Understudied proteins

Understudied proteins: Opportunities and challenges for functional proteomics

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

[1] Institute of Quantitative Biology, Biochemistry and Biotechnology, University of Edinburgh, Edinburgh, EH9 3FF, UK
[2] Wellcome Trust, London, NW1 2BE, UK
[3] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health System, Toronto, ON, Canada; Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, M5G 1X5
[4] Zhejiang Provincial Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China; Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China
[5] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, UK
[6] Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, California 92093, USA
[7] Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, CB2 1GA, UK
[8] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH-Royal Institute of Technology, Stockholm, Sweden
[9] Department of Molecular Biosciences, Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, Texas 78712, USA
[10] Department of Biochemistry, Charité University Medicine, Berlin, Germany; The Molecular Biology of Metabolism Laboratory, the Francis Crick Institute, London, UK
[11] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany; Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

[#] To whom correspondence should be addressed: georg.kustatscher@ed.ac.uk and juri.rappsilber@tu-berlin.de

Most research to date focuses on a limited set of increasingly well-known proteins, whereas the biological function of many others remains poorly understood. Functional proteomics could be instrumental in reducing the annotation gap by systematically associating uncharacterised proteins with proteins of known function and thereby laying the groundwork for detailed mechanistic studies.

Keywords: Functional proteomics - Uncharacterised genes - Annotation bias - Protein function initiative - Dark proteome

Protein annotation inequality impedes biomedical progress

Proteins receive very different levels of attention from scientists. The most popular 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 function of thousands of human proteins remains unexplored[2–5]. Indeed, this bias in the functional characterisation of the human proteome is massive: 95% of all life science publications are about an elite group of 5,000 especially well-studied human proteins[6]. The sequencing of the human genome was expected to be a crucial step towards reducing this bias, by identifying all human genes and thus offering researchers 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, in fact, 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 (**Figure 1**), a phenomenon also known as the street-light effect[8]. Meanwhile, many uncharacterised proteins are not subjected to functional studies despite strong evidence from "omics" studies for their association with human disease[2]. For example, the function 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 many 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 uncharacterised, as of 2015[11]. This bias extends to the ~3,000 proteins currently expected to be druggable: only 5–10% of the potentially druggable proteins are currently targeted by FDA approved pharmaceuticals[5].

Origins of protein annotation inequality

The reasons for this 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 towards 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, very small size is a strikingly common feature of under-studied proteins: 40% of the least well-annotated proteins in SwissProt are smaller than 15 kDa[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", which come from regions of the genome that were previously considered noncoding[15].

Other reasons for protein annotation inequality may reflect conceptual biases in the research system rather than protein properties. 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 chances to be cited and consequently also increases the chances for high-impact journal publications required for academic success[18]. However, large fields also tend to favour 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 standardisation of the experimental conditions. For example, under standard laboratory growth conditions, the deletion of about 20% of *S. 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" condition 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 1950's publication of the US Dept of Agriculture technical bulletin, which attempted to help farmers and biotechnologists to grow a wide variety of yeasts, for instance 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 current focus on hypothesis-driven, rather than question-driven, research[24,25]. It is difficult to formulate hypotheses on the mechanistic molecular function of an uncharacterised 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 towards the characterisation 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 characterisation" is typically interpreted as revealing molecular properties of a protein that are particularly relevant for drug development, e.g. 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 cell biology and basic research, it is equally important to study other levels of protein annotation, such as in which cellular processes, pathways and subcellular compartments a protein functions. 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 that is ideally suited to study this cell biological aspect of protein function, 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 also a 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 uncharacterised proteins are linked to well-studied factors and their biological functions[34–38].

