An open invitation to the Understudied Proteins Initiative

To the Editor — Much of life science research revolves around understanding the biological function of proteins. Some proteins, such as the tumor suppressor p53, have been studied extensively[1]. By contrast, thousands of human proteins remain 'understudied': their biological function is poorly understood and annotation of their molecular properties is scarce[2–6]. However, without a minimal amount of molecular annotation, it is difficult to formulate effective research questions and design experiments to investigate the function of these proteins in mechanistic detail[2].

The disparity in how much we know about individual proteins leads to a phenomenon known as the 'streetlight effect' or the 'rich-get-richer syndrome', in which research in a field preferentially targets proteins that are already well-studied[7]. There are many reasons for this, including practical considerations (for example, the abundance, solubility and size of a protein), the ease of designing a research plan that depends on available knowledge (for example, knockout phenotype, molecular interactions) and the availability of tools such as antibodies. In addition, working on proteins that already receive a lot of attention (for example, some disease-associated proteins) increases the chances of high-impact publications and funding. Hypothesis-driven (rather than question-driven) research may also contribute, as hypothesizing about the potential function of a completely uncharacterized protein is nearly impossible. Finally, some proteins may remain understudied because they are not expressed or required in standard laboratory conditions. Ironically, some of this problem is caused by the global desire to make research more reproducible through the standardization of experimental conditions.

One counter-argument is that the important proteins are being studied and the others are not as important to pursue. The evidence suggests otherwise: genome-wide studies show that research attention bias does not reflect the importance of genes for cellular processes and human disease[2,5]. For example, more than half of the host genes implicated in COVID-19 identified by genome-wide studies have not been pursued in more detail by targeted studies of the COVID-19 field[8]. Furthermore, the creation of a synthetic minimal bacterium required 149 proteins of unknown function[9]. If these proteins are crucial for the most-minimal cell possible to survive, they should be important to us.

