiens* and *A. thaliana*, without subsequent retraining for individual tissues. Consequently, a single model applies to any tissue, healthy or diseased, including conditions unseen during training. Despite the lack of tissue-specific customization, it consistently outperforms previous methods across all tissues.

This finding underscores the ability of convolutional front-end modules to learn robust sequence motifs and context-dependent combinations. By contrast, Hebditch et al. (2017), Eraslan et al. (2019), and Buric et al. (2025) relied on manually extracted features such as codon counts, amino acid composition, and short linear motifs identified by brute force, as well as more complex

features including secondary and tertiary structure predictions and physicochemical properties, all of which demand extensive manual engineering and domain-specific insight.

However, convolutional modules require large matched transcriptome-proteome datasets spanning thousands of genes to uncover and optimize relevant sequence features. When only a limited number of genes or tissues are available, the model may underperform compared with simpler or more narrowly tailored methods. With sufficient data, convolutional front-end modules can learn not only the most relevant motifs directly from raw sequences but also their variants. Similar motifs that differ at only a single nucleotide position can be captured within one convolutional filter, enabling the network to extract higher-order motifs that allow finer distinctions, especially regarding potential interaction partners. Additionally, such models promise a universal solution applicable to genes unseen during training. Because they learn species-specific factors, including codon usage weights and diverse motifs, previously unseen mutant sequences or even exogenous viral genes could be provided as input to predict their expected expression levels.

Detailed analysis of the learned parameters shows that simple features in the CDS, most notably codon counts and out-of-frame start and stop codons, drive much of the predictive power. This outcome is consistent with gradient descent optimization, where parameters that most reduce the loss are optimized first. Although more complex features are learned simultaneously, they require considerably more training steps and larger datasets to reach full optimization. Comparing the learned convolutional filters with published short linear motifs confirms that the network rediscovers binding sites for polypyrimidine tract binding proteins (Oberstrass et al., 2005), cleavage factors (Yang et al., 2010), and translational regulators (Kumari et al., 2007), as well as motifs associated with mRNA stability (Chen & Shyu, 1995; Savinov et al., 2021).

Although the training datasets aggregate protein and mRNA abundances across transcript isoforms, the model still detects the splicing-regulatory polypyrimidine tract, suggesting that it can infer the likelihood of alternative splicing. The model also appears more sensitive to motifs associated with mRNA decay and stability than to those linked to translational control. Following the logic of gradient descent, this finding implies that decay-related signals contribute more strongly to predictive performance. The confirmation that many learned filters correspond to known regulatory elements highlights the potential significance of the remaining, putative elements, particularly motifs of ten or more nucleotides and single-nucleotide variants of validated motifs. These may represent novel regulators and therefore warrant targeted *in vitro* validation.

In the extended model, network-based features were integrated with sequence-derived features to determine whether higher-order relationships between sequence motifs and potential interaction partners could be captured automatically. Because the concentrations of interacting proteins are typically unknown in real-world applications, mRNA abundances were used as proxies. Despite extensive hyperparameter tuning, no gain in predictive accuracy was achieved. A likely explanation is the limited size of the matched transcriptome-proteome training set. Only 29 *H. sapiens* tissues (Wang et al., 2019) and 30 *A. thaliana* tissues (Mergner et al., 2020) were available, each representing a single data point. Consequently, too few condition-specific observations were available to allow reliable estimation of condition-dependent parameters.

Similar network-based approaches have been explored with *H. sapiens* cancer datasets. Li et al. (2019) employed a random forest model that integrated mRNA concentrations of genes filtered by Gene Ontology annotations. Their strategy involved selecting candidate genes suspected to interact with or coregulate the target proteins, thereby capturing aspects of post-transcriptional regulation that single-gene analyses miss. The authors tested their approach on 182 tumor samples spanning multiple cancer subtypes and achieved a Pearson correlation coefficient (r) of 0.49 (r^2 = 0.24) when predicting protein levels. Srivastava et al. (2022) used a random forest model enhanced by brute-force selection of potentially influential genes. Thanks to a substantially larger dataset of 958 patient samples, they reported an r of 0.599 (r^2 = 0.36).

These findings underscore a positive relationship between sample size and predictive accuracy in network-based approaches. Yet, despite their strong performance, this benefit is tempered by the fact that both studies focus only on cancer cells, whose expression patterns can differ substantially from those in healthy tissues. To overcome the corresponding limitations for healthy tissues, additional input types, such as *in vitro* validated PPI and RNA-protein interaction data, together with ncRNA concentrations, such as miRNA, siRNA, tRNA, and rRNA, could yield more substantial gains in predictive accuracy. Although an estimated 98% of the human genome is transcribed into ncRNA (Mattick & Makunin, 2006) and interactions involving ncRNA play crucial roles in mRNA translation and thus in determining protein concentrations (Guil & Esteller, 2015), none of the approaches presented here include ncRNA concentrations.

The extended model attempted to infer these interactions from potential binding sites, but recent studies show that inferring them solely from sequence data remains a major challenge (Lannelongue & Inouye, 2024; Pan et al., 2019). Given the availability of extensive PPI networks for *H. sapiens* (Luck et al., 2020; Rolland et al., 2014; Rual et al., 2005) and *A. thaliana* (Altmann et al., 2020; Arabidopsis Interactome Mapping Consortium et al., 2011), as well as RNA-protein interaction networks (Caudron-Herger et al., 2021), a logical next step is to integrate these resources as additional input features. For both interaction types, graph convolutional networks

(GCNs) (Kipf & Welling, 2016) can meaningfully process the data by reducing their dimensionality, much like principal component analysis. Combining graph convolutional and regular convolutional front-end modules would require the model to learn fewer parameter combinations, making it easier to train even with limited condition-specific data.

Although the convolutional front-end modules have improved model accuracy, their receptive fields remain local to linear motifs, limiting their ability to capture three-dimensional protein structures that rely on combinations of distant motifs. Because these structural elements are crucial for facilitating protein interactions and strongly influence network-based approaches, future extensions of the model architecture could benefit from incorporating attention-based mechanisms (Vaswani et al., 2017). Proven effective in text processing, attention-based models can weight input information across long distances and are well suited to capturing distant dependencies in complex tertiary protein structures.

Another challenge in network-based approaches arises from the use of mRNA concentrations as initial proxies, even though they do not correlate perfectly with protein concentrations. To minimize the resulting inaccuracies, the architecture should be extended with recurrent layers that enable iterative predictions, refining the initial protein concentration estimate stepwise on the basis of previous outputs.

Overall, the sequence-based approach presented in publication A represents a substantial improvement in predictive power and generalizability over previous models for estimating protein concentrations. Nevertheless, considerable room for further improvement remains. Future research should focus on network-based expansions that integrate additional data types, such as experimentally validated interaction datasets and ncRNA concentrations. Addressing these inputs will also require fundamental architectural changes, including the incorporation of GCNs, attention mechanisms, and recurrent layers.

In publications B and C, two PPI networks, one between human and SARS-CoV-2 proteins and another between human and gut microbiome proteins, were generated using the Y2H system. My contribution included developing an ML application that automatically analyzes Y2H image data, generating and curating the final datasets, and conducting network analysis. The application combines conventional algorithms for detecting the 96-well plate pattern with a dilated CNN for evaluating yeast culture growth. Trained on previously collected and published data (Altmann et al., 2020), the model provides consistent assessments of yeast growth, eliminating user bias and greatly increasing analysis throughput.

Interaction network analysis in both cases reveals that the number of interaction partners follows an exponential distribution, a trend also observed in large PPI networks among *H. sapiens* proteins (Luck et al., 2020). A similar pattern appeared in the network-based analyses from publication A, where an exponential distribution was evident when correlating protein concentrations. Future *in vitro* experiments aimed at identifying PPIs can prioritize the proteins from publication A whose concentrations correlate strongly with those of multiple genes. These proteins are promising candidates for highly connected regulators of gene expression pathways. Conversely, the results of such experiments can serve as input data for future extensions of the CNN model described in publication A.

Despite the challenges encountered, this work demonstrates the utility of CNNs across multiple domains of network biology. CNNs have been applied effectively in predictive tasks, estimating protein abundances and assessing growth measurements, and in analytical tasks, extracting sequence features and quantifying correlations. Nevertheless, it is essential to acknowledge the limitations of these models and avoid their indiscriminate use in sensitive contexts. This work shows that continued advances in ML are both realistic and achievable and outlines concrete paths toward their realization. As progress continues toward a deeper understanding of gene expression regulation, these results can be applied in diverse contexts. For example, they can quantify the impact of individual nucleotide mutations on the protein concentrations of interacting genes. This not only advances scientific understanding but also holds significant potential for personalized medicine and crop improvement.

3. References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. https://doi.org/10.48550/arXiv.1603.04467

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422(6928)*, *198-207*. https://doi.org/10.1038/nature01511

Agarwal, V., & Shendure, J. (2020). Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep*, *31(7)*, *107663*. https://doi.org/10.1016/j.celrep.2020.107663

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, *33(8)*, *831-838*. https://doi.org/10.1038/nbt.3300

Alpaydin, E. (2014). *Introduction to Machine Learning* (Vol. 3). MIT press.

Altmann, M., Altmann, S., Rodriguez, P. A., Weller, B., Elorduy Vergara, L., et al. (2020). Extensive signal integration by the phytohormone protein network. *Nature*, *583(7815)*, *271-276*. https://doi.org/10.1038/s41586-020-2460-0

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5(4-5)*, *185-196*. https://doi.org/10.1016/0925-2312(93)90006-o

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *N Biotechnol*, *25(4)*, *195-203*. https://doi.org/10.1016/j.nbt.2008.12.009

Arabidopsis Interactome Mapping Consortium, Dreze, M., Carvunis, A.-R., Charloteaux, B., Galli, M., et al. (2011). Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science*, *333(6042)*, *601-607*. https://doi.org/10.1126/science.1203877

Armingol, E., Officer, A., Harismendy, O., & Lewis, N. E. (2021). Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet*, *22(2)*, *71-88*. https://doi.org/10.1038/s41576-020-00292-x

Bachmair, A., Novatchkova, M., Potuschak, T., & Eisenhaber, F. (2001). Ubiquitylation in plants: a post-genomic look at a post-translational modification. *Trends Plant Sci*, *6(10)*, *463-470*. https://doi.org/10.1016/s1360-1385(01)02080-5

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations 3 (ICLR 2015)*. https://doi.org/10.48550/arXiv.1409.0473

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294(5540)*, *93-96*. https://doi.org/10.1126/science.1065659

Barbosa, C., Peixeiro, I., & Romao, L. (2013). Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet*, *9(8)*, *e1003529*. https://doi.org/10.1371/journal.pgen.1003529

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, *116(2)*, *281-297*. https://doi.org/10.1016/s0092-8674(04)00045-5

Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., et al. (2015). Huntington disease. *Nat Rev Dis Primers*, *1(1)*, *15005*. https://doi.org/10.1038/nrdp.2015.5

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2013). *Stryer Biochemie* (Vol. 8). Springer. https://doi.org/10.1007/978-3-8274-2989-6

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS 2011)*. https://dl.acm.org/doi/10.5555/2986459.2986743

Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. *Proceedings of the 27th International Conference on International Conference on Machine Learning (IMCL10)*. https://dl.acm.org/doi/10.5555/3104322.3104338

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478(7369)*, *343-348*. https://doi.org/10.1038/nature10532

Browne, M. W. (2000). Cross-Validation Methods. *J Math Psychol*, *44(1)*, *108-132*. https://doi.org/10.1006/jmps.1999.1279

Buric, F., Viknander, S., Fu, X., Lemke, O., Carmona, O. G., et al. (2025). Amino acid sequence encodes protein abundance shaped by protein stability at reduced synthesis cost. *Protein Sci*, *34(1)*, *e5239*. https://doi.org/10.1002/pro.5239

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell*, *8(6)*, *679-698*. https://doi.org/10.1109/TPAMI.1986.4767851

Caudron-Herger, M., Jansen, R. E., Wassmer, E., & Diederichs, S. (2021). RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. *Nucleic Acids Res*, *49(D1)*, *D425-D436*. https://doi.org/10.1093/nar/gkaa1040

Chen, C. Y., & Shyu, A. B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci*, *20(11)*, *465-470*. https://doi.org/10.1016/s0968-0004(00)89102-1

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (elus). *4th International Conference on Learning Representations (ICLR 2016)*. https://doi.org/10.48550/arXiv.1511.07289

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat Rev Genet*, *10(3)*, *184-194*. https://doi.org/10.1038/nrg2537

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20(3)*, *273-297*. https://doi.org/10.1007/bf00994018

Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., et al. (2017). Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res*, *27(12)*, *2015-2024*. https://doi.org/10.1101/gr.224964.117

Deserno, T. M. (2011). Biomedical Image Processing. *Biological and Medical Physics, Biomedical Engineering*. https://doi.org/10.1007/978-3-642-15816-2

Deshaies, R. J. (1999). SCF and Cullin/Ring H2-based ubiquitin ligases. *Annu Rev Cell Dev Biol*, *15*, *435-467*. https://doi.org/10.1146/annurev.cellbio.15.1.435

Deutsche Welthungerhilfe, Concern Worldwide, & Institute for International Law of Peace and Armed Conflict. (2024). *Global Hunger Index*. https://web.archive.org/web/20250404181206/https://www.globalhungerindex.org/pdf/en/2024.pdf

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, *14(6)*, *671-683*. https://doi.org/10.1093/bib/bbs046

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., et al. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res*, *42*, *D259-266*. https://doi.org/10.1093/nar/gkt1047

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallstrom, B. M., et al. (2019). Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol Syst Biol*, *15(2)*, *e8513*. https://doi.org/10.15252/msb.20188513

Feng, T., Yamamoto, A., Wilkins, S. E., Sokolova, E., Yates, L. A., et al. (2014). Optimal translational termination requires C4 lysyl hydroxylation of eRF1. *Mol Cell*, *53(4)*, *645-654*. https://doi.org/10.1016/j.molcel.2013.12.028

Flinn, A. M., & Gennery, A. R. (2018). Adenosine deaminase deficiency: a review. *Orphanet J Rare Dis*, *13(1)*, *65*. https://doi.org/10.1186/s13023-018-0807-5

Frantz, C., Stewart, K. M., & Weaver, V. M. (2010). The extracellular matrix at a glance. *J Cell Sci*, *123(Pt 24)*, *4195-4200*. https://doi.org/10.1242/jcs.023820

Gao, M., Fritz, D. T., Ford, L. P., & Wilusz, J. (2000). Interaction between a poly(A)-specific ribonuclease and the 5' cap influences mRNA deadenylation rates in vitro. *Mol Cell*, *5(3)*, *479-488*. https://doi.org/10.1016/s1097-2765(00)80442-6

Gingold, H., & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Mol Syst Biol*, *7(1)*, *481*. https://doi.org/10.1038/msb.2011.14

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (PLMR9)*. https://api.semanticscholar.org/CorpusID:5575601

Guil, S., & Esteller, M. (2015). RNA-RNA interactions in gene regulation: the coding and noncoding players. *Trends Biochem Sci*, *40(5)*, *248-256*. https://doi.org/10.1016/j.tibs.2015.03.001

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *N Engl J Med*, *363(4)*, *301-304*. https://doi.org/10.1056/NEJMp1006304

Hanson, G., & Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol*, *19(1)*, *20-30*. https://doi.org/10.1038/nrm.2017.91

Hargrove, J. L., & Schmidt, F. H. (1989). The role of mRNA and protein stability in gene expression. *FASEB J*, *3(12)*, *2360-2370*. https://doi.org/10.1096/fasebj.3.12.2676679

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. *Alvey vision conference*, *15(50)*, *10-5244*. https://doi.org/10.5244/C.2.23

Harvard Medical School. (2025). https://web.archive.org/web/20250317153313/https://tcmp.hms.harvard.edu/rates

He, K. M., Zhang, X. Y., Ren, S. Q., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. https://doi.org/10.1109/Cvpr.2016.90

Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, *33(19)*, *3098-3100*. https://doi.org/10.1093/bioinformatics/btx345

Hershko, A., & Ciechanover, A. (1992). The ubiquitin system for protein degradation. *Annu Rev Biochem*, *61(1)*, *761-807*. https://doi.org/10.1146/annurev.bi.61.070192.003553

Hilborn, R. C. (2000). Chaos and Nonlinear Dynamics. https://doi.org/10.1093/acprof:oso/9780198507239.001.0001

Hinnebusch, A. G., Ivanov, I. P., & Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*, *352(6292)*, *1413-1416*. https://doi.org/10.1126/science.aad9868

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*, *9(8)*, *1735-1780*. https://doi.org/10.1162/neco.1997.9.8.1735

Huertas, A., & Medioni, G. (1986). Detection of intensity changes with subpixel accuracy using laplacian-gaussian masks. *IEEE Trans Pattern Anal Mach Intell*, *8(5)*, *651-664*. https://doi.org/10.1109/tpami.1986.4767838

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, *2(Volume 2, 2001)*, *343-372*. https://doi.org/10.1146/annurev.genom.2.1.343

Janmey, P. A., & McCulloch, C. A. (2007). Cell mechanics: integrating cell responses to mechanical stimuli. *Annu Rev Biomed Eng*, *9(Volume 9, 2007)*, *1-34*. https://doi.org/10.1146/annurev.bioeng.9.060906.151927

Joyce, A. R., & Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, *7(3)*, *198-210*. https://doi.org/10.1038/nrm1857

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596(7873)*, *583-589*. https://doi.org/10.1038/s41586-021-03819-2

Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nat Rev Genet*, *19(5)*, *299-310*. https://doi.org/10.1038/nrg.2018.4

Kawaguchi, R., & Bailey-Serres, J. (2005). mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Res*, *33(3)*, *955-965*. https://doi.org/10.1093/nar/gki240

Kelley, D. R. (2018). E3 Ubiquitin Ligases: Key Regulators of Hormone Signaling in Plants. *Mol Cell Proteomics*, *17(6)*, *1047-1054*. https://doi.org/10.1074/mcp.MR117.000476

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations (ICLR 2017)*. https://doi.org/10.48550/arXiv.1609.02907

Kitano, H. (2002a). Computational systems biology. *Nature*, *420(6912)*, *206-210*. https://doi.org/10.1038/nature01254

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, *295(5560)*, *1662-1664*. https://doi.org/10.1126/science.1069492

Komander, D., & Rape, M. (2012). The ubiquitin code. *Annu Rev Biochem*, *81*, *203-229*. https://doi.org/10.1146/annurev-biochem-060310-170328

Kraft, E., Stone, S. L., Ma, L., Su, N., Gao, Y., et al. (2005). Genome analysis and functional characterization of the E2 and RING-type E3 ligase ubiquitination enzymes of Arabidopsis. *Plant Physiol*, *139(4)*, *1597-1611*. https://doi.org/10.1104/pp.105.067983

Kuhn, U., & Wahle, E. (2004). Structure and function of poly(A) binding proteins. *Biochim Biophys Acta*, *1678(2-3)*, *67-84*. https://doi.org/10.1016/j.bbaexp.2004.03.008

Kumari, S., Bugaut, A., Huppert, J. L., & Balasubramanian, S. (2007). An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol*, *3(4)*, *218-221*. https://doi.org/10.1038/nchembio864

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., et al. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *PROTEOMICS*, *7(5)*, *655-667*. https://doi.org/10.1002/pmic.200600625

Lannelongue, L., & Inouye, M. (2024). Pitfalls of machine learning models for protein-protein interaction networks. *Bioinformatics*, *40(2)*. https://doi.org/10.1093/bioinformatics/btae012

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., et al. (2006). Machine learning in bioinformatics. *Brief Bioinform*, *7(1)*, *86-112*. https://doi.org/10.1093/bib/bbk007

Lebeer, S., Vanderleyden, J., & De Keersmaecker, S. C. (2010). Host interactions of probiotic bacterial surface molecules: comparison with commensals and pathogens. *Nat Rev Microbiol*, *8(3)*, *171-184*. https://doi.org/10.1038/nrmicro2297

Lecun, Y., & Bengio, Y. (1994). Word-Level Training of a Handwritten Word Recognizer Based on Convolutional Neural Networks. *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR 1994)*. https://doi.org/10.1109/ICPR.1994.576881

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86(11)*, *2278-2324*. https://doi.org/10.1109/5.726791

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, *11(5)*, *473-483*. https://doi.org/10.1093/bib/bbq015

Li, H., Siddiqui, O., Zhang, H., & Guan, Y. (2019). Joint learning improves protein abundance prediction in cancers. *BMC Biol*, *17(1)*, *107*. https://doi.org/10.1186/s12915-019-0730-9

Li, J., Van Vranken, J. G., Pontano Vaites, L., Schweppe, D. K., Huttlin, E. L., et al. (2020). TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat Methods*, *17(4)*, *399-404*. https://doi.org/10.1038/s41592-020-0781-4

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., et al. (2020). A reference map of the human binary protein interactome. *Nature*, *580(7803)*, *402-408*. https://doi.org/10.1038/s41586-020-2188-x

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *29th Conference on Neural Information Processing Systems (NIPS 2016)*. https://doi.org/10.48550/arXiv.1701.04128

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:16489696

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nat Biotechnol*, *21(3)*, *255-261*. https://doi.org/10.1038/nbt0303-255

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N Engl J Med*, *363(2)*, *166-176*. https://doi.org/10.1056/NEJMra0905980

Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Hum Mol Genet*, *15*, *R17-29*. https://doi.org/10.1093/hmg/ddl046

Merchante, C., Stepanova, A. N., & Alonso, J. M. (2017). Translation regulation in plants: an interesting past, an exciting present and a promising future. *Plant J*, *90(4)*, *628-653*. https://doi.org/10.1111/tpj.13520

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., et al. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature*, *579(7799)*, *409-414*. https://doi.org/10.1038/s41586-020-2094-2

Misra, B. B., Langefeld, C. D., Olivier, M., & Cox, L. A. (2018). Integrated Omics: Tools, Advances, and Future Approaches. *J Mol Endocrinol*, *62(1)*, *R21-R45*. https://doi.org/10.1530/JME-18-0055

Mizuno, H., Tsuyama, N., Harada, T., & Masujima, T. (2008). Live single-cell video-mass spectrometry for cellular and subcellular molecular detection and cell classification. *J Mass Spectrom*, *43(12)*, *1692-1700*. https://doi.org/10.1002/jms.1460

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., et al. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, *430(7001)*, *743-747*. https://doi.org/10.1038/nature02797

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, *5(7)*, *621-628*. https://doi.org/10.1038/nmeth.1226

Mugridge, J. S., Ziemniak, M., Jemielity, J., & Gross, J. D. (2016). Structural basis of mRNA-cap recognition by Dcp1-Dcp2. *Nat Struct Mol Biol*, *23(11)*, *987-994*. https://doi.org/10.1038/nsmb.3301

Munusamy, P., Zolotarov, Y., Meteignier, L. V., Moffett, P., & Stromvik, M. V. (2017). De novo computational identification of stress-related sequence motifs and microRNA target sites in untranslated regions of a plant translatome. *Sci Rep*, *7(1)*, *43861*. https://doi.org/10.1038/srep43861

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML10)*. https://dl.acm.org/doi/10.5555/3104322.3104425

Nakamura, Y., Ito, K., & Isaksson, L. A. (1996). Emerging understanding of translation termination. *Cell*, *87(2)*, *147-150*. https://doi.org/10.1016/s0092-8674(00)81331-8

Nakayama, K. I., & Nakayama, K. (2006). Ubiquitin ligases: cell-cycle control and cancer. *Nat Rev Cancer*, *6(5)*, *369-381*. https://doi.org/10.1038/nrc1881

National Cancer Institute. (2025). https://web.archive.org/web/20250604040444/https://crtp.ccr.cancer.gov/sf/pricing/

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., et al. (2011). Less label, more free: approaches in label-free quantitative mass spectrometry. *PROTEOMICS*, *11(4)*, *535-553*. https://doi.org/10.1002/pmic.201000553

Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Mol Syst Biol*, *2(1)*, *52*. https://doi.org/10.1038/msb4100095

O'Sullivan, B. P., & Freedman, S. D. (2009). Cystic fibrosis. *Lancet*, *373(9678)*, *1891-1904*. https://doi.org/10.1016/S0140-6736(09)60327-5

Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., et al. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, *309(5743)*, *2054-2057*. https://doi.org/10.1126/science.1114066

Oleksiak, M. F., Churchill, G. A., & Crawford, D. L. (2002). Variation in gene expression within and among natural populations. *Nat Genet*, *32(2)*, *261-266*. https://doi.org/10.1038/ng983

Page, A. M., & Hieter, P. (1999). The anaphase-promoting complex: new subunits and regulators. *Annu Rev Biochem*, *68*, *583-609*. https://doi.org/10.1146/annurev.biochem.68.1.583

Paik, I., Yang, S., & Choi, G. (2012). Phytochrome regulates translation of mRNA in the cytosol. *Proc Natl Acad Sci U S A*, *109(4)*, *1335-1340*. https://doi.org/10.1073/pnas.1109683109

Pan, X., Yang, Y., Xia, C. Q., Mirza, A. H., & Shen, H. B. (2019). Recent methodology progress of deep learning for RNA-protein interaction prediction. *Wiley Interdiscip Rev RNA*, *10(6)*, *e1544*. https://doi.org/10.1002/wrna.1544

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *33rd Conference on Neural Information Processing Systems (NIPS 2019)*. https://doi.org/10.48550/arXiv.1912.01703

Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., & Gygi, S. P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, *2(1)*, *43-50*. https://doi.org/10.1021/pr025556v

Pestova, T. V., Kolupaeva, V. G., Lomakin, I. B., Pilipenko, E. V., Shatsky, I. N., et al. (2001). Molecular mechanisms of translation initiation in eukaryotes. *Proc Natl Acad Sci U S A*, *98(13)*, *7029-7036*. https://doi.org/10.1073/pnas.111145798

Phizicky, E. M., & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, *59(1)*, *94-123*. https://doi.org/10.1128/mr.59.1.94-123.1995

Pickart, C. M., & Eddins, M. J. (2004). Ubiquitin: structures, functions, mechanisms. *Biochim Biophys Acta*, *1695(1-3)*, *55-72*. https://doi.org/10.1016/j.bbamcr.2004.09.019

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, *81-106*. https://doi.org/10.1007/bf00116251

Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA-protein interactions. *Nat Methods*, *16(3)*, *225-234*. https://doi.org/10.1038/s41592-019-0330-1

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *Int J Proteomics*, *2014(1)*, *147648*. https://doi.org/10.1155/2014/147648

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. *Lancet*, *376(9757)*, *2018-2031*. https://doi.org/10.1016/S0140-6736(10)61029-X

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., et al. (2014). A proteome-scale map of the human interactome network. *Cell*, *159(5)*, *1212-1226*. https://doi.org/10.1016/j.cell.2014.10.050

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, *437(7062)*, *1173-1178*. https://doi.org/10.1038/nature04209

Savinov, A., Brandsen, B. M., Angell, B. E., Cuperus, J. T., & Fields, S. (2021). Effects of sequence motifs in the yeast 3' untranslated region determined from massively parallel assays of random sequences. *Genome Biol*, *22(1)*, *293*. https://doi.org/10.1186/s13059-021-02509-6

Schmidhuber, J. (1992). Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. *Neural Computation*, *4(1)*, *131-139*. https://doi.org/10.1162/neco.1992.4.1.131

Schulman, B. A., & Harper, J. W. (2009). Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nat Rev Mol Cell Biol*, *10(5)*, *319-331*. https://doi.org/10.1038/nrm2673

Scott, D. E., Bayly, A. R., Abell, C., & Skidmore, J. (2016). Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nat Rev Drug Discov*, *15(8)*, *533-550*. https://doi.org/10.1038/nrd.2016.29

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577(7792)*, *706-710*. https://doi.org/10.1038/s41586-019-1923-7

Shalgi, R., Hurt, J. A., Krykbaeva, I., Taipale, M., Lindquist, S., & Burge, C. B. (2013). Widespread regulation of translation by elongation pausing in heat shock. *Mol Cell*, *49(3)*, *439-452*. https://doi.org/10.1016/j.molcel.2012.11.028

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, *26(10)*, *1135-1145*. https://doi.org/10.1038/nbt1486

Shinbrot, T., Grebogi, C., Wisdom, J., & Yorke, J. A. (1992). Chaos in a double pendulum. *American Journal of Physics*, *60(6)*, *491-499*. https://doi.org/10.1119/1.16860

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations (ICLR 2014)*. https://doi.org/10.48550/arXiv.1312.6034

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1409.1556

Somero, G. N. (2020). The cellular stress response and temperature: Function, regulation, and evolution. *J Exp Zool A Ecol Integr Physiol*, *333(6)*, *379-397*. https://doi.org/10.1002/jez.2344

Srivastava, H., Lippincott, M. J., Currie, J., Canfield, R., Lam, M. P. Y., & Lau, E. (2022). Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners. *PLoS Comput Biol*, *18(11)*, *e1010702*. https://doi.org/10.1371/journal.pcbi.1010702

Stanford Medicine. (2025). https://web.archive.org/web/20240709100916/https://med.stanford.edu/dgac/pricing.html

Stefanini, M., Lovino, M., Cucchiara, R., & Ficarra, E. (2023). Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. *Comput Methods Programs Biomed*, *234*, *107504*. https://doi.org/10.1016/j.cmpb.2023.107504

Stewart, M. D., Ritterhoff, T., Klevit, R. E., & Brzovic, P. S. (2016). E2 enzymes: more than just middle men. *Cell Res*, *26(4)*, *423-440*. https://doi.org/10.1038/cr.2016.35

Terzaghi, W. B., & Cashmore, A. R. (1995). Light-Regulated Transcription. *Annual Review of Plant Physiology and Plant Molecular Biology*, *46(1)*, *445-474*. https://doi.org/10.1146/annurev.pp.46.060195.002305

The World Bank. (2024). *World development indicators*. https://databank.worldbank.org/source/world-development-indicators

Tsonis, A. A., & Elsner, J. B. (1989). Chaos, Strange Attractors, and Weather. *Bulletin of the American Meteorological Society*, *70(1)*, *14-23*. https://doi.org/10.1175/1520-0477(1989)070%3C0014:CSAAW%3E2.0.CO;2

Turner, M., & Diaz-Munoz, M. D. (2018). RNA-binding proteins control gene expression and cell fate in the immune system. *Nat Immunol*, *19(2)*, *120-129*. https://doi.org/10.1038/s41590-017-0028-4

University of Florida. (2025). https://web.archive.org/web/20250211113712/https://biotech.ufl.edu/proteomics/pm-services-fees/

Valtonen, M., & Karttunen, H. (2009). The Three-Body Problem. https://doi.org/10.1017/cbo9780511616006

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

Voorhees, R. M., & Ramakrishnan, V. (2013). Structural basis of the translational elongation cycle. *Annu Rev Biochem*, *82(1)*, *203-236*. https://doi.org/10.1146/annurev-biochem-113009-092313

Wang, D., Eraslan, B., Wieland, T., Hallstrom, B., Hopf, T., et al. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol*, *15(2)*, *e8503*. https://doi.org/10.15252/msb.20188503

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2020). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9(2)*, *187-212*. https://doi.org/10.1007/s40745-020-00253-5

Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, *6(2)*, *109-118*. https://doi.org/10.1038/nrg1522

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10(1)*, *57-63*. https://doi.org/10.1038/nrg2484

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat*, *17(4)*, *263-270*. https://doi.org/10.1002/humu.22

Wek, R. C. (2018). Role of eIF2alpha Kinases in Translational Control and Adaptation to Cellular Stress. *Cold Spring Harb Perspect Biol*, *10(7)*. https://doi.org/10.1101/cshperspect.a032870

Williams, R. J., & Zipser, D. (1989). A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, *1(2)*, *270-280*. https://doi.org/10.1162/neco.1989.1.2.270

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Natural Language Processing and Chinese Computing*, *563-574*. https://doi.org/10.1007/978-3-030-32236-6_51

Yang, Q., Gilmartin, G. M., & Doublie, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proc Natl Acad Sci U S A*, *107(22)*, *10062-10067*. https://doi.org/10.1073/pnas.1000848107

Yu, F., & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. *4th International Conference on Learning Representations (ICLR 2016)*. https://doi.org/10.48550/arXiv.1511.07122

Zhang, J., Campbell, R. E., Ting, A. Y., & Tsien, R. Y. (2002). Creating new fluorescent probes for cell biology. *Nat Rev Mol Cell Biol*, *3(12)*, *906-918*. https://doi.org/10.1038/nrm976

Zhuang, Z., Shen, X., & Pan, W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics*, *35(17)*, *2899-2906*. https://doi.org/10.1093/bioinformatics/bty1050

Zrimec, J., Borlin, C. S., Buric, F., Muhammad, A. S., Chen, R., et al. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun*, *11(1)*, *6141*. https://doi.org/10.1038/s41467-020-19921-4

A. Inferring protein from transcript abundances using convolutional neural networks

Schwehn and Falter-Braun *BioData Mining* (2025) 18:18
https://doi.org/10.1186/s13040-025-00434-z

BioData Mining

RESEARCH Open Access

Inferring protein from transcript abundances using convolutional neural networks

Patrick Maximilian Schwehn[1] and Pascal Falter-Braun[1,2*]

*Correspondence:
pascal.falter-braun@helmholtz-munich.de

[1] Institute of Network Biology (INET), Molecular Targets and Therapies Center (MTTC), Helmholtz Munich, Neuherberg, Germany
[2] Microbe-Host Interactions, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

Abstract

Background: Although transcript abundance is often used as a proxy for protein abundance, it is an unreliable predictor. As proteins execute biological functions and their expression levels influence phenotypic outcomes, we developed a convolutional neural network (CNN) to predict protein abundances from mRNA abundances, protein sequence, and mRNA sequence in *Homo sapiens (H. sapiens)* and the reference plant *Arabidopsis thaliana (A. thaliana)*.

Results: After hyperparameter optimization and initial data exploration, we implemented distinct training modules for value-based and sequence-based data. By analyzing the learned weights, we revealed common and organism-specific sequence features that influence protein-to-mRNA ratios (PTRs), including known and putative sequence motifs. Adding condition-specific protein interaction information identified genes correlated with many PTRs but did not improve predictions, likely due to insufficient data. The integrated model predicted protein abundance on unseen genes with a coefficient of determination (r^2) of 0.30 in *H. sapiens* and 0.32 in *A. thaliana*.

Conclusions: For *H. sapiens,* our model improves prediction performance by nearly 50% compared to previous sequence-based approaches, and for *A. thaliana* it represents the first model of its kind. The model's learned motifs recapitulate known regulatory elements, supporting its utility in systems-level and hypothesis-driven research approaches related to protein regulation.

Keywords: Translational regulation, Protein-to-mRNA ratio, Convolutional neural networks, Regression analysis, Explainable AI

Background

For the predictive analysis of biological systems and modeling of molecular processes, it is essential to determine the context-dependent quantitative protein inventory. Nearly all biological processes, including metabolism, signaling, transport, mechanical functions, and immune responses are mediated by proteins. Hence, precise control of protein expression is fundamental for all organisms during development and to cope with environmental challenges [1, 2]. However, while mRNA concentrations can be readily measured by bulk or single-cell sequencing technologies, protein concentrations often correlate poorly with mRNA concentrations [3, 4]. Sensitivity analyses of quantitative

models [5], and evidence from genome-wide association studies, where variants altering gene expression can have major effects on signaling and disease, highlight the importance of understanding protein concentration changes [6]. Thus, accurate determination of protein concentrations in different cell types, in different conditions and of different genetic variants is critical for both mechanistic understanding and effective modeling of biological systems [7].

The experimental measurement of proteomes remains a technically challenging and costly process [8]. By contrast, obtaining systems-level transcriptomic data is both more affordable and common [9]. In the absence of direct protein abundance data, analytical and modeling approaches either rely on experimental determination of protein concentrations for the conditions of interest, which is limited by cost and throughput, or use approximations from transcript concentrations. In silico methods that more accurately model protein concentration are expected to improve the precision of computational analyses. Yet, the complexity of proteostatic regulation currently prevents the scaling of mechanism-based modeling approaches [10–15]. Recent advances in artificial intelligence and machine learning have enabled the development of quantitative predictive models for various challenging biological problems [16–21]. These achievements have been facilitated by the increasing availability of systematic large datasets for training and evaluation [22].

With matched transcriptome-proteome datasets becoming available for both *H. sapiens* [3] and *A. thaliana* [4], we sought to leverage these data to more accurately infer protein concentrations from mRNA concentrations for unseen genes. Earlier machine learning efforts for predicting protein concentrations relied on explicitly defined input features of the mRNA, such as start/stop-codon context, or the protein sequence, such as linear peptide motifs [17, 23]. To reduce assumptions and limit bias, as well as to streamline feature selection, we applied convolutional layers that learn sequence-based features directly. We experimentally optimized the CNN architecture and analyzed its learned weights to identify sequence features most influential for determining PTRs. To begin examining differences and similarities in translational regulation across large evolutionary distances, we developed our models in parallel for both *H. sapiens* and *A. thaliana*.

Methods

Datasets

We used expression data from Wang et al. [3] for *H. sapiens* and Mergner et al. [4] for *A. thaliana*. In the *H. sapiens* dataset, transcript abundances were originally normalized as fragments per kilobase million (FPKM), while in the *A. thaliana* dataset they were normalized as transcripts per million (TPM) [24]. Both studies applied a minimum threshold of 1 FPKM or 1 TPM, respectively. To ensure consistency between the datasets we converted the *H. sapiens* transcript data into TPM using the formula $TPM = FPKM / \Sigma(FPKM) * 10^6$ [25]. Proteome measurements in both datasets were given as intensity-based absolute quantification (iBAQ) [26] and were filtered using a minimum intensity threshold of 5,000. Because the *A. thaliana* data were $\log_2$-transformed, we also applied $\log_2$-transformations to the *H. sapiens* data.

For sequence data, we used the same database releases employed by the respective studies for RNA-seq and LC–MS/MS experiments, except that the *A. thaliana* untranslated regions (UTRs) were taken from a slightly larger release. Specifically, for *H. sapiens*, we obtained sequence data from Ensembl release 83 [27]. For *A. thaliana*, we used Araport11 release 2016–06 [28] for coding sequence (CDS) and the 2022–09–14 release for UTRs. All sequence data were processed by one-hot encoding and then padded to a uniform length, ensuring that each input to the model had the same dimensionality.