Proteomics approaches are particularly well suited for revealing functional associations at large scale. This includes 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 a subcellular space[52–55] (Box 1). For example, the majority of centrosomal proteins were considered to have been identified[56] before antibody-based proteomics identified hundreds of additional centrosomal proteins[57]. Note that here we focus on mass spectrometry and antibody-based proteomics, but powerful alternative proteomics approaches 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 mass spectrometry-based proteomics does not yet reach the gene coverage of genomic approaches, observing proteins directly can be especially informative when studying 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, *i.e.* 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 characterisation 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 for this include BioPlex[71] and hu.MAP[72], which use mass spectrometry 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 characterise 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 towards 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, 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 way 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 mass spectrometry[77]. Similarly, FDR is routinely calculated for all mass spectrometry 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 mass spectrometry[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 the predicted structures could be experimentally confirmed, *e.g.* by crosslinking mass spectrometry[84]. These and other in-cell techniques are particularly attractive as many proteins require assistance for folding or 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 will maximise 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 not detectable with either dataset alone[86]. Such computational tools could also accelerate science through providing data-driven hypothesis generation, *i.e.* 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 to suggest 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 and a better understanding of the extent to which proteins are capable of 'moonlighting'.

How to quantify progress of functional characterisation?

To develop, optimise 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 characterisation 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 on a molecular level, *e.g.* 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 GO coverage[3] and a system to classify proteins based on their development as drug targets[5]. Each of these metrics captures or emphasises slightly different aspects of the available annotations. They do not distinguish between original characterisation 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 towards 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 US American 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, *e.g.* from protein abundance and solubility. Therefore, to achieve a systematic reduction in the genome-wide annotation bias, it may be necessary to optimise 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 and 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 must 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, GFP 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 (**Figure 2**). Our “understudied proteins initiative” will include different data generation approaches, develop an integration framework and make the annotations available to researchers in an appropriate form. 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).

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, its implications for the goals of an understudied proteins initiative and how progress towards these goals could be monitored. This will set the frame for an open discussion on what technologies or developments may be able to systematically unlock the potential of currently uncharacterised proteins in biomedical research and therefore become part of a larger roadmap.

