Title

Data-driven design of a tool to promote the investigation of understudied genes

Authors

Reese AK Richardson[1,2], Heliodoro Tejedor Navarro[2,3], Luis A Nunes Amaral[2,3,4,5]*, Thomas Stoeger[2]*

[1]Interdisciplinary Biological Sciences, Northwestern University

[2]Department of Chemical and Biological Engineering, Northwestern University

[3]Northwestern Institute on Complex Systems, Northwestern University

[4]Department of Physics and Astronomy, Northwestern University

[5]Department of Molecular Biosciences, Northwestern University

*Correspondence to:

L.A.N.A. (amaral@northwestern.edu), T.S. (thomas.stoeger@northwestern.edu)

Abstract

Present-day gene research follows exploratory patterns established prior to the Human Genome Project. We test and quantify existing hypotheses on why understudied genes are not pursued further despite having been found as hits in genome-wide studies. We combine our insights to design a tool that allows scientists to engage with existing biases in gene selection and use them to their benefit.

Manuscript

Research into human genes concentrates on a subset of genes that were already frequently investigated prior to the completion of the Human Genome Project in 2003[1-5]. This concentration stems from historically acquired research patterns rather than present-day experimental possibilities[6,7]. For most human diseases, these patterns lead to little correlation between the volume of literature published on individual genes and the strength of supporting evidence from genome-wide approaches[8-13].

Although high-throughput -omics technologies can provide insights on numerous genes across the genome at a time and thus offer the promise to counter historically acquired research patterns[14-17], this discrepancy has persisted[9,18-22] even as the popularity of -omics technologies has risen[5,23,24]. We therefore sought to use bibliometric data to delineate where and why understudied human protein-coding genes are abandoned as research targets after an -omics experiment has been performed.

We considered 651 genome-wide association studies (GWAS, from studies indexed by the NHGRI-EBI GWAS catalog[25]), 461 studies using affinity capture mass spectrometry (Aff-MS, indexed by BioGRID[26]), 250 transcriptomic studies (indexed by the EBI Gene Expression Atlas[27]), and 21 genome-wide screens using CRISPR (indexed by BioGRID Open Repository of CRISPR Screens[26]) (see PRISMA diagrams in **Figures S1-S4**). We denote genes that are found as having statistically significant changes in expression or associations with a phenotype as 'hit' genes. Most studies had hit genes that had appeared in a similar number of articles as any randomly chosen human protein-coding gene. In contrast, hit genes that are highlighted in the title or abstract are strongly over-represented among the 20% highest-studied genes in all biomedical literature (**Figure 1A**). The reception of these -omics studies in later scientific literature either reproduced this initial selection of highly studied genes or slightly mitigated it. These trends are independent of significance threshold (**Figure S5**) and whether we considered the current scientific literature or literature published before 2003, before any of these articles had been published (**Figure S6**). Jointly, the above findings demonstrate that understudied genes become abandoned between the completion of -omics experiments and the reporting of results, rather than being abandoned by later research.

We performed a literature review to identify factors that may affect which hit genes are highlighted and which are abandoned (**Table S1**). While all factors could plausibly affect gene selection within individual domains of biomedical research, we used the -omics data described above to measure how much these factors align with the selective highlighting of hit genes in the title or abstract of GWAS, Aff-MS, transcriptomics, and CRISPR studies (**Table 1** and **Table S2**). We identified 44 factors that relate to genes and found 39 (17 out of 22 binary factors and 22 out of 22 continuous factors) associated with selection in at least one assay type at $p < 0.001$. Factors are heavily intercorrelated (**Figure S7** and **S8**).

A valid concern of biomedical researchers is that focusing on less-investigated genes will yield articles with lower citation impact[17], as observed around the turn of the millenium[28]. However, we found the opposite for present-day research on human genes; in biomedical literature at-large, articles focusing on less-investigated genes accumulate more citations, an effect that has held consistently since 2001 (**Figure 1B and 1C).** Across multiple disease-related fields, we similarly find that when new genes are introduced to the field's literature, articles that introduced previously lowly-studied genes received a higher number of citations than papers that introduced previously highly-studied genes (**Figure S9**). Articles that introduce new genes to a field's literature also tend to receive more citations than articles that publish on genes that have already been introduced to that field's literature (**Figure S10**). Notably, our findings resemble those of reports on scientific novelty outside of biomedical research[29].