Machine learning

All experiments were performed in TensorFlow 2.8 [29] using default parameters unless stated otherwise. Each experiment was repeated independently five times, and for each repetition, we applied tenfold cross-validation. We optimized the models with stochastic gradient descent without momentum. Learning rates were tuned based on the type of input features: for codon counts, nucleotide counts, amino acid counts, and start-/stop-codon context, we used a learning rate of $1*10^{-3}$; for single sequence inputs such as 5' UTR, 3' UTR, CDS, and protein sequence, we used $4*10^{-4}$; and for experiments that combined multiple features, we used $1*10^{-4}$. Single-feature experiments were trained for 256 epochs, whereas combined-feature experiments ran for 512 epochs, both with a batch size of 32. We employed a custom NaN-safe mean squared error (MSE) as the loss function to accommodate the varying number of valid data points per gene.

The codon, nucleotide, and amino acid count features were each processed through a $\log_2$-transformation followed by a dense layer. The start- and stop-codon context features were passed directly to a dense layer. For the 5' UTR, 3' UTR, CDS, and protein sequences, we employed a single convolutional layer containing 16 filters. The filter size was set to 8, 10, or 12 (corresponding to the experiments denoted as 8nt, 10nt, and 12nt, respectively) and was multiplied by the dimensionality of the one-hot encoding, which was 4 for the 5' UTR, 3' UTR, and CDS, and 20 for the protein sequences. Each convolutional layer output then passed through a series of activation and pooling operations, including a tanh activation, ReLU activation, sum-pooling, and a $\log_2$-transformation, followed by a final dense layer (Fig. 2A, bottom). In the combined-feature experiments, we introduced two intermediate dense layers, the first with 32 filters and the second with 16 filters, before the final dense layer.

In all experiments, the final dense layer consisted of two filters plus a bias term. These two outputs represented the parameters a and b in the equation $\log_2(\text{iBAQ}) = a * \log_2(\text{TPM}) + b$ (Fig. 2A, top). When additional input genes were included, we expanded the final dense layer to produce (2 + the number of input genes) outputs. This enabled the equation $\log_2(\text{iBAQ}) = a * \log_2(\text{TPM}) + b + \delta * \log_2(\text{additionalTPM})$, where additionalTPM is the TPM vector of the additional input genes. The Python implementation of these models is provided in the Supplementary Source Code.

Clustering

We performed clustering of the convolutional filters using scikit-learn 1.1 [30] to identify sequence motifs. Each experiment involved standardizing the filters independently. We then applied a cutoff of 0.2 on the standard deviation within each position of the filter to identify positions with sufficient variation in nucleotides or amino acids. After

this filtering step, we duplicated, padded, and shifted each filter to generate all possible offset combinations. We clustered the resulting matrices for each feature using OPTICS [31] with a Euclidean distance. For nucleotide input features, we used an xi value of 10^{-2}, and for amino acid input features, we used $5*10^{-3}$. After clustering, we sorted the clusters from largest to smallest and removed redundant filters, ensuring each filter was only represented once despite initial duplications and shifts. For visualization, we scaled the clusters so that their largest peak had a maximum value of 1. We then centered the clusters in this peak and removed nucleotides with values ≤ 0.1.

Cross-correlation

In the cross-correlation experiment (Fig. 3D), we conducted a tenfold cross-validation with 5 independent repeats. For each repeat and fold, we computed the linear regression and MSE for every possible gene–gene combination. We then averaged these MSE values across all repeats and folds. The resulting output matrices measured 18,200 by 18,200 for *H. sapiens* and 25,285 by 25,285 for *A. thaliana*. We replaced values involving gene pairs with fewer than 21 matched data points with NaN. Subsequently, we specifically searched for pairs with NaN-safe minimum values.

Gene ontology

We performed functional enrichment analysis using the Panther web service 18.0 [32] and the Gene Ontology database, accessed in May 2023 [33]. We set the annotation to 'GO biological process complete', used Fisher's exact test as the test type, and applied a False Discovery Rate correction.

Results & discussion

Predicting protein from RNA levels can be accessed at different levels of resolution. We first explored how well the relationship between transcript level and protein level can be predicted from sequence features. We analyzed two matched transcriptome-proteome datasets spanning 29 tissues for *H. sapiens* [3] and 30 tissues for *A. thaliana* [4] (Methods). We began by examining the distribution of data points in both species (Fig. 1A). For each organism, we grouped genes based on their number of matched mRNA-protein data points into two categories: genes with 20 or more data points and those with fewer than 20. To reduce statistical artifacts and enhance robustness, we focused subsequent analyses and training on genes with at least 21 matched data points. This subset comprised 7,606 *H. sapiens* and 11,230 *A. thaliana* genes. Genes represented by exactly 20 matched data points – 205 *H. sapiens* and 336 *A. thaliana* – were reserved as a hold-out set for independent testing.

We first determined which of four commonly used regression methods best captured the relationship between mRNA and protein concentrations. Using five independent repeats of a tenfold cross-validation that excluded tissues, we evaluated prediction quality with the coefficient of determination (r^2) (Supplementary Fig. 1A). By definition, r^2 ranges from negative infinity to 1, where 0 corresponds to the mean of the target distribution – in this case, the average protein concentrations across all tissues. Linear regression produced the most reliable predictions and thus served as the baseline for

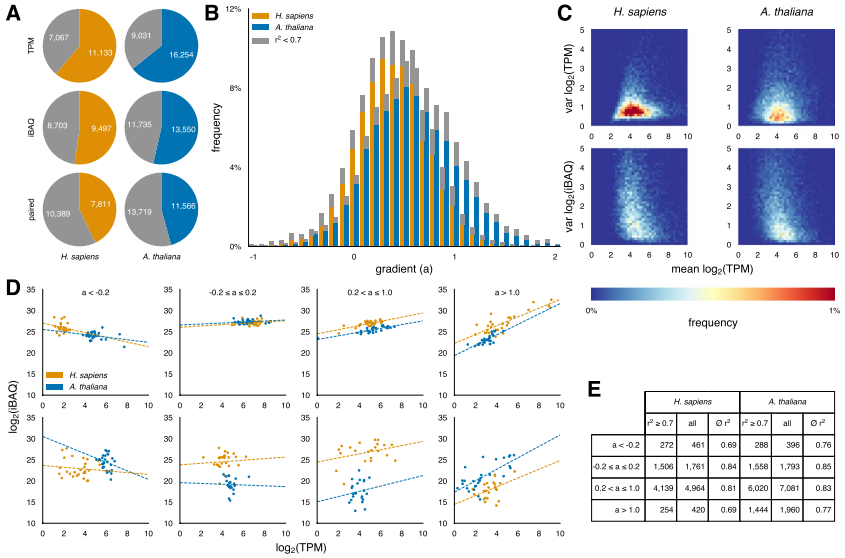


Fig. 1 Analysis of matched transcriptome-proteome datasets from *H. sapiens* and *A. thaliana*. **A** Genes grouped based on the number of valid data points: genes with ≥ 20 data points are colored, genes with fewer than 20 are shown in gray. **B** Distribution of linear regression gradients. Poor-quality regressions ($r^2 < 0.7$) are shaded in gray. **C** Average transcript abundances plotted against the variance of transcript and protein abundances (left: *H. sapiens*, right: *A. thaliana*). **D** Representative examples within the indicated gradient bin. The top panels show the 100th best-fitting gene, and the bottom panels show the 100th worst-fitting gene. **E** Histogram and average values for genes in the indicated gradient bin

subsequent analyses. More complex regressors, such as quadratic models, were less robust, a likely indication of overfitting due to limited data points.

We also examined the common assumption that changes in transcript abundances directly reflect protein abundance changes. To test this, we calculated a tissue-specific $\log_2$ iBAQ ($iBAQ_{SPEC}$) value as the product of the average $\log_2$-transformed protein abundance ($iBAQ_{AVG}$) and the ratio of tissue-specific (TPM_{SPEC}) to average (TPM_{AVG}) $\log_2$-transformed transcript abundance ($iBAQ_{SPEC} = iBAQ_{AVG} * TPM_{SPEC}/TPM_{AVG}$). This approach yielded negative r^2 values for both organisms (*H. sapiens*: $r^2 = -1.5$; *A. thaliana*: $r^2 = -3.4$), suggesting that simply normalizing protein abundances by changes in transcript abundances is inferior to using average protein abundances alone.

To investigate whether linear regression faces inherent limitations, we grouped genes based on their linear regression gradients and classified the resulting fits into two categories: good ($r^2 \geq 0.7$) and poor ($r^2 < 0.7$). The distribution of gradients was approximately Gaussian for both species and showed no systematic bias related to fit quality, as both good and poor fits were evenly represented across the range (Fig. 1B). For *H. sapiens*, this distribution was more narrowly centered, which may be due to technical differences in data processing or reflect distinct aspects of translational regulation compared to A. thaliana. Notably, we observed a subset of genes in both species that exhibited negative gradients yet still achieved high r^2 values.

Exploring how the number of matched data points affects extrapolation quality reveals that too few data points hinder the identification of consistent relationships, but once at least 20 are available, predictions become fairly robust (Supplementary

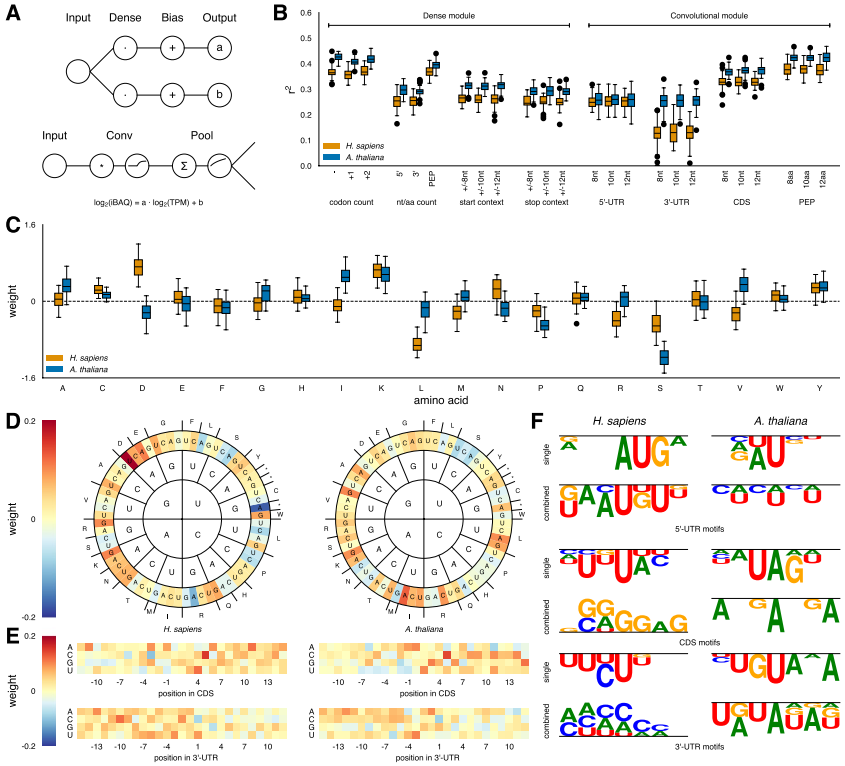
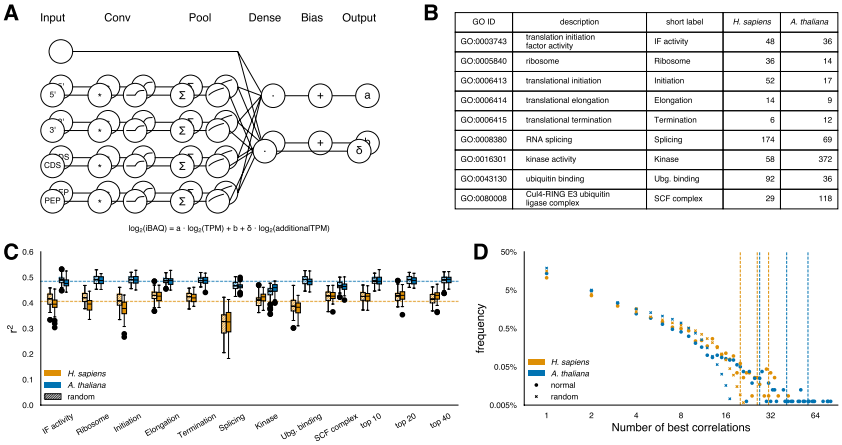


Fig. 2 Overview of sequence-based experiments. **A** Schematic representation of the two sequence-based front-end modules. **B** Predictive performance of each sequence input feature during cross-validation. Numbers indicate either the nucleotide shift of the codon count, the range of the context for start-/stop-context, or the filter size of the convolutional layers. **C** Learned weights for amino acid usage in the single-feature model. **D** Learned weights for codon usage in the combined-feature model (left: *H. sapiens*, right: *A. thaliana*). **E** Learned weights for start- and stop-codon context in the combined-feature model (top panels: start-codon context, bottom panels: stop-codon context, left: *H. sapiens*, right: *A. thaliana*). **F** Largest motif clusters identified in both the single-feature model and the combined-feature model for each sequence input feature (left: *H. sapiens*, right: *A. thaliana*)

Fig. 1B). Since the total count of matched data points depends on both the number per gene and how many genes populate each bin, excluding genes with fewer matched data points does not notably reduce the available dataset (Supplementary Fig. 1C). Overall, we used 213,444 of 256,551 (83%) available data points for *H. sapiens* and 322,197 of 379,611 (85%) for *A. thaliana* in our training and testing.

We next explored how average transcript abundances influence the variance in both transcript and protein abundances (Fig. 1C). In both *H. sapiens* and *A. thaliana*, the highest variance occurs at $\log_2$ TPM between 3 and 5. While the average $\log_2$ TPM variance is significantly lower in *H. sapiens* than in *A. thaliana* ($P < 2 * 10^{-9}$, Mann–Whitney U test), the protein variance does not differ substantially ($P = 0.25$, Mann–Whitney U test). Even though low-abundance transcripts were readily detectable, their corresponding proteins were not as easily measured at the lower end of the transcript scale, giving the variance distribution a truncated appearance (Fig. 1D, E).



GO ID	description	short label	*H. sapiens*	*A. thaliana*
GO:0003743	translation initiation factor activity	IF activity	48	36
GO:0005840	ribosome	Ribosome	36	14
GO:0006413	translational initiation	Initiation	52	17
GO:0006414	translational elongation	Elongation	14	9
GO:0006415	translational termination	Termination	6	12
GO:0008380	RNA splicing	Splicing	174	69
GO:0016301	kinase activity	Kinase	58	372
GO:0043130	ubiquitin binding	Ubq. binding	92	36
GO:0080008	Cul4-RING E3 ubiquitin ligase complex	SCF complex	29	118

Fig. 3 Overview of cell context-specific experiments. **A** Expanded sequence-based model architecture incorporating additional context-dependent input genes. **B** Histogram of manually selected GO terms of potential regulatory genes. **C** Predictive performance of each tested input feature during cross-validation. Dotted lines indicate the best-performing model from previous experiments. **D** Linear cross-correlation of all gene–gene combinations. Dotted lines indicate cutoffs for the top 10, 20, and 40 most correlated genes

For illustration, we plotted the 100th best- and worst-fitting gene within the indicated gradient bin for both species. Notably, several genes with a negative gradient ($a < -0.2$) are not statistical artefacts; instead, they show a clear trend across many data points. Although on average genes with negative gradients have the lowest coefficient of determination, more than half of them still achieve a very robust $r^2 \geq 0.7$. Surprisingly, nearly a quarter of all analyzed genes have gradients close to zero, indicating that increases in mRNA concentration do not substantially affect protein concentration under the measured steady-state conditions. About 85% of these genes display stable protein concentrations, while some poor fits resemble non-linear point clouds, suggesting more complex regulatory mechanisms at play.

Functionally, genes with negative gradient ($a < -0.2$) and robust fits ($r^2 \geq 0.7$) are associated with core cell regulatory processes in both species [32]. In *H. sapiens*, these genes are enriched in functions such as 'translation' (FDR = $2.5 * 10^{-46}$, Fisher's exact test), 'gene expression' (FDR = $6.9 * 10^{-31}$, Fisher's exact test), and 'metabolic process' (FDR = $8.6 * 10^{-24}$, Fisher's exact test). Similarly, in *A. thaliana*, the corresponding functions include 'protein metabolic process' (FDR = $4.1 * 10^{-7}$, Fisher's exact test), 'macromolecule metabolic process' (FDR = $9.6 * 10^{-7}$, Fisher's exact test), and 'intracellular transport' (FDR = $7.0 * 10^{-6}$, Fisher's exact test) (Supplementary Table 1).

To develop a linear regressor for predicting protein abundance – specifically the gradient (a) and offset (b) of the linear relationship between protein to RNA concentrations – we first sought to identify the most informative features. This helped us understand how different parts of the mRNA and the protein sequence contribute to the gradient (representing the PTR). At the same time, it allowed for individual hyperparameter optimization for each input feature.

We implemented two distinct front-end modules, each tailored to different input data types. For histogram-like and fixed-length sequence features (such as codon,

nucleotide, and amino acid counts, as well as start-/stop-codon context), we employed two parallel dense layers combined with a bias operator to predict a and b (the 'dense module') (Fig. 2A, top). For identifying linear motifs within continuous sequences (5' and 3' UTRs, coding sequences (CDS), and the translated protein sequence (PEP)), we applied a convolutional module that uses sliding windows (Fig. 2A, bottom, Methods), followed by two parallel dense layers combined with bias operators to predict a and b (Fig. 2A, top).

Figure 2B shows the predictive power of each sequence feature during cross-validation. Ultimately, we integrated the modules representing all features into a single model, optionally adding one or two additional dense layers. The combined-feature model with two dense layers achieved an average r^2 of 0.30 *for H. sapiens* and 0.32 for *A. thaliana* on independent test data. The difference in performance might stem from inconsistencies in data quality, as variations in methodology have been shown to significantly affect mRNA-protein correlations [34]. We also tested whether combining both datasets might leverage any shared patterns in translational regulation. However, training a joint model for both organisms reduced performance in each by more than two percentage points (not shown). Likewise, a minimal model that omitted potentially redundant information showed a lower r^2, indicating that all sequence features contribute valuable information (not shown).

Among the 'dense module' features, codon counts were most informative for *A. thaliana*, while amino acid counts performed best for *H. sapiens*. This finding aligns with previous studies, which identified simple sequence features, such as codon and nucleotide counts, as the most significant predictors of protein concentrations [35]. Somewhat surprisingly, out-of-frame codon counts were only slightly less predictive than in-frame codon counts. Given the strong predictive value of the amino acid counts, it seems unlikely that nucleotide identity alone, independent of coding potential, influences the PTR. A more plausible explanation is that even when the reading frame is shifted, the resulting codon counts still capture underlying trends in overall amino acid composition, albeit in a scrambled form. This interpretation is supported by the observation that a $+2$ shift in the reading frame outperforms a $+1$ shift. Shifting by $+2$ bases preserves the connection between the first two nucleotides of each codon, discarding mainly the less informative wobble position, and thus retains a closer approximation of the amino acid composition.

Among the "convolutional module" features, the coding sequence stands out as the most informative for predicting protein abundance, especially when examined in its translated form. While most features show slightly better performance in *A. thaliana*, the predictive power of 5' UTR motifs is virtually identical for both organisms. In contrast, the 3' UTRs differ substantially in their predictive utility, with human 3' UTRs proving more challenging (Fig. 2B). This may reflect the complexity introduced by the generally longer 3' UTRs in *H. sapiens*, which complicates identifying stable, informative motifs. Notably, the predictive values of the 5' and 3' UTR nucleotide counts from the dense module are indistinguishable, especially in *H. sapiens*.

To pinpoint which features drive protein abundance prediction, we extracted the learned weights for the gradient parameter (a) from both the combined and single-feature models across all repeats and folds, visualizing them as averaged heatmaps

(Fig. 2C-F, Supplementary Fig. 2A-D). In the combined-feature models, these weights are generally smaller than in the single-feature models. While the overall trends remain similar, some signals become weaker or even disappear in the combined-feature models. This suggests partial redundancy among features, such as between codon and amino acid counts, and that the final dense layer sums all weights without additional scaling. To highlight variations in codon usage and focus on the non-fixed parts of the start- and stop-codon contexts, we truncated the color scales at 1 for the single-feature model visualization.

For most features, we observed both commonalities and clear differences between *H. sapiens* and *A. thaliana*, helping to explain why modeling PTRs jointly for both species reduces predictive performance. Examining amino acid counts, we found that only a few amino acids were strongly informative, even in the single-feature module (Fig. 2C). In *H. sapiens*, the charged aspartate (D) and lysine (K) positively influence PTRs, while leucine (L) and serine (S) have a negative impact. By contrast, in *A. thaliana*, D exerts a negative effect, and isoleucine (I), which shows no effect in humans, has a strong positive impact. Overall, hydrophobic amino acids tend to contribute more positively to PTR predictions in plants than in humans. This may reflect the differing temperature ranges in which these proteins function organisms thrive – 37 °C for humans versus a variable range (4 °C to over 30 °C) for *A. thaliana* – imposing distinct biophysical constraints on protein stability and abundance.

Redundant codons also seem to enable finer tuning of PTRs. For example, both S and arginine (R) are encoded by six codons, and within these sets, certain codons increase PTR while others reduce it. In *H. sapiens*, AAG for R is positively weighted, whereas CGA is negatively weighted. Similarly, in *A.* thaliana, AGC for S is positive, while UCA reduces the PTR (Fig. 2D).

Next, we examined the sequence context of start- and stop-codons in the combined-feature model. In both organisms, the most prominent positive impact arises from a cytosine (C) at the +5-position relative to the start codon (Fig. 2E, top panels). Intriguingly, a study in yeast that investigated ribosome occupancy and translational efficiency also identified a +5 C and a weaker enrichment for uracil (U) at the +4 and +6 positions [36]. In our data, the effect of the +5 C is strong in both human and plants, with an additional contribution from U at the +4 position in humans (Fig. 2E). Furthermore, the presence of adenine (A) before the ATG codon in both species aligns with patterns observed in yeast, where efficiently translated genes also show purines (A or a guanine (G)), in the −3 position, known as the Kozak sequence (A/G)CCAUGG [37]. Our results suggest slight deviations from this canonical motif, with an A at the −1 position and a U at the +1 position exerting a stronger positive influence than the classic C and G at these positions.

Like the start context, we also see consistent patterns around the stop codon (Fig. 2E, bottom panels). In both organisms, a C at the +1 position following the stop codon and a G at the +3 position strongly reduce the predicted PTR. By contrast, substituting a U at the +1 position increases the PTR in humans.

To further understand the patterns of the convolutional filters, we clustered the learned weights from each repeat and fold, incorporating all possible shifted combinations. After clustering each experiment and species separately, we examined the

resulting motifs. Similar to the start- and stop-codon context, the convolutional module recovered numerous known motifs with well-documented biological significance and mechanisms. For example, the CUCUCU motif identified in the 5' UTRs of *A. thaliana* corresponds to a binding site for polypyrimidine tract-binding proteins, which are involved in various aspects of RNA metabolism [38]. In the 3' UTR of *A. thaliana*, the UGUA motif, visible in the single-feature model (Fig. 2F, right, panel 1), is known to bind the cleavage factor I_m [39]. Likewise, the UAUAUA motif found in *A. thaliana* 3' UTRs, was previously described in yeast as a stability regulator [40]. In humans, a G-quadruplex motif in the 5' UTR of oncogenes has been shown to modulate translation [41]. Interestingly, the strong weight assigned to this motif in the human CDS suggests that its functional impact may not be strictly confined to the 5' UTR, indicating some positional flexibility in its regulatory role.

Many of the motifs discovered by the convolutional module have been described previously, with their molecular mechanisms already established. Their recurrence here reinforces their importance and suggests that motifs not yet characterized could also be functionally relevant. However, it is crucial to consider that the extracted weights reflect patterns learned by the model rather than direct biological entities. For instance, the highest scoring 3' UTR motif found by the combined-feature model in *A. thaliana* includes both the UAUAUA motif and the UGUA motif identified by the single-feature model. In this way, the extracted weights may represent a blend of functional sequence motifs that influence PTR through diverse regulatory mechanisms.

Overall, we observe substantial variation in the predictive power of different features and only modest improvements when combining all features compared to using the single best feature alone. This finding suggests that we have largely captured the information provided by local one-dimensional sequence features. The remaining unexplained variance likely reflects more complex, context- and environment-dependent regulatory mechanisms, as well as interactions among distant regulatory elements within individual RNA molecules.

We next considered whether additional context-dependent information might improve protein abundance predictions. We modified our model architecture to compute a sequence-based vector δ, which was then used in a dot product with the mRNA concentrations of selected genes (Fig. 3A, Material and Methods). We hypothesized that these genes, potentially involved in proteostasis, or reflective of general cell state, such as kinases, could provide insight into PTR variance (Fig. 3B). As controls, we used randomly selected gene sets of the same size. Surprisingly, none of the tested gene sets improved the model's prediction in any meaningful way (Fig. 3C). Some sets slightly reduced performance, while the random controls remained unchanged.

Given these results, we asked if it was possible to identify sets of informative input genes analytically, without relying on prior knowledge. We performed linear regression for every possible gene–gene pair, using cross-validation and multiple repeats. For each target gene (output gene), we selected the input gene that best supported protein abundance prediction and recorded how often each input gene appeared as a top correlate across all targets. Plotting the resulting degree distribution (Fig. 3D), we found that while controls produced correlations for up to 18 and 24 genes in *H.*

sapiens and *A. thaliana*, respectively, several actual input genes correlated with far more target genes ($P = 1.1 * 10^{-20}$ for *H. sapiens* and $P = 1.2 * 10^{-6}$ for *A. thaliana*, Mann–Whitney U-test).

We then tested whether the top 10, 20, or 40 most correlated genes as input could improve predictions. Again, this had only minor impact compared to our previous best models (Fig. 3C). Despite the limited gains, we examined the gene ontology (GO) functions of these highly correlated (average $r^2 > 0.8$) genes. In *H. sapiens*, they were enriched for immunity-related processes, such as 'adaptive immune response' (FDR = $2.1 * 10^{-5}$, Fisher's exact test), 'regulation of immune system process' (FDR = $1.9 * 10^{-4}$, Fisher's exact test) and 'immune effector process' (FDR = $2.3 * 10^{-3}$, Fisher's exact test). By contrast, in *A. thaliana,* top functions included 'response to stimulus' (FDR = $7.2 * 10^{-7}$, Fisher's exact test) and specifically 'response to light stimulus' (FDR = $2.0 * 10^{-4}$, Fisher's exact test), suggesting that environmental factors strongly influence protein abundances (Supplementary Table 2). These results hint that the predominant challenges faced by each organism – environmental variability for plants and immune challenges for mammals – may shape the underlying regulatory landscape of protein homeostasis.

We suspect that insufficient data is limiting our ability to predict protein abundances from transcript abundances. Although we have thousands of matched transcript-protein pairs for identifying sequence features, the number of tissues is restricted to 29 for *H. sapiens* and 30 for *A. thaliana*. Even if traditional guidelines like the "rule of ten" (ten times more data points than variables) are not strictly applicable, a model built from only 29 conditions can effectively utilize at most 29 variables. In contrast, a proteins interaction network context involves roughly 20,000 other genes, and capturing condition-specific interaction network effects on protein abundance would therefore require thousands of condition-specific matched datasets – three orders of magnitude more than currently available.

Conclusions

We developed a CNN-based model to predict protein abundance and identify relevant sequence features, achieving $r^2 = 0.30$ for *H. sapiens* and $r^2 = 0.32$ for *A. thaliana*. It is important to note that our coefficient of determination was calculated independently for each gene, rather than across tissues, avoiding some of the misleading interpretations seen in previous work [42]. Despite the limitations of interaction network-based approaches, our model surpasses earlier sequence-based efforts by providing a framework that can be extended to account for alternative, previously unseen sequences, enabling more precise mRNA-dependent protein abundance predictions.

For *H. sapiens*, our model improves prediction performance by nearly 50% compared to previous sequence-based methods ($r^2 = 0.19$ [43] and $r^2 = 0.22$ [23]). Moreover, to our knowledge, our model is the first sequence-based predictor for *A. thaliana*, establishing a benchmark for plant studies. In *Saccharomyces cerevisiae*, previous attempts to predict protein abundances independent of transcript abundances have reached $r^2 = 0.48$ [44]. For *H. sapiens*, interaction network-based methods applied to cancer cells averaged Pearson correlation coefficients of $r = 0.49$ ($r^2 = 0.24$) [45], and a tailored ovarian cancer model reached $r = 0.60$ ($r^2 = 0.36$) [46]. Notably, this study reported that incorporating

the mRNA abundances of all protein interaction partners improved predictions – a factor we did not consider here.

Our combined sequence and abundance-based approach may be approaching the limit imposed by the current training data and might be susceptible to overfitting. Future work should explore integrating CNNs and other deep learning models with more extensive training data, as well as considering complex regulatory interactions beyond those examined here. Moreover, for the cross-validation genes were randomly distributed among training- and validation sets, and it is possible that homologous sequences cause data leakage resulting in slightly increased performance measures. The reidentification of experimentally validated sequence motifs suggest that not only correlated, but indeed causal sequence motifs were identified. Nonetheless, future expansion of this work will aim to eliminate this issue, e.g. as proposed [47]. At the same time, it is telling that various approaches plateau at similar performance levels, suggesting that current unknown factors and constraints will require new conceptual frameworks or additional data to achieve further improvements in predictive power.

The fact that our sequence-based model performs equally well on the more challenging, RNA-dependent PTR in multicellular *A. thaliana* underscores its robustness. It also supports the notion that approximately one-third of the variation in protein concentrations is determined by intrinsic sequence features, while the remaining two-thirds arise from higher-level regulatory mechanisms. To model these more complex factors effectively, additional training data or a better understanding of the underlying regulatory principles will be needed.

By examining the trained neural network, we identified key input features that not only improve prediction accuracy but also mirror known molecular regulatory mechanisms, such as those governing translation initiation or mRNA stability. To address the data limits in condition-specific predictions, it may be feasible to constrain gene sets based on protein–protein [5, 48, 49] or protein-RNA [50, 51] interaction networks. Such an approach could capture a reasonable subset of the regulatory landscape, thereby reducing the combinatorial complexity and improving predictive performance.

Abbreviations

A	Adenine
A. thaliana	*Arabidopsis thaliana*
C	Cytosine
CDS	Coding sequence
CNN	Convolutional neural network
D	Aspartate
FDR	False discovery rate
FPKM	Fragments per kilobase million
G	Guanine
GO	Gene ontology
H. sapiens	*Homo sapiens*
I	Isoleucine
iBAQ	Intensity based absolute quantification
K	Lysine
L	Leucine
NaN	Not-a-number
PEP	Protein sequence
PTR	Protein-to-mRNA ratio
R	Arginine
r	Pearson correlation coefficient
r^2	Coefficient of determination
S	Serine

TPM	Transcripts per million
U	Uracil
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00434-z.

Supplementary Material 1: Supplementary Fig. 1. Extended regression analysis. A Prediction scores of different regression methods. B Prediction scores of the linear regressor as a function of the number of matched data points. C Frequency distributions of genes and their matched data points.

Supplementary Material 2: Supplementary Fig. 2. Extended overview of sequence-based experiments. A Learned weights for amino acid usage in the combined-feature model. B Learned weights for codon usage in the single-feature model (left: H. sapiens, right: A. thaliana). C Learned weights for start- and stop-codon context in the single-feature model (top panels: start-codon context, bottom panels: stop-codon context, left: H. sapiens, right: A. thaliana). D Second largest motif clusters identified in both the single-feature model and the combined-feature model for each sequence input feature (left: H. sapiens, right: A. thaliana).

Supplementary Material 3: Supplementary Table 1. Functional analysis of genes with a negative gradient ($a < -0.2$) and robust fits ($r2 \geq 0.7$). A Raw linear regressions values for H. sapiens. B Raw linear regressions values for A. thaliana. C Gene Ontology enrichment analysis for H. sapiens. D Gene Ontology enrichment analysis for A. thaliana.

Supplementary Material 4: Supplementary Table 2. Functional analysis of genes with good average fits ($r2 \geq 0.8$). A Raw linear regressions values for H. sapiens. B Raw linear regressions values for A. thaliana. C Gene Ontology enrichment analysis for H. sapiens. D Gene Ontology enrichment analysis for A. thaliana.

Supplementary Material 5: Supplementary Table 3. Calculated linear parameters and predictions for: A H. sapiens. B A. thaliana.

Supplementary Material 6: Source Code.

Acknowledgements
Not applicable.

Authors' contributions
Conceptual idea: PFB; experimental design & interpretation: PS, PFB; implementation, coding: PS; figures: PS; manuscript writing: PS, PFB. All authors reviewed the manuscript and agree with publication.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported by HDHL-INTIMIC "Interrelation of the Intestinal Microbiome, Diet and Health" (BMBF 01EA1803 to P.F.-B.), the European Union's Horizon 2020 Research and Innovation Programme (Project ID 101003633, RiPCoN; Project ID 101137201, CLARITY (P.F.-B.); and the Free State of Bavaria's AI for Therapy (AI4T) Initiative through the Institute of AI for Drug Discovery (AID) (P.F.-B.). This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 April 2024 Accepted: 14 February 2025
Published online: 27 February 2025

References

1. Merchante C, Stepanova AN, Alonso JM. Translation regulation in plants: an interesting past, an exciting present and a promising future. Plant J. 2017;90(4):628–53.
2. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020;21(10):630–44.
3. Wang D, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol. 2019;15(2):e8503.

4. Mergner J, et al. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature. 2020;579(7799):409–14.
5. Yu H, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008;322(5898):104–10.
6. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
7. Aldridge BB, et al. Physicochemical modelling of cell signalling pathways. Nat Cell Biol. 2006;8(11):1195–203.
8. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
10. Paik I, Yang S, Choi G. Phytochrome regulates translation of mRNA in the cytosol. Proc Natl Acad Sci U S A. 2012;109(4):1335–40.
11. Shalgi R, et al. Widespread regulation of translation by elongation pausing in heat shock. Mol Cell. 2013;49(3):439–52.
12. Barbosa C, Peixeiro I, Romao L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. 2013;9(8):e1003529.
13. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016;352(6292):1413–6.
14. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 2018;19(1):20–30.
15. Wek RC. Role of eIF2alpha kinases in translational control and adaptation to cellular stress. Cold Spring Harb Perspect Biol. 2018;10(7):a032870.
16. Alipanahi B, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
17. Munusamy P, et al. De novo computational identification of stress-related sequence motifs and microRNA target sites in untranslated regions of a plant translatome. Sci Rep. 2017;7(1):43861.
18. Cuperus JT, et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. 2017;27(12):2015–24.
19. Zhuang Z, Shen X, Pan W. A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. Bioinformatics. 2019;35(17):2899–906.
20. Eraslan G, et al. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
21. Zrimec J, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11(1):6141.
22. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13(4):227–32.
23. Eraslan B, et al. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol Syst Biol. 2019;15(2):e8513.
24. Conesa A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13.
25. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281–5.
26. Schwanhausser B, et al. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337–42.
27. Martin FJ, et al. Ensembl 2023. Nucleic Acids Res. 2023;51(D1):D933–41.
28. Cheng CY, et al. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89(4):789–804.
29. Abadi M, et al. TensorFlow: a system for large-scale machine learning. OSDI. 2016;16:265–83.
30. Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
31. Ankerst M, et al. Optics. ACM SIGMOD Rec. 1999;28(2):49–60.
32. Thomas PD, et al. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022;31(1):8–22.
33. Carbon S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;5(2):288–9.
34. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165(3):535–50.
35. Vogel C, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010;6(1):400.
36. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol. 2011;7(1):481.
37. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986;44(2):283–92.
38. Oberstrass FC, et al. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science. 2005;309(5743):2054–7.
39. Yang Q, Gilmartin GM, Doublie S. Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3′ processing. Proc Natl Acad Sci U S A. 2010;107(22):10062–7.
40. Savinov A, et al. Effects of sequence motifs in the yeast 3′ untranslated region determined from massively parallel assays of random sequences. Genome Biol. 2021;22(1):293.
41. Kumari S, et al. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. Nat Chem Biol. 2007;3(4):218–21.
42. Fortelny N, et al. Can we predict protein from mRNA levels? Nature. 2017;547(7664):E19–20.
43. Hebditch M, et al. Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics. 2017;33(19):3098–100.
44. Ferreira M, et al. Protein abundance prediction through machine learning methods. J Mol Biol. 2021;433(22):167267.
45. Li H, et al. Joint learning improves protein abundance prediction in cancers. BMC Biol. 2019;17(1):107.
46. Srivastava H, et al. Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners. PLoS Comput Biol. 2022;18(11):e1010702.

47. Ferrer Florensa A, et al. SpanSeq: similarity-based sequence data splitting method for improved development and assessment of deep learning projects. NAR Genom Bioinform. 2024;6(3):lqae106.
48. Michaelis AC, et al. The social and structural architecture of the yeast protein interactome. Nature. 2023;624(7990):192–200.
49. Lambourne L, et al. Binary interactome models of inner- versus outer-complexome organisation. bioRxiv. 2022:2021.03.16.435663.
50. Brannan KW, et al. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. Nat Methods. 2021;18(5):507–19.
51. Shchepachev V, et al. Defining the RNA interactome by total RNA-associated protein purification. Mol Syst Biol. 2019;15(4):e8689.

Publisher's Note

B. A proteome-scale map of the SARS-CoV-2–human contactome

nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01475-z

A proteome-scale map of the SARS-CoV-2–human contactome

Received: 23 February 2022

Accepted: 15 August 2022

Published online: 10 October 2022

Check for updates

A full list of authors and their affiliations appears at the end of the paper.

