# Understudied proteins: opportunities and challenges for functional proteomics

Most research aiming at understanding the molecular foundations of life and disease has focused on a limited set of increasingly well-known proteins while the biological functions of many others remain poorly understood. We propose to form the Understudied Protein Initiative with the objective of reducing the annotation gap by systematically associating uncharacterized proteins with proteins of known function, thereby laying the groundwork for future detailed mechanistic studies.

Georg Kustatscher, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S. Lilley, Emma Lundberg, Edward M. Marcotte, Markus Ralser and Juri Rappsilber

Different proteins receive very different levels of attention from scientists. The most frequently studied protein in the human proteome is p53. On average, it is the subject of two publications per day[1]. At the same time, the biological functions of thousands of human proteins remain unexplored[2–5]. This bias in the functional characterization of the human proteome is massive: 95% of all life science publications focus on a group of 5,000 particularly well-studied human proteins[6]. The sequencing of the human genome was expected to be a crucial step toward reducing this bias: by identifying all human genes, researchers would be offered opportunities to study previously unknown genes. However, in 2011, a decade after the publication of the genome sequence, 75% of publications still focused on genes that were already being studied before the genome was mapped[7]. Annotation inequality has increased since then and has almost doubled since the human genome sequence was released[2].

Annotation inequality hinders biomedical progress because mechanistic investigations of gene–disease associations typically focus on proteins that are already well known (Fig. 1), a phenomenon also known as the street-light effect[8]. Meanwhile, many uncharacterized proteins are not subjected to functional studies despite strong evidence from omics studies for their association with human disease[2]. For example, the functions of many proteins involved in rare diseases (which are not rare collectively) are poorly understood[9]. Moreover, common diseases such as neurodevelopmental disorders and cancer are caused by collections of numerous rare genetic variants in different genes[10]. Remarkably, out of the 1,878 genes that are essential for proliferation in a human cell line, 330 (18%) remained uncharacterized as of 2015 (ref. [11]). This bias extends to the ~3,000 proteins currently expected to be druggable: only 5–10% of these potentially druggable proteins are currently targeted by FDA-approved pharmaceuticals[5].

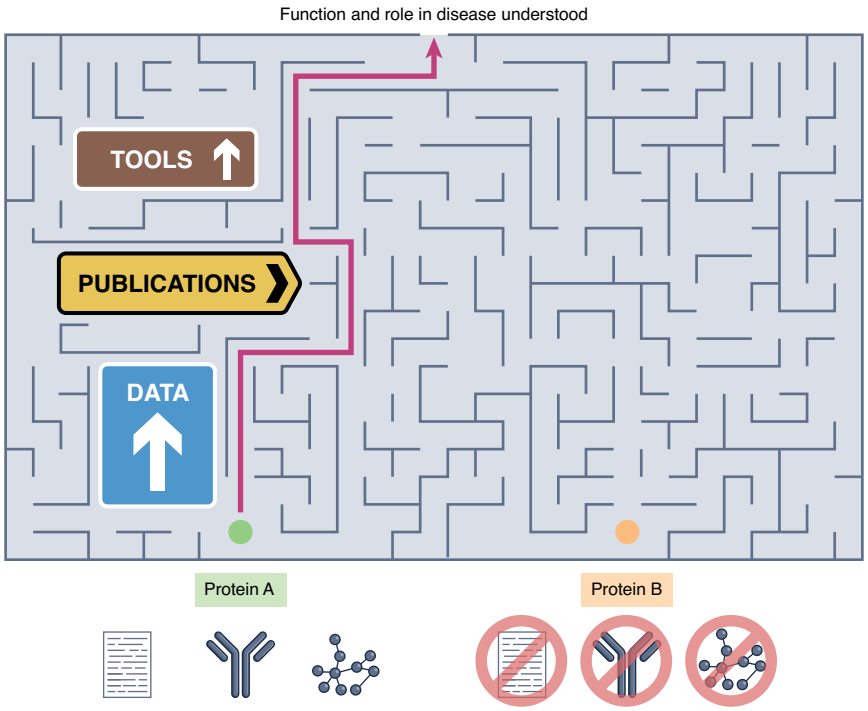


**Fig. 1 | Protein annotation inequality impedes biomedical progress.** The availability of prior publications, data and tools dictates the ease by which research questions involving a protein can be formulated and addressed. This reinforces annotation bias and the persistence of understudied proteins.

Functional proteomics could be instrumental in reducing the annotation gap by systematically associating uncharacterized proteins with proteins of known function and thereby assigning them to cellular processes. An important element of targeting uncharacterized proteins is to broaden the range of investigation beyond typical laboratory conditions and the limited set of laboratory model organisms' genetic backgrounds. With a focus on mass spectrometry (MS)-based methods, here we outline opportunities and challenges for a coordinated functional proteomics initiative that would lay the groundwork for future detailed mechanistic studies.

