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 remain 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

33 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 purification–mass spectrometry (AP-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 some genes are less studied because they are rarely identified as hits in 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**). This demonstrates, in line with earlier studies[8-13], that the lack of publications on some genes is not explained by a lack of underlying biological experimental evidence.

Evaluating the second possibility, we found that hit genes that are highlighted in the title or abstract are 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**). We also find that this effect holds when controlling for the number of genes in each title/abstract by only considering one randomly-chosen gene per title/abstract (**Figure S7**).

Understudied genes are least frequently highlighted in the title/abstract in transcriptomics experiments and most frequently highlighted in 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 studied in the biomedical literature than 75% of genes. Hit median gene highlighted in the title/abstract in GWAS studies was more studied in the biomedical literature than 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 typically accumulate more citations, an effect that has held consistently since 2001 (**Figure 2).** Further, since 1990, articles about the least popular genes have at times been 3 to 4 times more likely to be among the most cited articles than articles on the most popular genes whereas articles on the most popular genes have been slightly less to be highly cited than lowly cited (**Figure S8**).

To rule out that these macroscopic observations stem from us having aggregated over different diseases, we separately analyzed 602 disease-related MeSH terms (**Figure S9**). 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 again opposes the hypothesis that less-investigated genes will yield articles with lower impact.

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 S10**). 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 are reminiscent of 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

168 sections of the publication most accessible to other scientists. While, overall, understudied
169 genes (and high-throughput assays themselves[5]) correlate with increased citation impact, for
170 high-throughput studies any potential gain in citations is either absent or too small to be
171 significant. Thus, there may not be any incentive for authors of high-throughput studies to
172 highlight understudied genes.
173
174 *Identification of biological and experimental factors associated with selection of highlighted*
175 *genes*
176
177 To illuminate why understudied genes are abandoned between experimental results and the
178 write-up of results, we performed a literature review to identify factors that have been
179 proposed to limit studies of understudied genes (**Table S1**). These factors range from
180 evolutionary factors (e.g., whether a gene only has homologs in primates), to chemical factors
181 (e.g., gene length or hydrophobicity of protein product), to historical factors (e.g., whether a
182 gene's sequence has previously been patented) to materialistic factors affecting experimental
183 design (e.g., whether designed antibodies are robust for immunohistochemistry).
184
185 As any of these factors could plausibly affect gene selection within individual domains of
186 biomedical research, we returned to the -omics data described above (**Figure 1B**) and measured
187 how much these factors align with the selective highlighting of hit genes in the title or abstract
188 of GWAS, AP-MS, transcriptomics, and CRISPR studies.
189
190 We identified 45 factors that relate to genes and found 33 (12 out of 23 binary factors and 21
191 out of 22 continuous factors) associated with selection in at least one assay type at Benjamini-
192 Hochberg FDR < 0.001 (**Figure 3, Table S2,** and **Table S3**). Across the four assay types, the most
193 informative binary factor describes whether there is a plasmid available for a gene in the
194 AddGene plasmid catalog. We cautiously hypothesize that this might reflect on many different
195 research groups producing reagents surrounding the genes that they actively study. The most
196 informative continuous factor is the number of research articles about a gene (**Figure 1B**).
197
198 To better understand how all 45 factors are related, we performed a cluster analysis of the
199 collected factors (**Figure S11** and **Figure S12**). This clustering suggests that many factors
200 influencing the abandonment of understudied genes are not independent. For instance, we find
201 that the number of articles about a gene is heavily correlated with the number of annotations
202 for that gene in all surveyed databases. In another case, gene length is heavily correlated with
203 the number of GWAS annotations for a gene, as described before in terms of transcript length
204 and single-nucleotide polymoprhisms[37].
205
206 *Data-driven design of a tool to promote the investigation of understudied genes*
207
208 To promote the investigation of understudied genes, we combined all the above insights to
209 create a tool we denoted *find my understudied genes* (FMUG). Our literature review revealed
210 several tools and resources aiming to promote research of understudied genes by publicizing
211 understudied genes[38-45] or by providing information about hit genes[7,46-50]. However, we noted