Understanding the mechanisms of coronavirus disease 2019 (COVID-19) disease severity to efficiently design therapies for emerging virus variants remains an urgent challenge of the ongoing pandemic. Infection and immune reactions are mediated by direct contacts between viral molecules and the host proteome, and the vast majority of these virus–host contacts (the 'contactome') have not been identified. Here, we present a systematic contactome map of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with the human host encompassing more than 200 binary virus–host and intraviral protein–protein interactions. We find that host proteins genetically associated with comorbidities of severe illness and long COVID are enriched in SARS-CoV-2 targeted network communities. Evaluating contactome-derived hypotheses, we demonstrate that viral NSP14 activates nuclear factor κB (NF-κB)-dependent transcription, even in the presence of cytokine signaling. Moreover, for several tested host proteins, genetic knock-down substantially reduces viral replication. Additionally, we show for USP25 that this effect is phenocopied by the small-molecule inhibitor AZ1. Our results connect viral proteins to human genetic architecture for COVID-19 severity and offer potential therapeutic targets.

Despite over 200,000 SARS-CoV-2 publications in the past two years, fundamental questions remain about the molecular mechanisms of genetic risk factors for severe and fatal COVID-19, the cause of long-persisting disease symptoms (long COVID) and the challenge to identify therapeutic targets[1]. These issues remain urgent in light of incomplete vaccination rates, continuously emerging variants and anticipated future pathogens. Fundamentally, infections are initiated by physical contacts between viral proteins and cellular receptors that set off molecular rearrangements culminating in viral entry and unpacking, followed by cellular reprogramming and host defense response triggering. Each of these steps is mediated by contacts between viral and host molecules that determine functional consequences, including proteolytic cleavage or inflammatory signaling, and ultimately clinical manifestations (Fig. 1a). Therefore, understanding the mechanisms by which human genetic variation affects COVID-19, as well as the behavior of newly emerging virus variants such as Delta (δ) and Omicron (o), requires knowledge of these contacts to enable studies on how variants functionally alter virus–host interactions. For SARS-CoV-2, the contacts between the viral spike and human ACE2 proteins are documented by several hundred structures. In contrast, no direct interaction partners are known for many other viral proteins, precluding even domain-level contact models. Because co-complex assays predominantly detect indirect protein-associations[2], the virus–host contactome remains largely unexplored and unknown. To address this fundamental research gap, we systematically identified protein–protein contacts between SARS-CoV-2 and the human proteome.

Results

SARS-CoV-2–host contactome mapping

We used a multiassay screening and evaluation framework to generate a high-quality virus–host contactome map[2,3]. To increase detection sensitivity in the initial screening by yeast two hybrid (Y2H), we

e-mail: caroline.demeret@pasteur.fr; marc_vidal@dfci.harvard.edu; michael_calderwood@dfci.harvard.edu; fritz.roth@utoronto.ca; pascal.falter-braun@helmholtz-muenchen.de

https://doi.org/10.1038/s41587-022-01475-z

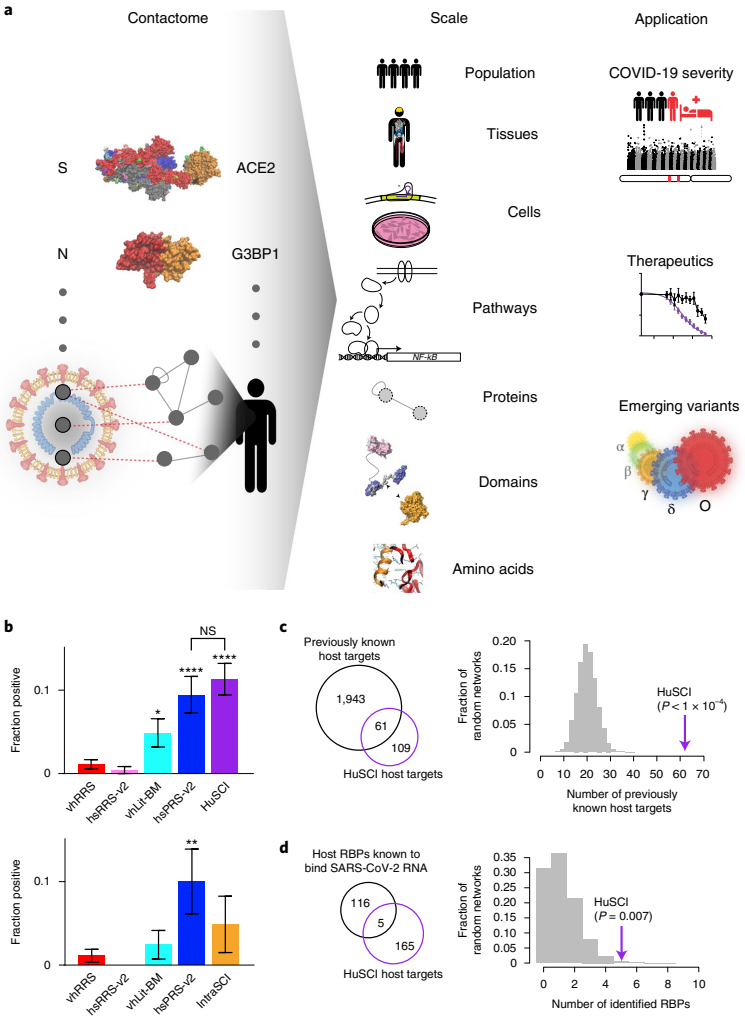


Fig. 1 | Generation and quality assessment of HuSCI. a, The contactome, the sum of physical contacts between viral and host macromolecules, mediates cellular perturbations that enable viral replication and cause disease manifestations. N, nucleocapsid protein; S, spike protein. **b**, Orthogonal validation; fraction of pairs that are yN2H-positive in HuSCI (top, $n = 282$ pair configurations representing 148 HuSCIs interaction pairs) and IntraSCI (bottom, $n = 41$ pair configurations for 25 IntraSCI interaction pairs), in the benchmark positive control sets hsPRS-v2 ($n = 180$ pair configurations for 60 interaction pairs) and vhLit-BM ($n = 164$ pair configurations for 40 interaction pairs), and negative control sets hsRRS-v2 ($n = 234$ pair configurations for 78 protein pairs) and vhRRS ($n = 360$ pair configurations for 178 protein pairs). Asterisks indicate significant differences from vhRRS benchmark (*$P = 0.023$; **$P = 0.005$; ****$P = 1.57 \times 10^{-5}$ hsPRS-v2, $P = 1.02 \times 10^{-7}$ HuSCI; NS, not significant; two-sided Fisher's exact test; center, proportion of positives; error bars, standard error of proportion). Precise P values for all dataset pairs, biological repeats and n for each test are shown in Supplementary Table 3. **c**, Overlap of SARS-CoV-2 targets identified in HuSCI with previously identified target proteins of other viruses (left) and actual overlap (arrow) compared to $n = 10{,}000$ randomized control networks (right) (one-sided, empirical $P < 0.0001$). **d**, Host targets identified in HuSCI overlap with RNA-binding proteins (RBPs) bound to SARS-CoV-2 RNA upon infection (left) and actual overlap (arrow) compared to $n = 10{,}000$ randomized control networks (right) (one-sided, empirical $P = 0.007$).

used two complementary assay versions (Extended Data Fig. 1a): (1) a plate-based version using 'bait' and 'prey' N-terminal fusion proteins encoded on low-copy plasmids and *GAL1-HIS3*-based growth selection ($Y2H_{HIS3}$)[2,3], and (2) a new system based on the Barcode Fusion Genetics (BFG)-Y2H technology[4], using a C-terminal fusion prey protein encoded from a high-copy plasmid and selecting cells expressing green fluorescent protein (*GAL1-GFP*) from a pooled liquid culture ($Y2H_{GFP}$) (unpublished). Using $Y2H_{HIS3}$, 26 viral open reading frames (vORFs; Supplementary Table 1) were screened against 17,472 human ORFs (covering 83% of all pairings of human and viral protein-coding genes, that is 83% 'search space completeness') in both orientations; that is, as bait and prey (Extended Data Fig. 1a). Human candidate interactors were pairwise retested in triplicate against every vORF, yielding 118 interactions involving 14 viral and 92 human proteins. We refer to this

$Y2H_{HIS3}$-based human SARS-CoV-2 interactome dataset as $HuSCI_{HIS3}$. Using $Y2H_{GFP}$, 28 vORFs were screened against 14,627 human ORFs (70% completeness) (Extended Data Fig. 1a). After stringent filtering and *HIS3*-based verification, this yielded 93 interactions involving 13 viral and 84 human host proteins. We refer to this dataset as $HuSCI_{GFP}$ and to the union with $HuSCI_{HIS3}$ as HuSCI (Supplementary Table 1). We also carried out a targeted screen with previously identified SARS-CoV-1 host interactors; of the 62 testable orthologous SARS-CoV-2–human pairs, six were found to interact ($HuSCI_{ORTH}$) (Supplementary Table 2). $Y2H_{GFP}$ also yielded an intraviral SARS-CoV-2 interactome of 25 binary interactions among 19 vORFs (IntraSCI; Supplementary Table 1). Having collectively identified a contactome of 204 direct virus–host and 25 intraviral interactions among 170 host and 19 viral proteins, we next assessed data quality.

Seven interactions were identified in both $HuSCI_{GFP}$ and $HuSCI_{HIS3}$. Albeit nominally low, this overlap is consistent with the complementary nature of the assays and pipelines. Specifically, the screens interrogated incompletely overlapping protein sets and were each 50%–60% saturated. Each version used for screening has an assay sensitivity of 20%–25%[5] (fraction of detectable interactions); thus, the overlap is consistent with known screening parameters[2] and a low false-discovery rate. Moreover, from these parameters we can estimate that HuSCI covers 15%–22% of the complete contactome between SARS-CoV-2 and host proteins (Methods).

To further assess data quality experimentally, we compared detection rates of our datasets in the yeast-based nanoluciferase complementation assay (yN2H)[6] to those of established human positive and random reference sets (hsPRS-v2 and hsRRS-v2)[5,6]. As additional benchmarks, we derived a set of 55 well-documented binary interactions between human and coronavirus proteins from the curated literature (virus–host literature binary multiple reference set; vhLit-BM) and a virus–host random reference set (vhRRS) (Supplementary Table 3). We tested HuSCI, IntraSCI and each benchmark set by yN2H (Fig. 1b and Extended Data Fig. 1b). At a stringent scoring threshold of 1% vhRRS, the validation rates of both HuSCI alone and the union of HuSCI with IntraSCI (UnionSCI) were statistically indistinguishable from the two positive control sets (hsPRS-v2, $P = 0.76$; vhLit-BM, $P = 0.06$; Fisher's exact test versus UnionSCI), and each was significantly higher than those of the negative control sets (hsRRS-v2, $P = 4 \times 10^{-7}$; vhRRS, $P = 1 \times 10^{-7}$; Fisher's exact test versus UnionSCI; Fig. 1b and Supplementary Table 3). Thus, the biophysical quality of our virus–host contactome map is at least on par with high-quality interactions supported by multiple experiments in the curated literature. Although IntraSCI is too small for a separate evaluation by yN2H, 5 of 25 interactions overlap with a previous study[7] ($P = 4.6 \times 10^{-3}$, empirical test; Extended Data Fig. 1c).

The biological relevance of our virus–host contactome map is suggested by the observations that the identified host proteins are enriched for (1) known targets of other viruses[8] ($P < 1 \times 10^{-4}$, empirical test; Fig. 1c), (2) proteins that change phosphorylation status upon SARS-CoV-2 infection[9,10] ($P < 1 \times 10^{-4}$, empirical test; Extended Data Fig. 1d), (3) proteins that directly interact with SARS-CoV-2 RNA[11] ($P = 0.007$, empirical test; Fig. 1d) and (4) proteins that change RNA-binding status during SARS-CoV-2 infection[11] ($P = 0.022$, empirical test; Extended Data Fig. 1e). These results demonstrate that IntraSCI and HuSCI (Fig. 2a) are of high biophysical quality and enriched for host proteins relevant to SARS-CoV-2 biology.

Complementarity of contactome and co-complex datasets

Previous studies investigating host and SARS-CoV-2 proteins used either affinity purification followed by mass spectrometry (AP-MS) to identify co-complex associations[9,12–15] or biotin identification (BioID) to find proteins in spatial proximity[16–18]. However, co-complex maps capture largely indirect associations in stable complexes that persist through affinity purification[2] and, likely due to experimental differences, the datasets exhibit limited agreement among each other (Extended Data Fig. 2a). For a subset of such co-complex associations, contacts can be computationally modeled[19]. In contrast, binary interactome maps provide direct contact partners and are enriched for regulatory interactions[2]. Despite these differences, 20 of the 204 HuSCI-interacting pairs were found in co-complex and BioID studies, and 58 (34%) of the 170 HuSCI host proteins were associated with a SARS-CoV-2 protein by these studies (Supplementary Table 1). Thus, the contactome map is consistent with previous indirect association datasets while providing substantial novelty.

Although SARS-CoV-2 primarily infects lung and airway tissues, it can spread to additional tissues and this expanded tropism is characteristic for COVID-19 and important for long COVID symptoms[20]. As previous SARS-CoV-2 interaction datasets could only detect host proteins expressed in the specific assay cell lines, we wondered whether HuSCI was also complementary in terms of the tissue specificity of identified host proteins. Using the Human Protein Atlas (HPA)[21], we defined 'tissue-specific' and 'common' human proteins (Supplementary Table 4). Whereas the AP-MS and BioID data are biased toward common host proteins, HuSCI is more representative of the human proteome and shows good coverage of proteins expressed in the diverse tissues in which SARS-CoV-2 RNA has been detected[22] (Fig. 2b, Extended Data Fig. 2b,c and Supplementary Table 4). Thus, the HuSCI contactome has unique advantages for understanding tissue-specific perturbations by SARS-CoV-2.

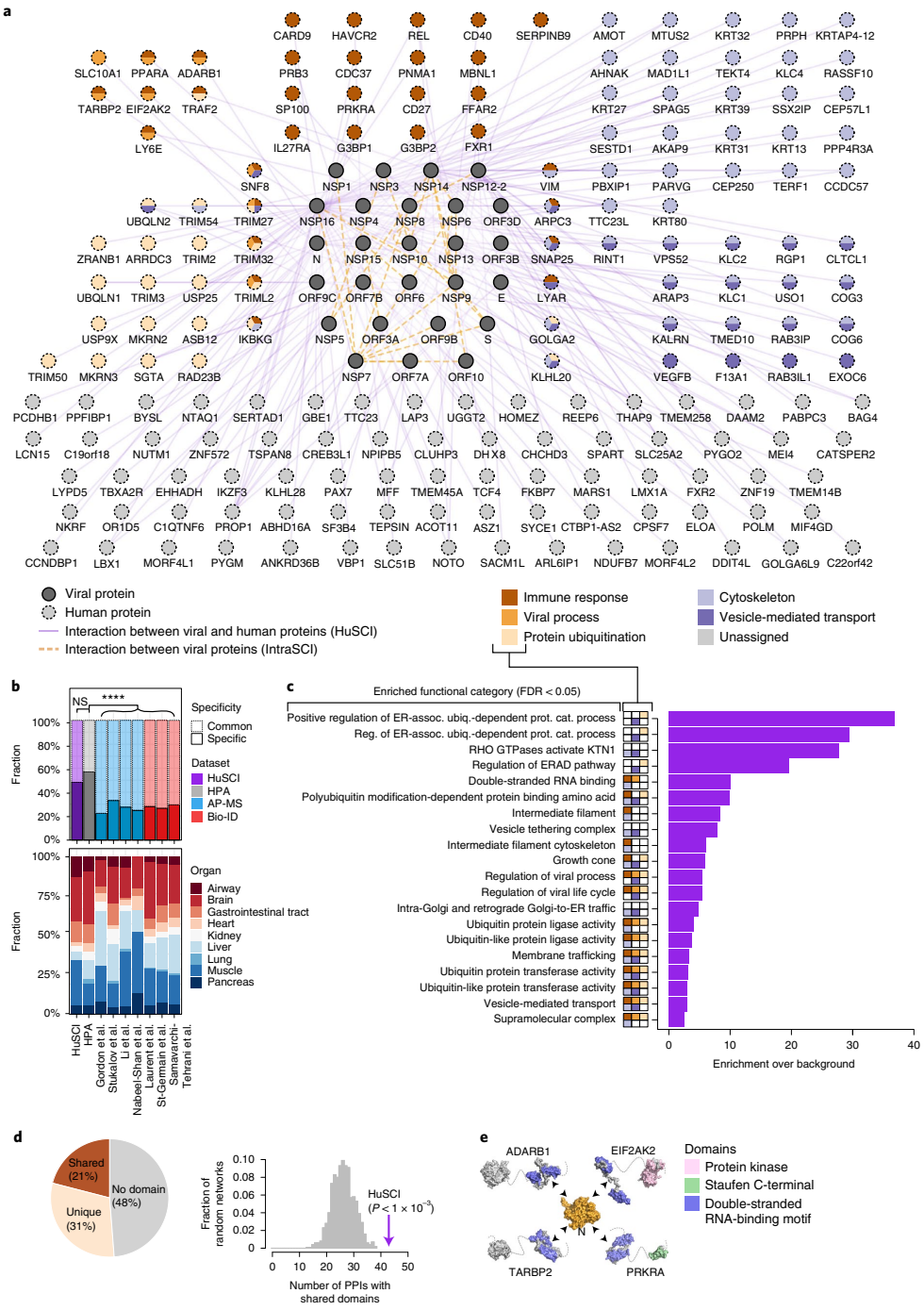
SARS-CoV-2 targeted functions

To understand which host functions are directly perturbed by the virus, we investigated SARS-CoV-2 targeted pathways. Broad functions enriched among host proteins include (1) immune response, (2) viral process, (3) protein ubiquitination, (4) cytoskeleton and (5) vesicle-mediated transport (Fig. 2c). These largely agree with functions identified in association and proximity datasets[9,12–18] (Supplementary Table 5). Focusing on immune pathways, we noticed that NSP9, NSP14 and NSP16 contact key regulators of cytokine production such as REL (c-REL proto-oncogene, NF-κB subunit), IKBKG (inhibitor of NF-κB kinase regulatory subunit gamma, also known as IKKγ or NEMO) and TRAF2 (tumor necrosis factor (TNF) receptor-associated factor 2). HuSCI interactors of the membrane-spanning NSP6 were enriched

Fig. 2 | Network representation and functional assessment of HuSCI. **a**, Combined HuSCI and IntraSCI networks. Node colors of human proteins represent broad enriched functions as indicated in legend. Node labels for human proteins correspond to approved HGNC symbols; accession identifiers and descriptions are listed in Supplementary Table 1. **b**, Proportion of host targets in common and specific expression groups in all (top) and in SARS-CoV-2 RNA-positive organs (bottom) across eight datasets: purple, HuSCI; gray, HPA[21]; blue, AP-MS datasets from Gordon et al.[12,13], Stukalov et al.[9], Li et al.[14] and Nabeel-Shah et al.[15]; red, BioID datasets from Laurent et al.[16], St-Germain et al.[17] and Samavarchi-Tehrani et al.[18]. Two-sided Fisher's exact test, Bonferroni adjusted $P < 0.0001$. Full statistical details and exact P values are listed in Supplementary Table 4. **c**, Functions enriched among host proteins found in HuSCI ($P = 0.05$, Fisher's exact test with FDR correction). Broad functional groups are indicated in small boxes according to legend in panel a. Full statistical details are listed in Supplementary Table 5. ER-assoc. ubiq.-dependent prot. cat., Endoplasmic reticulum-associated ubiquitin-dependent protein catabolic. **d**, Proportion of virus–host interactions in which the human protein has domains that are present in other interactors of the viral protein (shared), not present in other interactors of the viral protein (unique) or no annotated domains (left) and number of shared-domain interactions in HuSCI (arrow) compared to $n = 1{,}000$ randomized control networks (gray distribution) (right). One-sided empirical $P < 0.001$. PPI, protein-protein interactions. **e**, Exemplary 'shared-domain interaction' between the viral nucleocapsid protein and four interactors containing a double-stranded RNA-binding motif. Domain colors according to legend; gray parts lack domain annotations.

https://doi.org/10.1038/s41587-022-01475-z

for immune receptors ($P < 0.01$, empirical test), including CD40 and IL27RA (IL27 receptor subunit alpha). Intriguingly, NSP6 also directly interacts with LY6E, a host restriction factor that limits viral entry for SARS-CoV-2 and other coronaviruses[23]. Several other targets are RNA-binding proteins that function in innate immunity and response to viral infection[24]. MKRN2, together with G3BP1/2, has been suggested



to regulate olfactory signaling mRNAs[25], pointing to potential mechanistic links underlying anosmia in COVID-19. Thus, direct SARS-CoV-2 protein interactors function in immune pathways and viral processes relevant to COVID-19.

Viral proteins contact shared host-protein domains

The restricted size of viral genomes limits their coding potential. We therefore wondered to what extent this limitation yielded viral proteins that bind multiple human proteins via target-shared domains, thus offering opportunities for structure-based drug discovery. We sought domains shared by multiple human targets of each viral protein. In the contactome, SARS-CoV-2 proteins engaged in 43 interactions involving such shared domains (21% of HuSCI; $P < 0.001$, empirical test; Fig. 2d, Extended Data Fig. 2d and Supplementary Table 6), corresponding to 22 significant virus protein-domain pairs ($P < 0.001$, empirical test; Supplementary Table 6). Although the difference was not statistically significant, the 21% proportion of the virus–host contactome with shared-domain interactions was numerically higher than the corresponding 17% in the human reference interactome network (HuRI)[26]. Specific examples in HuSCI include four interactors of the nucleocapsid protein sharing the double-stranded RNA-binding motif ($P < 0.05$, Fisher's test; Fig. 2e) and the recently confirmed finding that viral nucleocapsid protein interacts with the NTF2 domains of G3BP1 and G3BP2[27]. Disease-causing mutations are located in the interaction interfaces of the enriched domains of several human proteins (for example, TNF receptor domain of CD40 (ref. [28]) or zf-CCCH in MKRN3 (ref. [29])). Thus, recurrent structural themes may reflect binding mechanisms that are subject to modulation by human coding variants affecting infection outcome[30,31] or by rationally designed drugs.

HuSCI links to COVID-19 risk loci

The severity of COVID-19 symptoms and outcomes are highly variable, and understanding the underlying molecular mechanisms may enable effective treatments. Recently, two independent meta-studies identified genetic loci that are associated with severe COVID-19 illness[32,33] (Fig. 3a and Extended Data Fig. 3a), but mechanistic links to viral infection remain unknown. Similarly, several preconditions increase the risk of severe COVID-19, but for these, the molecular links are also poorly understood. At least two models can help to conceptualize how this genetic variation relates to virally targeted host proteins. In a 'direct' model, genetic variation in targeted host proteins modulates disease outcome, exemplified by the interaction of adenovirus E1A oncoprotein with the tumor suppressor protein pRb[34]. In an alternative 'indirect' model, genetic variation in the network neighborhood of targeted host proteins modulates the downstream effects and thereby influences disease outcome. A precedent for this model was observed in a plant system, where pathogen-targeted host proteins tend to interact with proteins relevant to disease severity and fitness (encoded by highly variable genes under balancing selection)[35]. The availability of a high-quality contactome map enabled us to address this fundamental question for COVID-19. Because bias toward well-studied proteins in the SARS-CoV-2 literature[36] (Fig. 3b and Extended Data Fig. 3b) limits mechanistic understanding and can cause artifacts, we focused our analyses on systematic protein interaction datasets. The direct model was not supported, given that no targeted host protein from HuSCI was encoded from a critical illness associated locus[32,33] ('critical illness proteins'), and only one (HLA-G, associated with ORF3) was found in a single co-complex study[9]. Investigating the indirect model, we sought contacts between targeted host proteins and critical illness proteins, finding 20 ($P = 0.002$, empirical test; Fig. 3c)[32] and 8 ($P = 0.012$, empirical test; Extended Data Fig. 3a, c)[33] in the binary HuRI host network map. In contrast, the virus-associated host-protein sets from AP-MS studies[9,12,13] interact with no more critical illness proteins than expected by chance (Extended Data Fig. 3d). Functionally, the HuSCI host-target proteins linking critical illness to SARS-CoV-2 proteins are enriched in microtubule organization, membrane trafficking and TNF signaling annotations (Supplementary Table 7). Intriguingly, three of seven direct OAS1 interactors are targeted by NSP14 and NSP16, and all three have Golgi- and membrane trafficking-related functions, providing protein contacts that support the finding that the Neanderthal-derived protective OAS1 variant promotes degradation of viral RNA in endoplasmic reticulum- and Golgi-derived virus replication organelles[37]. These observations indicate that, consistent with the indirect model, clinically relevant genetic variation acts in the local network neighborhood of viral contact proteins.

To further explore the local subnetworks surrounding targeted host proteins and their links to human genetic variation, we identified 204 subnetwork communities in HuRI[26] (Fig. 3d) that were significantly targeted by SARS-CoV-2 (nominal $P < 0.05$, Fisher's exact test; Supplementary Table 8). Examples include community 28, enriched for 'negative regulation of viral transcription' (false discovery rate (FDR) = 0.0018; Fig. 3d) and community 52, enriched for 'Arp2/3 complex-mediated actin nucleation' (FDR = 0.0002; Supplementary Table 8). The Arp2/3 complex enables human respiratory viruses to spread among adjacent cells without forming virions[38], and ARPC3 scored among the top 50 in two CRISPR screens for SARS-CoV-2 host factors[39,40]. We then asked whether direct viral target proteins and proteins in each community are encoded by genes associated with human traits of 114 uniformly processed genome-wide association studies (GWASs)[41]. Variation in genes encoding direct viral targets was only associated with 'depression' (FDR = 0.03, MAGMA). In contrast, among the communities, genetic variation associated with severe COVID-19 illness was associated with ten virus-targeted communities, more communities than any other human trait. In contrast, host-protein sets from AP-MS studies were enriched in fewer communities (nominal $P < 0.05$, Fisher's exact test; Extended Data Fig. 3e and Supplementary Table 8), and only one host-protein-enriched community each from two AP-MS datasets was enriched for genetic variation associated with severe COVID-19 (refs. [13,14]) (Li et al. community 14 and Gordon et al. community 11; Extended Data Fig. 3f). Intriguingly, of the 14 human traits (from 15 studies) associated with 20 additional HuSCI-target-enriched communities, 8 traits (from 9 studies) are clinical risk factors for severe COVID-19 and long COVID, including high body mass index (BMI)[42], hypothyroidism[43] and schizophrenia[44] ($P = 0.01$, Fisher's exact test; Fig. 3d, Extended Data Fig. 3e,f and Supplementary Table 8). These links between viral targets and genetic variation associated with COVID-19 comorbidities open the possibility that this genetic variation may impact the course of infection and severity of COVID-19 independent of trait manifestation. Other traits associated with host-target-enriched communities, such as neuroticism, have not been linked to COVID-19 symptoms, possibly because the genetic influence is masked by confounding parameters such as behavior[45], and should be considered in the future. Together, these results suggest that the HuSCI contactome map is a powerful and unique resource for studying molecular mechanisms by which human genetics affect the outcome of SARS-CoV-2 infection.

Validation of pathways and host targets

We next explored specific hypotheses for viral proteins and human target functions. Both literature reports and our analyses suggest a role for NF-κB immune signaling in SARS-CoV-2 infection. Because we observed multiple interactions of viral proteins with different members of the NF-κB signaling pathway, we used reporter assays to determine whether and in which direction (that is, activating or inhibiting) viral factors modulate pathway activity. Transfection of NSP14, which interacts with multiple positive NF-κB regulators, resulted in dose-dependent transcriptional activation of NF-κB and even further augmented NF-κB activity following proinflammatory TNF-α stimulation in HEK293 cells (Fig. 4a,b, Extended Data Fig. 4a,b and Supplementary Table 9). This finding suggests that SARS-CoV-2 can induce a proinflammatory state during COVID-19 via direct interaction of NSP14 with NF-κB activators.

https://doi.org/10.1038/s41587-022-01475-z

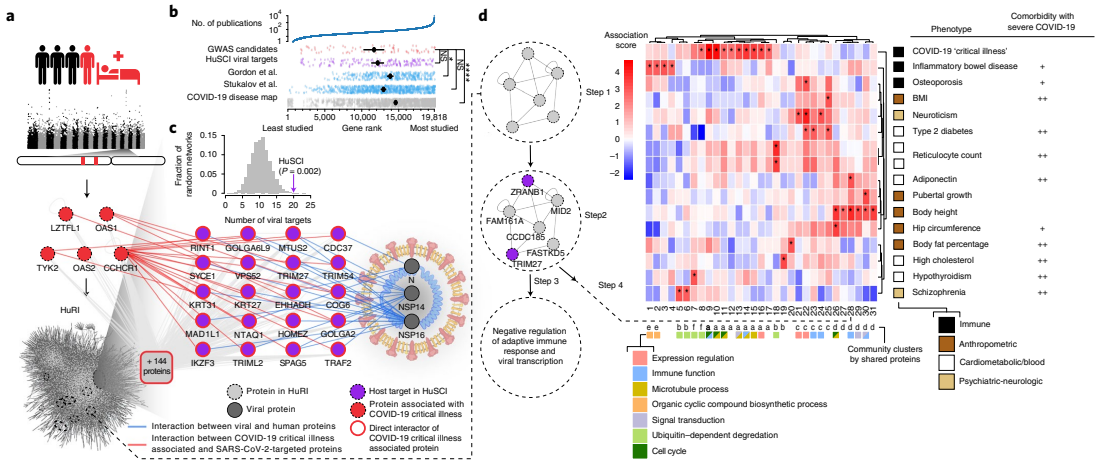


Fig. 3 | HuSCI host targets link to genetic variation for severe COVID-19. **a**, HuRI interactors (gray nodes) of COVID-19 'critical illness proteins' loci (seed, red nodes) include a significant ($P = 0.0009$, empirical test) number of direct SARS-CoV-2 targeted proteins (purple nodes). A total of 144 additional seed protein interactors are not resolved individually. Node and edge colors according to legend. **b**, Genes in indicated COVID-19 datasets ranked across the human genome by number of publications. Number of publications is indicated by the top panel on log scale. Asterisks indicate significant differences relative to COVID-19-associated genes[32,33]; NS, $P = 1$ (HuSCI) and $P = 0.36$ (Stuckalov et al.); *$P = 0.047$; ****$P = 0.000014$, two-sided Mann–Whitney U test, Bonferroni correction, from top to bottom: $n = 45, 170, 383, 876$ and 849; error bars are 95% confidence intervals of the mean, calculated by 1,000 bootstrap samples. **c**, Virus-interactor enrichment: number of direct SARS-CoV-2 protein interacting HuSCI proteins in the subnetwork formed by proteins encoded by seed proteins[32,33] and their first-level interactors (arrow) compared to $n = 10{,}000$ randomized control networks (gray distribution). **d**, Of 3,603 communities in Human Reference Interactome (HuRI) with ≥4 members (step 1), 204 are significantly targeted by SARS-CoV-2 (two-sided nominal $P < 0.05$; Fisher's exact test) (step 2); Gene Ontology (GO) enrichment identifies functions associated with each community (step 3); and MAGMA identifies 31 communities significantly associated with human traits (FDR < 0.05) (step 4), the great majority of which are COVID-19 comorbidities. Example community 28 is significantly targeted by SARS-CoV-2 in HuSCI (two-sided $P = 0.0078$; Fisher's exact test, uncorrected) and enriched for negative regulation of adaptive immune response and viral transcription. Functional descriptors in squared boxes according to legend (Supplementary Table 8); relation of indicated traits to COVID-19 is indicated in rightmost column as general link (+) (e.g., via immunity) and clinical evidence for modulation of diseases symptoms and risk for severe or long COVID (+ + ; Extended Data Fig. 3f). BMI, body mass index.

These results are corroborated by a study that implicates IMPDH2 in NF-κB pathway activation by NSP14 (ref. [46]). Moreover, transcriptional profiling experiments have demonstrated NF-κB activation in HEK293 cells and in patients following SARS-CoV-2 infection[47,48]. As TNF-α has a central role in the cytokine storm that contributes to many COVID-19 deaths[49], the observation that SARS-CoV-2 activates this system in a cell-intrinsic manner may have therapeutic implications.

We explored the role of the NSP14 interactor IKBKG/NEMO, an essential mediator of canonical NF-κB signaling[50], for transcriptional activation. We generated IKBKG HEK293 knockout (KO) clones (Extended Data Fig. 4) and checked for NF-κB activation in three independent clones after NSP14 transfection (Fig. 4c). IKBKG deficiency abolished NF-κB activation in response to TNF-α and severely impaired NSP14-induced NF-κB activation, providing evidence for a functional role of IKBKG in driving NF-κB activation by NSP14. Interestingly, the residual NF-κB reporter induction upon NSP14 expression in the KO cells indicates that other NSP14 interactors (for example, TRAF2 and REL) contribute to the full NF-κB transcriptional response.

We wondered whether NF-κB signaling proteins and virally targeted host proteins in enriched functional groups other than 'immune response' (Fig. 2a) are important for viral replication. After generating A549 alveolar basal epithelial adenocarcinoma cells that exogenously express human ACE2 (A549-ACE2), we quantified viral replication in the presence and absence of CRISPR-Cas9-mediated KO of viral-target-encoding genes. Of eight genes that were selected from enriched functional groups and successfully knocked out, deletion of five (63%) resulted in a significant decrease of viral replication (Fig. 4d). Intriguingly, deletion of three NSP14-interacting proteins of the NF-κB signaling system (REL, IKBKG and TRAF2) resulted in strong reduction of viral replication (Fig. 4d and Extended Data Fig. 4f,g). This finding is consistent with a model in which SARS-CoV-2 directly activates NF-κB via NSP14, with this activation being required for successful viral replication. Deletion of kinesin light chain 1 (KLC1), a cargo adaptor protein for microtubule mediated transport, caused reduction of replication by ~80% ($P < 0.0001$, Kruskal–Wallis test). Beyond this observation, deletion of ubiquitin-specific peptidase 25 (USP25), which has antiviral functions in influenza and herpes infections[51], resulted in essentially complete elimination of viral replication without impacting cell growth, suggesting that human USP25 is required by SARS-CoV-2 (Fig. 4d, Extended Data Fig. 4f,g and Supplementary Table 10).

Inspired by the strong effect on viral replication, we explored USP25 as an antiviral drug target using the small molecule AZ1, which effectively inhibits USP25 and USP28 enzymatic activity[52]. Using an infectious clone-derived SARS-CoV-2 (icSARS-COV-2) harboring a mNeonGreen marker[53], we showed that treatment with 10 μM AZ1 effectively inhibits SARS-CoV-2 replication in Vero E6 cells (Fig. 4e). Next, we used an independent icSARS-CoV-2 expressing nanoluciferase[54] for dose titrations. The AZ1 compound interfered with SARS-CoV-2 replication with half-maximum effective concentration (EC_{50}) values of 0.8 μM and 0.1 μM in HEK293-ACE2 and Vero E6 cells, respectively (Fig. 4f and Supplementary Table 11), on par with the effects of the clinically approved remdesivir (Extended Data Fig. 4h). Effective concentrations are in the range of the half-maximal inhibitory concentration determined for inhibition of USP25/28 enzymatic activities[52], further supporting that USP25 is necessary for SARS-CoV-2 replication. Although the antiviral activity of AZ1 was independently identified in

https://doi.org/10.1038/s41587-022-01475-z

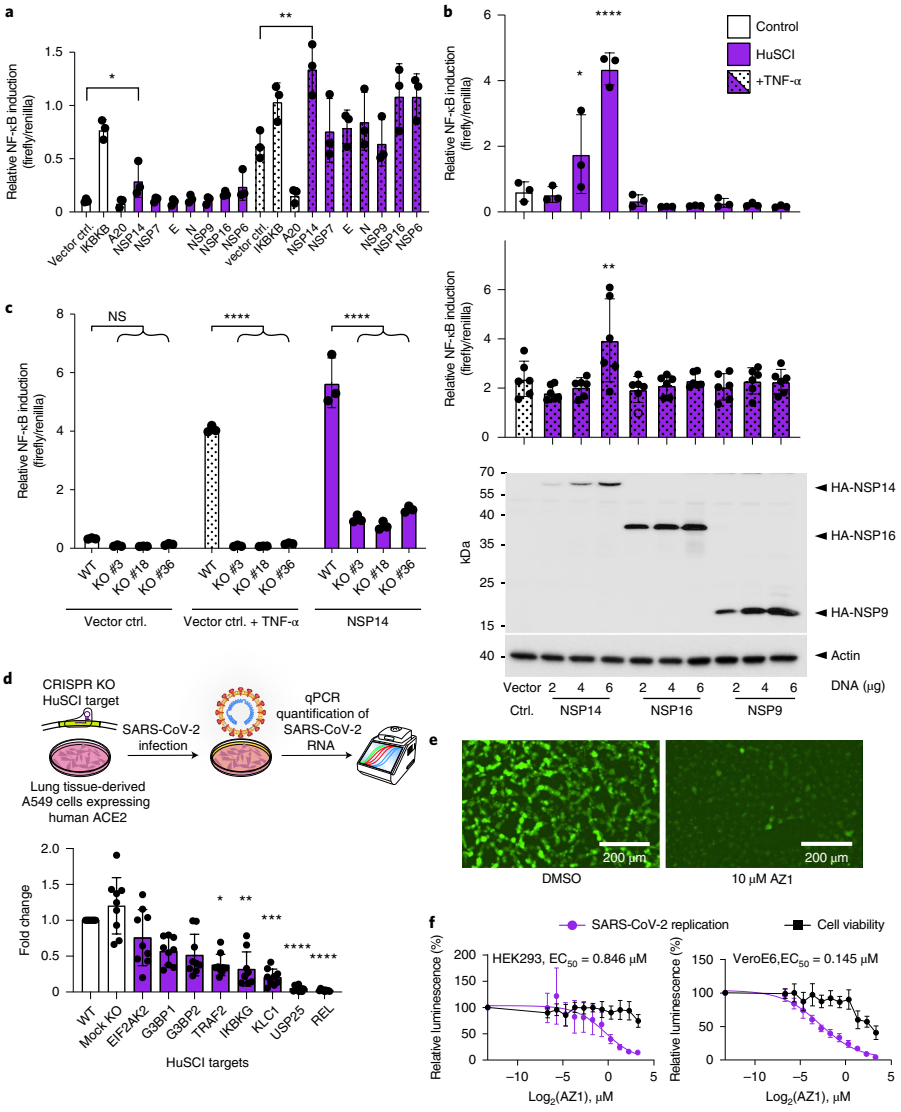


Fig. 4 | Validation of pathways and host targets. a, Relative NF-κB transcriptional reporter activity in unstimulated (left) and TNF-α-stimulated conditions (one-way analysis of variance (ANOVA) with Dunnett's multiple comparisons test, $P = 0.0395$ and $P = 0.0047$, respectively). Error bars represent standard deviation of the mean, $n = 3$. **b**, Relative NF-κB transcriptional reporter activity at different amounts of transfected viral protein-encoding plasmid in unstimulated (top) and TNF-α-stimulated conditions (middle) (one-way ANOVA with Dunnett's multiple comparisons test: $^{*}P = 0.0183$, $^{****}P < 0.0001$ and $^{**}P = 0.0012$, respectively). Error bars represent standard deviation of the mean, $n = 3$ (top) and $n = 6$ (middle). Representative anti-hemagglutinin (HA) western blot demonstrating levels of tagged viral protein in titration experiments relative to actin beta (ACTB) loading controls (bottom). **a** and **b**, Precise *P* values, biological repeats and *n* for each test are shown in Extended Data Fig. 4 and Supplementary Table 9. **c**, Relative NF-κB transcriptional reporter activity under unstimulated (left), TNF-α-stimulated (middle) and NSP14-induced conditions in wild-type (WT) and three independent IKBKG KO clones of HEK293 cells (two-way ANOVA with Dunnett's multiple comparisons test). Error bars represent standard deviation of the mean, $n = 3$. Precise *P* values, biological repeats and *n* for each test are shown in Extended Data Fig. 4. ctrl., control. **d**, Schematic of viral replication assay (top) and viral replication in wild-type, mock KO and CRISPR KOs of the indicated HuSCI host targets (bottom) (Kruskal–Wallis with Dunn's multiple comparisons test, $^{*}P = 0.031$, $^{**}P = 0.0047$, $^{***}P = 0.0003$, $^{****}P < 0.0001$, respectively). Error bars represent standard deviation of the mean, $n = 9$. Precise *P* values, biological repeats and *n* for each test are shown in Extended Data Fig. 4 and Supplementary Table 10. **e**, Fluorescence microscopy images showing replication of icSARS-CoV-2-mNeonGreen in infected Vero E6 cells treated with 10 μM AZ1 or solvent (DMSO, dimethylsulfoxide). **f**, Cell viability and relative replication of icSARS-CoV-2-nanoluciferase in HEK293 cells (left) and Vero E6 cells (right) at different concentrations of AZ1. The EC_{50} values were calculated with a variable slope model. Error bars represent standard deviation of the mean, $n = 8$ biological repeats and full analysis in Supplementary Table 11.

a small-molecule screen[55], our results inform mechanistic studies by identifying NSP16 as a viral interaction partner. NSP16 and associated complexes methylate viral RNA to prevent its detection and destruction by the innate immune system[56,57]. The stable recruitment of USP25 may protect this complex from ubiquitination and degradation by the host defense machinery. Although elucidating precise mechanisms will require further studies, these findings illustrate the high potential of the HuSCI contactome map in helping to understand and inhibit the SARS-CoV-2 life cycle.