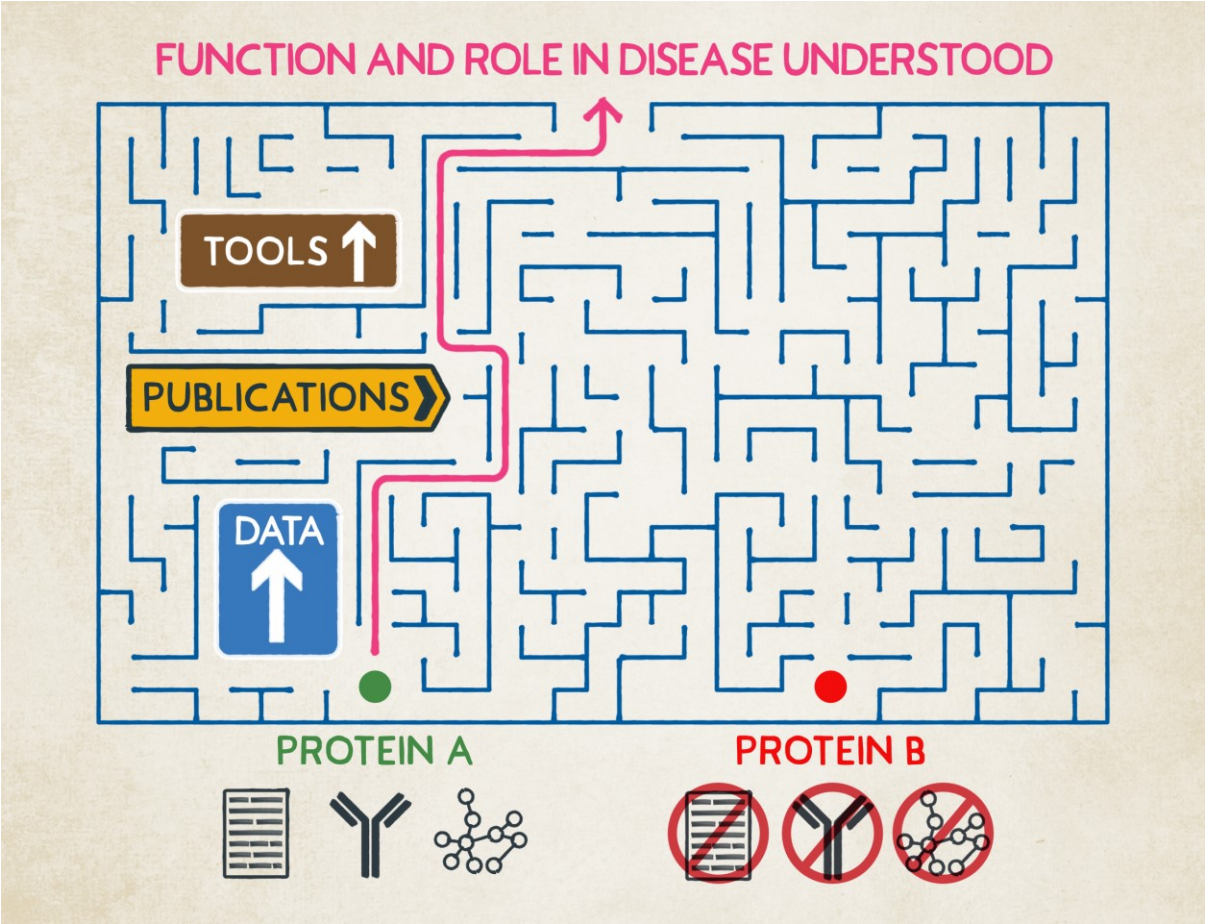

Figure 1: Protein annotation inequality impedes biomedical progress. The availability of prior publications, data and tools dictates the ease by which research questions involving a protein can be formulated and addressed. This reinforces annotation bias and the persistence of understudied proteins.

