Title

Meta-Research: understudied genes are lost in a leaky pipeline between genome-wide assays and reporting of results

Authors

Reese AK Richardson[1,2], Heliodoro Tejedor Navarro[2,3], Luis A Nunes Amaral[2,3,4,5]*, Thomas Stoeger[2,6,7]*

[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

[6]The Potocsnak Longevity Institute, Northwestern University

[7]Simpson Querrey Lung Institute for Translational Science, Northwestern University

*Correspondence to:

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

Abstract

Present-day publications on human genes primarily feature genes that already appeared in many publications prior to completion of the Human Genome Project in 2003. These patterns persist despite the subsequent adoption of high-throughput technologies, which routinely identify novel genes associated with biological processes and disease. Although several hypotheses for bias in the selection of genes as research targets have been proposed, their explanatory powers have not yet been compared. Our analysis suggests that understudied genes are systematically abandoned in favor of better-studied genes between the completion of -omics experiments and the reporting of results. Understudied genes are similarly abandoned by studies that cite these -omics experiments. Conversely, we find that publications on understudied genes may even accrue a greater number of citations. Among 45 biological and experimental factors previously proposed to affect which genes are being studied, we find

that 35 are significantly associated with the choice of hit genes presented in titles and abstracts of -omics studies. To promote the investigation of understudied genes we condense our insights into a tool, *find my understudied genes* (FMUG), that allows scientists to engage with potential bias during the selection of hits. We demonstrate the utility of FMUG through the identification of genes that remain understudied in vertebrate aging. FMUG is developed in Flutter and is available for download at fmug.amaral.northwestern.edu as a MacOS/Windows app.

Introduction

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]. For instance, we found that 44% of the genes identified as promising Alzheimer's disease targets by the U.S. National Institutes of Health (NIH) Accelerating Medicine Partnership for Alzheimer's Disease (AMP-AD) initiative have never appeared in the title or abstract of any publication on Alzheimer's disease[13]. Furthermore, when comparing gene-disease pairs, there is no correlation between the ranks of support by transcriptomics and occurrence in annotation databases[9].

Although -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 following -omics experiments. In the absence of any prior quantitative testing of existing hypotheses, it remains unclear whether policies to promote the exploration of a greater set of disease-related genes should focus on how experiments are conducted, how results are reported, or how these results are subsequently received by other scientists.

Data

We considered 450 genome-wide association studies (GWAS, from studies indexed by the NHGRI-EBI GWAS catalog[25]), 296 studies using affinity capture mass spectrometry (Aff-MS, indexed by BioGRID[26]), 148 transcriptomic studies (indexed by the EBI Gene Expression Atlas, EBI-GXA[27]), and 15 genome-wide screens using CRISPR (indexed by BioGRID Open Repository of CRISPR Screens, BioGRID ORCS[26]) (see PRISMA diagrams in **Figures S1-S4**). We denote genes that are found to have statistically significant changes in expression or associations with a phenotype as 'hit' genes.

As a surrogate for a given gene having been investigated closer, we consider whether it was reported in the title or abstract of a research article. We determined which genes were

mentioned in the title or abstract of articles using annotations from gene2pubmed[28] and PubTator[29]. We used NIH iCite v32 for citations[30]. For determining which gene properties were associated with selection as research targets, we synthesized quantitative measures from a variety of authoritative sources (see **Methods**).

Results

Understudied genes are abandoned at synthesis/writing stage

We sought to identify at which point in the scientific process understudied genes are ignored as research targets in investigations using -omics experiments (**Figure 1A**). To receive scholarly attention, a gene must travel through a pipeline from biological reality to experimental results to write-up of those results. These results must be extended by subsequent research by other scholars. Understudied genes do not progress all the way through the pipeline, but it is unclear where this leak primarily occurs. The first possibility is that seemingly understudied genes are, in fact, not understudied as they would rarely be identified through experiments. Prior studies have, however, shown that understudied genes are frequent hits in high-throughput experiments[8,9,31], suggesting that this is not the case. The second possibility is that understudied genes are frequently found as hits in high-throughput experiments but are not investigated further by the authors. The final possibility is that subsequent studies do not continue work on understudied genes revealed by the initial study.

Evaluating the first possibility, we found that understudied genes were frequently found as hits in high-throughput experiments. (**Figure 1B**, **Figure S5**, and **Figure S6**). This demonstrates, in line with earlier studies[8-13], that the lack of publications on some genes is not explained by underlying biological experimental evidence.

Evaluating the second possibility, we found that 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 1B**). These trends are independent of significance threshold (**Figure S5**) and (except for CRISPR screens) whether we considered the current scientific literature or literature published before 2003, before any of these articles had been published (**Figure S6**).

Understudied genes are least frequently elevated to the title/abstract in transcriptomics experiments and most frequently elevated to the title/abstract in CRISPR screens. GWAS studies tend to return better-studied genes as hits; the median hit gene in GWAS studies was more popular than 75% of genes. Hit genes promoted to the title/abstract in GWAS studies had a median popularity greater 85% of all protein coding genes. This may explain the prior observation that the total number of articles on individual genes partially correlates with the total number of occurrences as a hit in GWAS studies[32].

Evaluating the final possibility, we found that the reception of -omics studies in later scientific literature either reproduced authors' initial selection of highly studied genes or slightly mitigated it. Jointly, the above findings reinforce that understudied genes become abandoned

between the completion of -omics experiments and the reporting of results, rather than being abandoned by later research.

Subsequent reception by other scientists does not penalize studies on understudied genes

The abandonment of understudied genes could be driven by the valid concern of biomedical researchers that focusing on less-investigated genes will yield articles with lower impact[17], as observed around the turn of the millenium[33]. If this were the case, preemptively avoiding understudied hits would be the rational decision for authors of -omics studies.

We thus decided to complement our preceding analysis by an analysis explicitly focused on citation impact. Notably, we found that the concern of publications on understudied genes receiving fewer citations does not hold 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 2).** Important to human health, this also holds when only considering disease-related fields (**Figure S7**).

To rule out that these macroscopic observations stem from us having aggregated over different diseases, we separately analyzed 602 disease-related MeSH terms. We found 29 MeSH terms with a statistically significant Spearman correlation using Benjamini-Hochberg FDR < 0.01 (**Table S4**), of which 27 showed a negative association and only 2 a positive association. This result suggests that is may actually be rational for most scientists to pursue studies focusing on understudied genes, although most scientists specializing in a disease may also not receive more citations when focusing on understudied genes.