## Origins of protein annotation inequality

The reasons for the protein annotation bias are manifold. Some are of a practical nature, reflecting how easily a protein can be studied with widely available methods. For example, the availability of experimental tools such as antibodies, plasmids or curated reference data is a strong incentive to work on well-studied proteins[2,7]. The number of publications about a protein is also related to basic biological and biochemical properties, such as protein size, abundance, hydrophobicity and the sensitivity of its gene toward mutations[4]. The dynamic range of our detection devices does not yet match that of proteins in a cell. In fact, to date, 1,899 (9.6%) of the 19,733 human protein-coding genes lack credible support from any proteomics technology, some of which may constitute genome annotation errors[12].

In addition, having a very small size is a strikingly common feature among under-studied proteins: 40% of the least well-annotated proteins in SwissProt are smaller than 15 kDa (ref. [13]). This is despite the importance of microproteins, for example, as neuropeptides in brain development[14]. Moreover, what we currently consider to be the repertoire of understudied small proteins may just be the tip of an iceberg, as we are only beginning to uncover the array of 'alternative proteins' coming from genomic regions previously considered to be noncoding[15].

Other reasons for protein annotation inequality may reflect conceptual biases in the research system rather than properties of the proteins themselves. For example, it is often assumed that proteins studied by many people are functionally more important[7], although this is not supported by evidence such as genome-wide association studies or functional genomic screens[2,11,16,17]. In addition, scientists often prefer to explore a problem they already work on in more detail, in part because funding and peer-review systems are risk-averse[7]. Working in a large research field enhances the likelihood of being cited, and, consequently, also increases the possibility for high-impact journal publications, which are required for academic success[18]. However, large fields also tend to favor existing paradigms over new ideas, thus slowing scientific progress overall[4,19,20].

Equally important is the limited set of conditions studied in the laboratory, a situation that might paradoxically be a consequence of the desire to make research more reproducible through standardization of experimental conditions. For example, under standard laboratory growth conditions, the deletion of ~20% of *Saccharomyces cerevisiae* genes causes a lethal phenotype[21]. However, when the condition space is expanded, 97% of the genes are essential for optimal growth under at least one condition[22]. Indeed, the choice of 'standard' conditions often reflects historical reasons rather than the desire to capture the entirety of biological complexity. For instance, the most popular synthetic yeast medium in use today emerged from an early 1950s publication of the US Department of Agriculture technical bulletin which attempted to help farmers and biotechnologists to grow a wide variety of yeasts; for example, to start fermentation processes[23]. The problem is further compounded for multicellular organisms with specialized cell types; some tissues or cell types are much more studied than others.

Finally, protein annotation bias could reflect the focus on hypothesis-driven rather than question-driven research[24,25]. It is difficult to formulate hypotheses on the mechanistic molecular function of an uncharacterized protein. Intriguingly, the philosopher Francis Bacon, often credited as the father of the scientific method, argued in the early 1600s that experiments should not be driven by hypotheses for fear of introducing bias in the observer and stifling innovation[24,26]. In line with this, it has been suggested that strictly data-driven approaches could help to reduce protein annotation inequality[2,27].

## Accelerating drug discovery for understudied proteins

From a standpoint of drug discovery, fundamental advances toward the characterization of understudied proteins are being made by initiatives that improve our understanding of protein–small molecule interactions, such as the Structural Genomics Consortium[28], the Enzyme Function Initiative[29], the Illuminating the Druggable Genome program[5] and Open Targets[30]. In this context, 'functional characterization' is typically interpreted as revealing molecular properties of a protein that are particularly relevant for drug development; for example, its structure, ligands, inhibition by chemical probes and association with disease. Particular emphasis is placed on pharmacologically tractable protein families, such as ion channels, G-protein-coupled receptors and kinases[5,31,32].

From a perspective of understanding protein function, it is equally important to study other levels of protein annotation, such as cellular processes, pathways and subcellular compartments. In addition, many understudied proteins do not belong to a traditional druggable family, although the definition of a druggable protein is evolving over time as new approaches (such as PROTACs[33]) are developed. One set of methods ideally suited to study the cellular functions of proteins, and to do so on a comprehensive, proteome-wide scale, is functional proteomics.