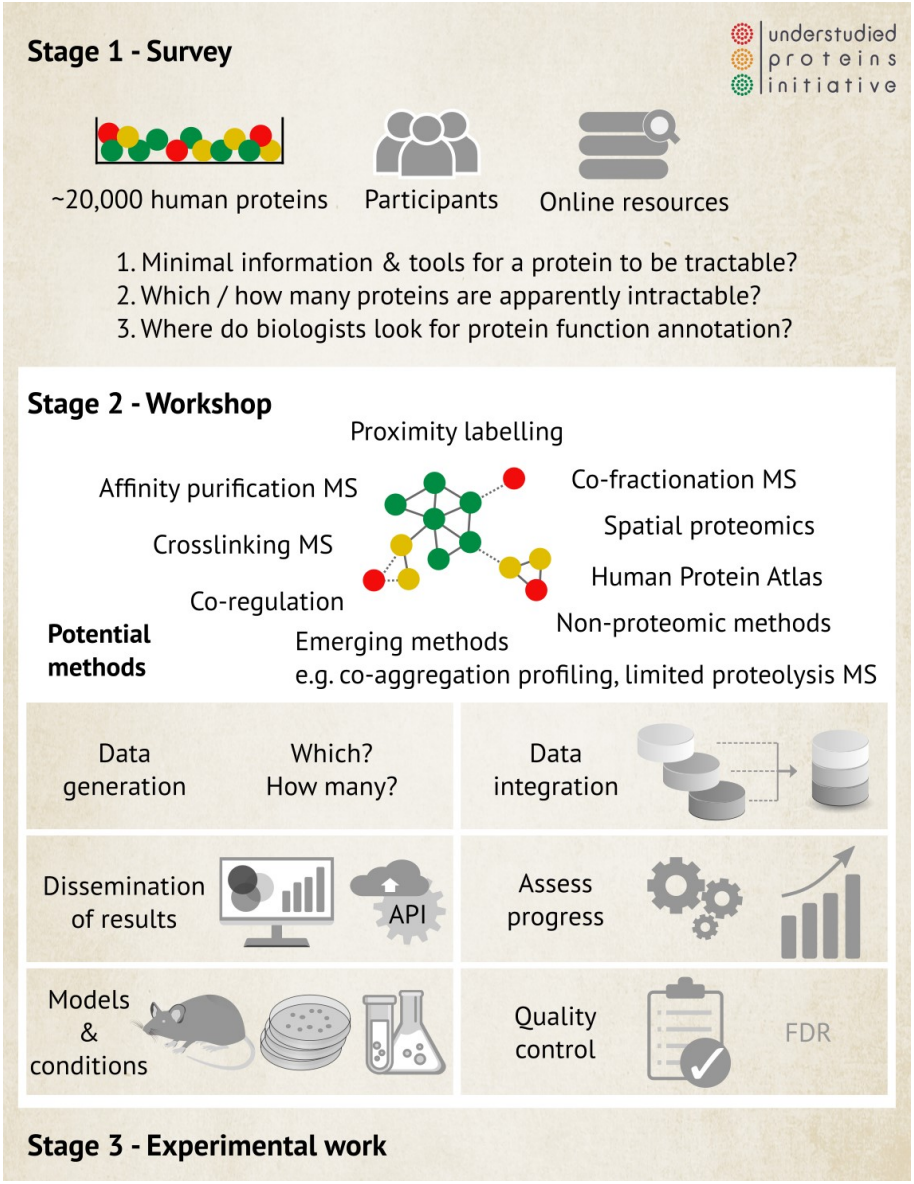

Figure 2: Roadmap of the understudied proteins initiative. A survey will help define the challenge and goals for the initiative. Then, a workshop will bring together experts from the large-scale data community to establish the initiative framework, covering six action areas to be discussed. Finally,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

Box 1: Proteomic approaches that reveal protein - protein associations

Mass spectrometry (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 (AP-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[43,98] and the subcellular localisation of proteins[54,55,80].

Proximity labelling MS: "Bait" proteins are fused to enzymes that enable biotinylation of "prey" proteins in living cells, which can subsequently be affinity-purified and quantified by MS[52,53]. Links reflect close spatial proximity of proteins.

Antibody-based proteomics: Subcellular localisation of proteins is revealed using antibodies[74]. The assays provide single cell resolution *in situ* and can detect multi-localizing proteins and 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-mass spectrometry-based methods that also reveal protein - protein associations[58,59], including binary assays such as Y2H[99], LUMIER[100], genetic interaction screening[16,17,62], and metabolic signature profiling[61].

References

1. Dolgin, E. The most popular genes in the human genome. *Nature* **551**, 427–431 (2017).
2. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018).
3. Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **9**, 180241 (2019).
4. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* **16**, e2006643 (2018).
5. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
6. Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, B. & Eisenhaber, F. Darkness in the Human Gene and Protein Function Space: Widely Modest or Absent Illumination by the Life Science Literature and the Trend for Fewer Protein Function Discoveries Since 2000. *Proteomics* **18**, e1800093 (2018).
7. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).
8. Dunham, I. Human genes: Time to follow the roads less traveled? *PLoS Biol.* **16**, e3000034 (2018).
9. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
10. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
11. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
12. Adhikari, S. *et al.* A high-stringency blueprint of the human proteome. *Nat. Commun.* **11**,

5301 (2020).

13. Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **37**, 1361–1371 (2019).
14. Bakos, J., Zatkova, M., Bacova, Z. & Ostatnikova, D. The Role of Hypothalamic Neuropeptides in Neurogenesis and Neuritogenesis. *Neural Plast.* **2016**, 3276383 (2016).
15. Cardon, T., Fournier, I. & Salzet, M. Shedding Light on the Ghost Proteome. *Trends Biochem. Sci.* **46**, 239–250 (2021).
16. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).
17. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
18. Fenner, M. What Can Article-Level Metrics Do for You? *PLoS Biol.* **11**, e1001687 (2013).
19. Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14569–14574 (2015).
20. Chu, J. S. G. & Evans, J. Too Many Papers? Slowed Canonical Progress in Large Fields of Science. (2018) doi:10.31235/osf.io/jk63c.
21. Winzeler, E. A. *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
22. Hillenmeyer, M. E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
23. Wickerham, L. J. No Title. *U.S. Dept. Agric.Tech. Bull.* **1029**, 1–56 (1951).
24. Glass, D. J. A critique of the hypothesis, and a defense of the question, as a framework for experimentation. *Clin. Chem.* **56**, 1080–1085 (2010).
25. Yanai, I. & Lercher, M. A hypothesis is a liability. *Genome Biol.* **21**, 231 (2020).
26. Bacon, F. *The Novum Organon, or a true guide to the interpretation of nature*. (Cambridge University Press, 2005).