212 the absence of tools enabling scientists to actively engage with factors that align with gene
213 selection. Although such factors are largely correlated when considering all genes (**Figure S11**
214 and **Figure S12**), some factors cluster together and the influence of specific factors could vary
215 across laboratories. For instance, scientists could vary in their ability to perform proteomics, or
216 ability to explore orthologous genes in *C. elegans*, or ability to leverage human population data,
217 or perform standardized mouse assays.
218
219 Our tool makes selection bias explicit, while acknowledging that different laboratories vary in
220 their techniques and capabilities for follow-up research. Rather than telling scientists about the
221 existence of biases, FMUG aims to prompt scientists to make bias-aware informed decisions to
222 identify and potentially tackle important gaps in knowledge that they are well-suited to
223 address. For this reason, we believe that FMUG will not be of value only to scientists engaging
224 in high-throughput studies, but also by scientists wishing to mine existing datasets for hit genes
225 that they would be well-positioned to investigate further.
226
227 FMUG takes a list of genes from the user (ostensibly a hit list from a high-throughput -omics
228 experiment) and provides the kind of information that will allow a user to select genes for
229 further study.
230
231 Users can employ filters that reflect the factors identified in our literature review and
232 supported by our analysis. The default information provided to users consists of factors that are
233 representative of the identified clusters (**Figure S11**) and strongly associated with gene
234 selection in high-throughput experiments (**Figure 3, Table S2,** and **Table S3**). In extended
235 options, users can select any factor that demonstrated a significant association with the
236 selection of genes. For instance, a user may need to decide whether loss-of-function intolerant
237 genes should be considered for further research or not, or whether there should be robust
238 evidence that a gene is protein-coding. Some of these filters are context aware. For instance, a
239 user may select genes that have already been studied in the general biomedical literature but
240 not yet within the literature of their disease of interest.
241
242 To provide real-time feedback, users are, in parallel, presented the number of articles about
243 genes in their initial input list and the number of articles about genes that passed their filters.
244 Users can then export their filtered list of genes. In the interest of researcher privacy, FMUG
245 keeps all information local to the user's machine. Usage of FMUG is illustrated in **Figure 4A** and
246 demonstrated in **Movie S1**. FMUG is developed in Flutter and is available for download at
247 fmug.amaral.northwestern.edu as a MacOS/Windows app. For the development of custom
248 software and analytical code, we provide the data underlying FMUG at
249 github.com/amarallab/fmug_analysis.
250
251 To determine the practical usefulness of FMUG to scientists we used an early prototype of
252 FMUG to identify understudied genes associated with aging. One of these genes was Splicing
253 factor, proline- and glutamine-rich (*Sfpq*), which had not yet been investigated toward its role in
254 biological aging. We found *Sfpq* to be transcriptionally downregulated during murine aging.
255 Others had shown *Sfpq* to be required for the transcriptional elongation of long genes[51]. This

led us to hypothesize that during vertebrate aging, the transcripts of long genes become downregulated in most tissues **(Figure 4B).** We found this hypothesis to be supported through a multi-species analysis which we published in December 2022 in Nature Aging[52], with another group publishing so in January 2023 in Nature Genetics[53], and a third group in iScience in March 2023[54].

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, we take a gene being mentioned in the title or abstract of an article as a proxy for a gene receiving attention by the article's authors. The title and abstract are space-limited and thus cannot accommodate discussion of large numbers of genes.

Fourth, 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[17,55], -omics serving research purposes beyond target gene identification[22], the absence of good protocols for mass spectrometry[56], the electronic distribution and reading of research articles[57], rates of reproducibility[17], career prospects of investigators[7,58], authors beginning manuscripts with something familiar before introducing something new[59], and the human tendency to fall back to simplifying heuristics when making decisions under conditions with uncertainty[60]. Fifth, 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.

Discussion

Efforts to address the gaps in detailed knowledge about most genes have crystallized as initiatives promoting the investigation of understudied sets of genes[16,61-65], an approach to gene scholarship recently termed 'unknomics'[66]. The insight that understudied genes are lost to titles and abstracts of research articles in a leaky pipeline between genome-wide assays and reporting of results, and FMUG, have already been useful in guiding our own unknomics research **(Figure 4B)**.