## Tackling annotation inequality with functional proteomics

Two different types of protein annotation efforts may be distinguished: original investigations and 'guilt-by-association' approaches. The original investigation of a novel biological function is an essential but time-consuming and costly effort involving many detailed mechanistic studies. For researchers to commit to such an effort, it is necessary for a protein to have a certain basal annotation level. Without this, hypotheses to probe a protein's function lack foundation. Here, annotation by 'functional association' can provide the lacking foundations through knowledge transfer, whereby previously uncharacterized proteins are linked to well-studied factors and their biological functions[34–38].

Proteomics approaches are particularly well suited to revealing functional associations on a large scale. Such approaches include techniques that identify protein–protein interactions, such as affinity purification MS[39–41], crosslinking MS[42] and co-fractionation MS[43]; approaches that identify which proteins are co-regulated[44–51]; and methods that reveal which proteins share subcellular space[52–55] (Box 1). For example, the majority of centrosomal proteins were considered to have been already identified[56], and then hundreds more were identified by antibody-based proteomics[57]. It is noteworthy that although we focus here on MS and antibody-based proteomics, powerful alternative proteomics approaches also exist that have been reviewed elsewhere[58,59]. There are also many functional genomics approaches that do not rely on measuring proteins for functional association, including gene expression profiling, whereby functionally related genes are linked on the basis of similar expression patterns[60], metabolic profiling[61] and genetic interaction screening[62]. Rapid advances in genome-wide CRISPR–Cas9 screening have accelerated the pace of functional annotation of proteins involved in susceptibility to therapeutic compounds, or those that become essential in a specific genetic context[63].

While MS-based proteomics does not yet reach the gene coverage of genomic approaches, observing proteins directly can be especially informative when studying

**Box 1 | Proteomic approaches that reveal protein–protein associations**

MS and antibody-based approaches that enable annotation transfer by identifying protein–protein interactions (PPIs) differ in the nature of the links they provide, their scalability and their biases. Each approach has strengths and weaknesses. The following is a non-exhaustive list of key technologies applied in recent years:

**Crosslinking MS:** Identifies PPIs by crosslinking proteins in vitro or in situ, followed by MS-based detection of crosslinked peptides. Links represent binary physical interactions between two proteins at amino acid residue resolution. Crosslinking MS is starting to be applied to complex mixtures, with the benefits of revealing protein interaction topology[42] and having a systematic error assessment[77].

**Affinity purification MS:** 'Bait' proteins are fused to affinity purification tags, expressed in cells and subsequently purified together with multiple 'prey' proteins that physically interact with the bait, either directly or indirectly[39–41]. Alternatively to epitope tags, antibodies or other specific affinity probes against the endogenous bait protein can be used.

**Co-fractionation MS:** Cell extracts are fractionated biochemically, typically using ultra-centrifugation, size exclusion chromatography or ion exchange chromatography, and protein co-fractionation patterns are identified by MS and compared by machine learning to identify protein complexes[43,99] and the subcellular localization of proteins[54,55,80].

**Proximity labeling MS:** 'Bait' proteins are fused to enzymes that enable biotinylation of 'prey' proteins in close spatial proximity in living cells, which can subsequently be affinity-purified and identified by MS[52,53].

**Antibody-based proteomics:** Subcellular localization of proteins is revealed using antibodies[74]. The assays provide single-cell resolution in situ, can detect multi-localizing proteins and may contribute to understanding pleiotropic effects.

**Protein co-regulation:** Protein abundance changes between different biological conditions, or in response to perturbations, are determined by MS and compared using correlation analysis or machine learning[13,44–51]. This improved on previous mRNA co-expression studies[13,64]. Unlike other methods listed here, protein co-regulation does not detect physical relationships but coordinated protein abundance changes, which are taken to reflect shared participation in a biological process.

**Emerging approaches:** Novel proteomics methods to study protein–protein interactions using MS are developed continuously. An example of a recent addition to the repertoire is thermal proteome profiling, which can detect shared membership in protein complexes[82].

Notably, there are a variety of non-MS-based methods that also reveal protein–protein associations[58,59], including binary assays such as yeast two-hybrid[100], LUMIER[101], genetic interaction screening[16,17,62] and metabolic signature profiling[61].

the function of (protein-coding) genes. For example, protein co-expression captures functional relationships considerably better than mRNA co-expression[13,64]. Protein-based analyses also have the potential to distinguish between proteoforms; that is, the individual molecular forms of expressed proteins[65], which, as a result of splicing and post-translational modifications, dramatically increase the functional diversity of the proteome[65]. Proteoform characterization may require the use of top-down[66,67] or middle-down[68] proteomics approaches. Proteomics is rapidly increasing in throughput, with methods emerging that allow for hundreds of proteomes to be recorded per day on a single mass spectrometer[69,70]. A new generation of functional proteomic studies will hence be able to generate a much more comprehensive spectrum of biological functionality.