27. Su, A. I. & Hogenesch, J. B. Power-law-like distributions in biomedical publications and research funding. *Genome Biol.* **8**, 404 (2007).
28. Williamson, A. R. Creating a structural genomics consortium. *Nat. Struct. Biol.* **7 Suppl**, 953 (2000).
29. Gerlt, J. A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950–9962 (2011).
30. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
31. Fedorov, O., Müller, S. & Knapp, S. The (un)targeted cancer kinome. *Nat. Chem. Biol.* **6**, 166–169 (2010).
32. Knapp, S. *et al.* A public-private partnership to unlock the untargeted kinome. *Nat. Chem. Biol.* **9**, 3–6 (2013).
33. Sun, X. *et al.* PROTACs: great opportunities for academia and industry. *Signal Transduct Target Ther* **4**, 64 (2019).
34. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
35. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* **21**, 697–700 (2003).
36. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
37. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
38. Gligorijevic, V., Barot, M. & Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* **34**, 3873–3881 (2018).
39. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12**, 1576–1590 (2012).

40. Meyer, K. & Selbach, M. Quantitative affinity purification mass spectrometry: a versatile technology to study protein-protein interactions. *Front. Genet.* **6**, 237 (2015).
41. Smits, A. H. & Vermeulen, M. Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol.* **34**, 825–834 (2016).
42. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
43. Salas, D., Stacey, R. G., Akinlaja, M. & Foster, L. J. Next-generation Interactomics: Considerations for the Use of Co-elution to Measure Protein Interaction Networks. *Mol. Cell. Proteomics* **19**, 1–10 (2020).
44. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013).
45. Kustatscher, G. *et al.* Proteomics of a fuzzy organelle: interphase chromatin. *EMBO J.* **33**, 648–664 (2014).
46. Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158**, 1415–1430 (2014).
47. Kustatscher, G., Grabowski, P. & Rappsilber, J. Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data. *Proteomics* **16**, 393–401 (2016).
48. Williams, E. G. *et al.* Systems proteomics of liver mitochondria function. *Science* **352**, aad0189 (2016).
49. Gupta, S., Turan, D., Tavernier, J. & Martens, L. The online Tabloid Proteome: an annotated database of protein associations. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkx930.
50. Singh, S. A. *et al.* Co-regulation proteomics reveals substrates and mechanisms of APC/C-dependent degradation. *EMBO J.* **33**, 385–399 (2014).
51. Kirchner, M. *et al.* Computational protein profile similarity screening for quantitative mass

spectrometry experiments. *Bioinformatics* **26**, 77–83 (2010).

52. Gingras, A.-C., Abe, K. T. & Raught, B. Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Curr. Opin. Chem. Biol.* **48**, 44–54 (2019).
53. Trinkle-Mulcahy, L. Recent advances in proximity-based labeling methods for interactome mapping. *F1000Res.* **8**, (2019).
54. Gatto, L., Breckels, L. M. & Lilley, K. S. Assessing sub-cellular resolution in spatial proteomics experiments. *Curr. Opin. Chem. Biol.* **48**, 123–149 (2019).
55. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).
56. Paz, J. & Lüders, J. Microtubule-Organizing Centers: Towards a Minimal Parts List. *Trends Cell Biol.* **28**, 176–187 (2018).
57. Danielsson, F. *et al.* Spatial Characterization of the Human Centrosome Proteome Opens Up New Horizons for a Small but Versatile Organelle. *Proteomics* **10**, 1900361 (2020).
58. Lam, M. H. Y. & Stagljar, I. Strategies for membrane interaction proteomics: no mass spectrometry required. *Proteomics* **12**, 1519–1526 (2012).
59. Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. *Sci Adv* **6**, eaax8978 (2020).
60. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
61. Mülleder, M. *et al.* Functional Metabolomics Describes the Yeast Biosynthetic Regulome. *Cell* **167**, 553–565.e12 (2016).
62. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, (2016).
63. le Sage, C., Lawo, S. & Cross, B. C. S. CRISPR: A Screener's Guide. *SLAS DISCOVERY:*