Our literature review also identified further factors that we could not test more directly because of absent access to fitting data. These are: experts' tendency to deepen their expertise[3], a perceived lack of accuracy of -omics studies[17,30], -omics serving research purposes beyond target gene identification[31], the absence of good protocols for mass spectrometry[32], the peer-review system[3,19,33,34], the electronic distribution and reading of research articles[35], rates of reproducibility[17], career prospects of investigators[7,33], and the human tendency to fall back to simplifying heuristics when making decisions under conditions with uncertainty[36].

Our literature review recovered several tools and resources to promote research of understudied genes[37-44] or to obtain more information about hit genes[7,45-49]. However, we noted an absence of tools that enable scientists to actively engage with biases affecting gene selection. We thus combined all the above data-driven insights to create "Find my Understudied Genes" (FMUG). This tool makes selection bias explicit, while acknowledging that different laboratories vary in their techniques and capabilities for follow-up research. Rather than telling scientists about the existence of biases, FMUG prompts scientists to make informed decisions that use awareness of these biases to identify and potentially tackle important gaps in knowledge that they are well-suited to address.

FMUG takes a list of genes from the user (ostensibly a hit list from a high-throughput -omics experiment) and provides information allowing the user to select genes for further study.

Users can employ filters that reflect the factors identified in our literature review and supported by our analysis. For instance, a user may need to decide whether genes that are specific to primates should be considered for further research or not, or whether there should be robust evidence that a gene is protein-coding. Some of these filters are context aware. For instance, a user may select genes that have already been studied in the general biomedical literature but not yet within the literature of their disease of interest.

To provide real-time feedback, users are, in parallel, presented the number of articles about genes in their initial input list and the number of articles about genes that passed their filters. Users can then export their filtered list of genes. In the interest of researcher privacy, FMUG keeps all information local to the user's machine. Usage of FMUG is summarized in **Figure 2** and demonstrated in **Movie S1**. FMUG is developed in Flutter and is available for download at fmug.amaral.northwestern.edu as a MacOS/Windows app. For the development of custom software and analytical code, we provide the data underlying FMUG at github.com/amarallab/fmug_analysis.

Efforts to address the gaps in detailed knowledge about most genes have crystallized as initiatives promoting the investigation of understudied sets of genes[16,50-54], an approach to gene scholarship recently termed 'unknomics'[42]. We hope that enabling scientists to consciously engage with bias in research target selection will enable more biomedical researchers to participate in unknomics, to the benefit of their own research impact and towards the advancement of our collective understanding of the entire human genome. These hopes are

grounded in experience; we used an early prototype of FMUG to identify understudied genes associated with aging. This led us to articulate a specific hypothesis on aging that we subsequently found supported through an analysis which we recently published in Nature Aging[55] **(Figure S11).**