Perturbed contactome in SARS-CoV-2 variants

Evaluating the impact of novel viral strains on the contactome has been largely restricted to spike protein interactions with ACE2 and antibodies[58]. Wondering if coding variants in other viral proteins perturb the contactome and thereby modulate viral effects, we explored the potential of 19 SARS-CoV-2 mutations in 14 variants of 9 proteins from the Alpha, Beta, Gamma and Delta strains to alter interactions with host contact targets in HuSCI (Supplementary Table 12). Indeed, some mutations resulted in perturbed interactions. The Alpha strain mutant combination D3L, S235F in the nucleocapsid protein reduced interaction with ARPC3, the SARS-CoV-2 host factor discussed above. Similarly, the Beta-strain mutation P71L in the envelope (E) protein diminished the interaction with BAG4, an antiapoptotic protein involved in TNF signaling (Extended Data Fig. 5). Although it is currently unknown whether the respective interactions promote viral replication or facilitate immune recognition, the observed changes demonstrate the plasticity of the contactome and, together with recent reports of increased replication of the Delta strain[59], strongly suggest that this dimension of viral evolution should also be monitored to assess the risk posed by emerging variants.

Discussion

In summary, we present a validated contactome map, HuSCI, which provides direct interactions between SARS-CoV-2 and human target proteins in pathways and tissues relevant to COVID-19. HuSCI enables identification of paths of direct contact between viral target proteins and proteins encoded from loci that modify the risk for critical COVID-19 illness and important comorbidities. Examining specific hypotheses for both viral and host proteins, we demonstrate that NSP14 activates the NF-κB pathway even beyond pathway activation by cytokines. Moreover, the majority of the virally targeted host proteins we evaluated, including key NF-κB regulators, are required for efficient SARS-CoV-2 replication. For one of these targeted host proteins, USP25, we confirm that a small-molecule inhibitor can dramatically reduce viral replication and implicate a mechanism for this potential therapeutic. Last, we demonstrate that coding changes in SARS-CoV-2 strains perturb the intracellular interactome. We anticipate that these findings and the contactome resource will stimulate important research toward characterizing new viral strains, understanding the mechanism of COVID-19 symptoms and developing therapies for current and future pandemics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01475-z.

References

1. Nalbandian, A. et al. Post-acute COVID-19 syndrome. *Nat. Med.* **27**, 601–615 (2021).
2. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
3. Altmann, M. et al. Extensive signal integration by the phytohormone protein network. *Nature* **583**, 271–276 (2020).
4. Yachie, N. et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12**, 863 (2016).
5. Braun, P. et al. An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2009).
6. Choi, S. G. et al. Maximizing binary interactome mapping with a minimal number of assays. *Nat. Commun.* **10**, 3907 (2019).
7. Li, Y. et al. SARS-CoV-2 induces double-stranded RNA-mediated innate immune responses in respiratory epithelial-derived cells and cardiomyocytes. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2022643118 (2021).
8. Orchard, S. et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358–D363 (2014).
9. Stukalov, A. et al. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246–252 (2021).
10. Bouhaddou, M. et al. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* **182**, 685–712.e19 (2020).
11. Kamel, W. et al. Global analysis of protein-RNA interactions in SARS-CoV-2 infected cells reveals key regulators of infection. *Mol. Cell* **81**, 2851–2867 (2021).
12. Gordon, D. E. et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370**, eabe9403 (2020b).
13. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020a).
14. Li, J. et al. Virus-host interactome and proteomic survey reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. *Med (N Y)* **2**, 99–112.e7 (2021).
15. Nabeel-Shah, S. et al. SARS-CoV-2 nucleocapsid protein binds host mRNAs and attenuates stress granules to impair host stress response. *iScience* **25**, 103562 (2022).
16. Laurent, E. M. N. et al. Global BioID-based SARS-CoV-2 proteins proximal interactome unveils novel ties between viral polypeptides and host factors involved in multiple COVID19-associated mechanisms. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.28.272955 (2020).
17. St-Germain, J. R. et al. A SARS-CoV-2 BioID-based virus-host membrane protein interactome and virus peptide compendium: new proteomics resources for COVID-19 research. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.28.269175 (2020).
18. Samavarchi-Tehrani, P. et al. A SARS-CoV-2–host proximity interactome. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.03.282103 (2020).
19. Wierbowski, S. D. et al. A 3D structural SARS-CoV-2–human interactome to explore genetic and drug perturbations. *Nat. Methods* **18**, 1477–1488 (2021).
20. Callard, F. & Perego, E. How and why patients made long covid. *Soc. Sci. Med.* **268**, 113426 (2021).
21. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
22. Dorward, D. A. et al. Tissue-specific immunopathology in fatal COVID-19. *Am. J. Respir. Crit. Care Med.* **203**, 192–201 (2021).
23. Zhao, X. et al. LY6E restricts entry of human coronaviruses, including currently pandemic SARS-CoV-2. *J. Virol.* **94**, e00562-20 (2020).
24. Garcia-Moreno, M. et al. System-wide profiling of RNA-binding proteins uncovers key regulators of virus infection. *Mol. Cell* **74**, 196–211 (2019).
25. Zanzoni, A., Spinelli, L., Ribeiro, D. M., Tartaglia, G. G. & Brun, C. Post-transcriptional regulatory patterns revealed by protein-RNA interactions. *Sci. Rep.* **9**, 4302 (2019).
26. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

27. Kruse, T. et al. Large scale discovery of coronavirus-host factor protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. *Nat. Commun.* **12**, 6761 (2021).
28. Ferrari, S. et al. Mutations of CD40 gene cause an autosomal recessive form of immunodeficiency with hyper IgM. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12614–12619 (2001).
29. de Vries, L., Gat-Yablonski, G., Dror, N., Singer, A. & Phillip, M. A novel MKRN3 missense mutation causing familial precocious puberty. *Hum. Reprod.* **29**, 2838–2843 (2014).
30. Zhong, Q. et al. An inter-species protein-protein interaction network across vast evolutionary distance. *Mol. Syst. Biol.* **12**, 865 (2016).
31. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
32. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
33. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
34. Whyte, P. et al. Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* **334**, 124–129 (1988).
35. Weßling, R. et al. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* **16**, 364–375 (2014).
36. Ostaszewski, M. et al. COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Mol. Syst. Biol.* **17**, e10387 (2021).
37. Soveg, F. W. et al. Endomembrane targeting of human OAS1 p46 augments antiviral activity. *eLife* **10**, e71047 (2021).
38. Cifuentes-Muñoz, N., Dutch, R. E. & Cattaneo, R. Direct cell-to-cell transmission of respiratory viruses: the fast lanes. *PLoS Pathog* **14**, e1007015 (2018).
39. Zhu, Y. et al. A genome-wide CRISPR screen identifies host factors that regulate SARS-CoV-2 entry. *Nat. Commun.* **12**, 961 (2021).
40. Daniloski, Z. et al. Identification of required host factors for SARS-CoV-2 infection in human cells. *Cell* **184**, 92–105.e16 (2021).
41. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol* **22**, 49 (2021).
42. Bliddal, S. et al. Acute and persistent symptoms in non-hospitalized PCR-confirmed COVID-19 patients. *Sci. Rep.* **11**, 13153 (2021).
43. Whiting, A., Reyes, J. V. M., Ahmad, S. & Lieber, J. Post-COVID-19 fatigue: a case of infectious hypothyroidism. *Cureus* **13**, e14815 (2021).
44. Mohan, M., Perry, B. I., Saravanan, P. & Singh, S. P. COVID-19 in people with schizophrenia: potential mechanisms linking schizophrenia to poor prognosis. *Front. Psychiatry* **12**, 666067 (2021).
45. VanderWeele, T. J. Genetic self knowledge and the future of epidemiologic confounding. *Am. J. Hum. Genet.* **87**, 168–172 (2010).
46. Li, T. et al. SARS-CoV-2 Nsp14 activates NF-κB signaling and induces IL-8 upregulation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.26.445787 (2021).
47. Hadjadj, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
48. Sun, G. et al. Comparative transcriptomic analysis of SARS-CoV-2 infected cell model systems reveals differential innate immune responses. *Sci. Rep.* **11**, 17146 (2021).
49. Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C. & Melguizo-Rodríguez, L. SARS-CoV-2 infection: the role of cytokines in COVID-19 disease. *Cytokine Growth Factor Rev.* **54**, 62–75 (2020).
50. Hayden, M. S. & Ghosh, S. Regulation of NF-κB by TNF family cytokines. *Semin. Immunol.* **26**, 253–266 (2014).
51. Lin, D. et al. Induction of USP25 by viral infection promotes innate antiviral responses by mediating the stabilization of TRAF3 and TRAF6. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11324–11329 (2015).
52. Wrigley, J. D. et al. Identification and characterization of dual inhibitors of the USP25/28 deubiquitinating enzyme subfamily. *ACS Chem. Biol.* **12**, 3113–3125 (2017).
53. Xie, X. et al. An infectious cDNA clone of SARS-CoV-2. *Cell Host Microbe* **27**, 841–848.e3 (2020).
54. Hou, Y. J. et al. SARS-CoV-2 reverse genetics reveals a variable infection gradient in the respiratory tract. *Cell* **182**, 429–446 (2020).
55. Grodzki, M. et al. Genome-scale CRISPR screens identify host factors that promote human coronavirus infection. *Genome Med.* **14**, 10 (2022).
56. Chang, L.-J. & Chen, T.-H. NSP16 2′-O-MTase in coronavirus pathogenesis: Possible prevention and treatments strategies. *Viruses* **13**, 538 (2021).
57. Alshiraihi, I. M., Klein, G. L. & Brown, M. A. Targeting NSP16 methyltransferase for the broad-spectrum clinical management of coronaviruses: managing the next pandemic. *Diseases* **9**, 12 (2021).
58. Li, Q. et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294.e9 (2020).
59. Syed, A. M. et al. Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles. *Science* **374**, 1626–1632 (2021).

Dae-Kyum Kim[1,2,3,4,5,25], Benjamin Weller[6,25], Chung-Wen Lin[6,25], Dayag Sheykhkarimli[1,2,3,4,25], Jennifer J. Knapp[1,2,3,4,25], Guillaume Dugied[7,8,9,25], Andreas Zanzoni[10], Carles Pons[11], Marie J. Tofaute[12], Sibusiso B. Maseko[13], Kerstin Spirohn[4,14,15], Florent Laval[4,13,14,15,16,17], Luke Lambourne[4,14,15], Nishka Kishore[1,2,3,4], Ashyad Rayhan[1,2,3,4], Mayra Sauer[6], Veronika Young[6], Hridi Halder[6], Nora Marín-de la Rosa[6], Oxana Pogoutse[1,2,3,4], Alexandra Strobel[6], Patrick Schwehn[6], Roujia Li[1,2,3,4], Simin T. Rothballer[6], Melina Altmann[6], Patricia Cassonnet[7,8,9], Atina G. Coté[1,2,3,4], Lena Elorduy Vergara[6], Isaiah Hazelwood[1,2,3,4], Betty B. Liu[1,2,3,4], Maria Nguyen[1,2,3,4], Ramakrishnan Pandiarajan[6], Bushra Dohai[6], Patricia A. Rodriguez Coloma[6], Juline Poirson[1,2,18], Paolo Giuliana[1,2,3,4], Luc Willems[16,17], Mikko Taipale[1,2,13], Yves Jacob[7,8,9], Tong Hao[4,14,15], David E. Hill[4,14,15,26], Christine Brun[10,19,26], Jean-Claude Twizere[4,13,16,26], Daniel Krappmann[12,26], Matthias Heinig[20,21,26], Claudia Falter[6,26], Patrick Aloy[11,22,26], Caroline Demeret[7,8,9,26] ✉, Marc Vidal[4,14,26] ✉, Michael A. Calderwood[4,14,15,26] ✉, Frederick P. Roth[1,2,3,4,23,26] ✉ and Pascal Falter-Braun[6,24,26] ✉

[1]Donnelly Centre for Cellular and Biomolecular Research (CCBR), University of Toronto, Toronto, Ontario, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3]Lunenfeld-Tanenbaum Research Institute (LTRI), Sinai Health System, Toronto, Ontario, Canada. [4]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA. [5]Department of Cancer Genetics and Genomics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. [6]Institute of Network Biology (INET), Molecular Targets and Therapeutics Center (MTTC), Helmholtz Zentrum München, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [7]Unité de Génétique Moléculaire des Virus à ARN, Département de Virologie, Institut Pasteur, Paris, France. [8]UMR3569, Centre National de la Recherche Scientifique, Paris, France. [9]Université de Paris, Paris, France. [10]Aix-Marseille Université, Inserm, TAGC, Marseille, France. [11]Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute for Science and Technology, Barcelona, Spain. [12]Research Unit Cellular Signal Integration, Institute of Molecular Toxicology and Pharmacology, Molecular Targets and Therapeutics Center (MTTC), Helmholtz Zentrum München, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [13]Laboratory of Viral Interactomes, GIGA Institute, University of Liège, Liège, Belgium. [14]Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [15]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [16]TERRA Teaching and Research Centre, University of Liège, Gembloux, Belgium. [17]Laboratory of Molecular and Cellular Epigenetics, GIGA Institute, University of Liège, Liège, Belgium. [18]Molecular Architecture of Life Program, Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada. [19]CNRS, Marseille, France. [20]Institute of Computational Biology (ICB), Computational Health Center, Helmholtz Zentrum München, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [21]Department of Informatics, Technische Universität München, Munich, Germany. [22]Institució Catalana de Recerca I Estudis Avaçats (ICREA), Barcelona, Spain. [23]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [24]Microbe-Host Interactions, Faculty of Biology, Ludwig-Maximilians-Universität (LMU) München, Planegg-Martinsried, Germany. [25]These authors contributed equally: Dae-Kyum Kim, Benjamin Weller, Chung-Wen Lin, Dayag Sheykhkarimli, Jennifer J. Knapp, Guillaume Dugied. [26]These authors jointly supervised this work: David E. Hill, Christine Brun, Jean-Claude Twizere, Daniel Krappmann, Matthias Heinig, Claudia Falter, Patrick Aloy, Caroline Demeret, Marc Vidal, Michael A. Calderwood, Frederick P. Roth, Pascal Falter-Braun. ✉e-mail: caroline.demeret@pasteur.fr; marc_vidal@dfci.harvard.edu; michael_calderwood@dfci.harvard.edu; fritz.roth@utoronto.ca; pascal.falter-braun@helmholtz-muenchen.de

Methods

Cloning SARS-CoV-2 ORFs

Two independent SARS-CoV-2 vORF collections were constructed in Gateway entry vectors. The $Y2H_{GFP}$ collection[60] includes all but one (NSP11 was too short for Gateway cloning) codon-optimized ORF of SARS-CoV-2, synthesized based on a published genome[61], which were cloned with and without stop codon, to enable C-terminal fusions. The $Y2H_{HIS3}$ entry clone collection is based on National Center for Biotechnology Information (NCBI) accession number NC_045512.2 and annotation[62]. $Y2H_{HIS3}$ vORFs were synthesized by Twist Bioscience without codon optimization and included 5´ and 3´ linkers with SfiI restriction sites. The 5´ linker incorporates a translational start ATG flanked by BamHI sites; the 3´ linker provides a stop codon flanked by PacI and AsiSI restriction sites. For $Y2H_{HIS3}$, vORFs were cloned into pENTR223.1 using SfiI restriction cloning, and the alternative ATG was removed by BamHI digest. A total of 28 vORFs were synthesized for $Y2H_{GFP}$ and 27 for $Y2H_{HIS3}$: NSP1-16 (except NSP11), S, E, M, N and ORFs 3A, 3B, 3D, 6, 7A, 7B, 8, 9B, 9C and 10[62–64] (Supplementary Table 1).

$Y2H_{HIS3}$ vORF entry clones were verified by full-length Sanger sequencing. As NSP10 had a one-base deletion, it was excluded from further experiments. vORFs were moved to the destination vectors pPC86 (N-terminal AD fusion, CEN origin)[3,65] and pHiDEST-DB (N-terminal DB fusion, CEN origin)[4] by Gateway cloning and confirmed by PCR. For $Y2H_{GFP}$, barcoded 'prey' (pAR068: C-terminal AD fusion, 2µ origin/pHiDEST-AD: N-terminal AD fusion, CEN origin), and 'bait' (pHiDEST-DB: N-terminal DB fusion, CEN origin) destination vectors were generated using published protocols[4], with the integration of the barcode locus at the SacI restriction site as described[26]. Single barcoded plasmid containing colonies were picked, arrayed into 384-well plates with 80 µl LB agar supplemented with 100 µg ml^{-1} carbenicillin and 35 µg ml^{-1} chloramphenicol (LB + Carb+CM) per well and incubated at 37°C for 16 h. Barcode sequences were identified using a modified Kiloseq procedure[66] using an Illumina NextSeq 500 and analyzed as previously described[4,26,66]. $Y2H_{GFP}$ vORFs and human ACE2 were moved by Gateway cloning into barcoded destination plasmids[4,26] pHiDEST-AD (N-terminal AD fusion, CEN origin (low copy number)) and pHiDEST-DB (N-terminal DB fusion, CEN origin (low copy number)) such that each ORF was linked to two to six barcodes in every configuration. Gateway cloning was performed individually and for ORF–barcode pairs using Sanger sequencing (TCAG, The Hospital for Sick Children) (Supplementary Table 13).

Generation of $HuSCI_{HIS3}$

The $Y2H_{HIS3}$ screening pipeline is essentially as previously described[65]. AD-Y and DB-X vORFs were transformed into yeast strains Y8800 (MATa) and Y8930 (MATα), respectively. NSP1 autoactivated as DB fusion and not screened in this orientation. DB-X vORFs were individually mated with 99 pools of ~188 AD-tagged human ORFs each, from human ORFeome v9.1 comprising 17,472 ORFs[26,67] (hORFeome9.1). For the reverse orientation, yeast with 27 AD-Y vORFs were pooled and mated against DB-X hORFeome9.1. Primary screening in both configurations was performed twice to increase sampling sensitivity. Unless otherwise noted, all yeast incubations are at 30°C, overnight without shaking.

For primary screening, saturated haploid AD-Y and DB-X yeast cultures were spotted on top of each other on yeast extract peptone dextrose (YEPD) agar (1%) plates and incubated for 24 h. Yeast were replica plated onto selective synthetic complete media lacking leucine, tryptophan and histidine (SC-Leu-Trp-His) + 1 mM 3-AT (3-amino-1,2,4-triazole)[3,65] (3-AT plates) and incubated for 72 h. From growing spots up to three colonies were picked and cultured in SC-Leu-Trp liquid medium for 2 d. For second phenotyping, cultures were spotted on diploid selection plates, incubated for 2 d and replica plated on 3-AT-plates and SC-Leu-His + 1 mM 3-AT + 1 mg per liter cycloheximide plates to identify spontaneous DB-X autoactivators[2]. Positive scoring colonies (growth on 3-AT-plates, no growth on cycloheximide plates) were picked, and ORFs were identified by Sanger sequencing[65]. For threefold verification, yeast strains corresponding to the identified human interaction partners were picked from archival glycerol stocks, cultured in liquid medium and mated (as described above) one-by-one against all vORFs, processed as described above and then scored. Colony growth was scored using a custom dilated convolutional neural network[68]. For training, previous datasets of more than 1,500 images of biochemically and functionally validated binary Y2H studies were used[3]. Each image was scaled to achieve equal pixel distance between the yeast spots of different images. The images were cropped and sliced, and the mean grayscale image of all spots on a plate was calculated. With this dataset, a simple front-end prediction module was trained consisting of six dilated convolutional layers with exponential increasing dilation rate and two dense layers at the end. After each layer except the last, a Leaky-ReLU activation was added[69]. The model was optimized with a combination of Softmax and Cross entropy and an Adam Optimizer[70]. The model achieved an accuracy >0.9 during all folds of a tenfold cross-fold validation. All positive scores were confirmed by a trained researcher. The verification step was done in triplicate and protein pairs scoring positive in at least two repeats were considered bona fide Y2H interactors. One representative colony of all interaction pairs was picked from selective plates to confirm the identities of X and Y by Sanger sequencing[65].

Generation of $HuSCI_{GFP}$

Barcoded ORFeomes. The barcoded human ORFeome consisting of 16,747 fully sequence-verified human ORFs with ~95% ORFs represented by two unique barcodes was previously described[26]. The barcoded bait and prey collections were arranged into a 10-by-10 screening matrix consisting of 10 DB and 10 AD groups, each containing ~1,400 ORFs with two distinct sets of unique barcodes, and ~200 ORFs with a single unique barcode set. Barcoded SARS-CoV-2 plasmids were transformed individually into RY3011 (AD plasmids) and RY3031 (DB plasmids) (genotypes in Supplementary Table 14). Transformed colonies were copied on fresh plates, incubated, scraped off and pooled to make glycerol stocks of all the barcoded SARS-CoV-2 ORFs plus the human ORF ACE2 in each plasmid configuration (with two or more barcodes per ORF).

Mating of pooled haploid yeast. Multiple pooled matings were performed using the frozen haploid pools. Each of the 10 human ORF pools (in C-terminal AD fusion plasmids with 2µ origin; pAR068) were separately mixed with the pool of SARS-CoV-2 ORFs plus human ACE2 (in N-terminal DB fusion plasmids with CEN origins; pHiDEST-DB). A separate mating was done between the SARS-CoV-2 pools in both AD and DB fusion, CEN origin plasmids (pHiDEST-AD, pHiDEST-DB). Negative controls were included in each mating and all matings were calculated to achieve >100× coverage of possible barcode combinations considering viability and mating efficiency. Procedurally, equal amounts of each haploid strain were mixed, the mixture was spread on 2x YEPD plus adenine agar plates (YPAD) and incubated for 24 h. Colonies on each mating plate were collected and re-spread across 20 15 cm SC-Leu-Trp plates supplemented with histidine (8 mM) and incubated for 72 h. These plates were then scraped off to make assay-ready pooled diploid glycerol stocks for each of the 11 groups.

Selection of yeast with interacting pair of DB-X and AD-Y by FACS. Pool of glycerol stocks were inoculated into 1-liter flasks with a starting vCFU of 30 M and incubated at 200 rpm for 24 h. Negative controls were started as 10 ml cultures and processed in parallel. 'Presort' cultures were prepared for each sample (2 × 10 ml cultures with OD_{600} 10) with doxycycline added (10 µg ml^{-1}) to these cultures to induce barcode swapping while these cultures were incubated for 24 h[4]. To prepare for fluorescence-activated cell sorting (FACS), cells were

concentrated by centrifugation (500 × *g*, 5 min) and resuspended in PBS to a final OD_{600} of 10. Propidium iodide (4 mg liter^{-1}) was added to identify dead yeast cells during FACS. Using the diploid negative control, the FACS gate for GFP-positive cells was set to capture 0.1% of GFP-negative cells, yielding a 0.01% false positive rate. Then, 100 million cells per group were sorted, and GFP-positive cells for each sample were plated on 10 SC-Leu-Trp+Ade+10x His (8 mM) plates and incubated for 72 h. Colonies were collected by scraping, centrifuged and resuspended into 2 × 10 ml cultures (OD_{600} = 10). Doxycycline (10 µg ml^{-1}) was added to induce barcode swapping, and cultures were incubated for 24 h, when plasmid DNA was extracted. Fused barcodes were PCR amplified with primers that attach modified Illumina i5 and i7 adapters to uniquely identify each sample. Following agarose gel analysis of PCR products, the bright band at ~350 bp was purified using a NucleoSpin Gel and PCR Clean-up kit. DNA concentrations were measured for each sample using a Qubit (Invitrogen, Q32851) and, guided by DNA concentration, samples were pooled to ensure equal sequencing depth relative to the number of protein pairs tested. After primer-dimer removal, DNA was quantified by qPCR, and the pooled NGS library was sequenced on an Illumina NextSeq using a mid- or high-output 150-cycles kit.

Read counting based on expected barcodes. The sequencing data were demultiplexed using bcl2fastq2 (v2.20.0.422) provided by Illumina with the following command: 'bcl2fastq -r 10 -p 20 -w 10 –no-lane-splitting –barcode-mismatches 1 –adapter-stringency 0.7 –ignore-missing-bcls –ignore-missing-filter –ignore-missing-positions'. After demultiplexing, the fastq files were aligned to the group specific reference files using bowtie2[71] with the following parameters:

For read 1: -q –norc –local –very-sensitive-local -t -p 23 –reorder.

For read 2: -q –nofw –local –very-sensitive-local -t -p 23 –reorder.

Reference files contained expected barcode sequences for the ORFs in each group. After alignments, reads with mapping quality scores <20 were removed. Following successful BFG barcode recombination[4], paired-end reads map to up-up or dn-dn when an interaction is present. The number of reads mapping to up-up and dn-dn were counted separately and merged as the final read count. The pipeline was implemented in Python v2.7.

Interaction scoring. For virus–host interactions, we used the product of marginal frequencies of bait and prey strains[4] to estimate the abundance of each diploid bait–prey strain in the presort condition ('PreSort'). The interaction score was defined by

$$
\begin{aligned}
IS_{ij} &= \frac{f_{ij}^{GFP}}{f_{ij}^{Pr\,eSort}} \\
f_{ij}^{Pr\,eSort} &= \sum_i c_{ij}^{Pr\,eSort} / \sum_j \left[\sum_i c_{ij}^{Pr\,eSort} \right] \\
f_{j}^{Pr\,eSort} &= \sum_j c_{ij}^{Pr\,eSort} / \sum_i \left[\sum_j c_{ij}^{Pr\,eSort} \right] \\
f_{j}^{Pr\,eSort} &= \max\left(f_i^{Pr\,eSort}, f_{AD}^{Floor}\right) \times \max\left(f_j^{Pr\,eSort}, f_{DB}^{Floor}\right) \\
f_{AD}^{Floor} &= 10^{-5} f_{DB}^{Floor} = 10^{-4} \\
f_{ij}^{GFP} &= c_{ij}^{GFP} / \sum_{ij} c_{ij}^{GFP}
\end{aligned}
$$

with the following variables: *c*, read count; *i*, AD barcode count; *j*, DB barcode count; *f*, frequency.

For every DB barcode, we used the 960 AD null barcodes to define the thresholds leading to a 1% false positive rate. An interaction was accepted as positive only if the ORF pair interaction score was above this threshold for two or more barcode pairs. For intraviral screening, we accepted as interactions those protein pairs for which the frequency of barcode pairs was 1,000 times greater than the median frequency of the corresponding DB barcode for three or more independent barcode pairs, similar to the scoring method previously used for BFG-Y2H with HIS3-based growth selection[4].

Pairwise retesting. Candidate interaction pairs for $HuSCI_{GFP}$ were verified in a pairwise HIS3 growth-based Y2H assay as described above ($Y2H_{HIS3}$ verification step), with minor modifications. Barcode replicates of candidate human AD-Y and viral DB-X were pooled prior to mating. vORFs NSP1 and NSP12 were omitted from this retesting due to DB autoactivation. After mating, colonies were replica plated on SC-Leu-Trp-His and 3AT-plates. After 72–96 h of yeast growth, these pairwise tests were scored according to the standardized scoring method used for the $Y2H_{HIS3}$ screen[3,65]. Interaction pairs scoring ≥3 were considered bona fide Y2H interactions.

Estimating completeness using the interactome framework

Assay sensitivity (S_a) is defined as the fraction of true interactions that can be detected by a given assay. Sampling sensitivity (S_s) is defined as a fraction of detectable true interactions that can be recovered by the pipeline used. Overall sensitivity of a given screen S can be calculated as $S = S_a \times S_s$. In pairwise settings $S_s = 1$ and the assay sensitivity is given by the fraction of hsPRS-v1/v2 pairs that score positive. $Y2H_{HIS3}$ was benchmarked previously[5] and has an assay sensitivity of $S_{a\text{-}HIS3} = 21.7\%$. Sampling sensitivity of $Y2H_{HIS3}$ after two repeats in two orientations has been shown to be $S_{s\text{-}HIS3} = \sim 60\%$[65], yielding a screening sensitivity of $S_{HIS3} = S_{a\text{-}HIS3} \times S_{s\text{-}HIS3} = 0.217 \times 0.6 = 13\%$. Given that $Y2H_{HIS3}$ screen had a search space completeness of 83% ($T_{HIS3} = 83\%$), the overall completion of $HuSCI_{HIS3}$ is $C_{HIS3} = T_{HIS3} \times S_{HIS3} = 0.83 \times 0.13 = 10.8\%$.

A different version of $Y2H_{GFP}$ using low-copy plasmids and N-terminally fused hybrid proteins ($lcnY2H_{GFP}$) was benchmarked using 84 pairs of hsPRS-v1 and 92 pairs of hsRRS-v1. Flow cytometry was used to score interactions based on percentage of singlets in GFP-positive gate, which was set using empty bait and prey constructs. In addition, $lcnY2H_{GFP}$ was benchmarked in a pooled setting using all possible combinations of proteins constituting 78 hsPRS-v2 and 77 hsRRS-v2 pairs supplemented with a set of 14 pairs of Y2H-positive controls defined as calibration set[4]. The experiment was carried out and interactions were scored as described above, except that no empirical null distribution was used. $lcnY2H_{GFP}$ recovered 12 out of 82 ($S_{a\text{-}lcnGFP} = 15\%$) hsPRS-v1 pairs when tested in a pairwise single bait–prey configuration and 8 of 92 (9%, $S_{s\text{-}lcnGFP} = 9/15 = 60\%$) hsPRS-v2 + calibration set pairs when tested in a pooled single bait–prey configuration, yielding $S_{lcnGFP} = S_{a\text{-}lcnGFP} \times S_{s\text{-}lcnGFP} = 0.15 \times 0.6 = 9\%$. It has been previously shown that using high-copy C-terminal fusions increases sensitivity by ~33% without affecting precision[26]. Thus, screening sensitivity of $Y2H_{GFP}$ was modeled from that of $lcnY2H_{GFP}$ as $S_{GFP} = S_{lcnGFP} \times 1.33 = 9\% \times 1.33 = 12\%$. Given that $Y2H_{GFP}$ covered 70% ($T_{GFP} = 70\%$) of all possible virus–human protein combinations, the completion level of the $Y2H_{GFP}$ dataset is $C_{GFP} = T_{GFP} \times S_{GFP} = 0.70 \times 0.12 = 8.4\%$. Only 4 out of 28 (14.2%) hsPRS-v1 pairs detected by the union of $Y2H_{HIS3}$ and $lcnY2H_{GFP}$ were detected with both methods, indicating a high degree of orthogonality (that is, different detection profiles of the methods used). In addition, $Y2H_{GFP}$ implemented in this study includes further differences such as high-copy and C-terminal fusion constructs for human proteins. Therefore, we conservatively estimate 90% orthogonality between $Y2H_{HIS3}$ and $Y2H_{GFP}$ (that is, ~90% of detected interactions are different: $O_{HIS3+GFP} = 90\%$). Thus, we estimate that the fraction of all true interactions captured by our merged interactome maps is $C_{HIS3+GFP} = (C_{HIS3} + C_{GFP}) \times O_{HIS3+GFP} \cong (0.108 + 0.084) \times 0.9 = 17.3\%$. Given the uncertainties associated with derivation of screening sensitivity, we estimate lower and higher bounds to be 15% ($S_{GFP} = 9\%$, excluding inferred gain in sensitivity due to high-copy C-terminal fusions) and 22% ($S_{GFP} = 13.5\%$, $S_{s\text{-}HIS3} = 70\%$ and $O_{HIS3+GFP} = 100\%$), respectively.

Pairwise Y2H testing of previously identified SARS-CoV-1 interactions

We identified 97 unique curated binary interactions with SARS-CoV-1 and human interaction partners[8] (Supplementary Table 2). For 77 of these, reagents to test interactions with SARS-CoV-2 orthologues were available in the barcoded human ORFeome. These involved 63 human proteins, 60 of which were covered by two barcode sets and three by a single barcode set. These were tested according to the 'pairwise retesting' protocol (above). Successful interactions were indicated by colony growth of both replicates in either condition.

Pairwise Y2H testing with SARS-CoV-2 variants

Lineage-defining mutations for the SARS-CoV-2 'variants of concern' as defined by the Centers for Disease Control and Prevention (Alpha, Beta, Gamma and Delta) were obtained from CoV-Spectrum[72,73] and mapped to the SARS-CoV-2 reference genome (NCBI accession number NC_045512.2). To generate variant ORFs, $Y2H_{HIS3}$ plasmids were used as template for mutation PCR (primers in Supplementary Table 12). Mutation PCR reaction products were transformed and sequence verified. Plasmids containing the desired mutation were directly transformed into yeast and processed in pairwise mating as described above. A complete list of mutations generated is shown in Supplementary Table 12. SARS-CoV-2 proteins for which interactions were identified in AD-fusions (N and E) were tested only against the identified interactors. All other variant proteins were tested against all HuSCI interactors. In total, 19 individual mutations in 14 unique variant proteins from 9 different viral proteins were tested. Four proteins with 8 cloned variants had interactors in $HuSCI_{HIS3}$, 1 protein with a single cloned variant had interactors in $HuSCI_{GFP}$ and 4 proteins with 5 variants had no HuSCI interactors.

yN2H validation

Using Gateway cloning, ORFs from the indicated subsets (Supplementary Table 3) were transferred into pDEST-N2H plasmids (pDEST-N2H-N1, -N2, -C1, and -C2) containing a *LEU2* (N1/C1 vectors) or a *TRP1* (N2/C2 vectors) auxotrophy marker and transformed into haploid *Saccharomyces cerevisiae* Y8800 (MATa) and Y8930 (MATα) strains. For cross-plate calibration, two protein pairs from the hsPRS-v2, with different N2H signal intensities, were included in duplicate on every plate (NCBP1/NCBP2 and SKP1/SKP2). Virus–human protein pairs were randomly distributed across the plates and tested together with hsPRS-v2/hsRRS-v2, which were in separate plates.

Overnight-grown haploid cultures were mated by mixing 5 µl of each haploid strain in 160 µl YEPD medium followed by overnight incubation. To measure background, all interactor ORFs were also mated with yeast with empty F1 or F2 plasmids. After mating, 10 µl culture each was inoculated into 160 µl SC-Leu-Trp and grown overnight, and then 50 µl was reinoculated into 1.2 ml SC-Leu-Trp and incubated for 24 h while shaking at 900 rpm. Cells were harvested (6,000 x *g*, 15 min), and the supernatant was discarded. Each yeast cell pellet was fully resuspended in 100 µl NanoLuc Assay solution[6]. Homogenized solutions were transferred into white flat-bottom 96-well plates and incubated in the dark (for 1 h at room temperature). Luminescence was evaluated for each sample with 2 s integration time. To score X–Y protein pairs, a normalized luminescence ratio (NLR) was calculated corresponding to the raw luminescence value of the tested pair (X-Y) divided by the maximum luminescence value from one of the two controls (X-Fragment 2 or Fragment 1-Y)[6]. The 1% RRS threshold was based on the vhRRS and determined using the R quantile function.