Nevertheless, protein annotation inequality is unlikely to be resolved exclusively by large-scale approaches. The first step in a concerted effort to address protein annotation bias could be to systematically provide the necessary minimal data foundation required for individual researchers conducting targeted experiments. Ongoing examples of this include BioPlex[71] and hu.MAP[72], which use MS for the large-scale identification of protein–protein interactions and protein complexes; the Human Protein Atlas[73,74], which uses antibodies to assign human proteins to different tissues and subcellular locations; and the neXt-CP50 project that aims to characterize 50 understudied proteins by proteomics[75].

## How to increase the impact of functional proteomics on mechanistic research

Some highly promising proteins remain ignored despite being perfectly amenable for detailed functional investigation[4]. Making protein–protein associations more accessible and usable for mechanistic follow-up studies will therefore be an important step toward reducing annotation inequality. Biologists can inspect molecular networks through a variety of powerful and user-friendly resources[76], including IntAct, BioGRID, NDEx and STRING. The fact that annotation bias is worsening[2] despite the wide availability of such resources could be the result of a number of factors. One may be a lack of awareness of such annotation portals among cell biologists. Others may be lack of trust in the available annotation, lack of annotations and lack of integration of different annotation types.

Cell biologists may hesitate to rely on data from large-scale projects due to a perceived lack of accuracy, which could be improved by better communication. Indeed, the possibility of treating error in a statistical sense is a particular strength of large-scale approaches. While error cannot be avoided, its size is a critical parameter to understand how reliable results are. One example of a functional proteomics technique where false discovery rate (FDR) calculation has been established is crosslinking MS[77]. Similarly, FDR is routinely calculated for all MS protein identifications[78,79]. In addition, in spatial proteomics, statistical frameworks are being developed to encapsulate confidence of assigning proteins to subcellular niches[80,81].

In addition to expanding the amount of available large-scale data, it will undoubtedly be necessary to develop new tools and techniques to provide additional, complementary links and fill systematic gaps left by current approaches. Examples of emerging functional proteomics technologies are crosslinking MS[42], coaggregation proteomics[82] and methods to study dynamic subcellular niches[52,55]. The large success by which protein structures can be predicted now[83] offers the exciting possibility to improve structure-based function prediction, especially when predicted structures could be experimentally