Returning to our observation that understudied hits from high-throughput assays are not promoted to the title and abstract of the resulting publication, we next tested if different experimental approaches demonstrated distinct associations between gene popularity and citations (**Figure S8**). Among 264 technique-related MeSH terms tested, there were 20 MeSH terms with a statistically significant Spearman correlation using Benjamini-Hochberg FDR < 0.01 (**Table S5**), of which 16 showed a negative association and only 4 a positive association. Notably, MeSH terms representing high-throughput techniques (e.g. D055106:Genome-Wide Association Study and D020869:Gene Expression Profiling) showed no significant association. This finding suggests that authors of high-throughput studies have little to gain or lose citation-wise by highlighting understudied genes.

To summarize, our investigations detail the previously described separation between “large-scale” and “small-scale” biological research[34-36]. Authors of high-throughput studies do not highlight understudied genes in the title or abstract of their publications, the sections of the publication most accessible to other scientists. While, overall, understudied genes (and high-throughput assays themselves[5]) correlate with increased citation impact, for high-throughput studies any potential gain in citations is either absent or too small to be significant. Thus, there may not be any incentive for authors of high-throughput studies to highlight understudied genes.

Identification of biological and experimental factors associated with selection of highlighted genes

To illuminate why understudied genes are abandoned between experimental results and the write-up of results, we performed a literature review to identify factors that have been proposed to limit studies of understudied genes (**Table S1**). These factors range from evolutionary factors (e.g., whether a gene only has homologs in primates), to chemical factors (e.g., gene length or hydrophobicity of protein product), to historical factors (e.g., whether a gene's sequence has previously been patented) to materialistic factors affecting experimental design (e.g., whether designed antibodies are robust for immunohistochemistry).

As any of these factors could plausibly affect gene selection within individual domains of biomedical research, we returned to the -omics data described above (**Figure 1**) and measured 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.

We identified 45 factors that relate to genes and found 35 (14 out of 23 binary factors and 21 out of 22 continuous factors) associated with selection in at least one assay type at $p < 0.001$ (**Figure 3, Table S2,** and **Table S3**). Across the four assay types, the most informative binary factor describes whether there is a plasmid available for a gene in the AddGene plasmid catalog. This might reflect that many different research groups produce reagents surrounding the genes that they actively study. The most informative continuous factor is the number of research articles about a gene, supporting the conclusion that gene popularity drives whether it is highlighted or not (**Figure 1**).

To better understand how all 45 factors are related, we performed a cluster analysis of the collected factors (**Figure 4, Figure S9,** and **Figure S10**). This clustering reflects the fact that many suggested factors influencing the abandonment of understudied genes are not independent. For instance, we find that the number of articles about a gene is heavily correlated with the number of annotations for that gene in all surveyed databases. In another case, gene length is heavily correlated with the number of GWAS annotations for a gene, as described before in terms of transcript length and single-nucleotide polymoprhisms[37].

Study limitations

Our study has several limitations. First, all analysis is subject to annotation errors in the various databases we employ. While these should be rare and not affect our overall findings, they may affect users who are interested in genes with discordant annotations. Second, we focus only on human genes. Different patterns of selection may exist for research on genes in other organisms. Third, 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[17,38], -omics serving research purposes beyond target gene identification[22], the absence of good protocols for mass

spectrometry[39], the electronic distribution and reading of research articles[40], rates of reproducibility[17], career prospects of investigators[7,41], and the human tendency to fall back to simplifying heuristics when making decisions under conditions with uncertainty[42]. Fourth, we cannot resolve further which specific step between the conduct of an experiment and the writing of a research article leads to the abandonment of hit genes. Finally, we interpret the results of high-throughput experiments based on their representation in the NHGRI-EBI GWAS, BioGRID, EBI-GXA and BioGRID ORCS databases. The authors of the original studies may have processed their data differently, obtaining different results.

As our present analysis is correlative, it also is tempting to propose controlled trials where published manuscripts on high-throughput studies randomly report hit genes in the abstract even if not investigated further by the authors.

Discussion

Efforts to address the gaps in detailed knowledge about most genes have crystallized as initiatives promoting the investigation of understudied sets of genes[16,43-47], an approach to gene scholarship recently termed 'unknomics'[48]. We believe that enabling scientists to consciously engage with bias in research target selection will enable more biomedical researchers to participate in unknomics, to the potential benefit of their own research impact and towards the advancement of our collective understanding of the entire human genome.

To achieve this goal, we combined all the above insights to create a tool we denoted *find my understudied genes* (FMUG). Our literature review revealed several tools and resources aiming to promote research of understudied genes by publicizing understudied genes[49-56] or by providing information about hit genes[7,57-61]. However, we noted the absence of tools enabling scientists to actively engage with factors that align with gene selection. Although such factors are largely correlated when considering all genes (**Figure 4, Figure S9,** and **Figure S10**), some cluster together and the influence of specific factors could vary across laboratories. For instance, scientists could vary in their ability to perform proteomics, or ability to explore orthologous genes in *C. elegans*, or ability to leverage human population data, or perform standardized mouse assays.

Our 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 aims to prompt scientists to make bias-aware informed decisions to identify and potentially tackle important gaps in knowledge that they are well-suited to address. For this reason, we believe that FMUG will not be of value only to scientists engaging in high-throughput studies, but also by scientists wishing to mine existing datasets for hit genes that they would be well-positioned to investigate further.

FMUG takes a list of genes from the user (ostensibly a hit list from a high-throughput -omics experiment) and provides the kind of information that will allow a 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. The default information provided to users consists of factors that are representative of the identified clusters (**Figure 4**) and strongly associated with gene selection in high-throughput experiments (**Figure 3, Table S2,** and **Table S3**). In extended options, users can select any factor that demonstrated a significant association with the selection of genes. For instance, a user may need to decide whether loss-of-function intolerant genes 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 illustrated in **Figure 5A** 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.

To determine the practical usefulness of FMUG to scientists we used an early prototype of FMUG to identify understudied genes associated with aging. One of these genes was Splicing factor, proline- and glutamine-rich (*Sfpq*), which had not yet been investigated toward its role in biological aging. We found *Sfpq* to be transcriptionally downregulated during murine aging. Others had shown *Sfpq* to be required for the transcriptional elongation of long genes[62]. This led us to hypothesize that during vertebrate aging, the transcripts of long genes become downregulated in most tissues **(Figure 5B).** We found this hypothesis to be supported through a multi-species analysis which we published in December 2022 in Nature Aging[63], with another group publishing so in January 2023 in Nature Genetics[64], and a third group in iScience in March 2023[65].

Materials and Methods

Genes information

Homo sapiens gene information was downloaded from NCBI Gene on Aug 16, 2022 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/All_Data.gene_info.gz]. Only genes with an unambiguous mapping of Entrez ID to Ensembl ID were used (n = 36,035). Number of gene synonyms, protein-coding status, and official gene symbol were derived from this dataset. A gene symbol was considered undefined if the gene's entry for HGNC gene symbol was "-".