Enrichment of previously known, phospho-regulated or RNA-binding host targets

From IntAct[8] (version: April 28, 2020), 2,151 human proteins reported to have binary interactions with any virus protein were defined as 'previously known host targets'. 2,005 of these ORFs were interrogated by our experiment, and further considered. HuSCI contained 61 previously known host targets. 2,254 human proteins that change phosphorylation changes upon SARS-CoV-2 infection were identified from A549 and Vero E6 cell lines[9,10], of which 2,007 were interrogated by our experiment and 37 are in HuSCI. 139 experimentally identified human proteins specifically bound to SARS-CoV-2 RNA (vRICs) and 335 human proteins with altered RNA-binding activity upon SARS-CoV-2 infection (cRICs) were obtained from a recent RNA-interactome study[11]. Then, 121 vRICs and 294 cRICs were interrogated by our experiment; 5 HuSCI proteins were vRICs, and 13 HuSCI proteins were cRICs. All the observations were tested for enrichment using Fisher's exact tests and by permutation tests with 10,000 permutations.

GO enrichment analysis

gProfiler[74] (database versions: Ensembl 104, Ensembl Genomes 51 and Wormbase ParaSite 15) was applied to identify enriched functional categories in HuSCI, AP-MS[9,12–15] and BioID studies[16–18]. The hORFeome9.1, which was used for contactome mapping, served as the background for HuSCI, otherwise the universal annotated human genes. 'Inferred from electronic annotations' annotations were excluded. Adjusted *P* values were calculated using the Benjamini–Hochberg procedure. Functional terms with a hypergeometric $P < 0.05$ and term size between 5 and 1,000 were collected and enrichment calculated as the ratio between observed and expected gene counts. To categorize HuSCI host proteins, five meta categories inspired by the functional enrichment analysis results were used, namely 'immune response' (GO:0006955), 'viral process' (GO:0016032), 'protein ubiquitination' (GO:0016567), 'cytoskeleton' (GO:0005856) and 'vesicle-mediated transport' (GO:0016192). Human proteins related to these categories were obtained from the AmiGO 2 (ref. [75]) (July 2021), and HuSCI host proteins were categorized based on their annotation to these meta categories.

Domain enrichment of host interacting proteins

Structural domains in human targets were identified from Pfam release 34.0 (ref. [76]) (March 2021). Interactions of viral proteins with human interactors that have common domains were defined as shared-domain interactions and counted for HuSCI. The procedure was repeated for 1,000 randomized HuSCI networks (degree-preserved random rewiring). The significance of every viral protein–human domain was assessed by Fisher's exact tests (Supplementary Table 6) using the number of V-D, V-!D, !V-D, and !V-D interacting pairs, in which V and D correspond to the viral protein and human domain of interest, and !V and !D to the rest of viral proteins and domains in the HuSCI network, respectively. We identified as enriched associations those with at least two V-D interactions and $P < 0.05$. We repeated the process for 1,000 randomized HuSCI networks (see above). Multiple domain copies in a given human protein were counted once.

NF-κB reporter assays

HEK293 (RRID: CVCL_0045, DSMZ) were cultured in complete DMEM (high glucose) supplemented with 10% fetal calf serum, 100 U ml^{-1} penicillin and 100 µg ml^{-1} streptomycin and maintained in humidified atmosphere at 5% CO_2 at 37°C. For the reporter assay, 1×10^6 HEK293 cells were seeded in a 60-mm cell culture dish one day before transfection. Transfection was done using the calcium phosphate protocol using 10 ng NF-κB reporter plasmid (6 × NF-κB firefly luciferase pGL2), 50 ng pTK reporter (Renilla luciferase) and expression vectors (Flag-IKKb (pRK5), Flag-A20 (pEF4) and SARS-CoV-2 constructs (pMH)) using a total of up to 6 µg DNA. Briefly, the DNA was diluted in 200 µl 250 mM $CaCl_2$ solution (Carl Roth, 5239.1), vortexed and added dropwise to 200 µl 2 × HBS (50 mM HEPES (pH 7.0), 280 mM NaCl, 1.5 mM $Na_2HPO_4 \times 2\,H_2O$, pH 6.93) while gently vortexing. After 15-min incubation at room temperature, the mix was added dropwise to cell culture dishes. Transfection media was replaced after 6-h incubation with complete DMEM. Then, 24 h after transfection cells were stimulated with 20 ng ml−1 TNF-α for 4 hours. Luciferase activity was measured

using the dual luciferase reporter kit (Promega, E1980) according to the manufacturer's protocol. The firefly and Renilla luminescence was determined with a luminometer (Berthold Centro LB960 microplate reader, software MikroWin 2010) and quantified in relative light units (RLU). NF-κB induction was specified as the ratio of firefly luminescence (RLU) to Renilla luminescence (RLU). Significance of relative NF-κB transcriptional activity was assessed via one-way ANOVA with Dunnett's multiple comparisons. Data evaluation was performed in GraphPad Prism v7.04.

Protein expression was verified by western blot of lysates. Briefly, proteins were separated by SDS-PAGE and transferred on polyvinylidene fluoride membranes. Membranes were blocked with 5% milk in 1 × PBS + 0.1% Tween-20 (PBS-T) for 1 h at room temperature. Primary antibodies in 2.5% milk in PBS-T were incubated overnight at 4°C, the membranes were washed three times with PBS-T and secondary antibodies were incubated (1.25% milk/PBS-T) for 1 h at room temperature. Anti-actin beta (SCBT, sc-47778), anti-FLAG M2 (Sigma-Aldrich, F3165) and anti-HA (Sigma-Aldrich, 11583816001, RRID:AB_514505) were used at a 1:1,000 dilution. Secondary antibody (Jackson ImmunoResearch, Jim-715-035-150) was used at a 1:10,000 dilution. For detection of horseradish peroxidase-catalyzed enhanced chemiluminescence, LumiGlo reagent (CST, 7003S) was used.

For generation of *IKBKG* KO HEK293 cells, oligonucleotides coding sgRNAs targeting exon 3 (5′-TGCATTTCCAAGCCAGCCAG-3′) or exon 2 (5′- GCTGCACCATCTCACACAGT-3′) were cloned into px458 (Addgene, 48138). HEK293 were transfected with 5 µg plasmid by standard calcium phosphate transfection. After one day, GFP-positive cells were sorted with a MoFlo cell sorter (Beckman Coulter, Cytomation) and seeded in 96-well plates at dilutions of 0.5–5.0 cells per well. Single-cell clones were expanded and screened for loss of IKBKG expression by western blot (RRID: AB_2124846). IKBKG-negative clones were verified by amplifying and sequencing a region of genomic DNA encompassing the sites targeted by PCR (exon 3: forward primer 5′-CTGGCCAACACGTACTTTTA-3′, reverse primer 5′-GGTTACGGTGAGCGAAGGCTC-3′; exon 2: forward primer 5′- CTGACATCTCCCTCCACAAAC-3′ and reverse primer 5′-GGAGCTGGAATGAACCTTCC-3′).

Functional effects on viral replication

Selection of host-target candidates. To evaluate if identified host targets are involved in viral replication, the following HuSCI proteins involved in host immune regulation[77] and viral life cycle regulation[51,78–80] by enriched GO terms in this study were selected: G3BP1, G3BP2, TRAF2, USP25, EIF2AK2, REL, IKBKG and KLC1.

Engineering of hACE2-expressing cells. A549 cells were seeded at 5 × 10^5 cells per well in six-well cell culture plates and cultured in DMEM with 10% FCS and 1% penicillin/streptomycin at 37°C and 5% CO_2 (standard media). After 24 h culture medium was replaced by fresh medium containing 4.5 × 10^7 transduction units hACE2 lentivirus per well and incubated for 4 hours at 37°C and 5% CO_2. The lentiviral inoculum was then replaced with 2 ml DMEM 10% FCS and 1% penicillin/streptomycin. After 24 h, the transduction was repeated with the same steps as above. Cell surface expression of hACE2 was monitored by FACS using the AttuneNxT Flow Cytometer (Thermo Fisher Scientific) and results were analyzed with FlowJo v10 Software (BD Life Sciences). The resulting cells are referred to as A549-hACE2.

Generation of KO cell lines. KO cells were generated using the target-specific CRISPR-Cas9-HDR (homology-directed recombination) KO directed technology developed by Santa Cruz Biotechnology, which enables selection of KO cells with puromycin and red fluorescent protein (Supplementary Table 15). Briefly, A549-hACE2 cells were seeded at 2.5 × 10^6 cells in T25 flasks and standard media. After 24 h, cells were cotransfected with 7.5 µg each of KO and HDR plasmids for the previously described targets and 15 µg KO plasmid for the mock KO, from Santa Cruz Biotechnology using FuGene (Promega, E2312). After 72 h, KO cells were selected with 2 µg/ml puromycin (Invivogen, ant-pr-1) for 3 d, and mock KO cells were treated with the same volume of Hepes solution (Sigma-Aldrich, 51558). One week later, red fluorescent protein-positive cells were sorted by flow cytometry. DNA from 2 × 10^6 cells was extracted and region of interest was amplified for each KO, except KLC1, in a 25-µl PCR using 50 ng genomic DNA and using one primer in the genomic DNA and one primer in the insert (primers are listed in Supplementary Table 15). KLC1 KO was verified by amplifying the sg-directed Cas9 region that had no corresponding HDR with one primer on each side of the region; the PCR product was purified using Nucleospin Gel and PCR Clean-up (Machery-Nagel, 11992242) and KO confirmed by Sanger sequencing.

Assessment of SARS-CoV-2 infection in A549-hACE2 KO versus wild-type cells. Wild-type and KO A549-hACE2 cells were seeded at 1 × 10^6 cells per well in 12-well plates and standard media. After 24 h, cells were infected at a multiplicity of infection (MOI) of 10^{-3}, with SARS-CoV-2 isolate hCoV19/France/GE1973/2020 (n = 3, biological replicates). Total RNA was extracted from infected cells at 72 h after infection, and SARS-CoV-2 replication was assessed by RT-qPCR using Orf1ab primers (5′-ATGAGCTTAGTCCTGTTG-3′; 3′-CTCCCTTTGTTGTGTTGT-5′) (n = 9, three technical replicates per biological replicate). GAPDH was used for normalization. Viral RNA was quantified according to the ΔΔCt standard method[81]. The effect of gene KO on viral replication was determined using the wild-type ORF1ab RNA level as a control as shown in the following equation: $2^{-(\Delta\Delta Ct)} = 2^{-(\Delta Ct\,KO - \Delta Ct\,WT)}$. Significance of the KO effect was calculated against the mock KO using an ordinary one-way nonparametric ANOVA Kruskal-Wallis with Dunn's multiple comparisons test using GraphPad Prism v9.

Assessment of the viability of the KO cell lines. A total of 8.0 × 10^5 cells of each KO cell line were seeded in a white 96-well plate and incubated at 37°C and 5% CO_2 for 24 h. Cell media was replaced with DMEM and incubated at 37 °C and 5% CO_2 for 72 h. Cell viability was measured using Cell Titer-Glo Luminescent Cell Viability Assay kit (Promega, G7750). Luminescence was measured on a Centro XS luminometer (Berthold; integration time, 0.5 s). Wild-type cells served as the reference and significance of cell viability was calculated against the mock KO using an ordinary one-way nonparametric ANOVA Kruskal–Wallis with Dunn's multiple comparisons test using GraphPad Prism v9.

Genes ranked by number of publications

Publication counts are derived from the gene2pubmed file from NCBI, downloaded on 16 November 2021. Only protein-coding genes were considered. For visualization, but not statistical assessment, of genes with equal numbers of publications, order was determined by random shuffling. P values were calculated by Mann–Whitney U test, with Bonferroni correction. Black dots indicate the mean; error bars represent the 95% confidence interval generated from 1,000 bootstrap samples.

Tissue specificity analysis

The Tissue Atlas dataset was obtained from the HPA database[21] (version 2021.04.09). The HPA categories 'tissue enriched', 'group enriched' and 'tissue enhanced' were combined with 'tissue-specific', 'low tissue specificity' was denoted as 'common' and the 'not detected' category was not included in this analysis. A total of 11,069 of 19,670 genes (56.3%) in the HPA dataset were defined as tissue specific, and 8,385 of 19,670 genes (42.6%) showed common expression profiles. Tissue distribution differences were determined using Fisher's exact test with Bonferroni correction.

SARS-CoV-2 organotropism data were obtained from post mortem examinations[22,82]. The RNA tissue-specific NX value (normalized transcripts per million) was extracted and used to denote whether the gene is specifically expressed in a given tissue. Tissues from the

Tissue Atlas were combined into organ systems and used to assess host-target tissues. Significance was evaluated by Fisher's exact test with Bonferroni correction.

Identification of genetic variation in host targets and network communities

Host network communities were identified using the OCG hierarchical community clustering algorithm on the Human Reference Interactome[26,83] as implemented in the linkcomm R package (V1.0-13) using 'centered cliques' as initial class system[84]. A total of 3,603 communities with a minimum size of 4 were found, of which 204 contained a significant number of virus interactors (that is, were significantly targeted) (nominal $P < 0.05$, Fisher's exact test; Supplementary Table 8). A community was annotated to a function if a GO term was enriched (FDR < 0.05) or if ≥20% or ≥30% of the annotated constituent proteins shared an annotation[85] (Supplementary Table 8). From AP-MS-based association studies[9,12–15], 57, 43, 18 and 17 significantly targeted communities were found, respectively (nominal $P < 0.05$, Fisher's exact test; Supplementary Table 8).

Uniformly processed GWAS summary statistics were downloaded for 114 traits from the GTEx GWAS analysis[41,86]. MAGMA[87] analysis was implemented in R 3.6.1 and consists of three steps: first, GWAS summary statistics across all single-nucleotide polymorphisms (SNPs) within a gene region are aggregated into a gene-level association P value. Next, the gene-level P value is transformed to a z-score (using the inverse normal cumulative distribution function). Finally, z-scores across all genes are modeled as a function of gene set membership and the default gene-level covariates (gene size in number of SNPs, the gene density (a measure of within-gene linkage disequilibrium), the inverse mean minor allele count) using a linear model. Association between gene set membership and GWAS z-scores is tested based on the null hypothesis beta = 0 for the coefficient associated with the gene set membership indicator variable. All targets, and the targeted network communities, were considered gene sets. Entrez gene IDs were used on the human genome assembly 38. Individual MAGMA analyses were performed for each trait based on summary statistics and linkage disequilibrium structure from the 1,000 genomes European reference panel always conditioning on default gene-level covariates (for example, gene length). For each gene set, standard error normalized beta coefficients constituted the association score, with larger values indicating greater chance of getting significant association. Following Benjamini–Hochberg multiple hypothesis correction, gene set–trait associations with FDR < 0.05 were selected. These pairs were subjected to follow-up analysis. SNPs localizing within genes of enriched gene sets were selected, and genes containing SNPs with GWAS $P < 5.0 \times 10^{-8}$ were selected for the enriched traits, which were considered 'GWAS hits'. As control the analysis was repeated for the 3,399 network communities that were not significantly targeted (Supplementary Table 8). For both targeted and non-targeted communities the probability of observing traits that are linked to COVID-19 outcomes was assessed. A literature survey identified 35 traits clinically linked to COVID-19 (score 2 in Supplementary Table 8), 18 'related to immune function' and 61 without connection. For the enrichment analysis we focused on the 'COVID-linked' traits; traits 'related to immune function' are also indicated in Fig. 3. Finally, Fisher's exact test was used to assess the significance traits being linked to COVID-19 (score 2) vs not (scores 0 and 1) in traits that are associated with not-virus- targeted communities ($P = 0.5$) vs virally targeted communities ($P = 0.01$). For the control analysis of AP-MS targeted communities, only genetic variation related to COVID-19 severity was evaluated. The contactome-targeted communities with significant GWAS trait associations were numbered 1–31.

Small-molecule inhibition

Remdesivir (Bio-Techne, 7226/10) and USP25/28 inhibitor AZ1 (Bio-Connect, HY-117370-5mg) were dissolved in DMSO. HEK293-ACE2 and Vero E6 (3×10^4 cells per well) were plated in white 96-well plates. After 24 h, cells were infected with SARS-CoV-2 (ref. [54]) (0.01 MOI) containing a nanoluciferase reporter and treated with the compounds in a 12-point twofold dilution series with 0–10 µM concentration. Each condition was done in triplicate, except for AZ1, which was done in quadruplicate for HEK293-ACE2 and one replicate for Vero E6. Cells were cultured for 24 h, and luminescence was quantified[88]. Cell viability was measured using the Cell Titer-Glo Luminescent Cell Viability Assay kit (Promega, G7750). EC_{50} values were calculated via the variable slope model in GraphPad Prism v9.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The protein-protein interaction (PPI) data from this publication have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct and assigned the identifier IM-28880 (ref. [89]). All data from the study are included in the article and associated files. Source data are provided with this paper.

The following data were obtained from the respective original publications: phosphorylation changes upon SARS-CoV-2 infection[9,10]; RNA-binding changes upon SARS-CoV-2 infection[11]; AP-MS virus–host association data: Gordon et al.[12,13], Stukalov et al.[9], Li et al.[14], Nabeel-Shah et al.[15]; BioID virus–host proximity data: Laurent et al.[16], St-Germain et al.[17] Samavarchi-Tehrani et al.[18]; human expression data: Human Proteome Atlas[21], SARS-CoV-2 organotropism[22,82]; human host interactome: HuRI[26]; GWAS data for severe COVID-19 illness[32,33]; and GWAS summary statistics for 114 traits: doi:10.5281/ZENODO.3518299. Interaction data for other viruses were downloaded from IntAct[8] (version: 28 April 2020). Publication counts were downloaded from gene2pubmed (NCBI) on 16 November 2021. Source data are provided with this paper.

Code availability

All source code related to this paper is available via GitHub (https://github.com/INET-HMGU/SARS-CoV-2-contactome)[90].

References

60. Kim, D.-K. et al. A comprehensive, flexible collection of SARS-CoV-2 coding regions. *G3* **10**, 3399–3402 (2020).
61. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
62. Wu, A. et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
63. Jungreis, I. et al. Conflicting and ambiguous names of overlapping ORFs in the SARS-CoV-2 genome: A homology-based resolution. *Virology* **558**, 145–151 (2021).
64. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
65. Altmann, M., Altmann, S., Falter, C. & Falter-Braun, P. High-quality yeast-2-hybrid interaction network mapping. *Curr. Protoc. Plant Biol.* **3**, e20067 (2018).
66. Weile, J. et al. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).
67. The ORFeome Collaboration. The ORFeome collaboration: a genome-scale human ORF-clone resource. *Nat. Methods* **13**, 191–192 (2016).
68. Fisher, Y. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. CoRR abs/1511.07122 (JMLR.org, 2016): n. pag.
69. Maas, A. L, Hannun, A. Y & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning*, **30** (Atlanta, GA, 2013).

70. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at *arXiv* https://arxiv.org/abs/1412.6980 (2014).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
72. Chen, C. et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants.*Bioinformatics* **38**, 1735–1737 (2021).
73. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
74. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193 (2007).
75. Carbon, S. et al. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288 (2009).
76. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412 (2021).
77. Shin, C. et al. MKRN2 is a novel ubiquitin E3 ligase for the p65 subunit of NF-κB and negatively regulates inflammatory responses. *Sci. Rep.* **7**, 46097 (2017).
78. Götte, B. et al. Separate domains of G3BP promote efficient clustering of alphavirus replication complexes and recruitment of the translation initiation machinery. *PLoS Pathog.* **15**, e1007842 (2019).
79. Hosmillo, M. et al. Noroviruses subvert the core stress granule component G3BP1 to promote viral VPg-dependent translation. *eLife* **8**, e46681 (2019).
80. Liu, S., Dominska-Ngowe, M. & Dykxhoorn, D. M. Target silencing of components of the conserved oligomeric Golgi complex impairs HIV-1 replication. *Virus Res.* **192**, 92–102 (2014).
81. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* **25**, 402–408 (2001).
82. Meinhardt, J. et al. Olfactory transmucosal SARS-CoV-2 invasion as a port of central nervous system entry in individuals with COVID-19. *Nat. Neurosci.* **24**, 168–175 (2020).
83. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **28**, 84–90 (2012).
84. Kalinka, A. T. & Tomancak, P. linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics* **27**, 2011–2012 (2011).
85. Chapple, C. E. et al. Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* **6**, 7412 (2015).
86. Barbeira, A. N. *et al*. GWAS and GTEx QTL integration. *Zenodo* https://doi.org/10.5281/ZENODO.3518299 (2019).
87. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
88. Coutant, E. P. et al. Bioluminescence profiling of NanoKAZ/NanoLuc luciferase using a chemical library of coelenterazine analogues. *Chemistry* **26**, 948–958 (2020).
89. Kim, D.K. et al. IM-28880. IMEx. https://www.ebi.ac.uk/legacy-intact/query/pubid:unassigned2933;jsessionid=E9D9D501AAC618B88078DBD0BD47AEFA?conversationContext=1 (2022).
90. Kim, D.K. et al. SARS-CoV-2-contactome. GitHub. https://github.com/INET-HMGU/SARS-CoV-2-contactome (2022).
91. Barron, E. et al. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. *Lancet Diabetes Endocrinol.* **8**, 813–822 (2020).
92. Leong, A. et al. Cardiometabolic risk factors for COVID-19 susceptibility and severity: a Mendelian randomization analysis. *PLoS Med.* **18**, e1003553 (2021).
93. Nikniaz, Z., Somi, M. H., Dinevari, M. F., Taghizadieh, A. & Mokhtari, L. Diabesity associates with poor COVID-19 outcomes among hospitalized patients. *J. Obes. Metab. Syndr.* **30**, 149–154 (2021).
94. Aung, N., Khanji, M. Y., Munroe, P. B. & Petersen, S. E. Causal inference for genetic obesity, cardiometabolic profile and COVID-19 susceptibility: a Mendelian randomization study. *Front. Genet.* **11**, 586308 (2020).
95. Freuer, D., Linseisen, J. & Meisinger, C. Impact of body composition on COVID-19 susceptibility and severity: a two-sample multivariable Mendelian randomization study. *Metabolism* **118**, 154732 (2021).
96. Wang, C. et al. Red cell distribution width (RDW): a prognostic indicator of severe COVID-19. *Ann. Transl. Med.* **8**, 1230 (2020).
97. Ouyang, S.-M. et al. Temporal changes in laboratory markers of survivors and non-survivors of adult inpatients with COVID-19. *BMC Infect. Dis.* **20**, 952 (2020).
98. Kearns, S. M. et al. Reduced adiponectin levels in patients with COVID-19 acute respiratory failure: a case-control study. *Physiol Rep.* **9**, e14843 (2021).
99. Hypothyroidism is associated with prolonged COVID-19-induced anosmia: a case-control study. *J. Neurol. Neurosurg. Psychiatry* **20**, jnnp–2021–326587 (2021).
100. Brancatella, A. et al. Subacute thyroiditis after SARS-CoV-2 infection. *J. Clin. Endocrinol. Metab.* **105**, dgaa276 (2020).
101. Nemani, K. et al. Association of psychiatric disorders with mortality among patients with COVID-19. *JAMA Psychiatry* **78**, 380–386 (2021).
102. Zhu, Z. et al. Association of obesity and its genetic predisposition with the risk of severe COVID-19: analysis of population-based cohort data. *Metabolism* **112**, 154345 (2020).
103. Derikx, L. A. A. P. et al. Clinical outcomes of COVID-19 in patients with inflammatory bowel disease: a nationwide cohort study. *J. Crohns. Colitis* **15**, 529–539 (2021).
104. Dar, H. Y., Azam, Z., Anupam, R., Mondal, R. K. & Srivastava, R. K. Osteoimmunology: the between bone and immune system. *Front. Biosci.* **23**, 464–492 (2018).

Acknowledgements

We thank P. Charneau for the hACE2 lentivirus. This work was supported by a Canadian Institutes for Health Research Foundation Grant (F.P.R.), the Canada Excellence Research Chairs Program (F.P.R.), the Thistledown Foundation (F.P.R.); the LabEx Integrative Biology of Emerging Infectious Diseases (10-LABX-0062; Y.J., C.D.) and Platform for European Preparedness Against (Re-)emerging Epidemics, EU (602525; Y.J. and C.D.), the European Union's Horizon 2020 Research and Innovation Programme (Project ID 101003633, RiPCoN; P.F.-B., C.B., P.A.), HDHL-INTIMIC 'Interrelation of the Intestinal Microbiome, Diet and Health' (BMBF Project ID 01EA1803; P.F.-B.), the Free State of Bavaria's AI for Therapy (AI4T) Initiative through the Institute of AI for Drug Discovery (AID) (P.F.-B.) and Fonds de la Recherche Scientifique (FRS-FNRS) grant PER-40003579 (J.-C.T., L.W.). F.L. was supported by a Belgian American Educational Foundation doctoral research fellowship, a Wallonia-Brussels International (WBI)-World Excellence fellowship and Fonds de la Recherche Scientifique (FRS-FNRS)-Télévie grant FC31747 (Crédit n° 7459421F). M.V. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, Wallonia-Brussels Federation, Belgium). C.P. was supported by a Ramon y Cajal fellowship (RYC-2017–22959). G.D. was supported by the Ministère de l'Education Nationale, de la Recherche et de l'Innovation with a fellowship from Université Paris Cité.

Author contributions

Conceptualization: D.-K.K., B.W., C.-W.L., D.S., J.J.K., C.F., F.P.R. and P.F.-B.; SARS-CoV-2 clone construction: D.-K.K., B.W., J.J.K., M.S., H.H., N.M.R., O.P. and P.C.; yeast two-hybrid screening: D.-K.K., B.W., D.S.,

J.J.K., N.K., A.R., V.Y., H.H., O.P., A.S., R.L., S.T.R., M.A., A.G.C., L.E.V., I.H., B.B.L., M.N., R.P., P.A.R.C. and P.G.; N2H validation: D.-K.K., B.W., J.J.K., K.S., F.L., P.S., L.W., Y.J., T.H., D.E.H., C.D., M.V. and M.A.C.; data analysis: B.W., C.-W.L., D.S., A.Z., C.P., L.L., P.S., B.D., C.B., M.H., C.F., P.A. and P.F.-B.; follow-up experiments: D.-K.K., B.W., J.J.K., G.D., M.J.T., S.B.M., J.P., M.T., J.-C.T., Y.J., D.K., C.D. and P.F.-B.; funding acquisition: J.-C.T., C.F., F.P.R. and P.F.-B.; supervision: C.B., J.-C.T. P.A., C.D., M.V., M.A.C., F.P.R. and P.F.-B.; writing: D.-K.K., B.W., C.-W.L., D.S., J.J.K., G.D., A.Z., C.P., M.T., A.R., O.P., F.L., B.B.L., M.N., J.-C.T., D.K., C.F., C.D., M.V., M.A.C., F.P.R. and P.F.-B.

Funding

Open access funding provided by Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH).

Competing interests

F.R. and M.V. are advisors and shareholders of SeqWell, Inc. (Beverly, MA, USA).

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01475-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01475-z.

Correspondence and requests for materials should be addressed to Caroline Demeret, Marc Vidal, Michael A. Calderwood, Frederick P. Roth or Pascal Falter-Braun.

Peer review information *Nature Biotechnology* thanks Ulrich Stelzl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

C. A gut meta-interactome map reveals modulation of human immunity by microbiome effectors

TITLE

A gut meta-interactome map reveals modulation of human immunity by microbiome effectors

AUTHORS

Veronika Young[1,15], Bushra Dohai[1,15], Thomas C. A. Hitch[2], Patrick Hyden[3], Benjamin Weller[1], Niels S. van Heusden[4], Deeya Saha[5], Jaime Fernandez Macgregor[5,6], Sibusiso B. Maseko[7], Chung-Wen Lin[1], Mégane Boujeant[5], Sébastien A. Choteau[5], Franziska Ober[8], Patrick Schwehn[1], Simin Rothballer[1], Melina Altmann[1], Stefan Altmann[1], Alexandra Strobel[1], Michael Rothballer[1], Marie Tofaute[8], Matthias Heinig[9,10], Thomas Clavel[2], Jean-Claude Twizere[7,11,12], Renaud Vincentelli[6], Marianne Boes[4], Daniel Krappmann[8], Claudia Falter[1], Thomas Rattei[3], Christine Brun[5,13], Andreas Zanzoni[5], Pascal Falter-Braun[1,14]✉

AFFILIATIONS

[1]Institute of Network Biology (INET), Molecular Targets and Therapeutics Center (MTTC), Helmholtz Munich; Neuherberg, Germany.

[2]Functional Microbiome Research Group, Institute of Medical Microbiology, University Hospital of RWTH Aachen; Aachen, Germany.

[3]Department of Microbiology and Ecosystem Science, Research Network: Chemistry meets Microbiology, University of Vienna; Vienna, Austria.

[4]Center for Translational Immunology, University Medical Center Utrecht; Utrecht, The Netherlands.

[5]Aix Marseille Univ, INSERM, TAGC, Turing Center for Living Systems; Marseille, France.

[6]Aix Marseille Univ, CNRS, AFMB, Turing Center for Living Systems; Marseille, France.

[7]Laboratory of Viral Interactomes, GIGA Institute, University of Liège; Liège, Belgium.

[8]Research Unit Signaling and Translation, Group Signaling and Immunity, Molecular Targets and Therapeutics Center (MTTC), Helmholtz Munich; Neuherberg, Germany.

[9]Institute of Computational Biology (ICB), Computational Health Center; Neuherberg, Germany.

[10]Department of Computer Science, TUM School of Computation, Information and Technology, Technical University of Munich; Garching, Germany.

[11]TERRA Teaching and Research Centre, University of Liège; Gembloux, Belgium.

[12]Laboratory of Algal Synthetic and Systems Biology, Division of Science, New York University Abu Dhabi; Abu Dhabi, United Arab Emirates.

[13]CNRS; Marseille, France.

[14]Microbe-Host Interactions, Faculty of Biology, Ludwig-Maximilians-Universität München; Planegg-Martinsried, Germany.

[15]These authors contributed equally

✉email: pascal.falter-braun@helmholtz-munich.de

40

41 **Correspondence and requests for materials** should be addressed to Pascal Falter-Braun.

42

43 **KEYWORDS**

44 microbiome, type-3-secretion system, virulence effectors, complex diseases, network biology,

45 protein-protein interactions, immune signaling, interactome

SUMMARY

The molecular mechanisms by which the gut microbiome influences human health remain largely unknown. Pseudomonadota is the third most abundant phylum in normal gut microbiomes. Several pathogens in this phylum can inject so-called virulence effector proteins into host cells. We report the identification of intact type 3 secretion systems (T3SS) in 5 - 20% of commensal Pseudomonadota in normal human gut microbiomes. To understand their functions, we experimentally generated a high-quality protein-protein meta-interactome map consisting of 1,263 interactions between 289 bacterial effectors and 430 human proteins. Effector targets are enriched for metabolic and immune functions and for genetic variation of microbiome-influenced traits including autoimmune diseases. We demonstrate that effectors modulate NF-κB signaling, cytokine secretion, and adhesion molecule expression. Finally, effectors are enriched in metagenomes of Crohn's disease, but not ulcerative colitis patients pointing toward complex contributions to the etiology of inflammatory bowel diseases. Our results suggest that effector-host protein interactions are an important regulatory layer by which the microbiome impacts human health.

MAIN

The host-associated microbiota influences human health in complex host genetics-dependent ways[1,2]. Especially intestinal microbes positively and negatively affect the risk for several complex diseases ranging from inflammatory bowel disease (IBD)[1] and asthma[3] to metabolic[4] and neurodegenerative diseases[5]. Members of the bacterial phylum Pseudomonadota (previously: Proteobacteria[6]) are prevalent in the human gut microbiome and their occurrence is influenced by dietary ingredients such as fat and artificial sweeteners[7]. Unique features of this phylum are the type-3, type-4, and type-6 secretion systems (TxSS) that enable the injection of bacterial proteins directly into the host cytosol. The presence of T3SS has been classically associated with pathogen virulence[8]. In the plant kingdom, however, important mutualistic microbes also communicate with the host via effector proteins to establish cohabitation and elicit host-beneficial effects[9]. We therefore wondered if commensal Pseudomonadota in the healthy human gut microbiome possess host-directed secretion systems.

T3SS are common in the normal human gut microbiome

Because of the higher quality and completeness of genome assemblies from cultured strains compared to metagenome-assembled genomes (MAGs), we first evaluated Pseudomonadota strains from gut and stool samples that were collected, among others, by the human microbiome project and were available from culture collections. Using EffectiveDB[10], a widely used tool for secretion system identification, we detected complete T3SS in 44 of the 77 reference strain genomes (Extended Data Table 1). To expand the scope, we analyzed genomes of 4,752 distinct strains, representing all major phyla from the human gut that had been isolated by the human gastrointestinal bacteria genome collection (HBC)[11], and the Unified Human Gastrointestinal Genome (UHGG) collection[12,13]. Of the 2,272 Gram-negative strains, 478 (21%) had complete T3SS (Fig. 1a); similar proportions have T4SS (527) and T6SS (719), both of which can also deliver effectors into host cells but also have other functions (Extended Data Fig. 1 and Extended Data Table 1)[14]. Together 729 of the 2,272 Gram-negative strains, *i.e.*, 34%, have at least one host-directed secretion system. Because culturing can bias the relative proportions of taxa, we sought to confirm the presence of T3SS in commensal microbiota using metagenome datasets. From 16,179 Pseudomonadota MAG bins with high or intermediate genome quality[15,16], 770, i.e., 5%, encoded complete T3SS (Fig. 1a and Extended Data Table 1). Notably, we only identified T3SS in Gammaproteobacteria, whereas no secretion systems were found in the Beta- or Epsilonproteobacteria in the datasets, except for a few *Helicobacter* strains. It is unclear if gut commensal strains in these orders lack T3SS, or if the systems differ from those of the better-characterized Gammaproteobacteria and they were missed by the algorithm. Across the analyses, T3SS were identified in strains of multiple

genera and were especially common among *Escherichia* (Fig. 1b and Extended Data Table 1). Notably, a recent *in vivo* profiling study of human digestive tracts using *in situ* sampling found *Escherichia* as the genus that was most significantly enriched in intestinal over stool samples[17]. Of the T3SS-positive (T3SS$^+$) species, 24 matched representatives in two cohorts of a dataset provided by the Weizmann Institute of Science (WIS cohorts)[18]. 59.4% of individuals in the Israeli cohort and 47.1% in the Dutch cohort had potentially T3SS$^+$ species in their gut microbiome, with relative abundances of 0.80% and 0.48%, respectively (Fig. 1c). The most common T3SS$^+$ species in both cohorts was *Escherichia coli*, appearing within 54% and 45% of individuals, respectively. Overall, T3SS$^+$ strains constitute a substantial proportion of commensal Pseudomonadota and are common in normal human gut microbiomes. We therefore aimed to understand the functions of T3SS-delivered effector proteins of commensal strains.

Commensal effectors are unrelated to known pathogen effectors

To identify gut microbiome-encoded effectors we used a combination of three complementary machine learning models[19-21] and considered 3,002 effector candidates from the 44 reference strains that were most confidently predicted by all tools (Extended Data Table 2). In addition, we identified 186 putative effectors in the 770 T3SS$^+$ MAGs (Extended Data Table 2). As T3SS and substrate effectors are best known for their role in supporting a pathogenic lifestyle, we investigated if the commensal bacterial effectors share sequence similarity with 1,638 known T3SS effectors from pathogens[22]. Only 17 of 3,002 (0.5%) effectors from strains and 6 of 186 (3%) from MAGs, respectively, showed extended high sequence similarity (≥ 90% sequence similarity across ≥ 90% length) to known pathogen effectors; lowering the thresholds to 50% similarity across 75% length only marginally increased the numbers to 34 (1%) and 7 (4%), respectively (Fig. 1d and Extended Data Table 2). The largest number of commensal effectors with similarity to pathogenic effectors were found in the genomes of *Escherichia albertii* (12 effectors with 67% to 98% identity) and *Yersinia enterocolitica* (10 effectors at > 98% identity). The fact that all such pathogen-similar commensal effectors were found in different species, of which some even belong to a different order than the respective pathogen, suggests that non-pathogenic microbes participate in the horizontal gene transfer of effectors[23,24]. This is supported by the observation that only a few pathogen-similar effectors were found among the approximately 20 - 80 effectors of each strain. Of the six pathogen-similar effectors found in MAGs, all but one matched the identified family of the pathogen from which they were initially identified (Extended Data Fig. 2 and Extended Data Table 2). Plausibly, these effectors originate from pathogens, or their relatives that were likely present in some samples. Jointly, the data show that effector complements of commensal bacteria are distinct from those of pathogens, thereby suggesting functions outside of the pathogen lifestyle.

A microbiome-host protein-protein meta-interactome map

To investigate the functions of commensal effectors, we cloned effector ORFs for experimental studies from 18 bacterial strains with diverse effector complements (Fig. 1e and Extended Data Fig. 1). We successfully PCR-cloned 786 ORFs for the 1,300 encoded effectors (60.2%) and 173 of 186 effector ORFs from MAG bins (meta-effectors) following chemical synthesis (Fig. 2a). Thus, 959 sequence-verified full-length effector ORFs were assembled as the human microbiome effector ORFeome (HuMEOme_v1) (Extended Data Table 2). With these, we conducted binary interactome (contactome) network mapping against the human ORFeome9.1 collection encoding 18,000 human gene products using a stringent multi-assay mapping pipeline[25]. In the main screen by yeast-2-hybrid (Y2H), we identified 1,071 interactions constituting the human-microbiome meta-interactome main dataset ($HuMMI_{MAIN}$) (Fig. 2b,c). To assess sampling sensitivity[26], i.e., saturation of the screen, we conducted three additional repeats of 290 randomly picked effectors and 1,440 human proteins, which yielded 39 verifiable interactions constituting the HuMMI repeat subset ($HuMMI_{RPT}$). The saturation curve indicates that the single main screen has a sampling sensitivity of ~32% (Fig. 2d). Last, to address how effector sequence similarity affects their interaction profiles we conducted a homolog screen. Effectors were grouped if they shared ≥ 30% sequence identity (Extended Data Table 2) and all effectors of one group were experimentally tested against the union of their human interactors. The resulting dataset ($HuMMI_{HOM}$) contains 398 verified interactions, of which 179 were not found in the other screens. Altogether, HuMMI contains 1,263 unique verified interactions between 289 effectors and 430 human proteins (Fig. 2b,c and Extended Data Table 3).