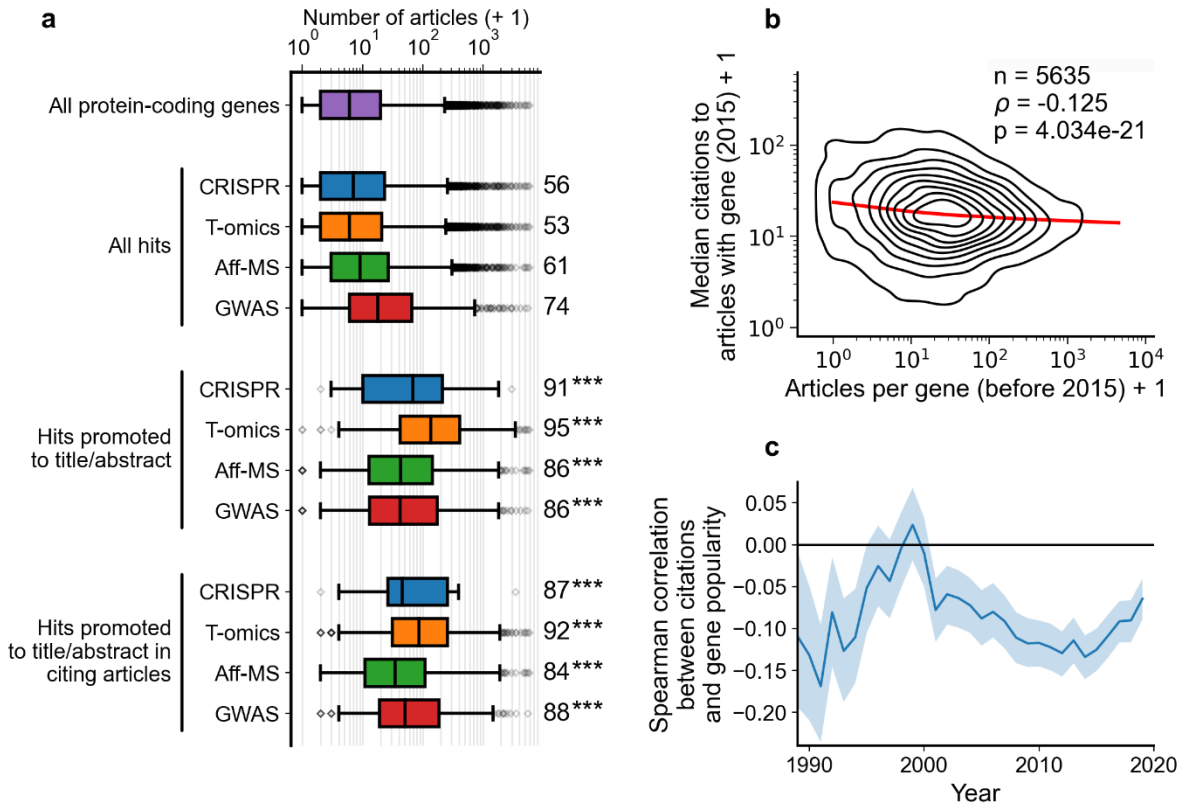

Figure 1: Prior popularity of genes positively correlates with selection by authors but negatively correlates with accrued citations. **a,** Bibliometric data reveals that understudied genes are frequently hits in -omics experiments but are not typically highlighted in the title/abstract of reporting articles. We considered articles reporting on genome-wide CRISPR screens (CRISPR, n=21 articles), transcriptomics (T-omics, n=250 articles), affinity capture – mass spectrometry (Aff-MS, n=461 articles), and GWAS (n=651 articles). Numbers to the right of each box plot indicate the percentile (in terms of number of articles about that gene) of all genes exceeded by the median gene in each box plot. *** denotes $p < 0.001$ by two-sided Mann-Whitney U test, comparing genes highlighted in title/abstract to genes present in hit lists. **b,** articles focusing on less popular genes tend to accrue more citations. Density plot shows correlation between articles per gene before 2015 and median citations to articles published in 2015. Contours correspond to deciles in density. Solid red line shows locally weighted scatterplot smoothing (LOWESS) regression. ρ is Spearman rank correlation and p the significance values of the Spearman rank correlation as described by Kendall and Stuart[56]. **c,** Spearman correlation of previous gene popularity (i.e. number of articles) to median citations per year since 1990. Solid blue line indicates nominal Spearman correlation, shaded region indicates bootstrapped 95% confidence interval (n=1,000). Only articles with a single gene in the title/abstract are considered, excluding the 30.4% of gene-focused studies which feature more than one gene in

the title/abstract.

Table 1: Association between factors with binary (True/False) identities and highlighting hits in title/abstract of reporting articles. Values represent the odds ratio between hits in the collected articles and hits mentioned in the title or abstract of collected articles (e.g. hits with a compound known to affect gene activity are 5.114 times as likely to be mentioned in the title/abstract in an article using transcriptomics). Collected articles are described in **Figure 1A** and **Figures S5** and **S6.** 95% confidence interval of odds ratio is shown in parentheses. * = p < 0.05, ** = p < 0.01, and *** = p < 0.001 by two-sided Fisher exact test. Continuous factors are evaluated in **Table S2**.