302 *Genes in title/abstract of primary research articles*
303
304 Homo sapiens gene information was downloaded from NCBI Gene on Aug 16, 2022
305 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz].
306 gene2pubmed was download from NCBI Gene on Aug 16, 2022
307 [ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz][28]. PubTator gene annotations were
308 downloaded from NIH-NLM on July 12, 2022
309 [https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/][28,66]. PubMed was downloaded on Dec
310 17, 2021 [https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/].
311
312 Only using PMIDs annotated as primary research articles, a human gene was considered as
313 mentioned in the title/abstract of the publication if gene was annotated as being in the
314 title/abstract by PubTator and the article appeared in gene2pubmed.
315
316 *CRISPR articles*
317
318 BioGRID ORCS[26] v1.1.6 was downloaded on April 25, 2022
319 [https://downloads.thebiogrid.org/BioGRID-ORCS/Release-Archive/BIOGRID-ORCS-1.1.6/]. Any
320 genome-wide CRISPR knockout screens in human with an associated PubMed ID in which hit
321 genes were mentioned in the title or abstract was considered (n = 15). 9,268 unique genes were
322 found as hits. Of these, 18 (0.19%) were elevated to titles/abstracts in the reporting articles and
323 19 (0.21%) were elevated to titles/abstracts in citing articles. A full list of PubMed IDs is
324 available in Supplementary File 2.
325
326 *Transcriptomics articles*
327
328 EBI-GXA[27] release 36 was downloaded on Sep 15, 2020
329 [https://web.archive.org/web/20201022184159/https://www.ebi.ac.uk/gxa/download]. This is
330 the most recent release of EBI-GXA available as a bulk download. Any transcriptomics
331 comparisons with an associated PubMed ID in which hit genes were mentioned in the title or
332 abstract was considered (n= 148). Analysis was restricted to protein-coding genes (some
333 screens featured non-protein-coding genes, but this was not common to all analyses). DE was
334 called at Benjamini-Hochberg FDR $q < 0.05$. 18,295 unique genes were found as hits. Of these,
335 161 (0.88%) were elevated to titles/abstracts in the reporting articles and 692 (3.78%) were
336 elevated to titles/abstracts in citing articles. A full list of PubMed IDs is available in
337 Supplementary File 2.
338
339 *Affinity capture – mass spectrometry articles*
340
341 BioGRID[26] v3.5.186 was downloaded on April 25, 2022
342 [https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.186/]. Any
343 interactions involving a human gene as the prey protein with an experimental evidence code of
344 'Affinity Capture-MS' labeled as 'High-Throughput' that had an associated PubMed ID in which
345 hit genes were mentioned in the title or abstract was considered (n= 296). Prey proteins in

346 these interactions were considered hits. 7,919 unique genes were found as hits. Of these, 311
347 (3.93%) were elevated to titles/abstracts in reporting articles and 407 (5.14%) were elevated to
348 titles/abstracts in citing articles. A full list of PubMed IDs is available in Supplementary File 2.
349
350 *GWAS articles*
351
352 The NHGRI-EBI GWAS catalog[25] (associations and studies) was download on Aug 17, 2022
353 [https://www.ebi.ac.uk/gwas/docs/file-downloads]. Any GWAS screens with an associated
354 PubMed ID in which hit genes were mentioned in the title or abstract was considered (n= 450).
355 Only SNPs occurring within a gene were considered hits. 1,043 unique genes were found as hits.
356 Of these, 413 (39.6%) were elevated to titles/abstracts in reporting articles and 319 (30.6%)
357 were elevated to titles/abstracts in citing articles. A full list of PubMed IDs is available in
358 Supplementary File 2.
359
360 *Citing articles*
361
362 NIH iCite v32 was downloaded on Aug 25, 2022[30]
363 [https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collectio
364 n_/4586573/32].
365
366 *Functional annotations*
367
368 Mapping of genes to Gene Ontology / Protein Interaction Database / WikiPathways / Reactome
369 / Kyoto Encyclopedia of Genes and Genomes / Human Phenotype Ontology / BioCarta
370 categories was derived from MSigDB v7.5 Entrez ID .gmt files, downloaded on Apr 12, 2022
371 [http://www.gsea-msigdb.org/gsea/downloads_archive.jsp].
372
373 *Between-species homology*
374
375 Homologene Build 68 was used to determine interspecies homology
376 [ftp.ncbi.nih.gov/pub/HomoloGene/build68/]. Human = taxid:9606, mouse = taxid:10090, rat =
377 taxid:10116, c. elegans = taxid:6239, d. melanogaster = taxid:7227, yeast = taxid:559292,
378 zebrafish = taxid:7955.
379
380 *Primate specificity*
381
382 Human genes were considered primate-specific if the only other members of their homology
383 group belonged to primate genomes. Primate taxonomy ids were downloaded from NCBI
384 Taxonomy on Sep 20, 2022
385 [https://www.ncbi.nlm.nih.gov/taxonomy/?term=txid9443[Subtree]].
386
387 *Number of publications in model organisms*
388

389 Gene information was downloaded from NCBI Gene on Aug 16, 2022
390 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/All_Data.gene_info.gz].
391
392 Only using PMIDs annotated as primary research articles, genes was considered as mentioned
393 in the title/abstract of the publication if gene was annotated as being in the title/abstract by
394 PubTator and the article appeared in gene2pubmed.
395
396 Genes in model organisms were mapped to human genes and the number of articles on those
397 mapping to human genes were counted. If a model organism's gene had homology to human
398 but no associated publications, the number of publications was resolved to zero. Otherwise,
399 counts were listed as NA.
400
401 *Mouse phenotype hits*
402
403 International Mouse Phenotyping Consortium data release 17.0 was downloaded on Aug 18,
404 2022 [https://www.mousephenotype.org/data/release]. Mouse genes were matched to human
405 genes with Homologene.
406
407 *Gene Expression Atlas (EBI-GXA)*
408
409 EBI-GXA release 36 was downloaded on Sep 15, 2020
410 [https://web.archive.org/web/20201022184159/https://www.ebi.ac.uk/gxa/download]. This is
411 the most recent release of EBI-GXA available as a bulk download. For probability of DE, only
412 RNA-seq comparisons were considered and DE was called at Benjamini-Hochberg $q < 0.05$.
413
414 *Global RNA expression*
415
416 RNA consensus tissue gene data from HPA release 21.1 was downloaded on Sep 20, 2022
417 [https://www.proteinatlas.org/about/download]. Global RNA expression was estimated by
418 taking the median expression (nTPM) across tissues for each gene and the proportion of tissues
419 with detectable (≥1 nTPM) expression for each gene.
420
421 *Expression in HeLa cells*
422
423 RNA cell line gene data from HPA release 21.1 was downloaded on Sep 20, 2022
424 [https://www.proteinatlas.org/about/download]. Expression is in nTPM.
425
426 *Previous patent activity*
427
428 Genes with patent activity were defined from Table S1 of Rosenfeld and Mason, 2013[67]. Genes
429 were mapped with their HGNC symbol. This analysis aligned sequences in patents to the human
430 genome to estimate patent coverage of human coding sequences. Although this does not
431 necessarily reflect whether the mapped genes were claimed directly by the patent holder, as

noted by others[68], this analysis remains the most comprehensive available for determining patent coverage of the human genome.