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

300 even if not investigated further by the authors. This intervention would need to be carefully
301 designed since abstracts are limited in their size. Further, the observed discrepancy between
302 the popularity of hits highlighted by GWAS versus other technologies suggests that some -omics
303 technologies may be more powerful than others for characterizing understudied genes. This
304 possibility merits further research and researchers participating in unknomics should consider
305 the relative strengths of each technology towards providing tractable results for follow-up.
306
307 We believe that enabling scientists to consciously engage with bias in research target selection
308 will enable more biomedical researchers to participate in unknomics, to the potential benefit of
309 their own research impact and towards the advancement of our collective understanding of the
310 entire human genome.
311
312 **Materials and Methods**
313
314 *Genes information*
315
316 Homo sapiens gene information was downloaded from NCBI Gene on Aug 16, 2022
317 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/All_Data.gene_info.gz]. Only genes with an
318 unambiguous mapping of Entrez ID to Ensembl ID were used (n = 36,035). Number of gene
319 synonyms, protein-coding status, and official gene symbol were derived from this dataset. A
320 gene symbol was considered undefined if the gene's entry for HGNC gene symbol was "-".
321
322 *Genes in title/abstract of primary research articles*
323
324 Homo sapiens gene information was downloaded from NCBI Gene on Aug 16, 2022
325 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz].
326 gene2pubmed was download from NCBI Gene on Aug 16, 2022
327 [ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz][28]. PubTator gene annotations were
328 downloaded from NIH-NLM on July 12, 2022
329 [https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/][28,67]. PubMed was downloaded on Dec
330 17, 2021 [https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/].
331
332 Only using PMIDs annotated as primary research articles, a human gene was considered as
333 mentioned in the title/abstract of the publication if gene was annotated as being in the
334 title/abstract by PubTator and the article appeared in gene2pubmed.
335
336 *CRISPR articles*
337
338 BioGRID ORCS[26] v1.1.6 was downloaded on April 25, 2022
339 [https://downloads.thebiogrid.org/BioGRID-ORCS/Release-Archive/BIOGRID-ORCS-1.1.6/]. Any
340 genome-wide CRISPR knockout screens in human with an associated PubMed ID in which hit
341 genes were mentioned in the title or abstract was considered (n = 15). 9,268 unique genes were
342 found as hits. Of these, 18 (0.19%) were highlighted in titles/abstracts in the reporting articles

343 and 19 (0.21%) were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is
344 available in **Table S6**.
345
346 *Transcriptomics articles*
347
348 EBI-GXA[27] release 36 was downloaded on Sep 15, 2020
349 [https://web.archive.org/web/20201022184159/https://www.ebi.ac.uk/gxa/download]. This is
350 the most recent release of EBI-GXA available as a bulk download. Any transcriptomics
351 comparisons with an associated PubMed ID in which hit genes were mentioned in the title or
352 abstract was considered (n= 148). Analysis was restricted to protein-coding genes (some
353 screens featured non-protein-coding genes, but this was not common to all analyses). DE was
354 called at Benjamini-Hochberg FDR $q < 0.05$. 18,295 unique genes were found as hits. Of these,
355 161 (0.88%) were highlighted in titles/abstracts in the reporting articles and 692 (3.78%) were
356 highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**.
357
358 *Affinity purification–mass spectrometry articles*
359
360 BioGRID[26] v3.5.186 was downloaded on April 25, 2022
361 [https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.186/]. Any
362 interactions involving a human gene as the prey protein with an experimental evidence code of
363 'Affinity Capture-MS' labeled as 'High-Throughput' that had an associated PubMed ID in which
364 hit genes were mentioned in the title or abstract was considered (n= 296). Prey proteins in
365 these interactions were considered hits. 7,919 unique genes were found as hits. Of these, 311
366 (3.93%) were highlighted in titles/abstracts in reporting articles and 407 (5.14%) were
367 highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**.
368
369 *GWAS articles*
370
371 The NHGRI-EBI GWAS catalog[25] (associations and studies) was download on Aug 17, 2022
372 [https://www.ebi.ac.uk/gwas/docs/file-downloads]. Any GWAS screens with an associated
373 PubMed ID in which hit genes were mentioned in the title or abstract was considered (n= 450).
374 Only SNPs occurring within a gene were considered hits. 1,043 unique genes were found as hits.
375 Of these, 413 (39.6%) were highlighted in titles/abstracts in reporting articles and 319 (30.6%)
376 were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in
377 **Table S6**
378
379 *Citing articles*
380
381 NIH iCite v32 was downloaded on Aug 25, 2022[30]
382 [https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collectio
383 n_/4586573/32].
384
385 *Functional annotations*
386

Mapping of genes to Gene Ontology / Protein Interaction Database / WikiPathways / Reactome / Kyoto Encyclopedia of Genes and Genomes / Human Phenotype Ontology / BioCarta categories was derived from MSigDB v7.5 Entrez ID .gmt files, downloaded on Apr 12, 2022 [http://www.gsea-msigdb.org/gsea/downloads_archive.jsp].