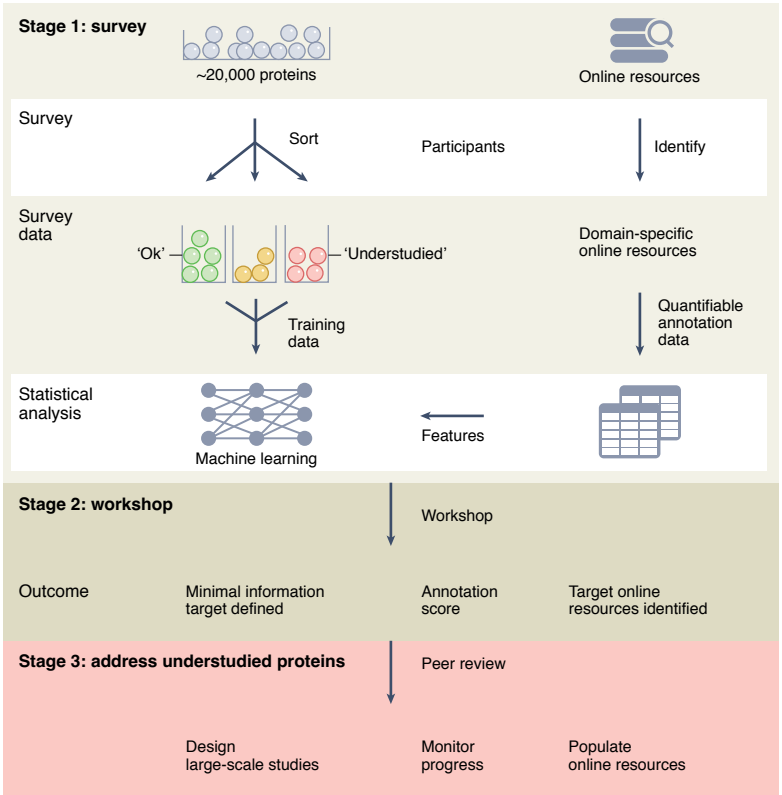

Fig. 1 | Roadmap of the Understudied Proteins Initiative. Stages 1 and 2 focus on defining the challenge and building a community. First, a survey among biomedical researchers (https://understudiedproteins.org/survey) will define the minimal information needed to counterbalance the current data bias that works against understudied proteins being included in mechanistic investigations. The survey will also reveal how many proteins are to be considered understudied and provide the data to train an algorithm to automatically assess annotation bias in the future. In addition, the survey will reveal at which locations researchers look for annotation and thus where new annotations should be added. In a second step, a workshop will bring together experts in different disciplines and technologies that provide large-scale data for systematic annotation of proteins to establish the framework of a coordinated understudied proteins initiative. The six action areas to be discussed are data generation, data integration, dissemination of results, assessment of progress, model systems and conditions to cover, and quality control. This will then lead to stage 3, the experimental work that will see a collaborative effort of many laboratories to tackle the problem of understudied proteins.

As current approaches to study proteins often reinforce the streetlight effect, we seek to pursue a different approach. We propose that a coordinated effort of the functional proteomics field could be an effective way to systematically advance the basic molecular characterization of understudied proteins, such that detailed studies become more feasible. With the goal of openly discussing, coordinating and initiating efforts to address these challenges, we established the Understudied Proteins Initiative[10], with participation of the Wellcome Trust (Fig. 1). In essence, for each understudied protein, we aim to provide enough molecular information (for example, protein interactions, colocalization or coexpression) that hypotheses about its putative function

can be made. Importantly, this should make it clear which field or laboratory with a particular research focus would be best placed to carry out further detailed studies of the protein. Thus, the giant task of characterizing the many understudied proteins is split into two parts: a large-scale precharacterization by omics laboratories, and a focused detailed investigation by molecular biology laboratories.

Choosing the right tools and experiments for such a large-scale data-generation effort requires critical input before data collection begins. As a first step, we have recently launched an openly accessible survey to allow us to better understand which human proteins remain understudied, what the minimal information is that would kick-start their inclusion in mechanistic investigations and where this information should be available (https://understudiedproteins.org/survey). Scientists who engage in mechanistic investigations are best placed to define this.

As a second step, we will then gather experimentalists and computational experts interested in large-scale approaches at a conference (https://understudiedproteins.org/conference) to discuss and identify ways to deliver this information. Ultimately, individual researchers stand to gain from the results of this initiative whenever they face new proteins in an ongoing study and need to prioritize novel targets for further investigation.

Survey participants will be shown a randomly selected human protein and asked to assign it to one of three annotation levels. In addition, they will declare which tools and resources were used for that assessment and what information they regard as important before starting experimental work with a new protein. We envision that respondents will need no more than five minutes per protein. Each protein will be presented to multiple participants, allowing us to average responses and capture the range of different interpretations and assessments of a protein's annotation level. In this way, the survey will deliver a manually curated assessment of the annotation level of human proteins. Although scores exist that express various aspects of protein annotation[3,6,11,12], our survey will return a score that specifically expresses how amenable a protein is to detailed mechanistic investigations.

Next, we will cross-reference this vote-based annotation score with the quantifiable annotation information available for the same protein and its homologs in publicly available resources named by participants and others, which could include PubMed, STRING, BioGRID, UniProt, Gene Cards, Wikipedia, Complex Portal and the Human Protein Atlas. This collated information will reveal key characteristics of understudied proteins, such as what type of quantifiable experimental evidence is available or lacking, and where it is accessible. Notably, this understanding is not limited to human proteins and guides the extension of our efforts toward other species.