Druggability

Druggable genes were identified from Table S1 of Finan et al., 2017[69]. Genes were mapped with their Ensembl identifier.

Gene length

GenBank was downloaded in spring 2017 (genome version GRCh38.p10). Gene length is defined here as the span of the longest transcript on the chromosome. This aligns with the model of gene length used in Stoeger et al., 2018[7].

Solubility

SwissProt protein sequences and mapping tables to Entrez GeneIDs were downloaded from Uniprot in spring 2017. Protein GRAVY score (ignoring Pyrrolysine and Selenocysteine) was estimated with BioPython[70].

Loss of function intolerance

Data was obtained from Karczewski et al.[71]. pLI scores > 0.9 on main transcripts, as flagged by authors, were considered as highly loss-of-function intolerant as described by Lek et al.[72].

Number of GWAS hits

EBI GWAS catalog[25] (associations and studies) was download on Aug 17, 2022 [https://www.ebi.ac.uk/gwas/docs/file-downloads]. Loci were mapped to the nearest gene.

Status as understudied protein

The Illuminating the Druggable Genome understudied protein list was downloaded on Sep 20, 2022 [https://github.com/druggablegenome/IDGTargets/blob/master/IDG_TargetList_CurrentVersion.json].

Human Protein Atlas

HPA release 21.1 was downloaded on Sep 20, 2022 [https://www.proteinatlas.org/search]. Evidence for a protein's existence, as determined by NeXtProt, HPA, or UniProt was resolved as True if the respective evidence entry was annotated as "Evidence at protein level". Status as a membrane protein was determined by whether the 'Protein class' column contained the string

‘membrane protein’. Antibodies were considered available for each protein if the protein’s entry in the ‘Antibody’ column was not null.

Availability of plasmids

The AddGene plasmid catalog was downloaded on Aug 12, 2022 [https://www.addgene.org/browse/gene/gene-list-data/?_=1666368044314].

Availability of compounds

The catalog of gene targets was downloaded from ChEMBL on Sep 20, 2022 [https://www.ebi.ac.uk/chembl/g/#browse/targets]. UniProt IDs were converted to Entrez IDs to identify which human genes were affected by any compound.

Mendelian inheritance

Autosomal dominant [https://hpo.jax.org/app/browse/term/HP:0000006] and autosomal recessive [https://hpo.jax.org/app/browse/term/HP:0000007] inherited disease-gene associations were downloaded from the Human Phenotype Ontology on Sep 20, 2022. Genes were considered to have evidence of Mendelian inheritance if they appeared in these lists of associations.

Code

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

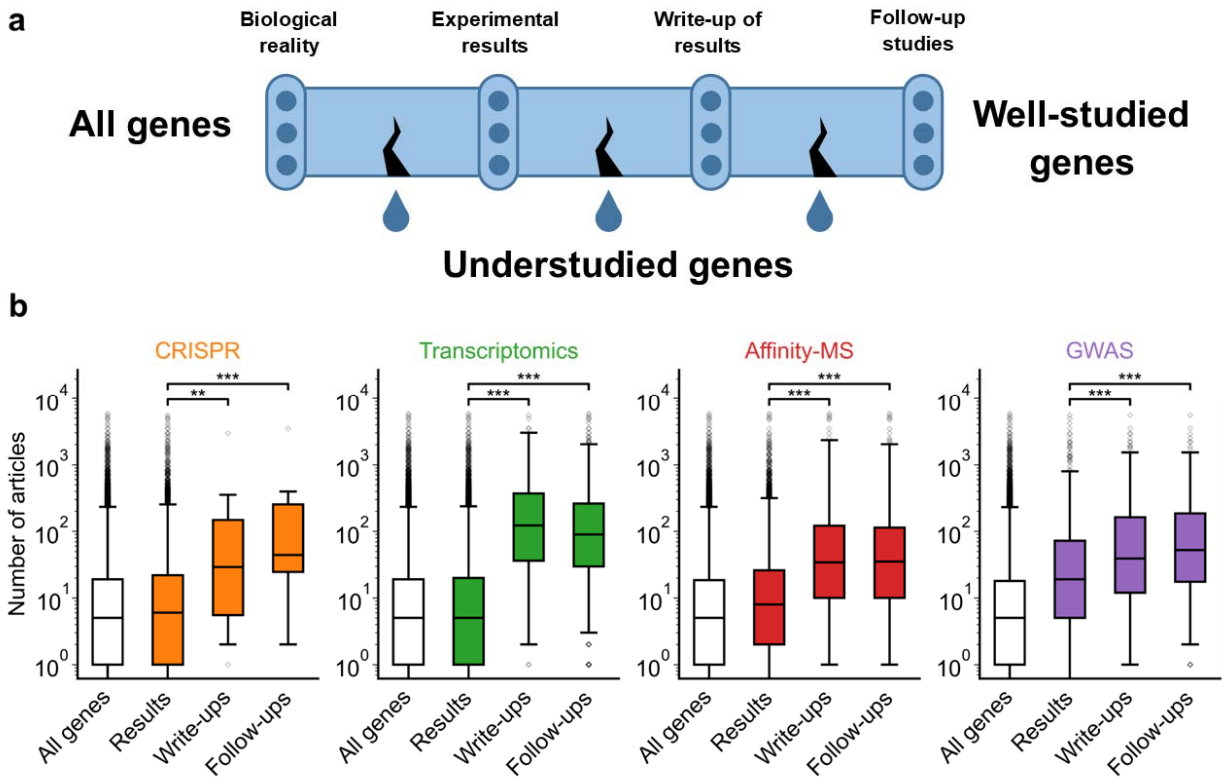

Figure 1: While understudied genes appear often as hits high-throughput -omics experiments, they are seldom highlighted by authors. a, Conceptual diagram depicting possible points of abandonment for understudied genes in studies using high-throughput -omics experiments. **b,** Bibliometric data reveals that understudied genes are frequently hits in -omics experiments but are not typically highlighted in the title/abstract of reporting articles, nor in the title/abstract or articles citing reporting articles. We considered articles reporting on genome-wide CRISPR screens (CRISPR, n=15 articles), transcriptomics (T-omics, n=148 articles), affinity capture – mass spectrometry (Aff-MS, n=296 articles), and GWAS (n=450 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 $p < 0.01$ and *** denotes $p < 0.001$ by two-sided Mann-Whitney U test, comparing genes highlighted in title/abstract to genes present in hit lists.