Between-species homology

Homologene Build 68 was used to determine interspecies homology [ftp.ncbi.nih.gov/pub/HomoloGene/build68/]. Human = taxid:9606, mouse = taxid:10090, rat = taxid:10116, c. elegans = taxid:6239, d. melanogaster = taxid:7227, yeast = taxid:559292, zebrafish = taxid:7955.

Primate specificity

Human genes were considered primate-specific if the only other members of their homology group belonged to primate genomes. Primate taxonomy ids were downloaded from NCBI Taxonomy on Sep 20, 2022 [https://www.ncbi.nlm.nih.gov/taxonomy/?term=txid9443[Subtree]].

Number of publications in model organisms

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 using PMIDs annotated as primary research articles, genes was considered as mentioned in the title/abstract of the publication if gene was annotated as being in the title/abstract by PubTator and the article appeared in gene2pubmed.

Genes in model organisms were mapped to human genes and the number of articles on those mapping to human genes were counted. If a model organism's gene had homology to human but no associated publications, the number of publications was resolved to zero. Otherwise, counts were listed as NA.

Mouse phenotype hits

International Mouse Phenotyping Consortium data release 17.0 was downloaded on Aug 18, 2022 [https://www.mousephenotype.org/data/release]. Mouse genes were matched to human genes with Homologene.

Gene Expression Atlas (EBI-GXA)

EBI-GXA release 36 was downloaded on Sep 15, 2020 [https://web.archive.org/web/20201022184159/https://www.ebi.ac.uk/gxa/download]. This is

430 the most recent release of EBI-GXA available as a bulk download. For probability of DE, only
431 RNA-seq comparisons were considered and DE was called at Benjamini-Hochberg $q < 0.05$.
432
433 *Global RNA expression*
434
435 RNA consensus tissue gene data from HPA release 21.1 was downloaded on Sep 20, 2022
436 [https://www.proteinatlas.org/about/download]. Global RNA expression was estimated by
437 taking the median expression (nTPM) across tissues for each gene and the proportion of tissues
438 with detectable (≥1 nTPM) expression for each gene.
439
440 *Expression in HeLa cells*
441
442 RNA cell line gene data from HPA release 21.1 was downloaded on Sep 20, 2022
443 [https://www.proteinatlas.org/about/download]. Expression is in nTPM.
444
445 *Previous patent activity*
446
447 Genes with patent activity were defined from Table S1 of Rosenfeld and Mason, 2013[68]. Genes
448 were mapped with their HGNC symbol. This analysis aligned sequences in patents to the human
449 genome to estimate patent coverage of human coding sequences. Although this does not
450 necessarily reflect whether the mapped genes were claimed directly by the patent holder, as
451 noted by others[69], this analysis remains the most comprehensive available for determining
452 patent coverage of the human genome.
453
454 *Druggability*
455
456 Druggable genes were identified from Table S1 of Finan et al., 2017[70]. Genes were mapped with
457 their Ensembl identifier.
458
459 *Gene length*
460
461 GenBank was downloaded in spring 2017 (genome version GRCh38.p10). Gene length is defined
462 here as the span of the longest transcript on the chromosome. This aligns with the model of
463 gene length used in Stoeger et al., 2018[7].
464
465 *Solubility*
466
467 SwissProt protein sequences and mapping tables to Entrez GeneIDs were downloaded from
468 Uniprot in spring 2017. Protein GRAVY score (ignoring Pyrrolysine and Selenocysteine) was
469 estimated with BioPython[71].
470
471 *Loss of function intolerance*
472