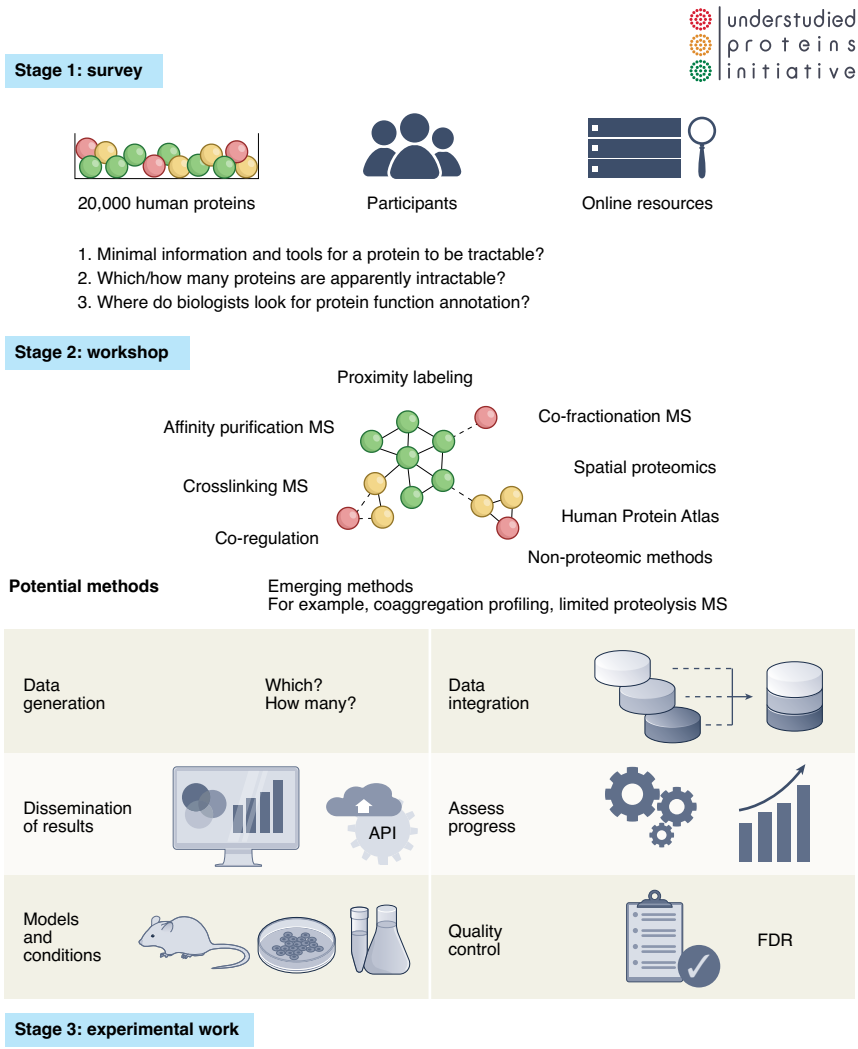


**Fig. 2 | Roadmap of the Understudied Proteins Initiative.** A survey will help define the challenge and goals for the initiative. Then a workshop will bring together experts from the large-scale data community to establish the initiative framework, covering six action areas to be discussed. Finally, a collaborative effort of many labs will experimentally tackle the problem of understudied proteins.

confirmed by, for example, crosslinking MS[84]. These and other intracellular techniques are particularly attractive, as many proteins require folding assistance, cofactors or post-translational modifications to function correctly and would therefore need to be studied in their native environment. In addition, it is becoming increasingly feasible to study proteomes of single cells, allowing the determination of cell-to-cell heterogeneity[85].

Finally, a key remaining challenge is the integration of different types of data across scales (time and space), which would maximize synergies between different types of omics data. An example for this is the integration of the Human Protein Atlas and BioPlex data, underpinning that the generation of a cellular hierarchy reveals many novel cellular systems undetectable by either dataset when used in isolation[86]. Such computational tools could also accelerate science through providing data-driven hypothesis generation; that is, opportunities for researchers to connect their data to big proteomics data.

Even where the function of a protein is well annotated, there is increasing evidence suggesting that a number of proteins have the capacity to carry out alternative, unrelated functions, reported in the literature as 'moonlighting'[87]. Historically, as researchers have assumed 'one-protein one-function', alternative functions have not been sought for most proteins. An additional benefit of the systems-wide interrogation of the functional proteome will be to provide alternative functional annotations even for well-studied proteins, as well as a better understanding of the extent to which proteins are capable of 'moonlighting'.

## How to quantify progress of functional characterization

To develop, optimize and evaluate strategies to tackle protein annotation inequality, one needs to be able to measure their impact in a robust and informative way. Measuring the degree of functional characterization is far from trivial, not least because the term itself can have different meanings. 'Protein function' may refer to the wider biological purpose of a protein, such as to which phenotype it associates, or to which metabolic pathway it belongs to. It could also refer to structural and mechanistic insights into how a protein fulfils these functions at a molecular level; for example, the enzymatic mechanism.

A number of approaches to determine protein annotation levels have been developed, including a literature score based on text mining[6], the UniProt annotation score[88], an assessment of Gene Ontology (GO) coverage[3] and a system to classify proteins based on their development as drug targets[5]. Each of these metrics captures or emphasizes slightly different aspects of the available annotations. They do not distinguish between original characterization and functional association. However, to systematically evaluate the performance of an annotation transfer system, it will be necessary to quantify it adequately. The McNamara fallacy[89] illustrates the danger of evaluating progress toward a complex goal on the basis of a single, easy-to-measure target variable without taking into account broader and more difficult to measure aspects of the challenge (McNamara's over-reliance on a single quantitative metric — number of enemy combatants killed or wounded — has been linked to the US failure in the Vietnam War).

## How to avoid exchanging one bias for another

We have argued that the proteome is a powerful layer for annotating gene function, but proteomics approaches are also susceptible to biochemical bias; for example, from protein abundance and solubility. Therefore, to achieve a systematic reduction in the genome-wide annotation bias, it may be necessary to optimize multiple individual functional proteomics methods and integrate their results in a concerted effort. One may also integrate proteomics data with data produced by other omic disciplines. Metabolomics, for instance, can capture a complementary functional spectrum[61,90].

Note that combining proteomics with genetics, functional genetics or metabolomics substantially improves the predictability of phenotypes[91,92].

Regardless of the approaches taken, however, the narrow window of standard laboratory conditions should probably be left behind. Recent multi-organism proteomics surveys[93,94] suggest that potentially many more proteins could be characterized by comparative proteomics, taking advantage of the broad evolutionary conservation of many proteins' functions and the differential accessibility of conserved proteins across organisms. The fact that many omic technologies can be directly applied to human cells, combined with the advent of genome editing, has raised concerns that funding for work on non-human organisms might be in decline[95,96], although in-depth statistics indicate that these concerns may be, at present, unfounded[97]. Studying a broad diversity of organisms has not only brought us penicillin, green fluorescent protein and CRISPR–Cas9, but may also help us to capture the functional spectrum of the human proteome.