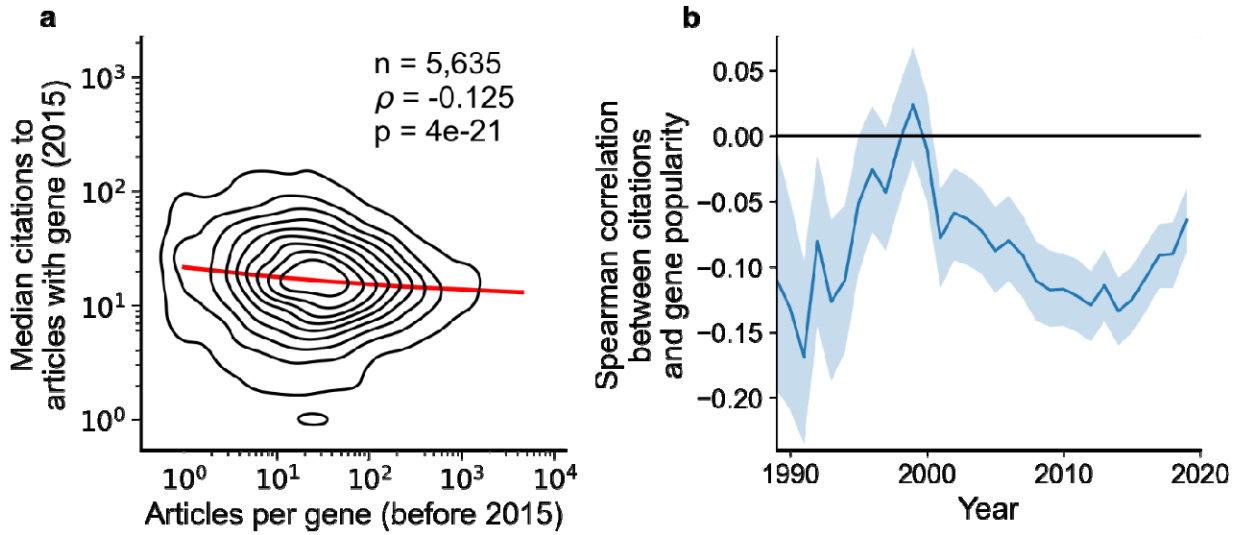

Figure 2: Articles focusing on less popular genes tend to accrue more citations. a, Density plot shows correlation between articles per gene before 2015 and median citations to articles published in 2015. Contours correspond to deciles in density. Solid red line shows locally weighted scatterplot smoothing (LOWESS) regression. ρ is Spearman rank correlation and p the significance values of the Spearman rank correlation as described by Kendall and Stuart[73]. **b,** 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.

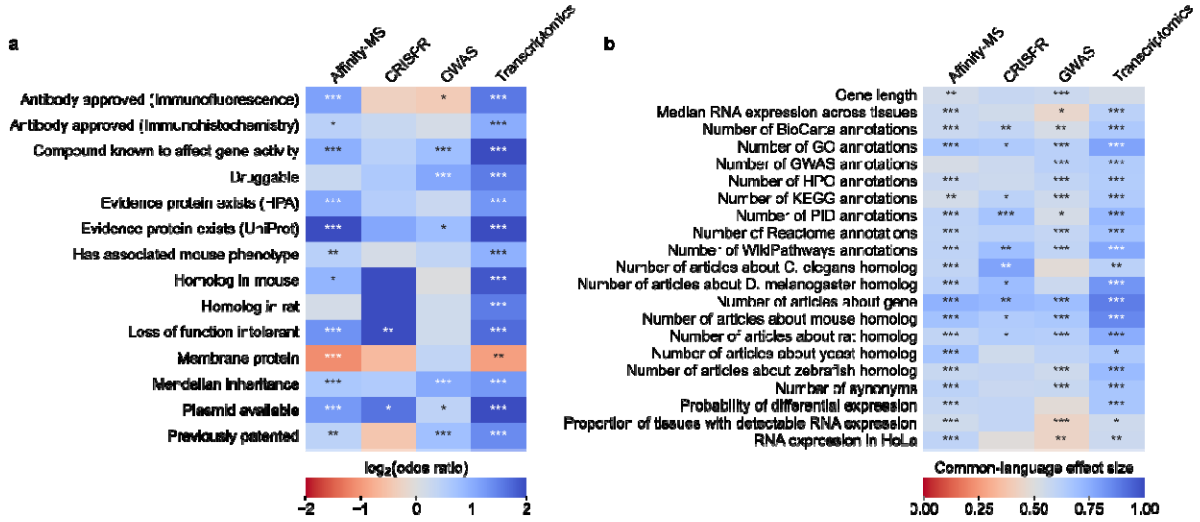

Figure 3: We evaluated which gene-related factors are associated with elevation to the title/abstract of an article featuring a high-throughput experiment. **a)** Association between factors with binary (True/False) identities and highlighting hits in title/abstract of reporting articles. Values represent 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 1B** 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. Results are shown numerically in **Table S2**. For consistency between studies, hits were restricted to protein-coding genes. Thus, status as a protein-coding gene could not be tested. †No genes without a defined HUGO symbol were found as hits in GWAS or transcriptomics studies. **b)** Association with factors with continuous identities and highlighting hits in title/abstract of reporting articles. Values represent F, the common-language effect size (equivalent to AUROC, where ~0.5 indicates little effect, >0.5 indicates positive effect and <0.5 indicates negative effect) of being mentioned in the titles/abstracts of the collected articles described in **Figure 1B** and **Figures S5** and **S6.** * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$ by two-sided Mann-Whitney U test. Results are shown numerically in **Table S3**.

