# An open invitation to the understudied proteins initiative

*Correspondence*

# An invitation to help define the challenge and goals for an understudied proteins initiative

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

**To the Editor —** Much of life science research revolves around understanding the biological function of the proteins. Some proteins, such as the tumour suppressor p53, have been studied extensively[1]. In contrast, thousands of human proteins remain "understudied", i.e. 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", whereby research in a field preferentially targets proteins that are already well-studied[7]. There are many reasons for this, including practical considerations such as the abundance, solubility and size of a protein, and the ease of designing a research plan which depends on available knowledge such as knock-out 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 for high-impact publications and funding. Hypothesis-driven (rather than question-driven) research may also contribute, as hypothesising about the potential function of a completely uncharacterised protein is nearly impossible. Finally, some proteins may remain understudied because they are not expressed or required in standard laboratory conditions. Paradoxically, some of the problem is hence caused by the global desire to make research more reproducible, through the standardisation of experimental conditions. So what, one might say. Maybe the important proteins are being studied, and the others are just not worth pursuing. Well, 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 COVID-19 host genes identified by genome-wide studies had not been pursued in more detail by targeted studies[8]. Furthermore, creation of a synthetic minimal bacterium required 149 proteins of unknown function[9]. If these proteins are crucial for the most minimal possible cell to survive, they should also be important to us. As our current approaches to study proteins often reinforce the "streetlight effect", we suggest pursuing a different approach. We propose that a coordinated effort of the functional proteomics field could be a strong way to systematically advance the basic molecular characterisation 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, with participation of the Wellcome Trust (*Kustatscher et al, accompanying manuscript*) (**Fig. 1**). In essence, for each understudied protein we aim to provide enough molecular information (such as protein interactions, co-localisation, co-expression) so that hypotheses about its putative function can be made. Importantly, this should make it clear which field or lab with a particular research focus would be best placed to carry out further detailed studies of the protein. Hence, the giant task of characterising the many understudied proteins is split into two parts, a large-scale pre-characterisation by omics labs and a focussed detailed investigation by molecular biology labs. To choose 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 is the minimal information that would kick-start their inclusion in mechanistic investigations, and where this information should be available (https://understudiedproteins.org/survey). Scientists that 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/conferencelink) 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 prioritise novel targets for further investigation.

Survey participants will be shown a randomly selected human protein and are asked to assign it to one of three annotation levels. In addition, they 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 the time spent on this to be no more than five minutes per

participant and 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. While scores exist that express various aspects of protein annotation[3,6,10,11], this 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 proteins in publicly available resources. These will include resources named by participants and others, if not already included, such as PubMed, STRING, BioGRID, UniProt, Gene Cards, Wikipedia, Complex Portal, Human Protein Atlas. This 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 towards other species. The free-text answers will allow us to cross-check if our data-based assessment agrees with what participants think on 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 the progress in protein annotation over time. Therefore, if everyone who reads this paper 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 studiable, we will discuss with funding agencies about setting up calls aimed at providing this information. A critical component will be the evaluation of the impact of different information sources by help of our automated annotation scoring. We will reveal the benefit of the respective data sets and approaches by monitoring the rate of annotation of understudied proteins. Measuring the impact 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 labs around the world will be able to add those currently understudied proteins that fall into their own interest and competence spheres to ongoing and future mechanistic investigations, to end the era of understudied proteins. Our initiative complements initiatives that have a strong emphasis either on bacterial proteins (COMBREX[12] and the Enzyme Function Initiative[13]) or on protein-small molecule interactions, such as the Structural Genomics Consortium[5,14], Open Targets[15] and the Illuminating the Druggable Genome (IDG) 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 characterisation of all proteins, the understudied proteins initiative will catalyse mechanistic investigations of understudied proteins, drive new biomedical research, and boost our understanding of the human proteome and its role in disease.