Advancing the Science of Drug Discovery **25**, 233–240 (2020).

64. Wang, J. *et al.* Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Mol. Cell. Proteomics* **16**, 121–134 (2017).
65. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
66. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).
67. Smith, L. M. & Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **359**, 1106–1107 (2018).
68. Sidoli, S. & Garcia, B. A. Middle-down proteomics: a still unexploited resource for chromatin biology. *Expert Rev. Proteomics* **14**, 617–626 (2017).
69. Bekker-Jensen, D. B. *et al.* A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Molecular & Cellular Proteomics* vol. 19 716–729 (2020).
70. Messner, C. B. *et al.* Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Systems* (2020) doi:10.1016/j.cels.2020.05.012.
71. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
72. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: Integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Cold Spring Harbor Laboratory* 2020.09.15.298216 (2020) doi:10.1101/2020.09.15.298216.
73. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
74. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).
75. Paik, Y.-K. *et al.* Launching the C-HPP neXt-CP50 Pilot Project for Functional

Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* **17**, 4042–4050 (2018).

76. Huang, J. K. *et al.* Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Syst* **6**, 484–495.e5 (2018).
77. Lenz, S. *et al.* Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *bioRxiv* 2020.05.25.114256 (2020) doi:10.1101/2020.05.25.114256.
78. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
79. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **7**, 40–44 (2008).
80. Kustatscher, G. & Rappsilber, J. Compositional Dynamics: Defining the Fuzzy Cell. *Trends Cell Biol.* **26**, 800–803 (2016).
81. Crook, O. M., Smith, T., Elzek, M. & Lilley, K. S. Moving Profiling Spatial Proteomics Beyond Discrete Classification. *Proteomics* **20**, e1900392 (2020).
82. Mateus, A. *et al.* Thermal proteome profiling for interrogating protein interactions. *Mol. Syst. Biol.* **16**, e9232 (2020).
83. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
84. Ryl, P. S. J. *et al.* In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human Mitochondria. *J. Proteome Res.* **19**, 327–336 (2020).
85. Labib, M. & Kelley, S. O. Single-cell analysis targeting the proteome. *Nature Reviews Chemistry* **4**, 143–158 (2020).
86. Qin, Y. *et al.* Mapping cell structure across scales by fusing protein images and interactions. 2020.06.21.163709 (2020) doi:10.1101/2020.06.21.163709.
87. Jeffery, C. J. Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc.*

Lond. B Biol. Sci. **373**, (2018).

88. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
89. O'Mahony, S. Medicine and the McNamara fallacy. *J. R. Coll. Physicians Edinb.* **47**, 281–287 (2017).
90. Allen, J. *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696 (2003).
91. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43**, 656–662 (2011).
92. Zelezniak, A. *et al.* Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Syst* **7**, 269–283.e6 (2018).
93. McWhite, C. D. *et al.* A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies. *Cell* **181**, 460–474.e14 (2020).
94. Müller, J. B. *et al.* The proteome landscape of the kingdoms of life. *Nature* **582**, 592–596 (2020).
95. Wangler, M. F., Yamamoto, S. & Bellen, H. J. Fruit flies in biomedical research. *Genetics* **199**, 639–653 (2015).
96. Warren, G. In praise of other model organisms. *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. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
99. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

100. Barrios-Rodiles, M. *et al*. High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307**, 1621–1625 (2005).