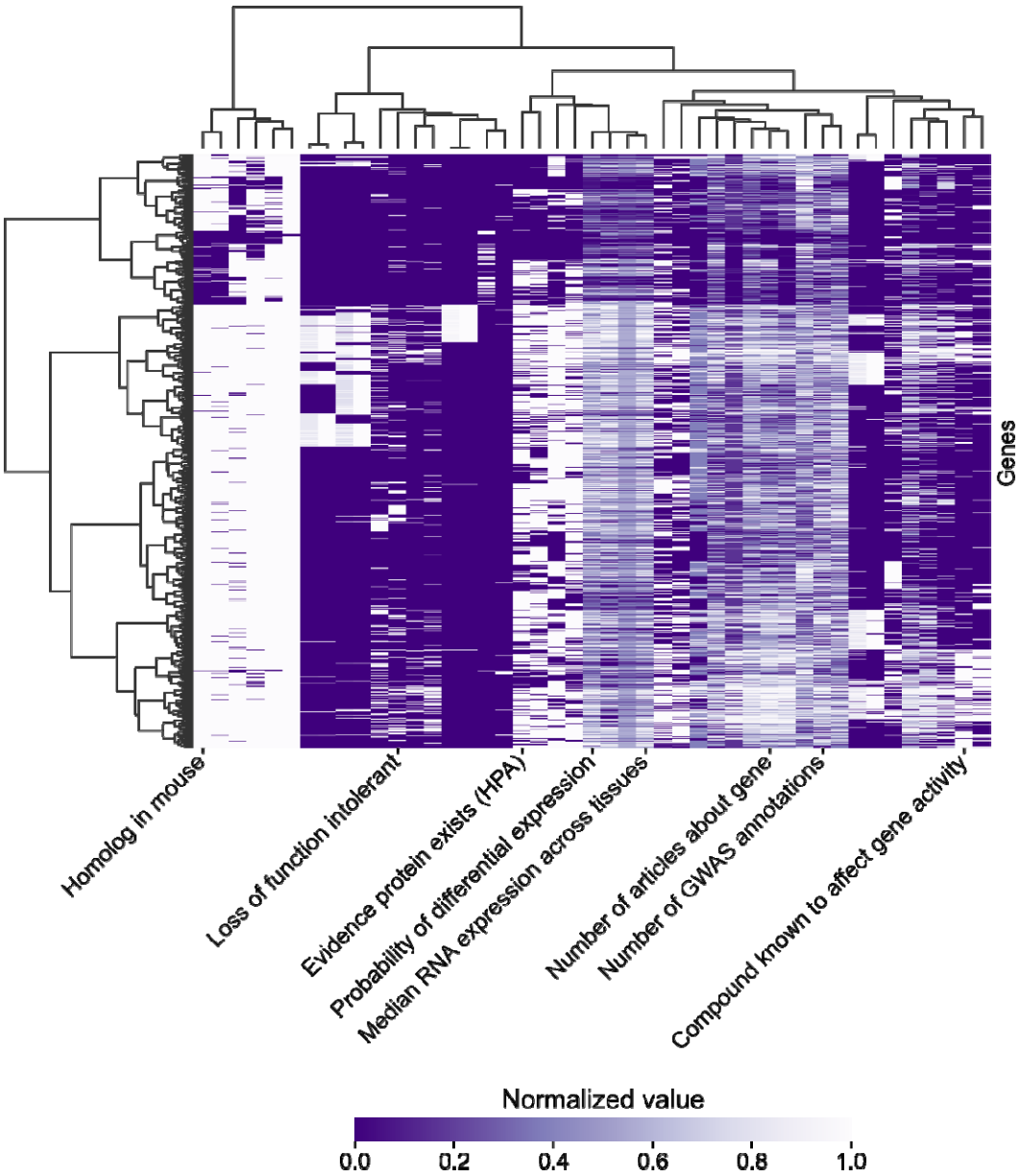

Figure 4: Clustermap showing collected factors across all human protein-coding genes. Factors are shown along the x axis, with genes along the y axis. Eight factors, representing the default factors we selected for FMUG, are shown (all factors are shown in **Figure S7** and **Figure S8**). Binary factors are coded to 0 (purple) and 1 (white), while continuous factors are ranked from 0 to 1 with ties resolved to minimum rank. Clustering was performed with Ward's method for hierarchical clustering[74].