To assess data quality, we assembled a positive control set of 67 well-documented manually curated binary interactions of bacterial (pathogen-) effectors with human proteins from the literature (bacterial human literature binary multiple – bhLit_BM-v1, Extended Data Table 3) and a corresponding negative control set of random bacterial and human protein pairs (bacterial host random reference set - bhRRS-v1). Benchmarking our Y2H assay in a single orientation with these and with the established human positive reference set (hsPRS-v2) and hsRRS-v2 indicated an assay sensitivity of ~13% and 17.5%, respectively, which is consistent with previous observations[27,28] (Fig. 2e and Extended Data Table 3). No negative control pair in either reference set scored positive, demonstrating the reliability of our system. In addition, we assessed the biophysical quality of HuMMI using the yeast nanoluciferase-2-hybrid assay (yN2H), which we benchmarked using the same four reference sets[25]. Notably, the retest rates of all sets involving bacterial proteins were lower than those of the human hsPRS-v2 and hsRRS-v2 across most of the scoring spectrum (Extended Data Fig. 2). Partly, this could be due to the nature of hsPRS-v2 pairs, which consist of very well-documented interaction pairs,

which may have been selected for good detectability. In addition, the fact that the RRS sets exhibit the same overall trend indicates that interactions with prokaryotic proteins are more challenging to reproduce in this eukaryotic assay system, which reinforces the necessity for bacterial protein-specific reference sets (Fig. 2f, Extended Data Fig. 2, and Extended Data Table 3). At thresholds where the control sets were well separated, the retest rate of 173 randomly selected HuMMI interactions was statistically indistinguishable from the positive control sets, and significantly different from those of the negative controls (Fig. 2f, Extended Data Fig. 2, and Extended Data Table 3), indicating that the biophysical quality of our dataset is comparable to those of well-documented interactions in the curated literature.

The degree distribution of $HuMMI_{MAIN}$ shows that numerous human proteins are targeted by multiple effectors (Fig. 2g and Extended Data Table 3), often from different species. Indeed, sampling analysis demonstrates that commensal effectors significantly converge on fewer host proteins than expected from a random process (Fig. 2h), thus suggesting selection for interactions with these targets. We had previously observed convergence of effectors from phylogenetically diverse pathogenic microbes on common proteins of their plant host[29,30]. In that system, we demonstrated with infection assays on genetic null mutant plant lines that the extent of convergence correlates with the importance of the respective host proteins for the outcome of the microbe-host interaction[29]. We therefore identified the human host proteins onto which commensal effectors converge. To this end, we sampled random effector targets for each strain and analyzed the distribution of repeatedly targeted proteins (Fig. 2i). While host proteins interacting with effectors from two strains are expected at high frequency by chance, targeting by four bacterial strains is unlikely to emerge by chance (Fig. 2i and Extended Data Table 3). Thus, the 60 human proteins targeted by effectors from four or even more commensal strains are subject to effector convergence and may be of general importance for human microbe-host interactions. Together with our recently published plant-symbiont interaction data[31], these data suggest that convergence has evolved as a universal feature of effector-host interactions independent of the microbial lifestyle and kingdom of the host organism.

Sequence features mediating effector-host interactions

The function of unknown proteins can often be inferred from better-studied orthologues, but convergence could also result from high sequence similarity among effectors. We therefore compared sequence- to interaction-similarity as a proxy for their function in host cells (Fig. 3a). Within the systematically retested $HuMMI_{HOM}$ clusters, both are poorly correlated and sequence similarity merely defines the upper limit for interaction similarity but does not imply it. This is illustrated by cluster 3, in which all seven effectors share over 90% mutual sequence

similarity while their pairwise interaction profile similarities range from identical to complementary (Fig. 3b and Extended Data Table 3).

Using $HuMMI_{MAIN}$ we also investigated if effectors without substantial sequence similarity share interaction similarity, which might indicate shared functions. In fact, clustering effectors by their pairwise interaction similarity identified substantial overlap outside the homology clusters (Extended Data Fig. 3), indicating that dissimilar effectors may have similar functions in the host. Both analyses indicate that effector function as measured by protein-interaction profiles is largely independent of overall sequence similarity.

Looking for structural correlates for interaction specificity, we wondered whether domain-domain or domain-short linear motif (SLiM) interfaces mediating the interactions can be identified (Fig. 3c). Using experimentally identified interaction templates[32], a putative interface was found for 52 interactions in the $HuMMI_{MAIN}$ screen (Extended Data Table 4). Of these, 43 interactions matched motif-domain templates passing one (Fig. 3d), and 22 passing two stringency criteria (Extended Data Fig. 3). Among the former, 23 interactions involve PDZ domains in the human protein, which recognize PDZ-binding motifs (PBM) in the C-terminus of interacting bacterial proteins. PDZ domain-containing proteins commonly mediate cell-cell adhesion, cellular protein trafficking, tissue integrity, as well as neuronal and immune signaling[33]. To experimentally validate these interfaces, individual and tandem PDZ domains from 13 human proteins and C-terminal peptides from 16 interacting bacterial effectors were tested via Holdup, a quantitative chromatographic *in vitro* interaction assay[34,35]. For 16 of 23 Y2H pairs (70%) at least one PDZ-peptide interaction was identified, all with affinities between 1 and 200 µM (Fig. 3e and Extended Data Table 4). In three instances two PDZ domains arranged in tandem were required to detect the interaction by Holdup, indicating that some Y2H pairs might have been missed because not all PDZ combinations of the proteins were tested. For human proteins with multiple PDZ domains, often different domains were the target for different effectors demonstrating both specificity and functional specialization of the effectors (Fig. 3e).

Because of their functioning in immune signaling and cell shape, PDZ domains are frequently targeted by viruses[36]. This opens the possibility that bacterial effectors and viral proteins compete for PDZ-binding and thus mutually influence their respective impact on the host. To gather support for this possibility, we identified viruses that can cause infections in the digestive tract, namely SARS-CoV-2[37], HPV16 and 18, which have a high prevalence in human guts and have been linked to colorectal cancer[38], and norovirus, a globally common cause of gastroenteritis and diarrhea[39]. We selected two hitherto unpublished interactions of Norovirus VP2 C-terminal peptide with DLG1 (domain 2) and MAGI1 (domain 4), and previously observed interactions between the C-terminal peptides of SARS-CoV-2 E with SHANK3, and

of HPV16 and 18 E6 with the PDZ domains of PICK1 and MAGI4 (domain 1), respectively[34]. Indeed, in fluorescent polarization assays the viral PBM peptides competed with those of the effectors Vfu_12, met_32, met_31, and met_46 (Fig. 3f and Extended Data Fig. 3). Similarly, the functionally well-characterized interaction of the C-terminus of HTLV1 Tax1 with DLG1[40] was competed off by the met_32 PBM peptide. Thus, viral and bacterial proteins may compete in the intracellular environment for binding partners and hence for influence on human cell function. Such competition could contribute to the previously observed mutual influence of microbiome and viral infection on each other[41].

Thus, while the overall sequence similarity of effectors does not correlate with their host-protein interaction profiles, several interfaces mediating the interactions can be identified. How these interactions compete with human and viral proteins to modulate the host network is an important question for future studies.

Effector-targeted functions and disease modules

To explore the potential roles of commensal effectors in the host we analyzed the functions of the targeted human proteins through gene ontology (GO) enrichment analysis (Fig. 4a, Extended Data Fig. 4, and Extended Data Table 5). Redundant parent-child GO-term pairs were grouped and are displayed by a representative term. Intriguingly, “response to muramyl-dipeptide (MDP)”, a bacterial cell wall-derived peptide that can be perceived by human cells, was among the most enriched functions, thus not only supporting the relevance of our interactions but indicating that effectors modulate cellular responses to their detection. Moreover, a key component of the MDP signaling pathway is NOD2, which is encoded by a major susceptibility gene for Crohn’s disease (CD)[42], an autoimmune disease with a strong etiological microbiome contribution[43]. In addition, several central immune signaling pathways are enriched among the targets, namely the NF-κB and the stress-activated protein kinase and Jun-N-terminal kinase (SAPK/JNK) pathways, supporting the notion that modulation of immune signaling is an important function of commensal effectors. Remarkably, five of the significantly targeted convergence-proteins belong to the NF-κB module (Extended Data Fig. 4), one of the evolutionarily oldest immune signaling pathways in animals that is already present in sponges[44]. This may reflect the long co-evolution between microbial effectors and this ancient immune coordinator. Relating to human disease, anti-TNF biologicals, which dampen NF-κB-driven immunity, are an important therapeutic for diverse autoimmune diseases including CD, psoriasis, and rheumatoid arthritis. Another highly enriched group of five terms relates to collagen production, which suggests that effectors may modulate the extracellular environment that hosts the microbes. Inflammation-independent fibrotic collagen production is an important clinical feature of CD, and the gut microbiota has been found to be a main driver[45]. As several metabolism-related terms were identified, we also tested directly

whether enzymes in the Recon3D[46] model of human metabolism were targeted. Indeed, we detected a significant enrichment of metabolic enzymes (P = 0.0001, Fisher's exact test) and nominally significant targeting of bile acid and glycerophospholipid metabolism, and fatty acid oxidation (Extended Data Table 5). Overall, however, despite the strong overall signal and general targeting of fatty acid metabolism, no individual metabolic subsystem stood out as being targeted by effectors from more than two strains or having more than two targeted proteins.

From a network perspective, proteins encoded by disease-genes (disease proteins) constitute nodes and form disease modules[47], whose functional perturbation promotes pathogenesis. Importantly, viruses can contribute to non-infectious disease etiology by binding to and similarly perturbing these disease proteins and modules[48]. Therefore, we wondered if bacterial effectors also target such network elements and may thereby influence human traits. We started with "causal genes/proteins" identified from genome-wide-association studies (GWAS) by the Open Targets initiative[49], and merged gene sets for traits identified as identical by their experimental factor ontology (EFO) terms (Extended Data Table 5). We first investigated direct effector targets. The strong enrichment of the "immunoglobulin isotype switching" trait among these is intriguing as the evolutionarily older IgA antibodies are emerging as having an important role in shaping the gut microbiome[50,51]. Effector-targeted proteins are further associated with diverse cancers and with diseases that have a strong immunological component, including asthma, psoriasis, allergies, and systemic lupus erythematosus (Fig. 4b, cutoff nominal P = 0.05, Fisher's exact test, Extended Data Table 5). While none of the identified diseases is currently known as an ailment of the gut it has emerged that the gut microbiome shapes immune homeostasis and contributes to lung and skin diseases like asthma[52] and psoriasis[53]. In addition, some of the disease-associated genes encode convergence proteins for effectors from multiple bacterial species (Fig. 2g). As such, it is plausible that proteins like REL or TCF4 are similarly targeted by effectors from Pseudomonadota in skin or lung microbiome communities and contribute to the identified diseases. Moreover, 26% of the effectors in HuMMI are also detectable in skin microbiome samples (Extended Data Table 5), indicating that commensal effectors are shared between different ecological niches.

A partly complementary explanation emerges from our previous studies of human and plant pathogen-host systems. In these evolutionary distant systems, we showed that genetic variation affecting the severity of infection does not reside in genes encoding direct targets but in interacting, i.e., neighboring proteins in the host network[25,29]. We, therefore, explored the network neighborhood of all effector-targets using short random walks in the human reference interactome (HuRI)[54]. We identified proteins that were significantly more often visited in HuRI

compared to degree-preserved randomly rewired networks, which we considered the 'neighborhood'. For each effector-targeted neighborhood, we assessed the enrichment of gene products associated with diverse human traits using Open Targets causal genes. Nominally significant associations were aggregated on a strain level and summarized for disease groups (Fig. 4c and Extended Data Table 5). Intriguingly, most disease groups for which susceptibility-gene products are enriched in the target neighborhoods represent traits that have been linked to the gut microbiome[55]. Apart from immunological traits, these include cardiovascular, metabolic, and neurological traits as well as multiple cancers, including colorectal cancer. Among the target neighborhoods for immunological diseases, we identified associations to CD (nominal $P = 8.5 * 10^{-5}$, Fisher's exact test) and inflammatory bowel disease (nominal $P = 0.0008$, Fisher's exact test) but not to ulcerative colitis (UC) (Fig. 4d and Extended Data Table 5). Neighborhoods harboring genetic susceptibility associated with psoriatic arthritis, asthma, and allergies were also significantly targeted, which recapitulates the observations for direct targets. Considering the importance of the microbiome for human metabolic disorders[55] it is noteworthy that network modules important for HDL and LDL cholesterol levels (nominal $P = 0.006$ and $P = 0.008$, respectively, Fisher's exact test), and several diabetes traits were significantly targeted albeit less recurrently than inflammatory diseases and cancers (Extended Data Table 5). Together, these results suggest that commensal effectors modulate their host's immune system and local metabolic and structural microenvironment. As genetic variation affecting the targeted proteins and their network neighborhood is linked to several human diseases, functional modulation of the same network neighborhoods by commensal effectors may contribute to disease etiology. The fact that the risk for several of the identified diseases is known to be modulated by the microbiome strengthens this hypothesis. We therefore investigated if commensal effectors, indeed, perturb some of the identified pathways and functions.

Effector function in human cells and disease

The NF-κB signaling module is enriched among the convergence proteins and all targets of commensal effectors (Fig. 4a and Extended Data Fig. 4). Because of its important role in many diseases, we chose a cell-based dual-luciferase assay[25] to test whether commensal effectors modulate NF-κB pathway activity in human cells. Indeed, five of 26 commensal effectors caused a significant increase in NF-κB pathway activity in the absence of exogenous stimulation suggesting pathway activation (Fig. 5a and Extended Data Table 6). Conversely, three effectors significantly reduced relative transcriptional NF-κB activity even in the presence of strong TNF stimulation (Fig. 5b, Extended Data Fig. 5, and Extended Data Table 6). Since some bacterial effectors also modulate NF-κB-independent induction of the thymidine kinase control promoter, we assessed the impact of selected effectors on endogenous expression of

NF-κB controlled human adhesion factor ICAM1 and cytokine secretion. We focused these experiments on two NF-κB activating (Kpn_9, met_7) and two NF-κB inhibiting (Pst_11, Cyo_12) bacterial effectors. ICAM1/CD54 is a glycoprotein that mediates intercellular epithelial adhesion and interactions with immune cells, specifically neutrophils. Epidemiologically, ICAM1 has been linked to CD such that increased ICAM1 expression is associated with higher disease risk[56] likely by facilitating recruitment and retention of inflammatory immune cells[57,58]. Interference with ICAM1-mediated neutrophil trafficking is currently being tested as a therapeutic approach to treat CD[59]. In colon carcinoma Caco-2 cells, expression of met_7 caused a significant increase of ICAM1 expression (P = 0.05, one-way ANOVA with Dunnett's multiple hypothesis correction, Extended Data Table 6) following stimulation with a pro-inflammatory cocktail. Expression of the inhibitory effectors Pst_11 and Cyo_12 did not significantly alter the induction of ICAM1 cell surface expression (Fig. 5c). We also investigated the effect of met_7 and Cyo_12 on cytokine secretion in unstimulated Caco-2 cells or following pro-inflammatory stimulation. In basal conditions, Cyo_12 reduced the secretion of several cytokines especially IL6 and IL8, whereas met_7 caused an increase in IL8 secretion in these conditions (Fig. 5d and Extended Data Table 6). Following proinflammatory stimulation, expression of Cyo_12 further reduced cytokine secretion. This effect was most pronounced for IL8, but also significant for IL6 and the pro-inflammatory IL1beta, IL18, and IL23. These cytokines are noteworthy as they are linked to IBD pathogenesis. IL23R has been associated to CD, and IL6 and IL23 stimulate the differentiation of Th17 cells, which have emerged as key players in CD[60,61]. IL8 is overexpressed in colonic tissue of IBD patients and has been suggested as a chemoattractant triggering neutrophil invasion[62,63]. In contrast, no significant impact of met_7 on cytokine secretion was detectable in the context of stimulation (Fig. 5e and Extended Data Fig. 5). Thus, commensal effectors can both stimulate and dampen intracellular immune signaling and this modulation can impact immune and tissue homeostasis via cell-cell adhesion and cytokine secretion.

As we identified both genetic and functional links between commensal effectors and IBD-related processes, we sought clinical evidence for a potential role of effectors in these diseases. We hypothesized that a potential role of effectors in IBD etiology may be reflected in altered effector prevalence in the microbiota of patients versus healthy controls. Analyzing a large dataset with > 800 IBD patient-derived and > 300 healthy control-derived metagenomes[64] we found 64 effectors that were significantly more prevalent in the metagenomes of CD patients compared to healthy controls (Fig. 5f and Extended Data Table 6). In metagenomes of UC patients only three effectors had a significantly different prevalence, and, intriguingly, these were less common compared to healthy controls (Extended Data Table 6). This trend was recapitulated when the prevalence distributions of all detected effectors

were analyzed. Whereas CD patients had a significantly higher load of effectors, the overall effector prevalence was lower in UC patients compared to healthy subjects (Fig. 5g and Extended Data Table 6). These opposing findings were unexpected as an increased abundance of Pseudomonadota has been reported both for CD and UC patients[65]. At the same time, many clinical features such as affected tissues and response to anti-TNF therapy differ between these two forms of IBD, rendering it plausible that effectors contribute differently to their etiology. Whether commensal effectors indeed causally contribute to disease etiology or acute flairs is an important question with potential therapeutic implications.

Discussion

The presence of T3SS in human commensal microbes has been noticed previously and was speculated to mediate crosstalk between the intestinal microbiota and the human host[66,67]. Here, we provide evidence that, analogous to the plant kingdom[31,68], also in the human gut T3SS and effectors function in commensal microbe-host interactions and modulate immune signaling. Thus, effector secretion appears to be used universally by Pseudomonadota to mediate interactions with multicellular eukaryotes independently of the lifestyle of the microbe.

Since, as we show, commensal effectors modulate immune signaling we hypothesized that this may affect the manifestation of human diseases, especially those involving the immune system. The influence of the microbiome on IBD etiology is well documented[1]. Therefore, it is noteworthy that IBD, especially CD, emerged in several of our analyses. Effectors target the “response to the muramyl-dipeptide” pathway which includes NOD2, a major CD-associated gene product[69]. Further, effectors target and regulate the NF-κB pathway, which is strongly activated by TNF, a key therapeutic target in CD[70]. Likewise, ICAM1 is a susceptibility gene for CD whereby high expression increases disease risk[56]. Secretion of IL6, IL8, and IL23 is significantly altered by effectors, and all have previously been linked to CD[61,63]. Thus, commensal effectors regulate several IBD-relevant pathways and can thus influence the establishment or maintenance of feedback loops during disease development[71]. This conclusion is strengthened by the observation that effectors are enriched in metagenomes of a CD patient cohort. Thus, multiple lines of evidence suggest that by modulating immune signaling, commensal effectors contribute to the etiology of CD.

Likely other microbial habitats of the human body, such as skin or lung, also host T3SS+ strains, and we identified effectors in a skin metagenome. It will be important to investigate this in the future to understand if those effectors have similar targets and effects on local cells. ICAM1, e.g., is the entry receptor for rhinovirus A[72], and an increased expression due to microbial effectors could increase the risk for infections and thus to develop asthma[73,74].

418 The broader question of how effectors influence the pathogenesis of IBD and other diseases
419 will be important to address in further detailed studies. Our molecular data show that different
420 effectors can have opposing impacts on immune pathways, analogous to genetic variants.
421 Thus, host genetics and effectors jointly impact on the molecular networks, and pathogenic
422 developments emerge from the interplay of protective and disease enhancing factors. For CD
423 specifically, however, our analyses suggest that effectors promote disease development.

424 In summary, we demonstrate that bacterial effector proteins constitute a hitherto unrecognized
425 regulatory layer by which the commensal microbiota communicates with host cells and
426 modulates human physiology. We anticipate that our findings and resources will open new
427 research directions towards understanding the host-genetics dependent mechanisms by which
428 the microbiome influences human health and exploring the potential of effectors for therapy
429 and prevention.

METHODS

Identification of T3SS+ strains in culture collections and MAGs

To collect reference genomes for strains available from culture collections, three large culture collections were queried for all Pseudomonadota strains: DSMZ via BacDive[75], ATCC (atcc.org) and BEI (beiresources.org). The strain numbers were looked up in GenBank (Release 229) from which 77 strains could be identified as perfect match.

MAGs that were at least 50% complete and less than 5% contaminated (as estimated by CheckM[76] from two different meta-studies were selected. 92,143 MAGs of Almeida et al.[15] and 9,367 Pseudomonadota MAGs from Pasolli et al.[16] were used as input for T3SS prediction scaled via massive parallel computing. The computational predictions presented have been achieved in part using the Vienna Scientific Cluster (VSC). The prediction performance of EffectiveDB[10] on incomplete and contaminated MAGs was assessed by 5-fold cross-validation with 5 repeats using 0 - 100% completeness and 0 - 50% contamination in 5% steps of simulated incompleteness/contamination, randomly sampling genes from test-set. In addition, T3SS were predicted for 4,753 strains isolated by the human gastrointestinal bacteria genome collection (HBC)[11], and the unified gastrointestinal genome (UHGG) collection[12,13]. A performance-improved re-implementation of the EffectiveDB classifier (https://github.com/univieCUBE/phenotrex, trained on EggNOG 4 annotations[77]) was used to predict functional T3SS present in MAGs and genomes of isolated strains. Threshold for positive prediction was defined as > 0.7.

Protein sequences were predicted from 44 T3SS-positive reference strains and MAGs using prodigal v2.6.3[76]. Of 770 MAGs a total of 474,871 representative protein sequences were identified using CD-HIT[78] (v4.8.1, parameters: `-c 1.0`). The identical procedure was performed for 44 genomes from culture collections resulting in 161,115 proteins. Machine-learning based tools were used to predict T3SS signals (EffectiveT3 v.2.0.1 and DeepT3 2.0[19]) or effector homology using pEffect[21] to extract potential effector proteins. The results of all three tools were combined using a 0 - 2 scoring scheme: 2 for perfect score (pEffect > 90, EffectiveT3 > 0.9999, DeepT3: both classifiers positive prediction), 1 for positive prediction as defined by default settings (pEffect > 50, EffetiveT3 > 0.95, DeepT3: one classifier) and 0 for negative prediction. Sequences with a sum score above 4 were regarded as potential effectors. Further, all sequences without start/stop-codon or trans-membrane region containing proteins (> 0 regions; predicted with TMHMM version 2.0) were excluded. Proteins were clustered using 90% sequence identity threshold (CD-HIT parameters `-c 0.9 -s 0.9`) to reduce redundancy. Effector-clusters with great diversity regarding T3SE-prediction scores were removed from the final set. Full data in Extended Data Table 1.

Identification of effector similarities and homology groups

Based on a mutual sequence identity of ≥ 30% over 90% of the common sequence length effectors were considered 'homologous' and included in the $HuMMI_{HOM}$ experiment to investigate the impact of sequence similarity on interaction similarity. Protein sequences were analyzed by global alignment using Needleman Wunsch algorithm implemented in the emboss package (Extended Data Table 2).

Commensal vs pathogen effector similarity

We gathered the sequences of 1,195 known pathogenic T3 effectors from the BastionHub database[79] (August 29th, 2022). We assessed the similarity between commensal and pathogenic effector sequences using BLAST (stand-alone, version 2.10[80]). For each commensal effector, the pathogen effector with the highest sequence similarity was considered as best match. Subsequently, we computed the alignment coverage over the pathogenic effector sequence. Full data in Extended Data Table 2.

Cohort analyses

Genomes of bacterial isolates from the human gut were gathered from multiple published datasets[11-13]. The presence of T3SS was predicted for each of these genomes as described above. GTDB-Tk (v2.1)[81] was used to assign the taxonomy to each of the genomes, and the concatenated bac120 marker proteins from this were used to generate a phylogenomic tree of the isolates, visualized in iTOL[82]. FastANI was used to match the T3SS positive genomes to the WIS representative genomes of the human gut[18] based on ANI values > 95%[83]. The relative abundance of the 10 matching representative genomes was then identified across 3,096 Israeli, and 1,528 Dutch patients[18].

Effector cloning

Bacterial strains from the ATCC collection were ordered from LGS Standard Standard (Wesel, Germany) or ATCC in the US (Manassas, Virginia). Bacterial strains from the DSMZ collection were obtained from the Leibniz-Institut DSMZ (Braunschweig, Germany) and strains from the BEI collection were ordered at BEI resources (Manassas, Virginia, USA) (Extended Data Table 2). Effectors identified from MAGs and effectors for the PRS were ordered at Twist Bioscience (San Francisco, CA, 660 USA). If no genomic DNA could be obtained strains were cultured according to the manufacturer's protocol and DNA was extracted using the NucleosSpin Plasmid (NoLid) Mini kit (Macherey-Nagel cat. No. 740499) with vortexing after addition of BufferA2 and BufferA3. A nested PCR was performed to add Sfi sites, the DNA was purified using magnetic beads (magtivio cat. no. MDKT00010075), followed by an Sfi digestion and another clean-up with magnetic beads. Digested PCR products were cloned into pENTR223.1

using T4 DNA Ligase (ThermoFisher ca. no. EL0011). Plasmids were propagated in DH5α *E. coli* and the plasmid DNA was extracted using the pipetting Bio Robot Universal System (Qiagen cat. no. 9001094) and the QIAprep 96 plus BioRobot kit (Qiagen cat. no. 962241). ORFs were verified by Sanger Sequencing. Effectors were cloned into the Y2H destination plasmid pDEST-DB (pPC97, Cen origin), the pDEST-N2H-N1 and -N2, or the mammalian expression vector pMH-FLAG-HA by an LR reaction of the Gateway System. After propagation in DH5α *E. coli* and DNA extraction plasmids were transformed into *S. cerevisiae* Y8930 (MATα mating type) as DB-X ORFs as described[84].

Meta-interactome mapping

A state-of-the-art high-quality Y2H screening pipeline was followed as previously described[25,85]. DB-X ORFs were tested for autoactivation by mating against AD-empty plasmids in Y8800 (MATa). 45 ORFs of the strains and 14 meta effectors tested positive and were excluded from subsequent steps. The remaining 900 ORFs were individually mated against pools of ~188 AD-Y human ORFs from the human ORFeome collection v9.1 including 17,472 ORFs[86]. During primary screening, haploid AD-Y and DB-X yeast cultures were spotted on top of each other and grown on yeast extract peptone dextrose (YEPD) agar (1%) plates. After incubation for 24 h, the clones were replica plated onto selective synthetic complete media lacking leucine, tryptophan and histidine (SC-Leu-Trp-His) + 1 mM 3-AT (3-amino-1,2,4-triazole) (3-AT plates) and replica cleaned after 24 h. 48 h later, three colonies were picked per spot and grown for 72h in SC-Leu-Trp liquid medium. For the secondary phenotyping, yeasts were spotted on SC-Leu-Trp plates and after incubation for 48 h replica plated and cleaned on 3-AT-plates and SC-Leu-His + 1 mM 3-AT + 1 mg per litre cycloheximide plates to identify spontaneous DB-X autoactivators. Clones growing on 3-AT plates, but not on cycloheximide plates were picked into yeast lysis and processed to generate a library for pair identification by Next Generation Sequencing using a modified KiloSeq procedure as previously described[25]. Identified DB-X and AD-Y pairs were mated individually during the fourfold verification, replica plated and cleaned after 24 hours and picked after another 48 h incubation. Growth scoring was performed using a custom dilated convolutional neural network as described[25]. Pairs scoring positive at least three out of the four repeats qualified as bona fide Y2H interactors. The AD-Y and DB-X constructs were identified once more by NGS. All interaction data are in Extended Data Table 3.

Assembling reference sets

To identify additional reliably documented interactions between bacterial effectors and human proteins for the positive control set (bhLit_BM-v1), we queried the IMEx consortium protein interaction databases[87] through the PSICQUIC webservice[88] (May 10th, 2021) using the T3

effectors UniprotKB accession numbers and fetched all the PubMed identifiers of the articles describing additional interactions. In total, we gathered 67 interactions between 29 T3 effectors and 64 human proteins, described in 13 distinct publications that underwent the manual curation step for inclusion in the PRS (Extended Data Table 3).

Y2H assay sensitivity

Effector ORFs from bhLit_BM-v1 and bhRRS-v1 (Extended Data Table 3) were transferred into pDEST-DB (DB-X) and transformed into *Saccharomyces cerevisiae* Y8930 (MATα). Yeast strains containing the corresponding AD-Y human ORF were picked from hORFeome9.1[86] and ORF identity verified by end-read Sanger sequencing of PCR products. Yeast strains harboring plasmids containing ORFs from hsPRS-v2/hsRRS-v2[89] were provided by the Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA. DB-X and AD-Y were mated fourfold with each other, as well as against yeast strains containing the corresponding DB-empty or AD-empty plasmid. Growth scoring was performed as described above for the fourfold verification. Pairs scoring positive at least three out of the four repeats qualified as bona fide Y2H interactors.

Interactome validation by yN2H

200 interactions were randomly picked from HuMMI and all ORFs from the indicated datasets (Extended Data Table 3) were transferred by Gateway LR reactions into pDEST-N2H-N1 and pDEST-N2H-N2 plasmids containing a *LEU2* or *TRP1* auxotrophy marker, respectively[89]. Successful cloning was monitored by PCR-mediated evaluation of insert size, and positive clones transformed into haploid *Saccharomyces cerevisiae* Y8930 (MATα) and Y8800 (MATa) strains, respectively. Protein pairs from all datasets were randomly distributed across matching 96-well plates.

5 µL of each haploid culture of opposite mating type grown to saturation was mated in 160 µL YEPD medium and incubated overnight. Additionally, each position was mated with yeast stains containing empty N1 or N2 plasmids, to measure background. 10 µL mated culture was inoculated in 160 µL SC-Leu-Trp and grown overnight. 50 µL of this overnight culture was reinoculated in 1.2 ml SC-Leu-Trp and incubated for 24 h at 1000 rpm. Cells were harvested 15 min at 3000 rpm, the supernatant discarded, and each cell pellet was fully resuspended in 100 µl NanoLuc Assay solution (Promega corp. Madison, WI, USA, cat# 1120). Homogenized solutions were transferred to white flat-bottom 96-well plates (Greiner Bio-One, Frickenhausen, Germany, cat# 655904) and incubated in the dark for 1 h at room temperature. Luminescence for each sample was measured on a SpectraMax ID3 (Molecular Devices, San Jose, CA, USA) with 2 s integration time. The normalized luminescence ratio (NLR) was calculated by dividing the raw luminescence of each pair (N1-X N2-Y) by the maximum luminescence value of one

of the two background measurements. All obtained NLR values were $\log_2$ transformed and the positive fraction for each dataset was determined at $\log_2$ NLR thresholds between –2 and 2, in 0.01 increments. Statistical results were robust across a wide range of stringency thresholds. Extended Data Table 3 reports the results at $\log_2$ NLR = 0. Reported *P* values were calculated by Fisher's exact test.

Interactome framework parameter calculation

Assay sensitivity (S_a), i.e., the fraction of detectable interactions was assessed employing the effector bhLit_BM-v1 (54 pairs) and bhRRS-v1 (73 pairs) as well as the human hsPRS-v2 (60 pairs) and hsRRS-v2 (78 pairs) for benchmarking. All reference sets were tested 4 times using the Y2H screening pipeline. To assess sampling sensitivity (S_s) a repeat screen was conducted. 288 bacterial effectors were screened 4 times against 5 pools comprising 1,475 human proteins. A saturation curve was calculated as described[85]. Briefly, all combinations of the number of interactions of the 4 repeats were assembled and the reciprocal values calculated. From these a linear regression was determined to obtain the slope and the intercept. Reciprocal parameters were calculated to find V_{max} and K_m and using the Michaelis-Menten-formula a saturation curve was predicted. *Overall sensitivity* emerges from both sampling and assay limitations and is calculated as $S_o = S_A * S_S$.

Sequence similarity and interaction profile

To investigate the relationship between the similarity of effector sequences and the similarity of their interaction profiles we calculated the pairwise Jaccard index, which measures the overlap between two effectors' interaction profiles. We calculated the Jaccard index of all possible effector pairs within a homology cluster. This index represents the ratio of number of human proteins targeted by both effectors to the total number of human proteins targeted by either of them. For our analysis, we only considered effector pairs where the total number of human proteins that are targeted by either effector was at least 3. We implemented the calculations described here as commands in R version 4.2.1.

Interface predictions

We used as input a representative set of effectors identified in isolated strains (2300 sequences clustered at 90% sequence identity) and all effectors identified in MAGs (186). We ran *mimic*INT as described in[32] and available at [https://github.com/TAGC-NetworkBiology/mimicINT]. Briefly, mimicINT performs domain searches in effector sequences with InterProScan[90] using the domain signatures from the InterPro database[91] retaining matches with an E-value below 10^{-5}. For host-like motif detection, mimicINT uses the SLiMProb tool from the SLiMSuite software package[92] by exploiting the motif definitions available in the ELM database[93]. Motifs are detected in disordered regions as defined by the IUPred

algorithm[94] using both short and long models (motif disorder propensity = 0.2, minimum size of the disordered region = 5). The interface inference step relies on the 3did database[95] (ii) the ELM database[93]. The workflow checks whether any of the effector proteins contains at least one domain or motif for which an interaction template is available. In this case, it infers the interaction between the given protein and all the host proteins containing the cognate domain (i.e., the interacting domain in the template). To control for false positive inference using motif-domain templates, mimicINT provides two scoring strategies. First, considering binding specificity of domains belonging to the same group (as PDZ or SH3)[96] an HMM-based domain score[97] is computed used to rank or filter the inferred interactions. Second, given the degenerate nature of motifs[98], mimicINT, using Monte-Carlo simulations, assesses the probability of a given SLiM to occur by chance in query sequences and, thus, can be used to filter false positives[99]. This statistical approach randomly shuffles the disordered regions of the input sequences to generate a large set of N randomized proteins.

Here, we first grouped effectors sequences by strain and effectors from MAGs were assigned to the closest strain. In the first experiment, disordered regions were shuffled 100,000 times using as background the effector sequences from the same strain (within-strain shuffling). In the second, regions were shuffled 100,000 times using as disorder background the full set of effector sequences (inter-strain shuffling). Subsequently, the occurrences of each detected motif in each effector sequence were compared to the occurrences observed in the corresponding set of shuffled sequences. We considered as significant all the motif occurrences having an empirical *P* value lower than 0.1. To evaluate whether the number of interface-resolved interactions inferred by mimicINT is significantly different from chance, we generated 10,000 random networks by sampling human proteins from the interaction search space in a degree-controlled manner. We then counted how many randomly generated networks mimicINT inferred a higher number of interfaces than for the one observed in the main screen network. Results and statistical details are in Extended Data Table 3.

Holdup assay

Domain production: 54 human PDZ domains and the 11 tandem constructs were recombinantly expressed as His_6-MBP-PDZ constructs in *E. coli* BL21(DE3) pLysS in NZY auto-induction LB medium (nzytech, MB17901)[100]. PDZ domains were purified by Ni^{2+}-affinity with a 96-tip automated liquid-handling system (Tecan Freedom Evoware) using 800 µl of Ni^{2+} Beads (Chelating Sepharose Fast Flow immobilized metal affinity chromatography, Cytiva) for each target. The domains were eluted in 2.5 ml of elution buffer: 250 mM imidazole, 300 mM NaCl, 50 mM Tris, pH 8.0 buffer, and then desalted using PD10 columns (GE healthcare, 17085101) into 3.5 ml of 50 mM Tris, pH 8.0, 300 mM NaCl, 10 mM Imidazole buffer. Concentration of desalted His_6-MBP-PDZ was determined using absorption at 280 nm on a

PHERAstar FSX plate reader (BMG LABTECH). Stock solutions were diluted to 4 µM and frozen at -20°C. To assess purity and confirm the concentrations, proteins were further analyzed by SDS-PAGE (LabChip™ GXII, Perkin Elmer). Peptides: 10-mers corresponding to the C-terminal sequences of effectors were ordered as synthetic biotinylated peptides from GenicBio Limited (Shanghai, China); the N-terminal biotin was attached via a 6-aminohexanoic acid linker, which we showed does not alter the peptide's binding or structural properties[34]. Purity was assessed by HPLC and mass spectrometry; all peptides were >95% pure. Depending on the amino acid composition and charge peptides were solubilized in dH_2O, 1.4% ammonia or 5% acetic acid, aliquoted at 10 mM concentration and stored at -20°C.

For the hold-up assay we followed published procedures[34,35]. Briefly, 2.5 µl of Streptavidin resin (Cytiva, 17511301) were incubated for 15 min with 20 µl of a 42 µM biotinylated peptide solution, in each well of a 384-well MultiScreenHTS™ filter plate (Millipore, MZHVN0W10). The resin was washed with 10 resin volumes (resvol) of hold-up buffer (50 mM Tris HCl, 300 mM NaCl, 10 mM imidazole, 5 mM DTT), and depleted by incubation for 15 min with 5 resvol of a 1 mM biotin solution, and three washes with 10 resvol of hold-up buffer. A single PDZ domain was then added to each well, incubated for 15 min with the peptide bound to the resin and the unbound PDZ was recovered by centrifugation into 384-well black assay plates for fluorescence readout. The concentration is quantified by intrinsic Trp fluorescence, fluorescein/mCherry was used for peak normalization. Binding affinities and equilibrium dissociation constants (k_D) were calculated as in[34], using the mean PBM concentration for k_D calculations. Raw values and statistical analysis are in Extended Data Table 3.