## The Understudied Proteins Initiative

We envisage that the time is right for a coordinated effort to reduce annotation inequality across the human genome and proteome (Fig. 2). Our Understudied Proteins Initiative will include different data generation approaches, develop an integration framework and make the annotations available to researchers via an appropriate platform. The project will aim to address not only the technical but also the biomedical reasons for missing gene functions, such as narrowly defined growth conditions, single time-point studies and the focus on very few laboratory models with low genetic variability. This protein function moonshot may also stimulate methodological developments in functional proteomics and may extend to other species.

As a first step, the goal must be defined clearly. If the contribution of functional proteomics is to stimulate mechanistic studies of under-characterized proteins, then what is the minimum information that scientists require to start such work? This question can only be answered by those that illuminate the cellular function of individual proteins in molecular and mechanistic detail. Ultimately, it is the sum of their individual subjective decisions as laboratory scientists and reviewers that decide what proteins are being studied in detail. We recently launched a survey to capture their views (https://understudiedproteins.org/survey)[98].

As a second step, a community of interested scientists must be built. This will be started at an upcoming meeting supported by the Wellcome Trust (https://understudiedproteins.org/conference). The meeting will discuss the outcome of the survey and its implications for the goals of an Understudied Proteins Initiative, and how progress toward these goals could be monitored. This will set the framework for an open discussion on what technologies or developments may be able to systematically unlock the potential of currently uncharacterized proteins in biomedical research, and therefore become part of a larger roadmap. ❐

Georg Kustatscher[1]✉, Tom Collins[2], Anne-Claude Gingras[3,4], Tiannan Guo[5,6], Henning Hermjakob[7], Trey Ideker[8], Kathryn S. Lilley[9], Emma Lundberg[10,11,12,13], Edward M. Marcotte[14], Markus Ralser[15,16] and Juri Rappsilber[17,18]✉

*[1]Institute of Quantitative Biology, Biochemistry and Biotechnology, University of Edinburgh, Edinburgh, UK. [2]Wellcome Trust, London, UK. [3]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health System, Toronto, Ontario, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [5]Zhejiang Provincial Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, China. [6]Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, China. [7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [8]Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, CA, USA. [9]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, UK. [10]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH-Royal Institute of Technology, Stockholm, Sweden. [11]Department of Bioengineering, Stanford University, Stanford, CA, USA. [12]Department of Pathology, Stanford University, Stanford, CA, USA. [13]Chan Zuckerberg Biohub, San Francisco, CA, USA. [14]Department of Molecular Biosciences, Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, TX, USA. [15]Department of Biochemistry, Charité University Medicine, Berlin, Germany. [16]The Molecular Biology of Metabolism Laboratory, the Francis Crick Institute, London, UK. [17]Bioanalytics, Technische Universität Berlin, Berlin, Germany. [18]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, UK.*
✉*e-mail: georg.kustatscher@ed.ac.uk; juri.rappsilber@tu-berlin.de*