Factor	**Aff-MS**	**CRISPR**	**GWAS**	**Transcriptomics**
Antibody approved (immunofluorescence)	2.177*** (1.662, 2.852)	1.214 (0.513, 2.874)	0.849 (0.667, 1.082)	2.44*** (1.808, 3.295)
Antibody approved (immunohistochemistry)	1.554*** (1.256, 1.924)	1.791 (0.757, 4.241)	1.184 (0.929, 1.508)	1.979*** (1.519, 2.578)
Antibody available	2.656** (1.363, 5.176)	1.171 (0.277, 4.954)	1.057 (0.702, 1.591)	1.982** (1.211, 3.244)
Compound known to affect gene activity	2.400*** (2.009, 2.867)	2.617* (1.228, 5.574)	1.764*** (1.396, 2.23)	5.114*** (4.014, 6.516)
Defined HGNC nomenclature†	-	-	-	-
Druggable	1.438*** (1.170, 1.767)	2.261 (1.034, 4.946)	2.132*** (1.682, 2.701)	3.37*** (2.663, 4.266)
Evidence of protein's existence (HPA)	2.237*** (1.754, 2.851)	1.638 (0.692, 3.878)	1.46** (1.139, 1.871)	2.54*** (1.913, 3.371)
Evidence of protein's existence (neXtProt)	1.94 (0.792, 4.750)	1.198 (0.162, 8.846)	1.256 (0.648, 2.432)	5.751*** (1.841, 17.97)
Evidence of protein's existence (UniProt)	4.554*** (2.148, 9.656)	3.417 (0.463, 25.207)	1.735* (1.114, 2.701)	8.25*** (3.671, 18.542)
Has associated mouse phenotype	1.376** (1.134, 1.670)	1.646 (0.719, 3.764)	1.242 (0.961, 1.606)	1.813*** (1.401, 2.346)
Homolog in C. elegans	0.91 (0.740, 1.119)	1.868 (0.838, 4.165)	0.773 (0.561, 1.065)	0.923 (0.670, 1.273)
Homolog in D. melanogaster	1.031 (0.856, 1.242)	2.033 (0.951, 4.349)	0.722* (0.540, 0.966)	0.842 (0.627, 1.130)
Homolog in mouse	1.415 (0.923, 2.17)	inf (NA, inf)	0.845 (0.547, 1.303)	3.062*** (1.673, 5.604)

Homolog in rat	1.121 (0.805, 1.562)	inf (NA, inf)	0.978 (0.661, 1.446)	2.643*** (1.593, 4.384)
Homolog in yeast	1.026 (0.782, 1.347)	0.361 (0.049, 2.664)	0.922 (0.539, 1.577)	1.28 (0.839, 1.952)
Homolog in zebrafish	1.47*** (1.172, 1.842)	2.668 (0.922, 7.721)	1.089 (0.849, 1.395)	1.528** (1.174, 1.989)
Loss-of-function intolerant	2.358*** (1.960, 2.837)	2.671* (1.185, 6.023)	1.143 (0.870, 1.502)	2.75*** (2.137, 3.539)
Membrane protein	0.518*** (0.393, 0.683)	0.715 (0.27, 1.89)	1.487** (1.163, 1.901)	0.635** (0.473, 0.853)
Mendelian inheritance	1.581*** (1.314, 1.901)	2.118 (0.981, 4.569)	2.098*** (1.659, 2.652)	2.383*** (1.874, 3.031)
Plasmid available	2.816*** (2.316, 3.423)	4.312*** (1.739, 10.692)	1.415** (1.124, 1.782)	7.563*** (5.508, 10.386)
Previously patented	1.437*** (1.202, 1.718)	1.049 (0.492, 2.233)	1.613*** (1.270, 2.048)	2.689*** (2.085, 3.467)
Protein coding†	-	-	-	-
Specific to primates	0.689 (0.303, 1.569)	0.0 (0.0, NA)	0.737 (0.284, 1.913)	0.267* (0.085, 0.834)
Understudied according to IDG	0.162* (0.023, 1.164)	2.739 (0.369, 20.324)	0.925 (0.345, 2.481)	0.000* (0.0, NA)

† for consistency between studies, hits were restricted to protein-coding genes. Matching between gene identifiers required that each gene have a defined HUGO symbol. Thus, these binary factors could not be tested.