The free-text answers from survey respondents will allow us to cross-check whether our data-based assessment agrees with what participants think regarding the minimal information that makes a protein a viable target of study, and where and how annotation should be accessible. In addition, on the basis of the annotation score and the cross-referenced quantifiable annotation information, we will train a machine-learning algorithm to automate the annotation scoring. An automated annotation scoring system allows us to keep scores up-to-date, assess proteins of other species and transparently monitor progress in protein annotation over time. Therefore, if a sizeable proportion of the community who reads this Correspondence and the paper in *Nature Methods*[10] participates in the survey and shares it with colleagues, then we will build a community-driven foundation for the Understudied Proteins Initiative.

With a clear understanding of what constitutes the experimental information that would make an understudied protein amenable to study, we will then start a discussion with funding agencies on how to set up calls aimed at providing this information. A critical component will be the evaluation of the effect of different information sources, facilitated by our automated annotation scoring. We will reveal the benefit of the respective datasets and approaches by monitoring the rate of annotation of understudied proteins. Measuring the effect of large-scale data will inform the effective use of funding, but also highlight where technology developments are needed to fill any systematic gaps left by current tools. Instead of lots of data, we aim to generate meaningful data. Eventually, thousands of laboratories around the world will be able to add those currently understudied proteins that fall into their own fields of interest to ongoing and future mechanistic investigations, thereby ending the era of understudied proteins. Our initiative complements those that have a strong emphasis either on bacterial proteins (COMBREX[13] and the Enzyme Function Initiative[14]) or on protein–small molecule interactions, such as the Structural Genomics Consortium[5,15], Open Targets[16] and the Illuminating the Druggable Genome program[6], which aims to improve our understanding of uncharacterized proteins within the three most commonly drug-targeted protein families (G-protein-coupled receptors, ion channels and protein kinases).

By providing a basic molecular characterization of all proteins, the Understudied Proteins Initiative will catalyze mechanistic investigations of understudied proteins, drive new biomedical research, and boost our understanding of the human proteome and its role in disease. We invite the community to get involved by participating in the survey and spreading the word. ❐

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[1,17,18]✉

[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, Institute of Biotechnology, 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: 9 May 2022
https://doi.org/10.1038/s41587-022-01316-z

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. Stoger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. *PloS Biol.* **16**, e2006643 (2018).
5. Edwards, A. M. et al. *Nature* **470**, 163–165 (2011).
6. Oprea, T. I. et al. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
7. Dunham, I. *PLoS Biol.* **16**, e3000034 (2018).
8. Stoeger, T. & Nunes Amaral, L. A. *eLife* **9**, e61981 (2020).
9. Hutchison, C. A. III et al. *Science* **351**, aad6253 (2016).
10. Kustatscher, G. *Nat. Methods* https://doi.org/10.1038/s41592-022-01454-x (2022).
11. Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, B. & Eisenhaber, F. *Proteomics* **18**, e1800093 (2018).
12. UniProt Consortium. *Nucleic Acids Res.* **47**, D506–D515 (2019).
13. Anton, B. P. et al. *PLoS Biol.* **11**, e1001638 (2013).
14. Gerlt, J. A. et al. *Biochemistry* **50**, 9950–9962 (2011).
15. Williamson, A. R. *Nat. Struct. Biol.* **7**, 953 (2000).
16. Koscielny, G. et al. *Nucleic Acids Res.* **45**, D985–D994 (2017).

Competing interests

T.G. is a shareholder of Westlake Omics Inc. T.I. is a cofounder of Data4Cure, is on the 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 cofounder, shareholder and scientific board member of Erisyon, Inc. G.K., T.C., A.-C.G., H.H., K.S.L., M.R. and J.R. declare no competing interests.