Published online: 09 May 2022
https://doi.org/10.1038/s41592-022-01454-x

References

1. Dolgin, E. *Nature* **551**, 427–431 (2017).
2. Haynes, W. A., Tomczak, A. & Khatri, P. *Sci. Rep.* **8**, 1362 (2018).
3. Wood, V. et al. *Open Biol.* **9**, 180241 (2019).
4. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. *PLoS Biol.* **16**, e2006643 (2018).
5. Oprea, T. I. et al. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
6. Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, B. & Eisenhaber, F. *Proteomics* **18**, e1800093 (2018).
7. Edwards, A. M. et al. *Nature* **470**, 163–165 (2011).
8. Dunham, I. *PLoS Biol.* **16**, e3000034 (2018).
9. Nguengang Wakap, S. et al. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
10. Leiserson, M. D. M. et al. *Nat. Genet.* **47**, 106–114 (2015).
11. Wang, T. et al. *Science* **350**, 1096–1101 (2015).
12. Adhikari, S. et al. *Nat. Commun.* **11**, 5301 (2020).
13. Kustatscher, G. et al. *Nat. Biotechnol.* **37**, 1361–1371 (2019).
14. Bakos, J., Zatkova, M., Bacova, Z. & Ostatnikova, D. *Neural Plast.* **2016**, 3276383 (2016).
15. Cardon, T., Fournier, I. & Salzet, M. *Trends Biochem. Sci.* **46**, 239–250 (2021).
16. Blomen, V. A. et al. *Science* **350**, 1092–1096 (2015).
17. Tsherniak, A. et al. *Cell* **170**, 564–576 (2017).
18. Fenner, M. *PLoS Biol.* **11**, e1001687 (2013).
19. Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. *Proc. Natl. Acad. Sci. USA* **112**, 14569–14574 (2015).
20. Chu, J. S. G. & Evans, J. A. *Proc. Natl Acad. Sci. USA* **118**, e2021636118 (2021).
21. Winzeler, E. A. et al. *Science* **285**, 901–906 (1999).
22. Hillenmeyer, M. E. et al. *Science* **320**, 362–365 (2008).
23. Wickerham, L. J. *Bull. US Dep. Agric.* **1029**, 1–56 (1951).
24. Glass, D. J. *Clin. Chem.* **56**, 1080–1085 (2010).
25. Yanai, I. & Lercher, M. *Genome Biol.* **21**, 231 (2020).
26. Bacon, F. *The Novum Organon, or a True Guide to the Interpretation of Nature* (Cambridge Univ. Press, 2005).
27. Su, A. I. & Hogenesch, J. B. *Genome Biol.* **8**, 404 (2007).
28. Williamson, A. R. *Nat. Struct. Biol.* **7**(Suppl), 953 (2000).
29. Gerlt, J. A. et al. *Biochemistry* **50**, 9950–9962 (2011).
30. Koscielny, G. et al. *Nucleic Acids Res.* **45**(D1), D985–D994 (2017).
31. Fedorov, O., Müller, S. & Knapp, S. *Nat. Chem. Biol.* **6**, 166–169 (2010).
32. Knapp, S. et al. *Nat. Chem. Biol.* **9**, 3–6 (2013).
33. Sun, X. et al. *Signal Transduct. Target. Ther.* **4**, 64 (2019).
34. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. *Nature* **402**, 83–86 (1999).
35. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. *Nat. Biotechnol.* **21**, 697–700 (2003).
36. Sharan, R., Ulitsky, I. & Shamir, R. *Mol. Syst. Biol.* **3**, 88 (2007).
37. Radivojac, P. et al. *Nat. Methods* **10**, 221–227 (2013).
38. Gligorijevic, V., Barot, M. & Bonneau, R. *Bioinformatics* **34**, 3873–3881 (2018).
39. Dunham, W. H., Mullin, M. & Gingras, A.-C. *Proteomics* **12**, 1576–1590 (2012).