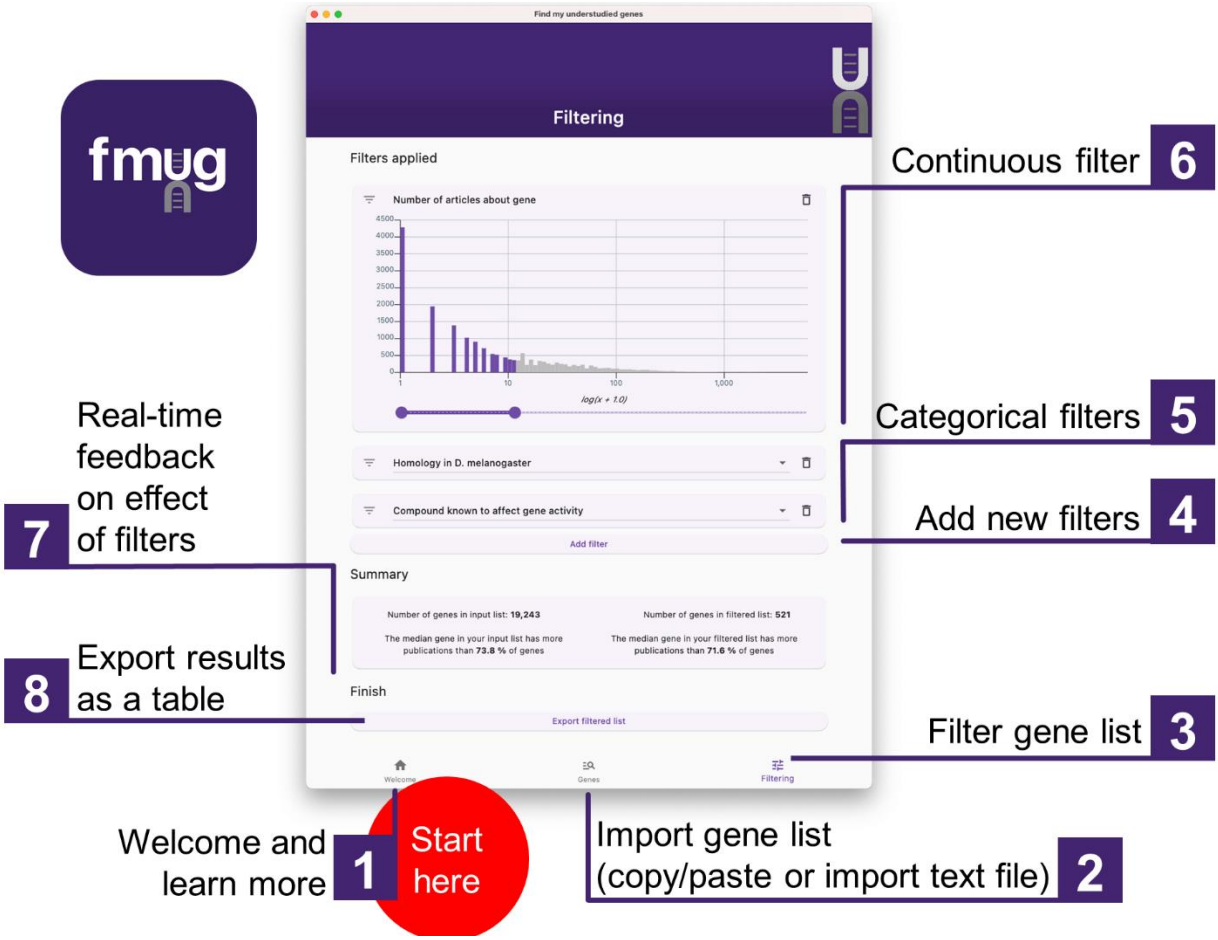

Figure 2: Diagram describing use of FMUG.

Data Availability

All underlying data for figures are available at github.com/amarallab/fmug_analysis.

Code Availability

Code for analysis is available at github.com/amarallab/fmug_analysis. Code for FMUG is available at github.com/amarallab/fmug.

Acknowledgements

We thank Xiaojing Sui for testing FMUG and Northwestern Information Technology for technical assistance. RR was supported in part by the National Institutes of Health Training Grant (T32GM008449) through Northwestern University's Biotechnology Training Program. LANA was supported by NSF 1956338, NIH U19AI135964 and Simons Foundation DMS-1764421. TS was supported by NIH K99AG068544. We thank Alexander Misharin, Richard Morimoto, and Scott

Budinger for feedback on an early prototype of FMUG which we used as part of our shared research into the biology of aging.