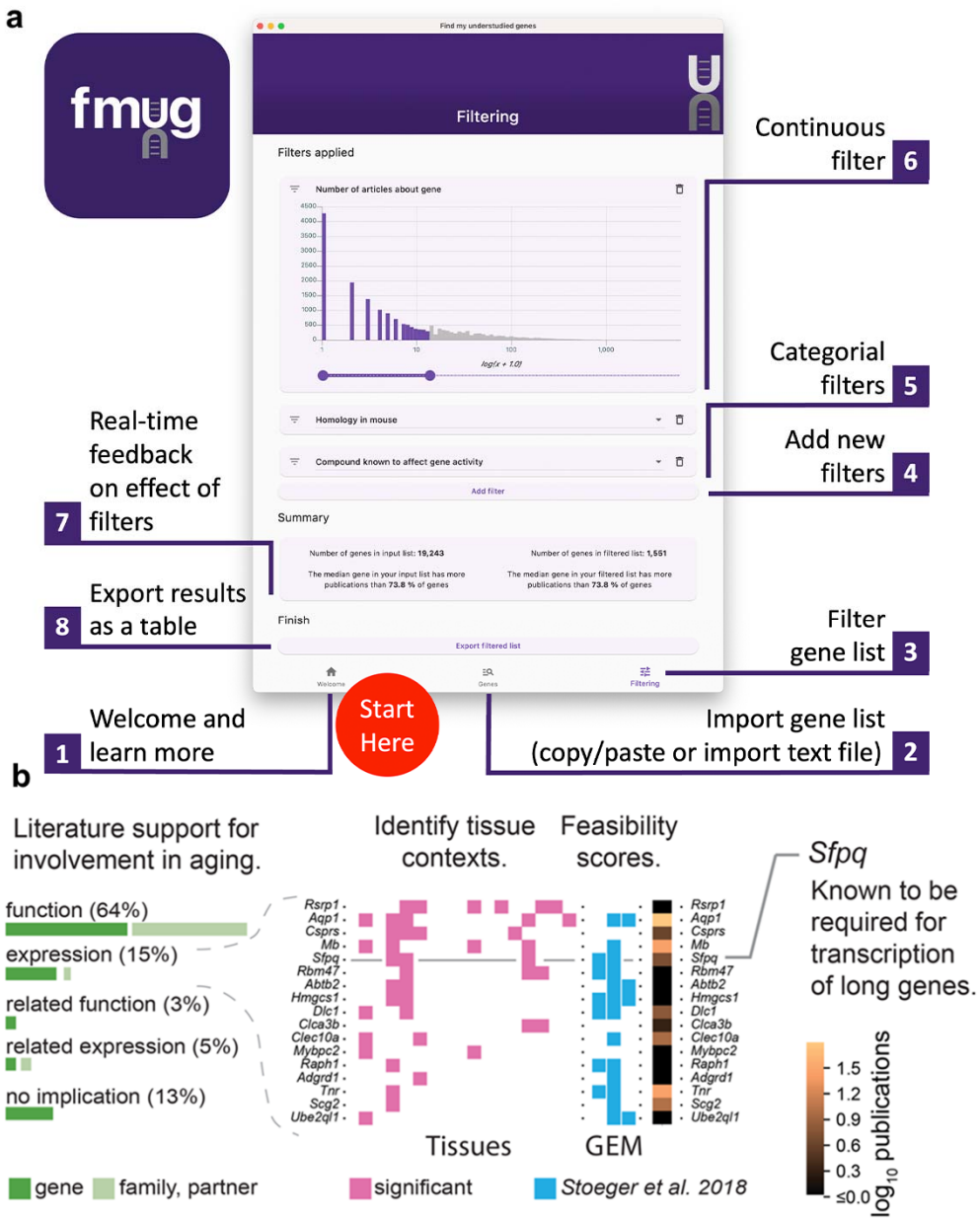

Figure 5: We created FMUG to help researchers identify understudied genes among their genes of interest and characterize their tractability for future research. a, Diagram describing use of FMUG. **b,** An early prototype of FMUG led us to the hypothesis that transcript length negatively correlates with up-regulation during aging. First, we identified genes that strongly associate with age-dependent transcriptional change across multiple cohorts. We then performed a literature review for each of these genes to identify the most direct way the genes (or evolutionally closely related genes or functionally closely related partner proteins) had been studied in aging. 64% had been functionally investigated in aging, 15% shown to change a measure of gene expression, 3% functionally investigated in a biological domain close to aging (such as senescence), and 5% shown to change a measure of gene expression in a biological

566 domain close to aging. For genes reported by others to change expression with age, we
567 identified tissues in which transcripts of the genes change during aging. We computed
568 'feasibility scores' scientific strategies (GEM: G: strong genetic support, E: and experimental
569 potential, M: homolog in invertebrate model organism) as described by Stoeger et al.[7] and total
570 number of publications in MEDLINE. Splicing factor, proline- and glutamine-rich (*Sfpq*) had
571 previously been demonstrated by Takeuchi et al. to be required for the transcriptional
572 elongation of long genes[62]. When performing a data-driven analysis of factors that could
573 possibly explain age-dependent changes of the entire transcriptome, we thus included gene
574 and transcript lengths, and subsequently found them to be more informative than transcription
575 factors or microRNAs[63].
576
577 **Data Availability**
578
579 All underlying data for figures are available at github.com/amarallab/fmug_analysis.
580
581 **Acknowledgements**
582
583 We thank Xiaojing Sui for testing FMUG and Northwestern Information Technology for technical
584 assistance. RAKR was supported in part by the National Institutes of Health Training Grant
585 (T32GM008449) through Northwestern University's Biotechnology Training Program. RAKR also
586 acknowledges support from the Dr. John N. Nicholson fellowship from Northwestern University
587 and Moderna Inc., "Identifying bias and improving reproducibility in RNA-seq
588 computational pipelines". LANA was supported by NSF 1956338, NIH U19AI135964 and Simons
589 Foundation DMS-1764421. TS was supported by NIH K99AG068544. We thank Alexander
590 Misharin, Richard Morimoto, and Scott Budinger for feedback on an early prototype of FMUG
591 which we used as part of our shared research into the biology of aging.
592
593 **References**
594
595 1 Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79-81,
596 doi:10.1016/S0168-9525(02)00014-8 (2003).
597 2 Su, A. I. & Hogenesch, J. B. Power-law-like distributions in biomedical publications and
598 research funding. *Genome Biol.* **8**, 404, doi:10.1186/gb-2007-8-4-404 (2007).
599 3 Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163-165,
600 doi:10.1038/470163a (2011).
601 4 Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology
602 annotations over time. *Bioinformatics* **29**, 476-482, doi:10.1093/bioinformatics/bts727
603 (2013).
604 5 Stoeger, T. & Nunes Amaral, L. A. The characteristics of early-stage research into human
605 genes are substantially different from subsequent research. *PLoS Biol* **20**, e3001520,
606 doi:10.1371/journal.pbio.3001520 (2022).
607 6 Grueneberg, D. A. *et al.* Kinase requirements in human cells: I. Comparing kinase
608 requirements across various cell types. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16472-16477,
609 doi:10.1073/pnas.0808019105 (2008).

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, doi:10.1371/journal.pbio.2006643 (2018).

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

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

10 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).

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

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

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, doi:10.1093/nar/gkac1139 (2022).

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, doi:10.1038/nature01626 (2003).

15 Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45-57, doi:10.1016/j.cell.2019.02.003 (2019).

16 Lloyd, K. C. K. *et al.* The Deep Genome Project. *Genome Biol* **21**, 18, doi:10.1186/s13059-020-1931-9 (2020).

17 Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods* **19**, 774-779, doi:10.1038/s41592-022-01454-x (2022).

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, doi:10.1186/s12920-017-0293-y (2017).

19 Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* **17**, 317-332, doi:10.1038/nrd.2018.14 (2018).

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, doi:10.1002/pmic.201800093 (2018).

21 Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol* **9**, 180241, doi:10.1098/rsob.180241 (2019).

22 Donohue, C. & Love, A. Perspectives on the Human Genome Project and genomics. *Minnesota Studies in the Philosophy of Science* **23** (in press).

23 Pena-Castillo, L. & Hughes, T. R. Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**, 7-14, doi:10.1534/genetics.107.074468 (2007).

24 Ellens, K. W. *et al.* Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res* **45**, 11495-11514, doi:10.1093/nar/gkx937 (2017).

654 25 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
655 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-
656 D1012, doi:10.1093/nar/gky1120 (2019).
657 26 Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of
658 curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187-200 (2021).
659 27 Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple
660 studies and organisms. *Nucleic acids research* **46**, D246-D251 (2018).
661 28 Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered
662 information at NCBI. *Nucleic acids research* **35**, D26-D31 (2007).
663 29 Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept
664 annotation for biomedical full text articles. *Nucleic acids research* **47**, W587-W593
665 (2019).
666 30 Hutchins, B. I., Santangelo, George. *iCite*, <https://doi.org/10.35092/yhjc.c.4586573 >
667 (2019).
668 31 Stoeger, T. & Nunes Amaral, L. A. COVID-19 research risks ignoring important host genes
669 due to pre-established research patterns. *Elife* **9**, e61981 (2020).
670 32 Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation
671 of the reasons why potentially important genes are ignored. *PLoS biology* **16**, e2006643
672 (2018).
673 33 Pfeiffer, T. & Hoffmann, R. Temporal patterns of genes in scientific publications. *Proc.*
674 *Natl. Acad. Sci. U. S. A.* **104**, 12052-12056, doi:10.1073/pnas.0701315104 (2007).
675 34 Knorr Cetina, K. *Epistemic Cultures*. (Harvard University Press, 1999).
676 35 Alberts, B. M. Limits to growth: In biology, small science is good science. *Cell*
677 *(Cambridge)* **41**, 337-338 (1985).
678 36 Richardson, S. S. a. S., Hallam. (Duke University Press, 2015).
679 37 Lopes, I., Altab, G., Raina, P. & De Magalhães, J. P. Gene size matters: an analysis of gene
680 length in the human genome. *Frontiers in Genetics* **12**, 559998 (2021).
681 38 Brown, L. A. & Peirson, S. N. Improving Reproducibility and Candidate Selection in
682 Transcriptomics Using Meta-analysis. *Journal of Experimental Neuroscience* **12**,
683 1179069518756296, doi:10.1177/1179069518756296 (2018).
684 39 Cesar-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478-
685 487, doi:10.1016/j.cell.2015.07.022 (2015).
686 40 Evans, J. A. Electronic publication and the narrowing of science and scholarship. *Science*
687 **321**, 395-399, doi:10.1126/science.1150473 (2008).
688 41 Alberts, B., Kirschner, M. W., Tilghman, S. & Varmus, H. Rescuing US biomedical
689 research from its systemic flaws. *Proc Natl Acad Sci U S A* **111**, 5773-5777,
690 doi:10.1073/pnas.1404402111 (2014).
691 42 Gilovich, T., Griffin, D. W. & Kahneman, D. *Heuristics and biases : the psychology of*
692 *intuitive judgment*. (Cambridge University Press, 2002).
693 43 Gerlt, J. A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950-9962,
694 doi:10.1021/bi201312u (2011).
695 44 Carter, A. J. *et al.* Target 2035: probing the human proteome. *Drug Discov Today* **24**,
696 2111-2115, doi:10.1016/j.drudis.2019.06.020 (2019).