473 Data was obtained from Karczewski et al.[72]. pLI scores > 0.9 on main transcripts, as flagged by
474 authors, were considered as highly loss-of-function intolerant as described by Lek et al.[73].
475
476 *Number of GWAS hits*
477
478 EBI GWAS catalog[25] (associations and studies) was download on Aug 17, 2022
479 [https://www.ebi.ac.uk/gwas/docs/file-downloads]. Loci were mapped to the nearest gene.
480
481 *Status as understudied protein*
482
483 The Illuminating the Druggable Genome understudied protein list was downloaded on Sep 20,
484 2022
485 [https://github.com/druggablegenome/IDGTargets/blob/master/IDG_TargetList_CurrentVersio
486 n.json].
487
488 *Human Protein Atlas*
489
490 HPA release 21.1 was downloaded on Sep 20, 2022 [https://www.proteinatlas.org/search].
491 Evidence for a protein's existence, as determined by NeXtProt, HPA, or UniProt was resolved as
492 True if the respective evidence entry was annotated as "Evidence at protein level". Status as a
493 membrane protein was determined by whether the 'Protein class' column contained the string
494 'membrane protein'. Antibodies were considered available for each protein if the protein's
495 entry in the 'Antibody' column was not null.
496
497 *Availability of plasmids*
498
499 The AddGene plasmid catalog was downloaded on Aug 12, 2022
500 [https://www.addgene.org/browse/gene/gene-list-data/?_=1666368044314].
501
502 *Availability of compounds*
503
504 The catalog of gene targets was downloaded from ChEMBL on Sep 20, 2022
505 [https://www.ebi.ac.uk/chembl/g/#browse/targets]. UniProt IDs were converted to Entrez IDs
506 to identify which human genes were affected by any compound.
507
508 *Mendelian inheritance*
509
510 Autosomal dominant [https://hpo.jax.org/app/browse/term/HP:0000006] and autosomal
511 recessive [https://hpo.jax.org/app/browse/term/HP:0000007] inherited disease-gene
512 associations were downloaded from the Human Phenotype Ontology on Sep 20, 2022. Genes
513 were considered to have evidence of Mendelian inheritance if they appeared in these lists of
514 associations.
515
516 *Code*