References

1 Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79-81 (2003). https://doi.org:10.1016/S0168-9525(02)00014-8

2 Su, A. I. & Hogenesch, J. B. Power-law-like distributions in biomedical publications and research funding. *Genome Biol.* **8**, 404 (2007). https://doi.org:10.1186/gb-2007-8-4-404

3 Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163-165 (2011). https://doi.org:10.1038/470163a

4 Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics* **29**, 476-482 (2013). https://doi.org:10.1093/bioinformatics/bts727

5 Stoeger, T. & Nunes Amaral, L. A. The characteristics of early-stage research into human genes are substantially different from subsequent research. *PLoS Biol* **20**, e3001520 (2022). https://doi.org:10.1371/journal.pbio.3001520

6 Grueneberg, D. A. *et al.* Kinase requirements in human cells: I. Comparing kinase requirements across various cell types. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16472-16477 (2008). https://doi.org:10.1073/pnas.0808019105

7 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). https://doi.org:10.1371/journal.pbio.2006643

8 Riba, M. *et al.* Revealing the acute asthma ignorome: characterization and validation of uninvestigated gene networks. *Sci Rep* **6**, 24647 (2016). https://doi.org:10.1038/srep24647

9 Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci Rep* **8**, 1362 (2018). https://doi.org:10.1038/s41598-018-19333-x

10 Stoeger, T. & Nunes Amaral, L. A. COVID-19 research risks ignoring important host genes due to pre-established research patterns. *Elife* **9** (2020). https://doi.org:10.7554/eLife.61981

11 Zhang, D. *et al.* Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science advances* **6**, eaay8299 (2020).

12 Border, R. *et al.* No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry* **176**, 376-387 (2019).

13 Byrne, J. A. *et al.* Protection of the human gene research literature from contract cheating organizations known as research paper mills. *Nucleic Acids Research* **50**, 12058-12070 (2022). https://doi.org:10.1093/nar/gkac1139

14 Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S. & Institute, U. S. N. H. G. R. A vision for the future of genomics research. *Nature* **422**, 835-847 (2003). https://doi.org:10.1038/nature01626

15 Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45-57 (2019). https://doi.org:10.1016/j.cell.2019.02.003
16 Lloyd, K. C. K. *et al.* The Deep Genome Project. *Genome Biol* **21**, 18 (2020). https://doi.org:10.1186/s13059-020-1931-9
17 Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods* **19**, 774-779 (2022). https://doi.org:10.1038/s41592-022-01454-x
18 Rodriguez-Esteban, R. & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics* **10**, 59 (2017). https://doi.org:10.1186/s12920-017-0293-y
19 Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* **17**, 317-332 (2018). https://doi.org:10.1038/nrd.2018.14
20 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). https://doi.org:10.1002/pmic.201800093
21 Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol* **9**, 180241 (2019). https://doi.org:10.1098/rsob.180241
22 Ratti, E. & Stoeger, T. (2021).
23 Pena-Castillo, L. & Hughes, T. R. Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**, 7-14 (2007). https://doi.org:10.1534/genetics.107.074468
24 Ellens, K. W. *et al.* Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res* **45**, 11495-11514 (2017). https://doi.org:10.1093/nar/gkx937
25 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019). https://doi.org:10.1093/nar/gky1120
26 Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187-200 (2021).
27 Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research* **46**, D246-D251 (2018).
28 Pfeiffer, T. & Hoffmann, R. Temporal patterns of genes in scientific publications. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12052-12056 (2007). https://doi.org:10.1073/pnas.0701315104
29 Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical combinations and scientific impact. *Science* **342**, 468-472 (2013).
30 Brown, L. A. & Peirson, S. N. Improving Reproducibility and Candidate Selection in Transcriptomics Using Meta-analysis. *Journal of Experimental Neuroscience* **12**, 1179069518756296 (2018). https://doi.org:10.1177/1179069518756296
31 Donohue, C. & Love, A. Perspectives on the Human Genome Project and genomics. *Minnesota Studies in the Philosophy of Science* **23** (in press).
32 Cesar-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478-487 (2015). https://doi.org:10.1016/j.cell.2015.07.022