Fluorescent polarization

All FITC labelled peptides were synthesized as 10-mers by Biomatik, Canada, as acetate salts of >98% purity. The FP experiments were performed with the His_6-MBP-PDZ proteins in 50 mM Tris, 300 mM NaCl, 1 mM DTT, pH 7.5 buffer in 384-well plates (Corning 3544). For direct binding the His_6-MBP fused PDZ domains were two-fold serially diluted with 12 dilutions, and a final volume of 10 µl. These were then incubated with 50 nM of the FITC labelled viral peptides and the plates were then read out after 1 h in FlexStation 3 (Molecular Devices) at 23°C, using 485 nm excitation and 520 nm emission. For competition experiments, the PDZ domain and FITC peptide were kept constant at 6 µM and 50 mM, respectively. The bacterial effectors peptides in 1% ammonia buffer were added to the PDZ in a four-fold dilution, (5 concentrations: 0 to 31.25 µM) and incubated at room temperature for 2 h. The FITC peptides were then added and further incubated for 1 h at RT. The plates were then read as above. Statistical analysis was performed using the Kruskal-Wallis test with Dunn's test followed by an FDR-correction. Raw values and statistical analysis are in Extended Data Table 3.

675 **Effector convergence**

676 To estimate the significance of effector convergence, we performed a permutation test by
677 randomly sampling 'target' nodes (n = 979) from Y2H identifiable proteins from the human
678 reference interactome map, HuRI[86], as the sampling space (n = 8,274). We used sampling
679 with replacement to allow repeatedly picking a protein. In each iteration, the number of
680 distinctly targeted proteins was counted. The resulting distribution from 10,000 random
681 permutations was used to calculate the z-score of the experimentally observed targets (n =
682 349). The *P* value is the area under the curve for the standard normal distribution up to a given
683 z-score. We calculated the *P* value as implemented in the "pnorm()" R function using the z-
684 score as input. To account for the two-tailed test, the *P* value was multiplied by 2. To avoid
685 artifacts due to differential sampling we only considered interactions in the $HuMMI_{MAIN,}$
686 excluding those human proteins targeted by effectors of the unknown strains and targets
687 outside HuRI. The rationale for the latter is that a substantial proportion of proteins that are not
688 in HuRI may not be suitable for Y2H analysis. Thus, restricting the analysis to the HuRI subset
689 increases the stringency.

690 To estimate the significance of the convergence of effectors from different strains (interspecies
691 convergence), we used a conditional permutation test that preserves the strain contribution.
692 For each iteration, we generated 18 samples, where for each sample, we randomly picked the
693 number of proteins equivalent to the observed targets of each strain (Extended Data Table 3).
694 From the full list of random picks that are assigned to all strains, the frequency of selecting a
695 protein was recorded. This frequency is the convergence value which indicates the number of
696 targeting strains. Using the convergence value distribution obtained from 10,000 iterations, we
697 identified the statistically significant number of strains sharing a target. The observed
698 convergence value ranges from 2 to 15 strains. We calculated the z-scores using the
699 convergence value distribution obtained from the conditional permutation test and the
700 associated *P* values as implemented in the "*pnorm()*" R function. The significant convergence
701 value (*P* value < 0.004) starts at 4 strains. We considered any target that is in common between
702 at least 4 strains to be subject to interspecies convergence.

703 **Function enrichment analysis**

704 We used the "*gost()*" function from the gprofiler2 version 0.2.1 R package[101] to identify enriched
705 functions in effector targets. This function implements a hypergeometric test to estimate the
706 significance of the abundance of genes considering the frequency of the genes in the function
707 annotation databases. The main input argument for this function is the gene list ("*query*"). The
708 function allows the user to optionally set input arguments, including the background
709 ("*custom_bg*"), evidence codes ("*evcodes*"), annotation databases ("sources"), methods for

correcting the hypergeometric test *P* values ("*correction_method*"), and other arguments that were set to their default options. We used the target official symbol identifiers as the "query" argument. The list of HuRI proteins was the "*custom_bg*" argument. The annotations inferred from electronic annotations were excluded by setting the "*exclude_iea*" argument to "*TRUE*". The hypergeometric test *P* values were corrected using Benjamin-Hochberg method by setting the "correction_method" argument to "*fdr*". The argument ("*sources*") was set to a vector ("*GO:BP*", "*KEGG*","*REAC*"), which encodes the search space across three function annotation databases: gene ontology biological process terms ("*GO:BP*")[102], Kyoto encyclopedia of genes and genomes ("*KEGG*") pathways[103], and Reactome pathway database ("*REAC*")[104]. After plugging in these inputs into the "*gost()*" function, the output is a named list where "result" is a data frame that tabulates the enrichment analysis results. We calculated the odds ratio and the fold enrichment to estimate the effect size of each tested function. The odds ratio was calculated for each function as the odds in the target set divided by the odds in the HuRI set. The odds in the target set are the number of function-annotated target proteins divided by that of the function-unannotated target proteins. Similarly, the odds in the HuRI set are the number of function-annotated HuRI proteins divided by that of function-unannotated HuRI proteins. The fold enrichment was calculated for each function by comparing the number of function-annotated target proteins to that of the expected. The expected value represents the number of function-annotated target proteins that is expected randomly based on the HuRI background. It is the product of the total number of targets (n = 349) by the rarity. The rarity is the number of function annotated HuRI proteins divided by the sum of annotated HuRI proteins. The total HuRI proteins annotated for GO:BP, KEGG, and REAC, are 6988, 3250, and 4592, respectively. Statistical details are in Extended Data Table 5.

Metabolic subsystem analysis

Several metabolism-related functions were significantly enriched in target proteins; therefore, we tested the abundance of targeted enzymes in metabolic subsystems using the human genome-scale metabolic model Recon3D[46]. To focus on metabolic enzymes as opposed to signaling enzymes, we excluded ligases and kinases from Recon3D analyses. We performed the hypergeometric test using the R function "*phyper()*" for each subsystem annotated in Recon3D (n = 95). The inputs to this function are: the number of subsystem-annotated targeted enzymes, the number of subsystem-annotated Recon3D enzymes, the number of subsystem-unannotated Recon3D enzymes, and the number of targeted enzymes (n = 16*)*. The nominal *P* values were corrected using Benjamin-Hochberg. We calculated the odds ratio and the fold enrichment using the same calculations described above for functional enrichments.

Random walk-based determination of commensal effector network neighborhoods

We have implemented a network propagation protocol based on a Random Walk with Restart (RWR) algorithm RWR-MH[105] to explore the network vicinity of the commensal effectors in HuRI[54], which contains 338 target proteins ($HuMMI_{MAIN}$ screen) of 243 commensal effectors. We used the human effector targets as seeds for the random walk and set the restart probability to the default value of 0.7. In this way, we obtained a ranked list of proteins in the network: the ones with the higher scores are more proximal to the seeds than those with lower scores. To assign statistical significance to the computed RWR scores, we implemented a normalization strategy based on degree-preserving network randomizations[106]. We thus generated 1,000 random networks from HuRI and ran the RWR algorithm to compute 1,000 scores for each network protein. We then computed an empirical *P* value for each protein in the network keeping as neighbor proteins only those with an empirical *P* value < 0.01.

Disease enrichment analysis

We tested the association of all target proteins, or those subject to convergence, with human diseases by performing a two-sided Fisher's exact test. We used the disease-causal genes identified by the Open Targets genetic portal, which prioritizes genes at GWAS loci based on variant-to-gene distance, molecular QTL colocalization, chromatin interaction, and variant pathogenicity[107]. This machine-learning approach assigns a locus to gene (l2g) score to identify the most likely causal gene for the genetic variation signal of any marker SNP. We considered a score of 0.5 or more as a threshold, as recommended by the authors[108]. The Fisher's exact test was performed using the function "*fisher.test()*" from "*stats*" R package version 4.2.2 with its default inputs whenever applicable. The input to this function is a 2 x 2 contingency table, where columns represent the query set and the background set, and rows denote the absence or presence of causal genes in the respective set. HuRI proteins were used as the background set, and the query set was either the target proteins or those subject to convergence. The calculated nominal *P* values from this function were then corrected using the Benjamin-Hochberg method as implemented in the "*p.adjust()*" function. The odds ratio and fold enrichment values were calculated as described in the functional enrichment section. Statistical details are in Extended Data Table 5.

Association with human traits and phenotype in network neighborhoods

For each set of significant neighborhood-proteins we tested for enrichment of Open Targets causal genes for human traits that had been investigated by 3 or more studies and for which the Open Targets initiative identified 3 or more causal genes (l2g ≥ 0.5). We used a two-sided Fisher's exact test to assess whether a given strain neighborhood is enriched in protein associated with a human trait or phenotype followed by Benjamini-Hochberg multiple testing correction. This yielded no significant association (FDR < 0.05). We therefore focused on 400

associations with a nominal *P* value < 0.01 and an OR > 3. Some disease categorizations were adjusted to better reflect etiology. Thus, Sjogren syndrome, eczema and psoriasis were considered an 'immunological' rather than eye or skin traits, and osteoarthritis was labeled as a disease of "musculoskeletal or connective tissue" rather than metabolic. For Fig. 4d some closely related traits were merged, i.e., three asthma terms and three psoriasis terms. Statistical details are in Extended Data Table 5.

NF-κB activation assay

HEK 293 (RRID: CVCL_0045, DSMZ) were maintained in DMEM with 10% FBS and 100 U/mL penicillin and 100 U/mL streptomycin at 37°C and 5% CO2. IKKβ (in pRK5 with a Flag-tag) served as positive control whereas A20 (in pEF4 with a Flag-tag) as the negative control. In a 60 mm cell culture dish 1 x 10^6 cells were seeded in 3 ml Medium. After 24 h cells were transfected using 10 ng NF-κB reporter plasmid (6 × NF-κB firefly luciferase pGL2), 50 ng pTK reporter (renilla luciferase) and 2 µg bacterial ORF in pMH-FLAG-HA. The DNA was added to 200 µl 250 mM $CaCl_2$ solution (Carl Roth cat. no. 5239.1), vortexed and added dropwise to 200 µl 2 × HBS (50 mM HEPES (pH 7.0) (Carl Roth cat. no. 9105.4), 280 mM NaCl (Carl Roth cat. no. 3957.2), 1.5 mM Na_2HPO_4 × 2 H_2O (Carl Roth cat. no. 4984.1, pH 6.93) which was vortexed. After 15 min incubation, the mixture was added dropwise to the cells. Medium was changed after 6 h incubation. To assess NF-κB inhibition, cells were treated for 4 h with 20 ng/ml TNF (Sigma-Aldrich cat. no. SRP3177) 24 h after transfection. Samples were washed, lysed, centrifuged and the supernatant was measured using the dual luciferase reporter kit (Promega, E1980) with a luminometer (Berthold Centro LB960 microplate reader, Software: MikroWin 2010). NF-κB induction was determined as Firefly luminescence to Renilla luminescence. *P* values were calculated using the Kruskal-Wallis test with Dunn's correction followed by an FDR-correction. Raw values and statistical analysis are in Extended Data Table 6.

Protein expression levels were checked by Western Blots. Proteins were separated by SDS-PAGE and transferred on polyvinylidene fluoride membranes, and after transfer blocked with 5% milk in 1 × PBS + 0.1% Tween-20 (PBST) for 1 h at room temperature. Primary antibodies were added in 2.5% BSA in PBS-T buffer at 4°C overnight. After 3 x 15min washes with PBS-T anti-mouse secondary antibody was added at a 1:10,000 dilution for 1 h at RT (Jackson ImmunoResearch Labs cat. no. 715-035-150, RRID:AB_2340770). Primary antibodies: anti-Actin beta (SCBT cat. no. sc-47778, RRID:AB_626632) at a 1:10,000 dilution, anti-FLAG M2 (Sigma Aldrich cat. no. F3165, RRID:AB_259529) at a 1:500 dilution and anti-HA (Sigma-Aldrich cat. no. 11583816001, RRID:AB_514505) at a 1:1,000 dilution. For detection the LumiGlo reagent (CST cat. no. 7003S) and a chemiluminescence film (Sigma-Aldrich cat. no. GE28-9068-36) were used.

ICAM1 assay

Caco-2 cells were maintained in DMEM Glutamax medium (Gibco) supplemented with 10% FBS, 1% Pen/Strep at 37°C in a humidified 5% CO2 incubator. Medium was refreshed twice a week. Caco-2 cells were plated in both 24- and 96-well plates 24 h before transfection. Six hours prior to transfection, culture medium was replaced with supplement-free DMEM. Co-transfections were performed using 40,000 MW linear polyethylenimine (PEI MAX®) (Polysciences, Warrington, USA) at a ratio of 1:5 pDNA:PEI. Equimolar ratios of the eGFP-plasmid and effector-plasmid were used to ensure equimolar representation of relevant ORFs. In total, 250 ng and 1 µg pDNA was added per well of the 96- and 24-well plates, respectively. pDNA-PEI complexes were formed by incubating pDNA and PEI at RT for 15 minutes, followed by the addition of supplement-free DMEM and another incubation of 15 minutes at RT. Cells were then exposed to the transfection mixture for 16 h, washed, and rested for 6 h in complete DMEM. Subsequently, cells were stimulated using an activation mix containing 200 ng/ml PMA (P8139-1MG, Sigma-Aldrich), 100 ng/ml LPS (L6529-1MG, Sigma-Aldrich), and 100 ng/ml TNF (130-094-014, Miltenyi Biotec). In 24-well plates, cells were stimulated for 24 h and detached from the plate using ice-cold PBS. In the 96-well plate, cells were stimulated for 48 h, treated with BD GolgiStop™ (554724, BD Biosciences) in the final 6 h of stimulation, and detached using trypsin/EDTA. Cells were washed twice and ICAM1 was stained using an anti-ICAM1 PE (#MHCD5404-4, Invitrogen) antibody. The mean fluorescent intensity of the GFP+ cell population was measured on a FACSFortessa™ flow cytometer (BD) and the data was analyzed using FlowJo V10.8.1 (BD). After positive tests for normal data distribution, significance was assessed using a one-way ANOVA with Dunnett's multiple comparisons test. Raw values and statistical analysis are in Extended Data Table 6.

Cytokine assays

Caco-2 cells were plated in 100 mm cell culture dishes three days prior to transfection. The transfection protocol was identical to that described above, however, a total of 20 µg pDNA was used per dish. Upon overnight transfection, cells were detached using Trypsin/EDTA and resuspended in cell sorting buffer (PBS + 2% FBS + 2mM EDTA). GFP+ cells were sorted into ice-cold FBS using a BD FACSAria III cell sorter (BD) and transferred to a 96-well plate at 30,000 cells per well. Upon a 24 h rest-period, cells were activated for 48 h using the activation mix described above. During cell stimulation, cell proliferation was monitored through longitudinal imaging of cell confluency in the Incucyte S3 Live cell analysis system (Essen BioScience). Cytokine levels were determined using the human inflammation panel 1 LEGENDplex™ kit (Biolegend) following the manufacturer's instructions. Cell culture supernatant of the above samples was used to analyze IL1beta. To this end, IL1beta ELISAs were performed using the ELISA MAX™ Deluxe Set Human IL1beta kit (437015, Biolegend)

following the protocol provided by the manufacturer. Statistical significance was evaluated using Kruskal-Wallis test with uncorrected Dunn's test. Raw values and statistical analysis are in Extended Data Table 6.

Protein ecology

Metagenomic assemblies from the Inflammatory Bowel Disease Multi'omics DataBases (IBDMBD)[64] and from the skin metagenome[109] were downloaded, and each samples protein repertoire predicted using Prodigal (options; -p meta)[110]. Effector proteins were compared to the metagenomic protein repertoires using DIAMOND (options; >90% query length, >80% identity). For analyses in Fig. 5, samples were grouped into patients with UC (n = 304), CD (n = 508), and controls without IBD (n = 334). The annotations were then converted into binarised vectors of presence and absence of each effector across the sample and the Fischer exact test, implemented within scipy python module, was used to determine if the prevalence of each effector occurring within CD or UC patient metagenomes compared to controls. Significance was then corrected using the Benjamini-Hochberg method. The significance of differences in prevalence distributions between healthy and either patient cohort were estimated by Wilcoxon rank-sum test, implemented in the "*wilcox.test()*" R function. Statistical details in Extended Data Table 6.

Statistics and reproducibility

Data were subjected to statistical analysis and plotted to Microsoft Excel 2010 or python or R scripts. For comparison of normally distributed values we used one-way ANOVA, for assessment of overlap for comparison of values not passing the normality tests we used Kruskal-Wallis test with Dunn's corrected as appropriate and indicated in the figure legends and methods. Enrichments were calculated using Fisher's exact test with Bonferroni FDR correction. All statistical evaluations were done as two-sided tests. Generally, a corrected *P* value < 0.05 was considered significant. GO, KEGG, and Reactome functional enrichments were calculated using profiler with the respectively indicated background gene sets. For the disease target enrichments and neighborhood associations no associations were significant after multiple hypothesis correction, which is why nominally significant associations calculated by Fisher's exact tests were used for Fig. 4c,d. All raw values, n, and statistical details are presented in supplementary tables as indicated in the Figure legends and methods sections.

AUTHOR CONTRIBUTIONS

Project conception: PFB
T3SS and effector analyses: PH, TH, SA, CB, AZ, TR, PFB
ORF cloning: VY, MR, MA, AS, PFB
Interactome mapping and validation: VY, SR, BW, AS, PFB
Interaction curation: VY, MA, MB, AZ, CF, PFB
Data analyses: BD, VY, DS, CWL, MB, SAC, PS, CB, AZ, PFB
Interface identification and validation: SAC, AZ, JFM, SBM, JCT, RV
Effector ecology: TH, TC
Cell-based assays: VY, NvdH, FO, PFB, DK, MB
Visualization: VY, BD, JFM, AZ, PFB
Funding acquisition: PFB, CF, AZ, CB, TR, DK
Manuscript writing and editing: PFB, VY, BD, BW, TH, CF, AZ

ACKNOWLEDGEMENTS

Plasmids and strains for hsPRS/RRS_v2 were kindly provided by Marc Vidal, David E. Hill, and Mike Calderwood, CCSB, Dana-Farber Cancer Institute, Boston, MA. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high-performance computing resources.

REPORTING SUMMARY

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

DATA AVAILABILITY

All sequence, interaction, and functional data generated in this study are available as supplementary information. The effectors identified and cloned for interactome mapping are presented in Extended Data Table 1. All protein-protein interaction data acquired in this study can be found in Extended Data Table 2 and Extended Data Table 3. The data for functional validation assays can be found in Extended Data Table 6. The protein interactions from this publication have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct[111] and assigned the identifier IM-29849. New effector sequences have been submitted to GenBank: BankIt2727690: OR372873 - OR373035 and OR509516 - OR509528.

CODE AVAILABILITY

All source code related to this paper is available as a zip file.

COMPETING INTERESTS

The authors declare no competing interests.

EXTENDED DATA TABLES:

Extended Data Table 1: T3SS in strains of the commensal human microbiome

Extended Data Table 2: Effector identification and cloning

Extended Data Table 3: Effector host interaction map

Extended Data Table 4: Interface identification and validation

Extended Data Table 5: Functional and disease enrichment

Extended Data Table 6: Functional assay data and IBD prevalence

FIGURES

Figure 1

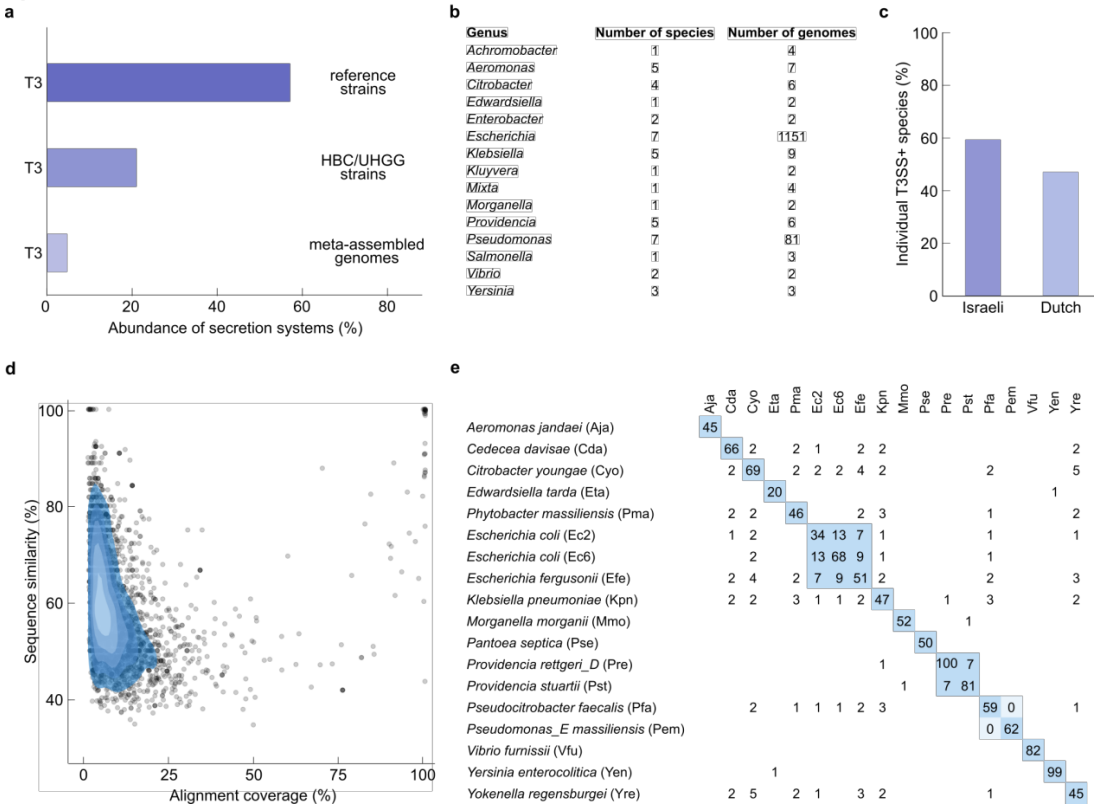


Fig. 1 | T3SS in commensal bacterial species in the gut microbiome. a, Proportion of Pseudomonadota genomes encoding complete T3SS among 77 reference strains of human intestinal and stool samples, in a collection of 4,475 strains isolated from normal human guts, and in meta-assembled genomes (MAG) of normal human guts. **b**, Most abundant genera and identified number of species and genomes encoding complete T3SS from the samples in **a**. **c**, Proportion of individuals in two human cohorts containing T3SS encoding microbial species. **d**, Similarity of 3,002 candidate effector-substrates for T3SS identified from commensal reference strains with 1,195 effectors from pathogenic microbes across the range of alignment coverages. **e**, Selection of 18 commensal Pseudomonadota strains with dissimilar effector complements used for subsequent functional analyses. Numbers indicate the count of shared effectors at >90% mutual sequence similarity across 90% common sequence length among the indicated strains. Full data for all panels in Extended Data Table 1.

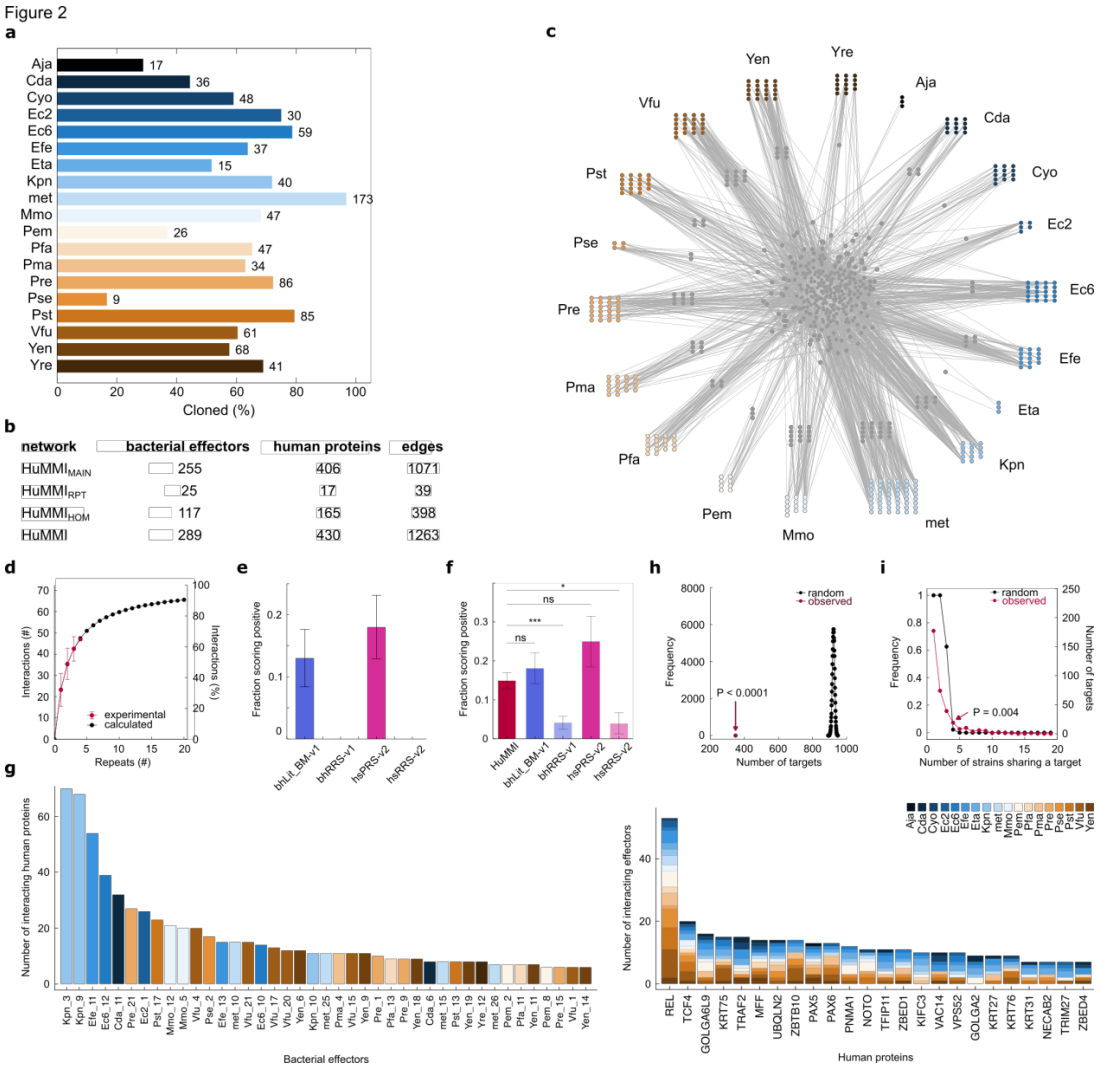


Fig. 2 | Meta-interactome network map of bacterial effectors with human proteins. **a**, Success rates of effector ORF cloning for each strain, and number of sequence verified ORFs (right). **b**, Number of interactions and involved proteins in the HuMMI subsets. **c**, Verified human microbiome meta-interactome (HuMMI) map. Grey nodes: human proteins; outer layer human proteins targeted only by the nearest strain; central human proteins by effectors from multiple strains. **d**, Sampling sensitivity: saturation curve calculated from the repeat experiment: red dots represent average of verifiable interactions found in any combination of indicated number of repeat screens; black dots and line: modeled saturation curve. **e**, Assay sensitivity: percentage of identified interactions from bhLit_BM-v1 (n = 54 pairs), bhRRS-v1 (n = 73 pairs), hsPRS-v2 (n = 60 pairs), hrRRS-v2 (n = 78 pairs) in our Y2H. Error bars present the standard error (SE) of proportion. **f**, Validation rate of a random sample of HuMMI interactions (n = 295 pair configurations) compared to four reference sets in the yN2H validation assay: bhLit_BM-v1 (n = 94 pair configurations), bhRRS-v1 (n = 145 pair

configurations), hsPRS-v2 (n = 44 pair configurations), hrRRS-v2 (n = 51 pair configurations). * P = 0.04; *** P = 0.0006; ns “no significant difference” (Fisher exact test; Extended Data Table 3). Error bars present SE of proportion. **g**, Left: degree distribution for the most connected effectors; right: effector-degree distribution for most targeted human proteins. Colors represent strains according to legend. **h**, Observed number of total effector targets in the human reference interactome (HuRI), compared to random expectation (exp. P < 0.0001; n = 10,000 randomizations). (**I**) Frequency distribution of human proteins targeted by effectors from the indicated number of different strains (red), compared to random expectation (black; n = 10,000). Targeting by effectors from four strains or more occurs significantly more often than expected by chance (exp. P = 0.004; n = 10,000).

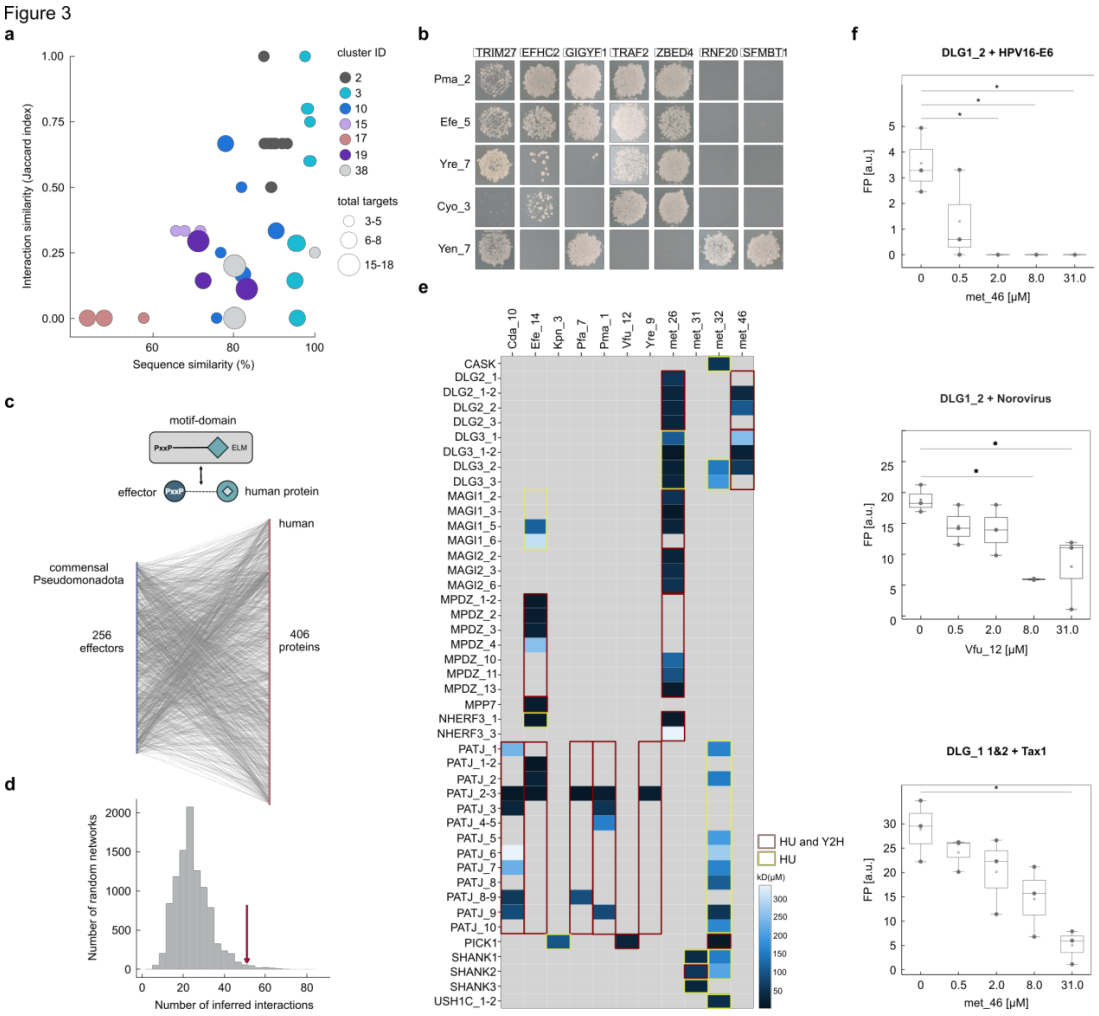


Fig. 3 | Interaction specificity and interaction motifs. a, Scatter plot of sequence- and Jaccard-interaction similarity for all effector pairs within indicated homology groups of $HuMMI_{HOM}$ with ≥ 3 interactors and effectors. Node size indicates union of human proteins targeted by effector-pair according to legend. **b**, Y2H data for one of four repeats for homology cluster 3. **c**, Schematic of interaction motif-domain interface identification in the effector-host interaction. **d**, Count of motif-domain pairs matching at least one stringency criteria identified in $HuMMI_{MAIN}$ (arrow) compared to random expectation (experimental *P* value, n = 10,000). **e**, Interaction strength of PDZ domains of human proteins with C-terminal 10 amino acid peptides of the effectors indicated on top. Calculated K_D according to legend. Overlap between HU and Y2H is indicated by colored frames. **f**, Competition of the interaction between human PDZ domains and viral PBM peptides by the indicated effector peptides. * $P < 0.05$ (Kruskal Wallis with Dunn's correction, n = 3). Boxes represent interquartile range (IQR), with the bold black line representing mean; whiskers indicate highest and lowest data point within 1.5 IQR.

Figure 4

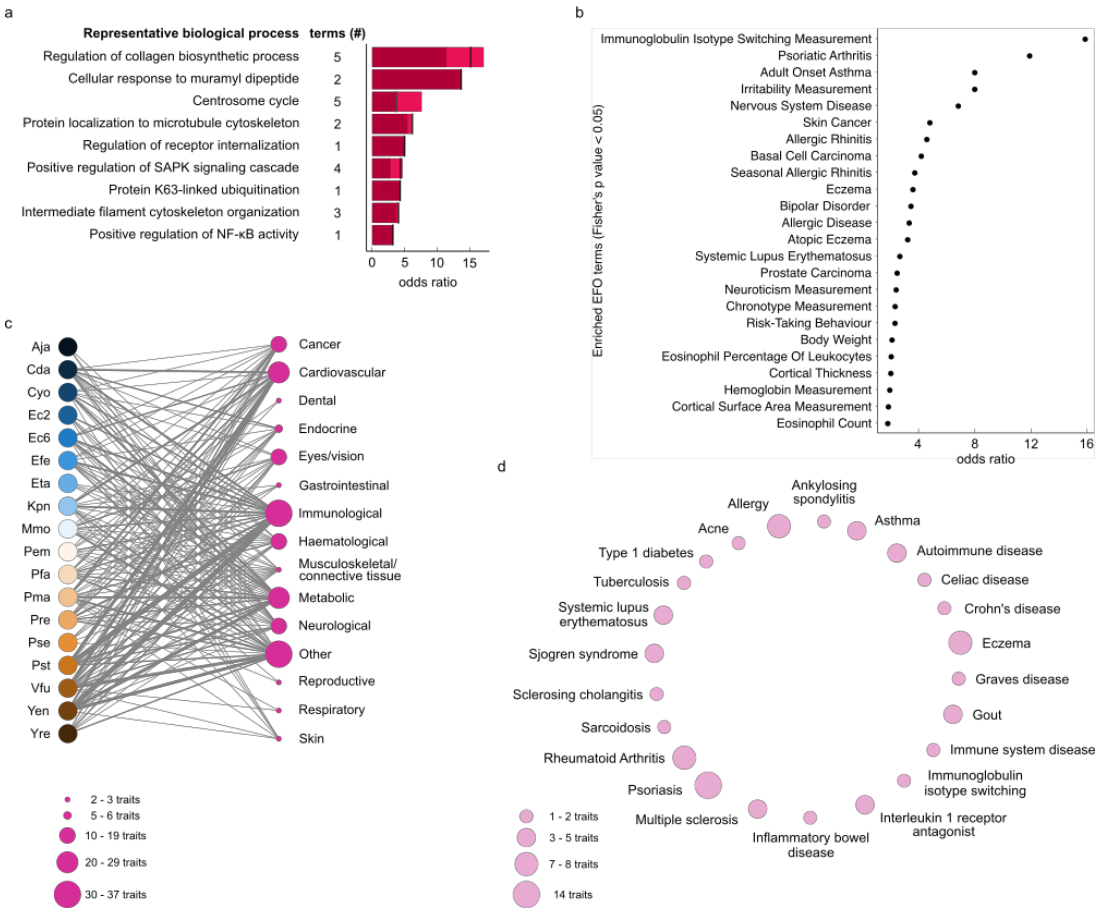


Fig. 4 | Function and disease association of microbially targeted human proteins. a, Odds ratios (OR) of representative functional annotations enriched among effector targeted human proteins (FDR < 0,05, Fisher's exact test with Bonferroni FDR correction). The number of represented terms is shown by terms (#). The lowest and highest OR observed for the represented group are indicated by light shaded area in each bar. Black line indicates OR for representative term. Full data and precise FDR and OR values in Extended Data Table 5. **b**, Genetic predisposition for traits and diseases enriched among human genes encoding effector targets in HuRI (cutoff FDR = 0.05, Fisher's exact test, n = 349). **c**, Disease groups for which genetic predisposition is enriched in network neighborhoods of effectors from the indicated strains. Trait node size corresponds to number of significantly targeted traits in that group according to legend. Stroke of strain-group edge reflects number of underlying significant effector-trait links (α < 0.01 and OR > 3, Fisher's exact test). **d**, Specific diseases underlying the 'immunological' group in c. Node size reflects the number of underlying effector-trait associations according to legend.