697 45 Rodgers, G. *et al.* Glimmers in illuminating the druggable genome. *Nat Rev Drug Discov*
698 **17**, 301-302, doi:10.1038/nrd.2017.252 (2018).
699 46 Kustatscher, G. *et al.* An open invitation to the Understudied Proteins Initiative. *Nat*
700 *Biotechnol* **40**, 815-817, doi:10.1038/s41587-022-01316-z (2022).
701 47 *EUbOPEN*, <https://www.eubopen.org/> (
702 48 Rocha, J. J. *et al.* Functional unknomics: Systematic screening of conserved genes of
703 unknown function. *PLoS biology* **21**, e3002222 (2023).
704 49 Duek, P., Gateau, A., Bairoch, A. & Lane, L. Exploring the Uncharacterized Human
705 Proteome Using neXtProt. *J Proteome Res* **17**, 4211-4226,
706 doi:10.1021/acs.jproteome.8b00537 (2018).
707 50 Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential
708 gene expression. *Proc Natl Acad Sci U S A* **116**, 6491-6500,
709 doi:10.1073/pnas.1802973116 (2019).
710 51 Perdigao, N. & Rosa, A. Dark Proteome Database: Studies on Dark Proteins. *High*
711 *Throughput* **8**, doi:10.3390/ht8020008 (2019).
712 52 Essegian, D., Khurana, R., Stathias, V. & Schurer, S. C. The Clinical Kinase Index: A
713 Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer.
714 *Cell Rep Med* **1**, 100128, doi:10.1016/j.xcrm.2020.100128 (2020).
715 53 Sheils, T. K. *et al.* TCRD and Pharos 2021: mining the human proteome for disease
716 biology. *Nucleic Acids Res* **49**, D1334-D1346, doi:10.1093/nar/gkaa993 (2021).
717 54 Rocha, J. *et al.* Functional unknomics: closing the knowledge gap to accelerate
718 biomedical research. *bioRxiv* (2022).
719 55 Higgins, D. P., Weisman, C. M., Lui, D. S., D'Agostino, F. A. & Walker, A. K. Defining
720 characteristics and conservation of poorly annotated genes in Caenorhabditis elegans
721 using WormCat 2.0. *Genetics* **221**, doi:10.1093/genetics/iyac085 (2022).
722 56 Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to
723 uncharacterized genes. *Nat Genet* **53**, 638-649, doi:10.1038/s41588-021-00840-z (2021).
724 57 Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: a novel functional
725 genomics compendium with automated data mining and query reformulation support.
726 *Bioinformatics* **14**, 656-664, doi:10.1093/bioinformatics/14.8.656 (1998).
727 58 Tan, J. *et al.* ADAGE signature analysis: differential expression analysis with data-defined
728 gene sets. *Bmc Bioinformatics* **18**, 512, doi:10.1186/s12859-017-1905-4 (2017).
729 59 Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification
730 of protein functions. *Nat Biotechnol* **37**, 1361-1371, doi:10.1038/s41587-019-0298-5
731 (2019).
732 60 Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
733 *Innovation (Camb)* **2**, 100141, doi:10.1016/j.xinn.2021.100141 (2021).
734 61 Jiang, J. *et al.* Systematic illumination of druggable genes in cancer genomes. *Cell Rep*
735 **38**, 110400, doi:10.1016/j.celrep.2022.110400 (2022).
736 62 Takeuchi, A. *et al.* Loss of Sfpq Causes Long-Gene Transcriptopathy in the Brain. *Cell Rep*
737 **23**, 1326-1341, doi:10.1016/j.celrep.2018.03.141 (2018).
738 63 Stoeger, T. *et al.* Aging is associated with a systemic length-associated transcriptome
739 imbalance. *Nature Aging* **2**, 1191-1206, doi:10.1038/s43587-022-00317-6 (2022).

740 64 Gyenis, A. *et al.* Genome-wide RNA polymerase stalling shapes the transcriptome during
741 aging. *Nat Genet* **55**, 268-279, doi:10.1038/s41588-022-01279-6 (2023).
742 65 Ibañez-Solé, O., Barrio, I. & Izeta, A. Age or lifestyle-induced accumulation of
743 genotoxicity is associated with a length-dependent decrease in gene expression.
744 *iScience*, 106368, doi:https://doi.org/10.1016/j.isci.2023.106368 (2023).
745 66 Wei, C. H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept
746 annotation for biomedical full text articles. *Nucleic Acids Res* **47**, W587-W593,
747 doi:10.1093/nar/gkz389 (2019).
748 67 Rosenfeld, J. A. & Mason, C. E. Pervasive sequence patents cover the entire human
749 genome. *Genome Med.* **5**, 27, doi:10.1186/gm431 (2013).
750 68 Tu, S. *et al.* Response to 'pervasive sequence patents cover the entire human genome'.
751 *Genome medicine* **6**, 1-3 (2014).
752 69 Finan, C. *et al.* The druggable genome and support for target identification and
753 validation in drug development. *Science translational medicine* **9**, eaag1166 (2017).
754 70 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
755 biology and bioinformatics. *Bioinformatics* **25**, 1422-1423,
756 doi:10.1093/bioinformatics/btp163 (2009).
757 71 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
758 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
759 72 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,
760 285-291, doi:10.1038/nature19057 (2016).
761 73 Kendall, M. G. & Stuart, A. Inference and Relationship *The Advanced Theory of Statistics*
762 **2** (1973).
763 74 Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the*
764 *American statistical association* **58**, 236-244 (1963).
765