33 Alberts, B., Kirschner, M. W., Tilghman, S. & Varmus, H. Rescuing US biomedical research from its systemic flaws. *Proc Natl Acad Sci U S A* **111**, 5773-5777 (2014). https://doi.org:10.1073/pnas.1404402111
34 Teplitskiy, M., Peng, H., Blasco, A. & Lakhani, K. R. Is novel research worth doing? Evidence from peer review at 49 journals. *Proc Natl Acad Sci U S A* **119**, e2118046119 (2022). https://doi.org:10.1073/pnas.2118046119
35 Evans, J. A. Electronic publication and the narrowing of science and scholarship. *Science* **321**, 395-399 (2008). https://doi.org:10.1126/science.1150473
36 Gilovich, T., Griffin, D. W. & Kahneman, D. *Heuristics and biases : the psychology of intuitive judgment*. (Cambridge University Press, 2002).
37 Duek, P., Gateau, A., Bairoch, A. & Lane, L. Exploring the Uncharacterized Human Proteome Using neXtProt. *J Proteome Res* **17**, 4211-4226 (2018). https://doi.org:10.1021/acs.jproteome.8b00537
38 Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential gene expression. *Proc Natl Acad Sci U S A* **116**, 6491-6500 (2019). https://doi.org:10.1073/pnas.1802973116
39 Perdigao, N. & Rosa, A. Dark Proteome Database: Studies on Dark Proteins. *High Throughput* **8** (2019). https://doi.org:10.3390/ht8020008
40 Essegian, D., Khurana, R., Stathias, V. & Schurer, S. C. The Clinical Kinase Index: A Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer. *Cell Rep Med* **1**, 100128 (2020). https://doi.org:10.1016/j.xcrm.2020.100128
41 Sheils, T. K. *et al.* TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res* **49**, D1334-D1346 (2021). https://doi.org:10.1093/nar/gkaa993
42 Rocha, J. *et al.* Functional unknomics: closing the knowledge gap to accelerate biomedical research. *bioRxiv* (2022).
43 Higgins, D. P., Weisman, C. M., Lui, D. S., D'Agostino, F. A. & Walker, A. K. Defining characteristics and conservation of poorly annotated genes in Caenorhabditis elegans using WormCat 2.0. *Genetics* **221** (2022). https://doi.org:10.1093/genetics/iyac085
44 Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nat Genet* **53**, 638-649 (2021). https://doi.org:10.1038/s41588-021-00840-z
45 Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* **14**, 656-664 (1998). https://doi.org:10.1093/bioinformatics/14.8.656
46 Tan, J. *et al.* ADAGE signature analysis: differential expression analysis with data-defined gene sets. *Bmc Bioinformatics* **18**, 512 (2017). https://doi.org:10.1186/s12859-017-1905-4
47 Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification of protein functions. *Nat Biotechnol* **37**, 1361-1371 (2019). https://doi.org:10.1038/s41587-019-0298-5
48 Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021). https://doi.org:10.1016/j.xinn.2021.100141

49 Jiang, J. *et al.* Systematic illumination of druggable genes in cancer genomes. *Cell Rep* **38**, 110400 (2022). https://doi.org:10.1016/j.celrep.2022.110400
50 Gerlt, J. A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950-9962 (2011). https://doi.org:10.1021/bi201312u
51 Carter, A. J. *et al.* Target 2035: probing the human proteome. *Drug Discov Today* **24**, 2111-2115 (2019). https://doi.org:10.1016/j.drudis.2019.06.020
52 Rodgers, G. *et al.* Glimmers in illuminating the druggable genome. *Nat Rev Drug Discov* **17**, 301-302 (2018). https://doi.org:10.1038/nrd.2017.252
53 Kustatscher, G. *et al.* An open invitation to the Understudied Proteins Initiative. *Nat Biotechnol* **40**, 815-817 (2022). https://doi.org:10.1038/s41587-022-01316-z
54 *EUbOPEN*, <https://www.eubopen.org/> (
55 Stoeger, T. *et al.* Aging is associated with a systemic length-associated transcriptome imbalance. *Nature Aging* **2**, 1191-1206 (2022). https://doi.org:10.1038/s43587-022-00317-6
56 Kendall, M. G. & Stuart, A. Inference and Relationship *The Advanced Theory of Statistics* **2** (1973).