Figure 5

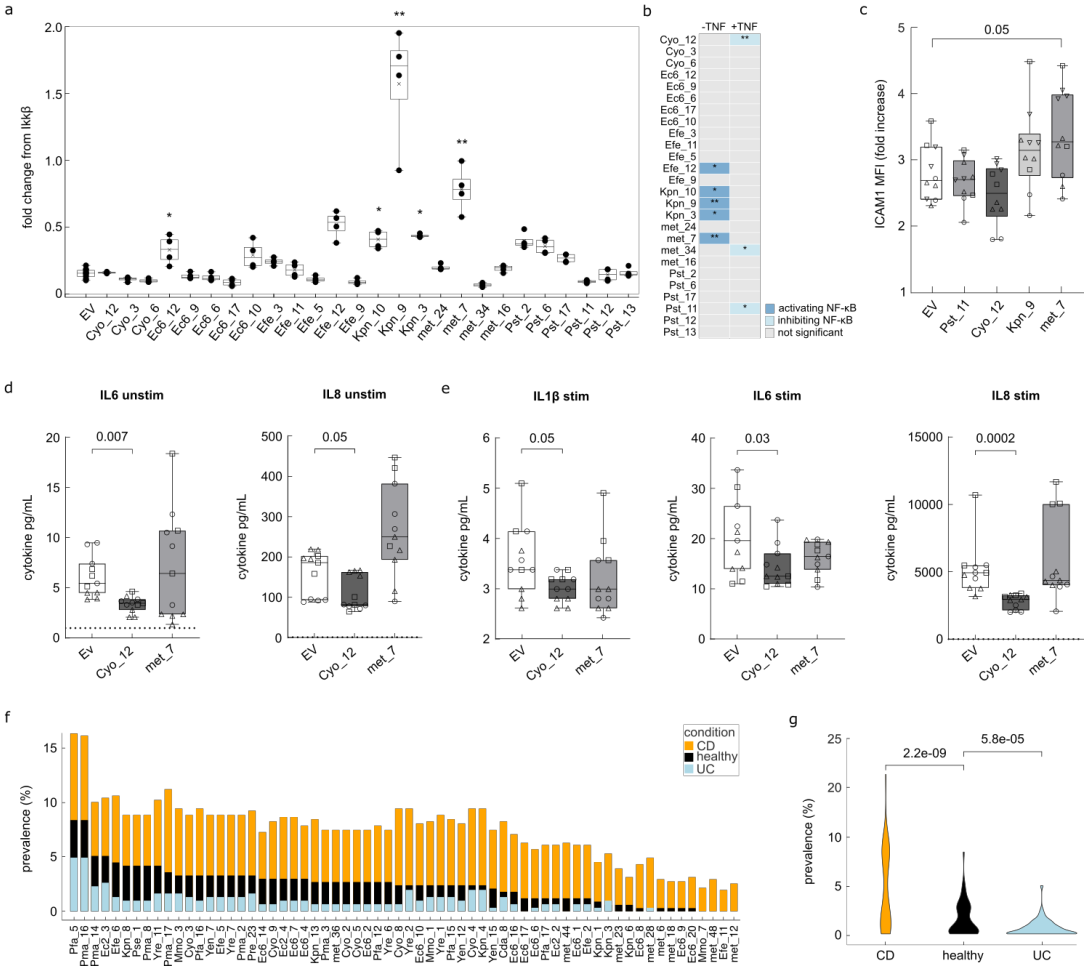
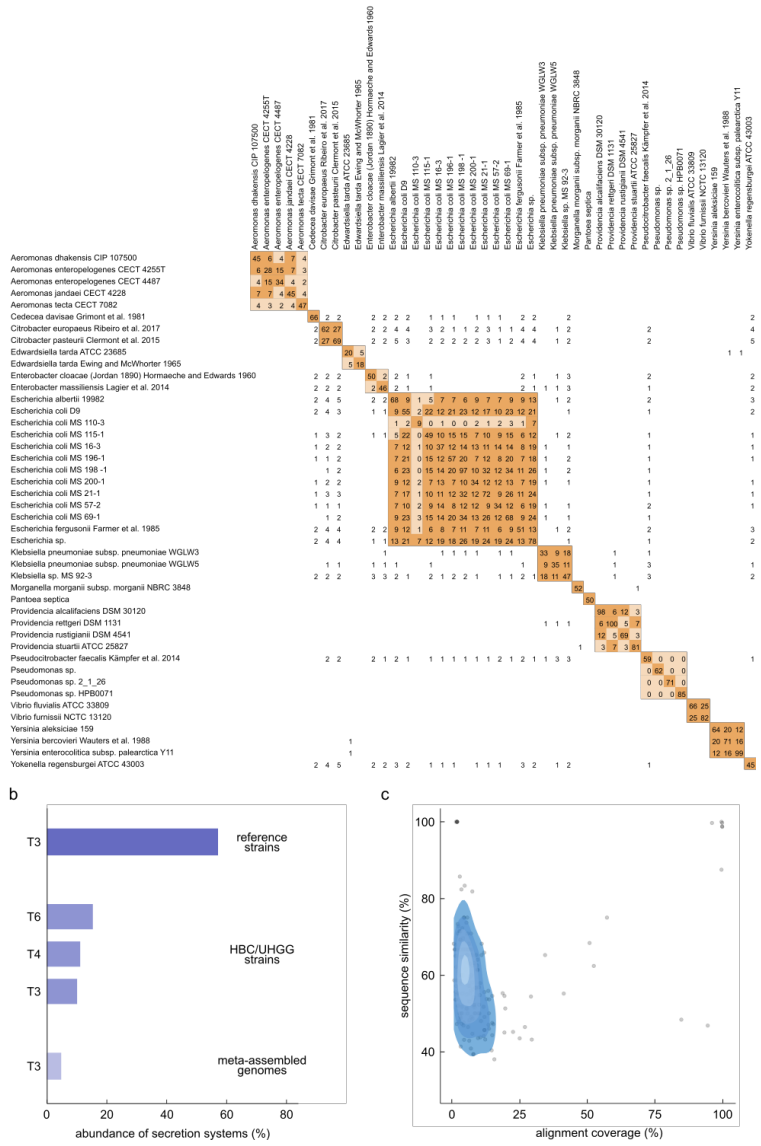


Fig. 5 | Effector impact on human cell function and clinical prevalence in IBDs. **a**, Relative NF-κB transcriptional reporter activity of HEK293 cells expressing the indicated effectors or empty vector (EV) in unstimulated conditions (Kruskal-Wallis test with Dunn's correction, * $P < 0.05$, ** $P < 0.01$, n = 4). Boxes represent IQR, black line indicates the mean, whiskers indicate highest and lowest data point within 1.5 IQR. **b**, Summary of significant impact of effectors on normalized NF-κB transcriptional reporter activity in baseline conditions and after TNF stimulation (Kruskal-Wallis test with Dunn's correction, * $P < 0.05$, ** $P < 0.01$, n = 4). **c**, Fold-induction of ICAM1 expression following pro-inflammatory stimulation of Caco-2 cells transfected with the indicated effectors (one-way ANOVA with Dunnett's multiple comparison test, n = 10). **d**, Concentration of cytokines secreted by Caco-2 cells in basal conditions transfected with the indicated effectors. EV indicates empty vector mock control. *P* values calculated by Kruskal-Wallis test (n = 11). Dashed line indicates detection limit of assay. **e**, Concentration of cytokines secreted by Caco-2 cells stimulated by a pro-inflammatory cocktail transfected with the indicated effectors. EV indicates empty vector mock control. Indicated *P*

964 values calculated by Kruskal-Wallis test (n = 11). Dashed line: detection limit of assay. C – E
965 Boxes represent IQR, black line indicates the mean, whiskers indicate highest and lowest data
966 point. **f**, Effector prevalence in metagenomes of CD (n = 504), and UC patients (n = 302)
967 compared to healthy controls. Effectors are significantly more prevalent in CD patient
968 metagenomes (FDR < 0.01; Fisher exact test, Benjamini-Hochberg correction). **g**, Effector
969 prevalence distribution among the indicated cohorts. *P* values calculated by Wilcoxon rank-
970 sum test, n as in f.

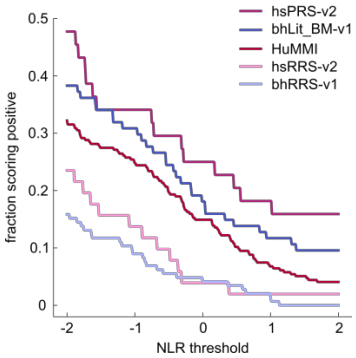
EXTENDED DATA FIGURES

Extended Figure 1



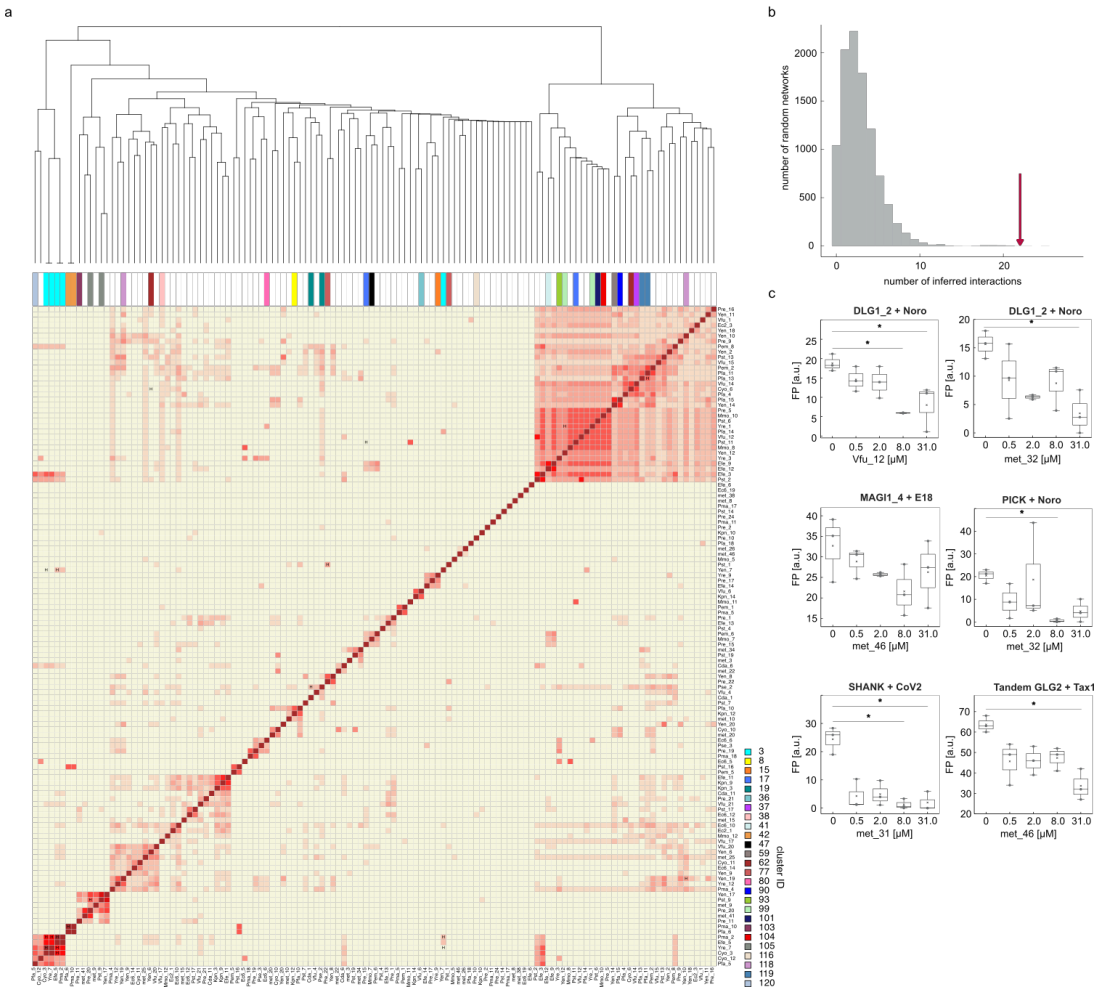
Extended Data Fig. 1 | T3SS in strains of the commensal gut microbiome. a, Effector-complement comparison of the 44 T3SS+ Pseudomonadota reference strains. Numbers indicate the count of shared effectors at >90% mutual sequence similarity across 90% common sequence length among the indicated strains. **b**, Abundance of secretion systems in Pseudomonadota genomes among the 77 reference strains of human intestinal and stool samples, in a collection of 4,475 strains isolated from normal human guts (HBC/UHGG strains) and in meta-assembled genomes (MAG) of normal human guts. **c**, Similarity of identified 186 candidate effectors from the 770 T3SS+ MAGs with 1,195 effectors from pathogenic microbes across the range of alignment coverages. Full data for all panels in Extended Data Table 1.

Extended Figure 2



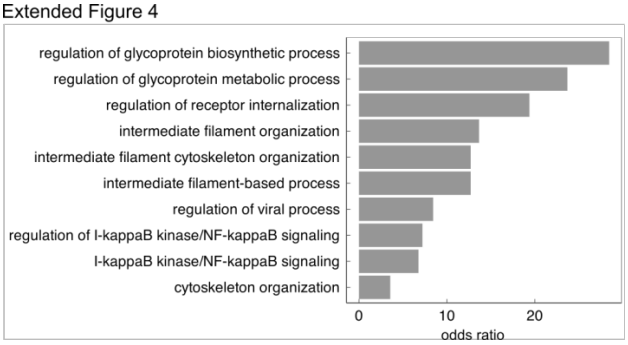
Extended Data Fig. 2 | Detection rates of protein pairs in different sets across varying thresholds in yN2H. Fractions scoring positive of the HuMMI dataset and benchmarking datasets (hsPRS-v2, bhLit_BM-v1, hsRRS-v2, bhRRS-v1) depending on the threshold of the normalized luminescence ratio (NLR). Full data in Extended Data Table 3.

Extended Figure 3



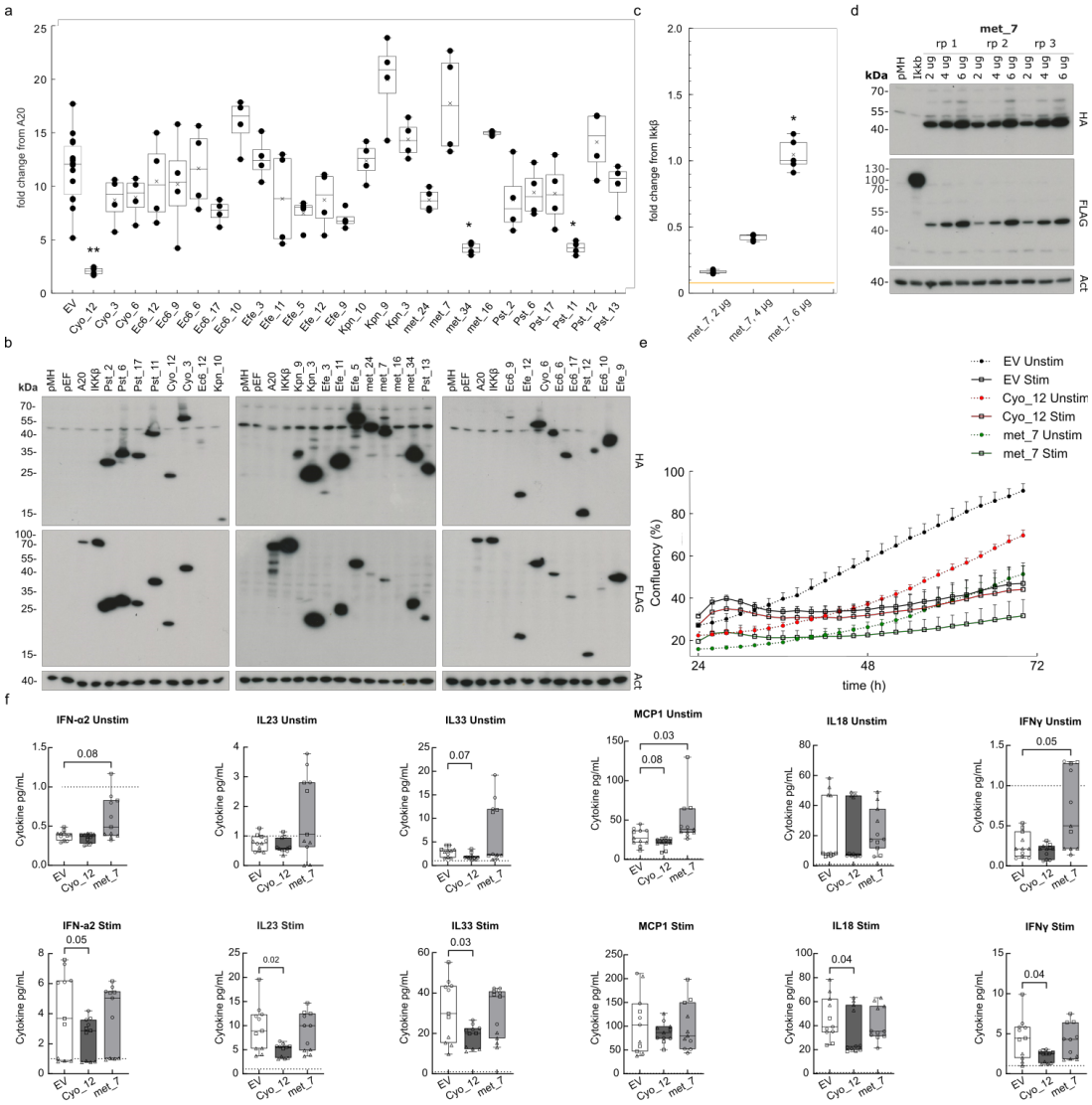
Extended Data Fig. 3 | Interaction specificity and interaction motifs. a, Jaccard-interaction similarity of all interacting effector-pairs with at least 3 shared human interactors. Color-intensity correlates with Jaccard-index. Effector pairs marked with "H" share the same homology cluster. Clusters are color-coded according to legend. **b**, Count of motif-domain pairs matching at least two stringency criteria identified in $HuMMI_{MAIN}$ (arrow) compared to n = 10,000 randomized control networks (empirical $P = 0.0003$). **c**, Competition of the interaction between human PDZ domain and viral PBM peptide by indicated C-terminal effector peptides. * $P < 0.05$ (Kruskal Wallis with Dunn's correction, n = 3). Boxes indicate IQR, black line represents mean, whiskers indicate highest and lowest data point within 1.5 IQR. Precise *P* values and n for each test are shown in Extended Data Table 4.

Extended Figure 4



Extended Data Fig. 4 | GO enrichment for convergence proteins. OR for functional annotations enriched among effector-targeted human proteins that are subject of convergence (FDR < 0.05, Fisher's exact test with Bonferroni FDR correction). Full data and precise FDR and OR values in Extended Data Table 5.

Extended Figure 5



Extended Figure 5 | Effector impact on human cell function. a. Relative NF-κB transcriptional reporter activity of HEK293 cells expressing the indicated effectors under TNF-stimulated conditions (Kruskal-Wallis test with Dunn's correction, * $P < 0.05$, ** $P = 0.01$, n = 4). Boxes represent IQR, with the bold black line representing the mean; whiskers indicate highest and lowest data point within 1.5 IQR. **b**, Representative anti-Hemagglutinin (HA) and anti-Flag (FLAG) western blots showing expression of transfected effector proteins relative to actin control (ACT). Empty pMH-Flag-HA (pMH), empty pEF4 (pEF). **c**. Titration of met_7 shows a concentration dependent specific increase of NF-κB reporter activity. Yellow line represents the empty vector value. (Kruskal-Wallis test with Dunn's correction, * $P < 0.05$, error bars: standard deviation of the mean, n = 5). Boxes represent IQR, with the bold black line representing the mean; whiskers indicate highest and lowest data point within 1.5 IQR. **d**,

Representative anti-Hemagglutinin (HA) and anti-Flag (FLAG) western blots for experiment in c showing expression of transfected effector proteins relative to actin control (ACT). **e**, Representative proliferation curves of Caco-2 cells transfected with empty vector (EV), Cyo_12 or met_7 in basal conditions (unstim) or following pro-inflammatory stimulation (stim) over 72 h after sorting. **f**, Concentration of cytokines secreted by Caco-2 cells transfected with the indicated effectors in basal conditions (Unstim) or following pro-inflammatory stimulation (Stim). EV indicates empty vector mock control. Indicated *P* values calculated by Kruskal-Wallis test with Dunn's multiple hypothesis correction (n = 11). Boxes represent IQR, with the bold black line representing the mean; whiskers indicate highest and lowest data point. Raw measurements, n, and precise *P* values for all panels in Extended Data Table 6.

REFERENCES

1 Plichta, D. R., Graham, D. B., Subramanian, S. & Xavier, R. J. Therapeutic Opportunities in Inflammatory Bowel Disease: Mechanistic Dissection of Host-Microbiome Relationships. *Cell* **178**, 1041-1056, doi:10.1016/j.cell.2019.07.045 (2019).

2 Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392-400, doi:10.1038/nm.4517 (2018).

3 Depner, M. *et al.* Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nat Med* **26**, 1766-1775, doi:10.1038/s41591-020-1095-x (2020).

4 Paun, A., Yau, C. & Danska, J. S. Immune recognition and response to the intestinal microbiome in type 1 diabetes. *J Autoimmun* **71**, 10-18, doi:10.1016/j.jaut.2016.02.004 (2016).

5 Keshavarzian, A. *et al.* Colonic Bacterial Composition in Parkinson's Disease. *Mov. Disord.* **30**, 1351-1360, doi:10.1002/mds.26307 (2015).

6 Oren, A. & Garrity, G. M. Valid publication of the names of forty-two phyla of prokaryotes. *Int J Syst Evol Microbiol* **71**, doi:10.1099/ijsem.0.005056 (2021).

7 Shin, N. R., Whon, T. W. & Bae, J. W. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol* **33**, 496-503, doi:10.1016/j.tibtech.2015.06.011 (2015).

8 Deng, W. Y. *et al.* Assembly, structure, function and regulation of type III secretion systems. *Nature Reviews Microbiology* **15**, 323-337, doi:10.1038/nrmicro.2017.20 (2017).

9 Miwa, H. & Okazaki, S. How effectors promote beneficial interactions. *Current Opinion in Plant Biology* **38**, 148-154, doi:10.1016/j.pbi.2017.05.011 (2017).

10 Eichinger, V. *et al.* EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res* **44**, D669-674, doi:10.1093/nar/gkv1269 (2016).

11 Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* **37**, 186-192, doi:10.1038/s41587-018-0009-7 (2019).

12 Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**, 1442-1452, doi:10.1038/s41591-019-0559-3 (2019).

13 Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053-2067 e2018, doi:10.1016/j.cell.2021.02.052 (2021).

14 Yang, X. B., Pan, J. F., Wang, Y. & Shen, X. H. Type VI Secretion Systems Present New Insights on Pathogenic Yersinia. *Frontiers in Cellular and Infection Microbiology* **8**, doi:10.3389/fcimb.2018.00260 (2018).

15 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

16 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662 e620, doi:10.1016/j.cell.2019.01.001 (2019).

17 Shalon, D. *et al.* Profiling the human intestinal environment under physiological conditions. *Nature*, doi:10.1038/s41586-023-05989-7 (2023).

18 Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat Commun* **13**, 3863, doi:10.1038/s41467-022-31502-1 (2022).

19 Jing, R. *et al.* DeepT3 2.0: improving type III secreted effector predictions by an integrative deep learning framework. *NAR Genom Bioinform* **3**, lqab086, doi:10.1093/nargab/lqab086 (2021).

20 Arnold, R. *et al.* Sequence-based prediction of type III secreted proteins. *PLoS Pathog* **5**, e1000376, doi:10.1371/journal.ppat.1000376 (2009).

21 Goldberg, T., Rost, B. & Bromberg, Y. Computational prediction shines light on type III secretion origins. *Scientific reports* **6**, 34516, doi:10.1038/srep34516 (2016).

22 Wang, J. W. *et al.* BastionHub: a universal platform for integrating and analyzing substrates secreted by Gram-negative bacteria. *Nucleic Acids Research* **49**, D651-D659, doi:10.1093/nar/gkaa899 (2021).

23 Ma, W. B., Dong, F. F. T., Stavrinides, J. & Guttman, D. S. Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *Plos Genetics* **2**, 2131-2142, doi:10.1371/journal.pgen.0020209 (2006).

24 Rohmer, L., Guttman, D. S. & Dangl, J. L. Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. *Genetics* **167**, 1341-1360, doi:10.1534/genetics.103.019638 (2004).

25 Kim, D. K. *et al.* A proteome-scale map of the SARS-CoV-2-human contactome. *Nat Biotechnol*, doi:10.1038/s41587-022-01475-z (2022).

26 Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90, doi:10.1038/nmeth.1280 (2009).

27 Braun, P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. *Proteomics* **12**, 1499-1518, doi:10.1002/pmic.201100598 (2012).

28 Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods* **6**, 91-97 (2009).

29 Wessling, R. *et al.* Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell host & microbe* **16**, 364-375, doi:10.1016/j.chom.2014.08.004 (2014).

30 Mukhtar, M. S. *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333**, 596-601, doi:10.1126/science.1203659 (2011).

31 Osborne, R. *et al.* Symbiont-host interactome mapping reveals effector-targeted modulation of hormone networks and activation of growth promotion. *Nat Commun* **14**, 4065, doi:10.1038/s41467-023-39885-5 (2023).

32 Choteau, S. A. *et al.* mimicINT: a workflow for microbe-host protein interaction inference. *bioRxiv*, 2022.2011.2004.515250, doi:10.1101/2022.11.04.515250 (2022).

33 Gutierrez-Gonzalez, L. H. *et al.* Peptide Targeting of PDZ-Dependent Interactions as Pharmacological Intervention in Immune-Related Diseases. *Molecules* **26**, doi:10.3390/molecules26216367 (2021).

34 Gogl, G. *et al.* Quantitative fragmentomics allow affinity mapping of interactomes. *Nat Commun* **13**, 5472, doi:10.1038/s41467-022-33018-0 (2022).

35 Vincentelli, R. *et al.* Quantifying domain-ligand affinities and specificities by high-throughput holdup assay. *Nat Methods* **12**, 787-793, doi:10.1038/nmeth.3438 (2015).

36 Javier, R. T. & Rice, A. P. Emerging Theme: Cellular PDZ Proteins as Common Targets of Pathogenic Viruses. *J Virol* **85**, 11544-11556, doi:10.1128/jvi.05410-11 (2011).

37 Zhang, H. *et al.* Pathogenesis and Mechanism of Gastrointestinal Infection With COVID-19. *Front Immunol* **12**, 674074, doi:10.3389/fimmu.2021.674074 (2021).

38 Damin, D. C., Ziegelmann, P. K. & Damin, A. P. Human papillomavirus infection and colorectal cancer risk: a meta-analysis. *Colorectal Dis* **15**, e420-428, doi:10.1111/codi.12257 (2013).

39 Karst, S. M. & Tibbetts, S. A. Recent Advances in Understanding Norovirus Pathogenesis. *Journal of Medical Virology* **88**, 1837-1843, doi:10.1002/jmv.24559 (2016).

40 Maseko, S. B. *et al.* Identification of small molecule antivirals against HTLV-1 by targeting the hDLG1-Tax-1 protein-protein interaction. *Antiviral Res*, 105675, doi:10.1016/j.antiviral.2023.105675 (2023).

41 Gonzalez, R. & Elena, S. F. The Interplay between the Host Microbiome and Pathogenic Viral Infections. *Mbio* **12**, doi:10.1128/mBio.02496-21 (2021).

42 Girardin, S. E. *et al.* Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *Journal of Biological Chemistry* **278**, 8869-8872, doi:10.1074/jbc.C200651200 (2003).

43 Knights, D. *et al.* Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med* **6**, 107, doi:10.1186/s13073-014-0107-1 (2014).

44 Brennan, J. J. & Gilmore, T. D. Evolutionary Origins of Toll-like Receptor Signaling. *Mol. Biol. Evol.* **35**, 1576-1587, doi:10.1093/molbev/msy050 (2018).

45 D'Alessio, S. *et al.* Revisiting fibrosis in inflammatory bowel disease: the gut thickens. *Nature Reviews Gastroenterology & Hepatology* **19**, 169-184, doi:10.1038/s41575-021-00543-0 (2022).

46 Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol* **36**, 272-281, doi:10.1038/nbt.4072 (2018).

47 Vidal, M., Cusick, M. E. & Barabasi, A.-L. Interactome Networks and Human Disease. *Cell* **144**, 986-998, doi:10.1016/j.cell.2011.02.016 (2011).

48 Gulbahce, N. *et al.* Viral Perturbations of Host Networks Reflect Disease Etiology. *Plos Comp. Biol.* **8**, doi:e1002531

10.1371/journal.pcbi.1002531 (2012).

49 Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research* **49**, D1302-D1310, doi:10.1093/nar/gkaa1027 (2021).

50 Bunker, J. J. *et al.* Natural polyreactive IgA antibodies coat the intestinal microbiota. *Science* **358**, eaan6619, doi:doi:10.1126/science.aan6619 (2017).

51 Pabst, O. & Slack, E. IgA and the intestinal microbiota: the importance of being specific. *Mucosal Immunology* **13**, 12-21, doi:10.1038/s41385-019-0227-4 (2020).

52 Alcazar, C. G. *et al.* The association between early-life gut microbiota and childhood respiratory diseases: a systematic review. *Lancet Microbe* **3**, e867-e880, doi:10.1016/S2666-5247(22)00184-7 (2022).

53 Mahmud, M. R. *et al.* Impact of gut microbiome on skin health: gut-skin axis observed through the lenses of therapeutics and skin diseases. *Gut Microbes* **14**, 2096995, doi:10.1080/19490976.2022.2096995 (2022).

54 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408, doi:10.1038/s41586-020-2188-x (2020).

55 Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *New England Journal of Medicine* **375**, 2369-2379, doi:10.1056/NEJMra1600266 (2016).

56 de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* **49**, 256-261, doi:10.1038/ng.3760 (2017).

57 Vainer, B., Nielsen, O. H. & Horn, T. Comparative studies of the colonic in situ expression of intercellular adhesion molecules (ICAM-1, -2, and -3), beta2 integrins (LFA-1, Mac-1, and p150,95), and PECAM-1 in ulcerative colitis and Crohn's disease. *Am J Surg Pathol* **24**, 1115-1124, doi:10.1097/00000478-200008000-00009 (2000).

58 Vainer, B. Intercellular adhesion molecule-1 (ICAM-1) in ulcerative colitis: presence, visualization, and significance. *Inflamm Res* **54**, 313-327, doi:10.1007/s00011-005-1363-8 (2005).

59 Biswas, S., Bryant, R. V. & Travis, S. Interfering with leukocyte trafficking in Crohn's disease. *Best Practice & Research Clinical Gastroenterology* **38-39**, 101617, doi:https://doi.org/10.1016/j.bpg.2019.05.004 (2019).

60 Brand, S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. *Gut* **58**, 1152-1167, doi:10.1136/gut.2008.163667 (2009).

61 Zhao, J. *et al.* Th17 Cells in Inflammatory Bowel Disease: Cytokines, Plasticity, and Therapies. *J Immunol Res* **2021**, 8816041, doi:10.1155/2021/8816041 (2021).

62 Mitsuyama, K. *et al.* IL-8 as an important chemoattractant for neutrophils in ulcerative colitis and Crohn's disease. *Clin Exp Immunol* **96**, 432-436, doi:10.1111/j.1365-2249.1994.tb06047.x (1994).

63 Herrero-Cervera, A., Soehnlein, O. & Kenne, E. Neutrophils in chronic inflammatory diseases. *Cell Mol Immunol* **19**, 177-191, doi:10.1038/s41423-021-00832-3 (2022).

64 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-+, doi:10.1038/s41586-019-1237-9 (2019).

65 Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* **4**, 293-305, doi:10.1038/s41564-018-0306-4 (2019).

66 Kelly, D., Conway, S. & Aminov, R. Commensal gut bacteria: mechanisms of immune modulation. *Trends Immunol* **26**, 326-333, doi:https://doi.org/10.1016/j.it.2005.04.008 (2005).

67 Büttner, D. Protein export according to schedule: architecture, assembly, and regulation of type III secretion systems from plant- and animal-pathogenic bacteria. *Microbiol Mol Biol Rev* **76**, 262-310, doi:10.1128/mmbr.05017-11 (2012).

68 Rodriguez, P. A. *et al.* Systems Biology of Plant-Microbiome Interactions. *Mol. Plant.* **12**, 804-821, doi:10.1016/j.molp.2019.05.006 (2019).

69 Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603, doi:10.1038/35079107 (2001).

70 Lichtenstein, G. R. *et al.* ACG Clinical Guideline: Management of Crohn's Disease in Adults. *American Journal of Gastroenterology* **113**, 481-517, doi:10.1038/ajg.2018.27 (2018).

71 Wera, O., Lancellotti, P. & Oury, C. The Dual Role of Neutrophils in Inflammatory Bowel Diseases. *Journal of Clinical Medicine* **5**, doi:10.3390/jcm5120118 (2016).

72 Basnet, S., Palmenberg, A. C. & Gern, J. E. Rhinoviruses and Their Receptors. *Chest* **155**, 1018-1025, doi:https://doi.org/10.1016/j.chest.2018.12.012 (2019).

73 Hayashi, Y. *et al.* Rhinovirus Infection and Virus-Induced Asthma. *Viruses* **14**, doi:10.3390/v14122616 (2022).

74 Zhang, Y. *et al.* The ORMDL3 Asthma Gene Regulates ICAM1 and Has Multiple Effects on Cellular Inflammation. *Am J Respir Crit Care Med* **199**, 478-488, doi:10.1164/rccm.201803-0438OC (2019).

75 Reimer, L. C. *et al.* BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res* **47**, D631-D636, doi:10.1093/nar/gky879 (2019).

76 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

77 Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-293, doi:10.1093/nar/gkv1248 (2016).

78 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

79 Wang, J. *et al.* BastionHub: a universal platform for integrating and analyzing substrates secreted by Gram-negative bacteria. *Nucleic Acids Res* **49**, D651-D659, doi:10.1093/nar/gkaa899 (2021).

80 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

81 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927, doi:10.1093/bioinformatics/btz848 (2019).

82 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

83 Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).

84 Altmann, M., Altmann, S., Falter, C. & Falter-Braun, P. High-Quality Yeast-2-Hybrid Interaction Network Mapping. *Curr Protoc Plant Biol* **3**, e20067, doi:10.1002/cppb.20067 (2018).

85 Altmann, M. *et al.* Extensive signal integration by the phytohormone protein network. *Nature* **583**, 271-276, doi:10.1038/s41586-020-2460-0 (2020).

86 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408, doi:10.1038/s41586-020-2188-x (2020).

87 Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods* **9**, 345-350, doi:10.1038/nmeth.1931 (2012).

88 del-Toro, N. *et al.* A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res* **41**, W601-606, doi:10.1093/nar/gkt392 (2013).

89 Choi, S. G. *et al.* Maximizing binary interactome mapping with a minimal number of assays. *Nat. Commun.* **10**, 3907, doi:10.1038/s41467-019-11809-2 (2019).

90 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

91 Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344-D354, doi:10.1093/nar/gkaa977 (2021).

92 Edwards, R. J., Paulsen, K., Aguilar Gomez, C. M. & Perez-Bercoff, A. Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol Biol* **2141**, 37-72, doi:10.1007/978-1-0716-0524-0_3 (2020).

93 Kumar, M. *et al.* ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res* **48**, D296-D306, doi:10.1093/nar/gkz1030 (2020).

94 Dosztanyi, Z. Prediction of protein disorder based on IUPred. *Protein Sci* **27**, 331-340, doi:10.1002/pro.3334 (2018).

95 Mosca, R., Ceol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* **42**, D374-379, doi:10.1093/nar/gkt887 (2014).

96 Gfeller, D. *et al.* The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol* **7**, 484, doi:10.1038/msb.2011.18 (2011).

97 Davey, N. E. *et al.* Attributes of short linear motifs. *Mol Biosyst* **8**, 268-281, doi:10.1039/c1mb05231d (2012).

98 Davey, N. E. *et al.* SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* **40**, 10628-10641, doi:10.1093/nar/gks854 (2012).

99 Hagai, T., Azia, A., Babu, M. M. & Andino, R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell reports* **7**, 1729-1739, doi:10.1016/j.celrep.2014.04.052 (2014).

100 Duhoo, Y. *et al.* High-Throughput Production of a New Library of Human Single and Tandem PDZ Domains Allows Quantitative PDZ-Peptide Interaction Screening Through High-Throughput Holdup Assay. *Methods Mol Biol* **2025**, 439-476, doi:10.1007/978-1-4939-9624-7_21 (2019).

101 Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res* **9**, doi:10.12688/f1000research.24956.2 (2020).

102 Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, doi:10.1093/genetics/iyad031 (2023).

103 Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587-D592, doi:10.1093/nar/gkac963 (2023).

104 Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* **50**, D687-D692, doi:10.1093/nar/gkab1028 (2022).

105 Valdeolivas, A. *et al.* Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **35**, 497-505, doi:10.1093/bioinformatics/bty637 (2019).

106 Biran, H., Kupiec, M. & Sharan, R. Comparative Analysis of Normalization Methods for Network Propagation. *Front Genet* **10**, 4, doi:10.3389/fgene.2019.00004 (2019).

107 Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527-1533, doi:10.1038/s41588-021-00945-5 (2021).

108 Barrio-Hernandez, I. *et al.* Network expansion of genetic associations defines a pleiotropy map of human cell biology. *Nat Genet* **55**, 389-398, doi:10.1038/s41588-023-01327-9 (2023).

109 Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59-64, doi:10.1038/nature13786 (2014).

110 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

111 Del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* **50**, D648-D653, doi:10.1093/nar/gkab1006 (2022).

Eigenständigkeitserklärung

Hiermit versichere ich an Eides statt, dass die vorliegende Dissertation mit dem Titel

Inferring protein from transcript abundances using convolutional neural networks

von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, wurden in jedem Fall unter Angabe der Quellen (einschließlich des World Wide Web und anderer elektronischer Text- und Datensammlungen) kenntlich gemacht. Weiterhin wurden alle Teile der Arbeit, die mit Hilfe von Werkzeugen der künstlichen Intelligenz de novo generiert wurden, durch Fußnote/Anmerkung an den entsprechenden Stellen kenntlich gemacht und die verwendeten Werkzeuge der künstlichen Intelligenz gelistet. Die genutzten Prompts befinden sich im Anhang. Diese Erklärung gilt für alle in der Arbeit enthaltenen Texte, Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

München, den 23.02.2026

(Ort / Datum)

Patrick Schwehn — P. Schwehn

(Vor und Nachname in Druckbuchstaben) (Unterschrift)

Affidavit

Herewith I certify under oath that I wrote the accompanying Dissertation myself.

Title: Inferring protein from transcript abundances using convolutional neural networks

In the thesis no other sources and aids have been used than those indicated. The passages of the thesis that are taken in wording or meaning from other sources have been marked with an indication of the sources (including the World Wide Web and other electronic text and data collections). Furthermore, all parts of the thesis that were de novo generated with the help of artificial intelligence tools were identified by footnotes/annotations at the appropriate places and the artificial intelligence tools used were listed. The prompts used were listed in the appendix. This statement applies to all text, graphics, drawings, sketch maps, and pictorial representations contained in the Work.

München, den 23.02.2026

(Location/date)

Patrick Schwehn — P. Schwehn

(First and last name in block letters) (Signature)