The GA4GH Phenopacket schema defines a computable representation of clinical data

To the Editor — Despite great strides made in the development and wide acceptance of standards for exchanging structured information about genomic variants, progress in standards for computational phenotype analysis for translational genomics has lagged behind. Phenotypic features (signs, symptoms, laboratory and imaging findings, results of physiological tests, etc.) are of high clinical importance, yet exchanging them in conjunction with genomic variation information is often overlooked or even neglected. In the clinical domain, substantial work has been dedicated to the development of computational phenotypes[1]. Traditionally, these approaches have largely relied on rule-based methods and large sources of clinical data to identify cohorts of patients with or without a specific disease[2–5]. However, they were not developed to enable deep phenotyping of abnormalities, to facilitate computational analysis of interpatient phenotypic similarity or to support computational decision support. To address this, the Global Alliance for Genomics and Health[6] (GA4GH) has developed the Phenopacket schema, which supports the exchange of computable longitudinal case-level phenotypic information for diagnosis of, and research on, all types of disease, including Mendelian and complex genetic diseases, cancers and infectious diseases. A Phenopacket characterizes an individual person or biosample, linking that individual to detailed phenotypic descriptions, genetic information, diagnoses and treatments (Fig. 1). The Phenopacket software is available at https://github.com/phenopackets/.

The 'PhenotypicFeature' is the central element of the Phenopacket schema. A 'PhenotypicFeature' can be used to describe any phenotypic characteristic, including signs and symptoms, laboratory findings, histopathology findings, and imaging and electrophysiological results, along with modifier and qualifier concepts. Each phenotypic feature is described using an ontology term. Although the Phenopacket schema does not mandate which ontology to use, it provides recommendations, such as the Human Phenotype Ontology[7] (HPO) for rare diseases and the National Cancer Institute Thesaurus (NCIT) for transmission of information about a cancer specimen (for example, pathological staging or more detailed information about histology or tumor markers)[8]. Within the schema, it is possible to indicate whether an abnormality was excluded during the diagnostic process (for example, whether a morphological cardiac defect was excluded by echocardiography) or to use other optional HPO terms to denote the severity, frequency (for example, number of occurrences of seizures per week), laterality (for example, unilateral) or other pattern of a phenotypic feature in the patient being described. Finally, the onset (and, if applicable, the resolution) of specific features can be indicated.

Other key elements of the schema are 'Measurement', which is used to capture quantitative (i.e., numerical), ordinal (for example, absent/present) or categorical measurements; 'Biosample', a description of biological material obtained from the individual represented in the Phenopacket and used for phenotypic, genotypic or other -omics analysis; and 'MedicalAction', which includes a hierarchical representation of medical actions, including medications, procedures and other actions taken for clinical management. The 'Treatment' element is a subelement of 'MedicalAction' and represents the administration of a pharmaceutical agent, broadly defined as prescription and over-the-counter medicines, vaccines and other therapeutic agents, such as monoclonal antibodies or chimeric antigen receptor (CAR)-T-cell therapy.

The 'Interpretation' element specifies interpretations of genomic findings. This element leverages complementary resources developed by the GA4GH Genomic Knowledge Standards Work Stream: the Variation Representation Specification (VRS) and VRS Added Tools for Interoperable Loquacious Exchange (VRSATILE)[6]. Further information on this and other elements is available in the online documentation (https://phenopacket-schema.readthedocs.io/).

The Phenopacket schema was designed to support several use cases. Phenotype-driven rare-disease genomic diagnostic software has previously used bespoke formats to represent phenotypic data (generally in the form of a list of HPO terms) and pedigree information. Phenopacket provides a standard input format for these tools that will simplify computational analysis pipelines, and the additional clinical information will enable analysis pipelines and algorithms to leverage other data, such as age of onset and excluded abnormalities. A number of databases have adopted the standard to represent the clinical data of individuals in the context of rare-disease genomics (European Genome-phenome Archive), registries (European Joint Programme on Rare Diseases and Western Australian Register of Developmental Anomalies), biosamples (EMBL-EBI BioSamples database) and biobanks (the Japanese Agency for Medical Research and Development Tohoku Medical Megabank project and National Center Biobank Network). In addition, Phenopackets can be used to store a computational representation of a case report, and we envision that authors could submit representations of patients as phenopackets to accompany published case reports