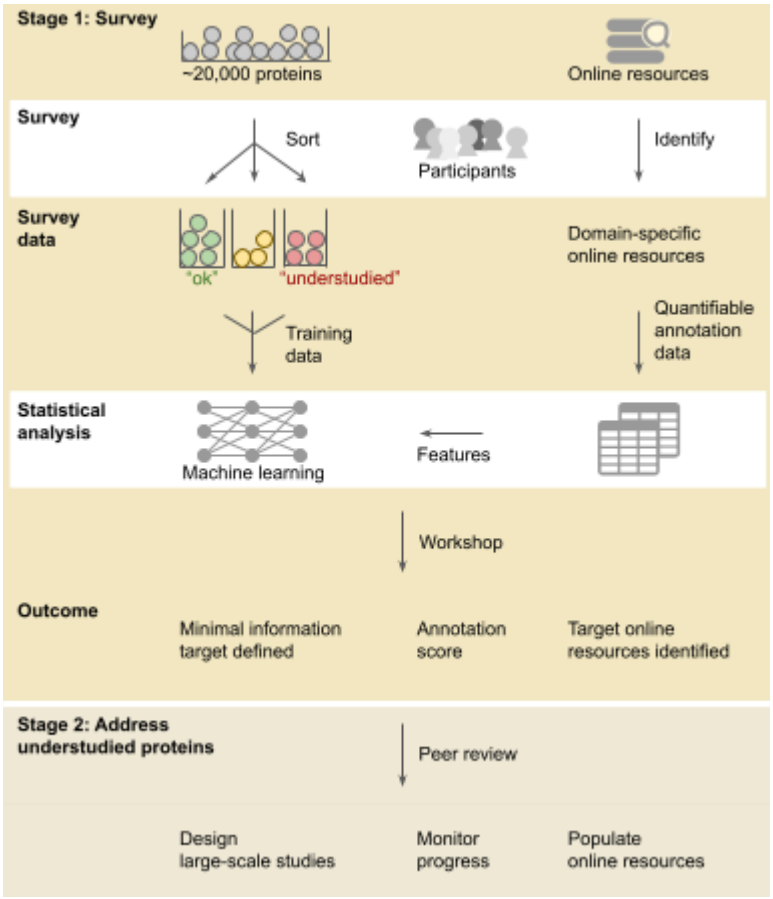


**Figure 1. First steps of an Understudied Proteins Initiative.** Most life sciences research focuses on a limited set of increasingly well-studied proteins, while thousands of human proteins remain understudied (the "streetlight effect"). To break this cycle, we believe that there is a need for the community to come together and identify the minimal information necessary to make understudied proteins "studiable". All molecular biologists are invited to help reveal understudied proteins through participating in a survey (https://understudiedproteins.org/survey). Sorting proteins into three categories (green, amber, red) will then allow us to define the magnitude of the challenge and build an automatic scoring tool to monitor the effect of measures to tackle understudied proteins. Participants will also reveal those resources that they use in their assessment of protein function and that should thus be used by the initiative to extract quantifiable information and ultimately to host newly generated information. Data generating labs will then converge at a workshop to conclude what minimal information is requested by molecular biologists and how to best obtain it (https://understudiedproteins.org/conference). The outcome of the survey and the workshop will be published to support grant applications for the subsequent activities, including data collection, at funding agencies worldwide.

## 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. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).
6. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
7. Dunham, I. Human genes: Time to follow the roads less traveled? *PLoS Biol.* **16**, e3000034 (2018).
8. Stoeger, T. & Nunes Amaral, L. A. COVID-19 research risks ignoring important host genes due to pre-established research patterns. *Elife* **9**, (2020).
9. Hutchison, C. A., 3rd *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
10. 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).
11. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
12. Anton, B. P. *et al.* The COMBREX project: design, methodology, and initial results. *PLoS Biol.* **11**, e1001638 (2013).
13. Gerlt, J. A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950–9962 (2011).
14. Williamson, A. R. Creating a structural genomics consortium. *Nat. Struct. Biol.* **7 Suppl**, 953 (2000).
15. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).