40. Meyer, K. & Selbach, M. *Front. Genet.* **6**, 237 (2015).
41. Smits, A. H. & Vermeulen, M. *Trends Biotechnol.* **34**, 825–834 (2016).
42. O'Reilly, F. J. & Rappsilber, J. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
43. Salas, D., Stacey, R. G., Akinlaja, M. & Foster, L. J. *Mol. Cell. Proteomics* **19**, 1–10 (2020).
44. Wu, L. et al. *Nature* **499**, 79–82 (2013).
45. Kustatscher, G. et al. *EMBO J.* **33**, 648–664 (2014).
46. Wu, Y. et al. *Cell* **158**, 1415–1430 (2014).
47. Kustatscher, G., Grabowski, P. & Rappsilber, J. *Proteomics* **16**, 393–401 (2016).
48. Williams, E. G. et al. *Science* **352**, aad0189 (2016).
49. Gupta, S., Turan, D., Tavernier, J. & Martens, L. *Nucleic Acids Res.* **46**, D581–D585 (2018).
50. Singh, S. A. et al. *EMBO J.* **33**, 385–399 (2014).
51. Kirchner, M. et al. *Bioinformatics* **26**, 77–83 (2010).
52. Gingras, A.-C., Abe, K. T. & Raught, B. *Curr. Opin. Chem. Biol.* **48**, 44–54 (2019).
53. Trinkle-Mulcahy, L. *F1000research* https://doi.org/10.12688/f1000research.16903.1 (2019).
54. Gatto, L., Breckels, L. M. & Lilley, K. S. *Curr. Opin. Chem. Biol.* **48**, 123–149 (2019).
55. Lundberg, E. & Borner, G. H. H. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
56. Paz, J. & Lüders, J. *Trends Cell Biol.* **28**, 176–187 (2018).
57. Danielsson, F. et al. *Proteomics* **20**, e1900361 (2020).
58. Lam, M. H. Y. & Stagljar, I. *Proteomics* **12**, 1519–1526 (2012).
59. Timp, W. & Timp, G. *Sci. Adv.* **6**, eaax8978 (2020).
60. Hughes, T. R. et al. *Cell* **102**, 109–126 (2000).
61. Mülleder, M. et al. *Cell* **167**, 553–565 (2016).
62. Costanzo, M. et al. *Science* **353**, aaf1420 (2016).
63. le Sage, C., Lawo, S. & Cross, B. C. S. *SLAS Discov.* **25**, 233–240 (2020).
64. Wang, J. et al. *Mol. Cell. Proteomics* **16**, 121–134 (2017).
65. Aebersold, R. et al. *Nat. Chem. Biol.* **14**, 206–214 (2018).
66. Toby, T. K., Fornelli, L. & Kelleher, N. L. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **9**, 499–519 (2016).
67. Smith, L. M. & Kelleher, N. L. *Science* **359**, 1106–1107 (2018).
68. Sidoli, S. & Garcia, B. A. *Expert Rev. Proteomics* **14**, 617–626 (2017).
69. Bekker-Jensen, D. B. et al. *Mol. Cell. Proteomics* **19**, 716–729 (2020).
70. Messner, C. B. et al. *Cell Syst.* **11**, 11–24.e4 (2020).
71. Huttlin, E. L. et al. *Cell* **162**, 425–440 (2015).
72. Drew, K., Wallingford, J. B. & Marcotte, E. M. *Mol. Syst. Biol.* **17**, e10016 (2021).
73. Uhlén, M. et al. *Science* **347**, 1260419 (2015).
74. Thul, P.J. et al. *Science* **356**, aal3321 (2017).
75. Paik, Y.-K. et al. *J. Proteome Res.* **17**, 4042–4050 (2018).
76. Huang, J. K. et al. *Cell Syst.* **6**, 484–495 (2018).
77. Lenz, S. et al. *Nat. Commun.* **12**, 3564 (2021).
78. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. *Nat. Methods* **4**, 787–797 (2007).
79. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. *J. Proteome Res.* **7**, 40–44 (2008).
80. Kustatscher, G. & Rappsilber, J. *Trends Cell Biol.* **26**, 800–803 (2016).
81. Crook, O. M., Smith, T., Elzek, M. & Lilley, K. S. *Proteomics* **20**, e1900392 (2020).
82. Mateus, A. et al. *Mol. Syst. Biol.* **16**, e9232 (2020).
83. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
84. Ryl, P. S. J. et al. *J. Proteome Res.* **19**, 327–336 (2020).
85. Labib, M. & Kelley, S. O. *Nat. Rev. Chem.* **4**, 143–158 (2020).
86. Qin, Y. et al. *Nature* **600**, 536–542 (2021).
87. Jeffery, C. J. *Phil. Trans. R. Soc. Lond. B* **373**, 20160523 (2018).
88. UniProt Consortium. *Nucleic Acids Res.* **47**, D506–D515 (2019).
89. O'Mahony, S. *J. R. Coll. Physicians Edinb.* **47**, 281–287 (2017).
90. Allen, J. et al. *Nat. Biotechnol.* **21**, 692–696 (2003).
91. Szappanos, B. et al. *Nat. Genet.* **43**, 656–662 (2011).
92. Zelezniak, A. et al. *Cell Syst.* **7**, 269–283 (2018).
93. McWhite, C. D. et al. *Cell* **181**, 460–474 (2020).
94. Müller, J. B. et al. *Nature* **582**, 592–596 (2020).
95. Wangler, M. F., Yamamoto, S. & Bellen, H. J. *Genetics* **199**, 639–653 (2015).
96. Warren, G. *J. Cell Biol.* **208**, 387–389 (2015).
97. Lauer, M. A look at trends in NIH's model organism research support. https://nexus.od.nih.gov/all/2016/07/14/a-look-at-trends-in-nihs-model-organism-research-support/ (2016).
98. Kustatscher, G. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01316-z (2022).
99. Havugimana, P. C. et al. *Cell* **150**, 1068–1081 (2012).
100. Luck, K. et al. *Nature* **580**, 402–408 (2020).
101. Barrios-Rodiles, M. et al. *Science* **307**, 1621–1625 (2005).



## Acknowledgements

The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (grant no. 203149).


## Author contributions

G.K. and J.R. wrote the manuscript with input from all authors.

## Competing interests

T.G. is a shareholder of Westlake Omics Inc. T.I. is a co-founder of Data4Cure, is on its Scientific Advisory Board and has an equity interest. T.I. is on the Scientific Advisory Board of Ideaya BioSciences and has an equity interest. E.L. is advisor for Pixelgen Technologies and Moleculent. E.M.M. is a co-founder, shareholder and scientific board member of Erisyon, Inc. G.K., T.C., A.-C.G., H.H., K.L., M.R. and J.R. declare no competing interests.

## Additional information

**Peer review information** *Nature Methods* thanks Susanne Gräslund, Lydie Lane and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.