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

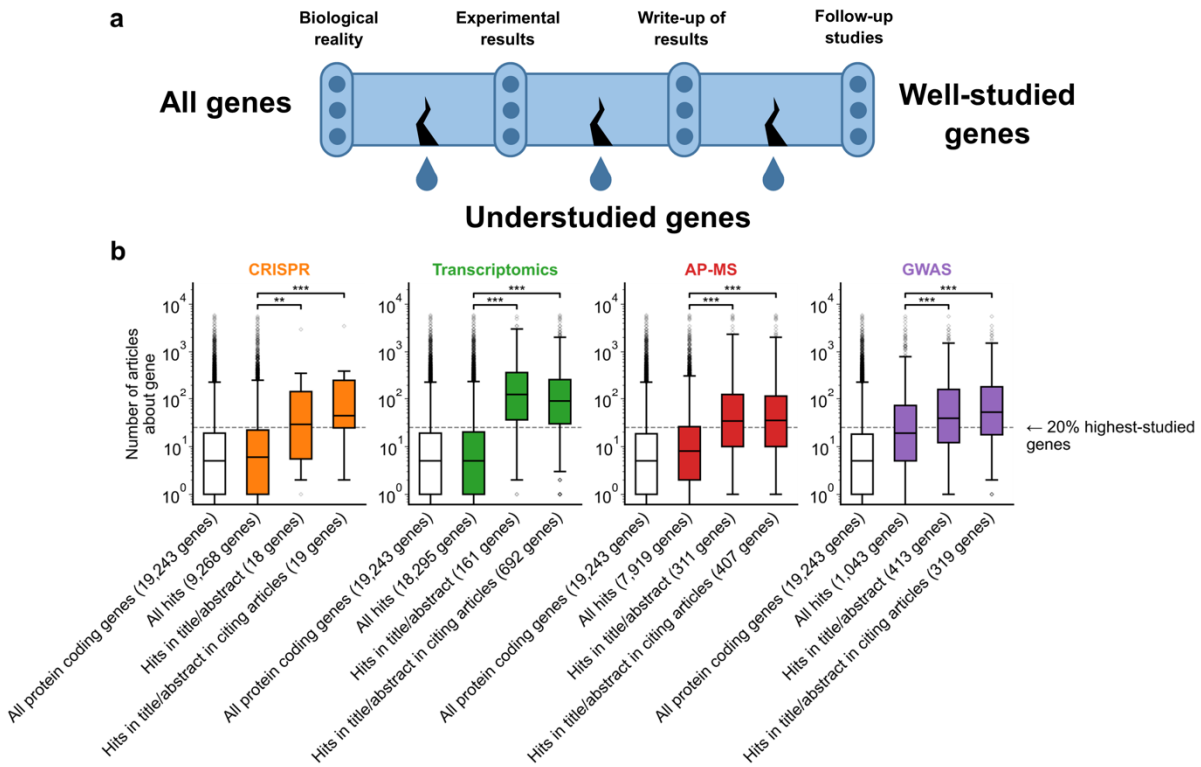


Figure 1: A shift in focus towards well-studied genes occurs during the summarization and write-up of results and remains in subsequent studies. **a,** Conceptual diagram depicting possible points of abandonment for understudied genes in studies using high-throughput -omics experiments. **b,** We identified articles reporting on genome-wide CRISPR screens (CRISPR, 15 focus articles and 18 citing articles), transcriptomics (T-omics, 148 focus articles and 1,678 citing articles), affinity purification–mass spectrometry (AP-MS, 296 focus articles and 1,320 citing articles), and GWAS (450 focus articles and 3,524 citing articles). Focusing only on protein-coding genes (white box plot), we retrieved data uploaded to repositories describing which genes came up as "hits" in each experiment (first colored box plot). We then retrieved the hits mentioned in the titles and abstracts of those articles (second colored box plot) and hits mentioned in the titles and abstracts of articles citing those articles (third colored box plot). Unique hit genes are only counted once. 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. ** 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.

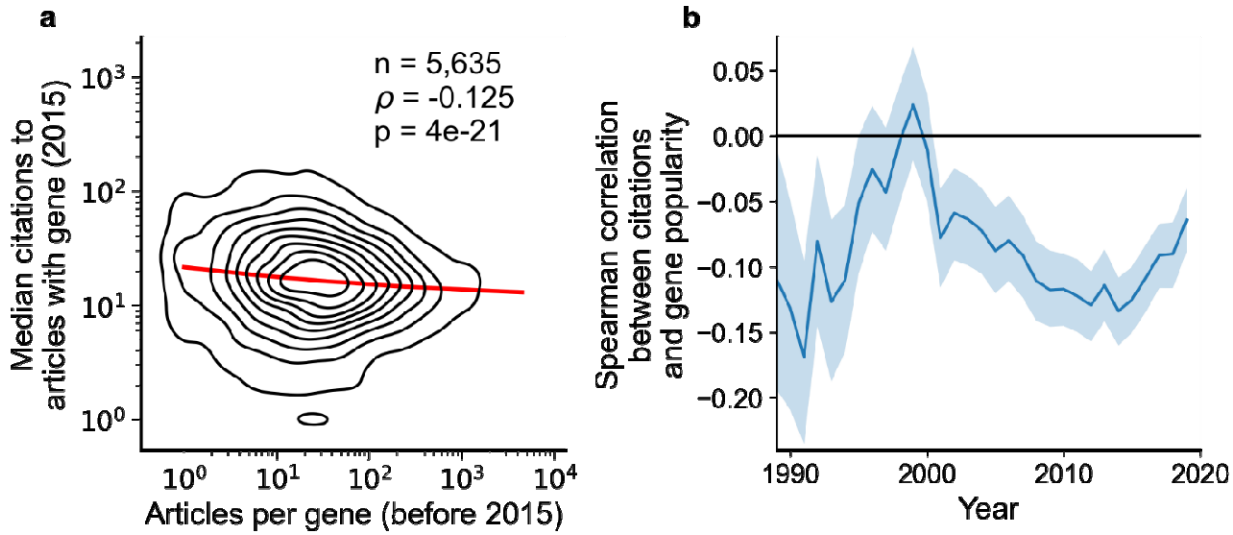


Figure 2: Articles focusing on less popular genes tend to accrue more citations. a, Density plot shows correlation between articles per gene before 2015 and median citations to articles published in 2015. Contours correspond to deciles in density. Solid red line shows locally weighted scatterplot smoothing (LOWESS) regression. ρ is Spearman rank correlation and p the significance values of the Spearman rank correlation as described by Kendall and Stuart[74]. We forgo depicting more recent years than 2015 to allow for citations to accumulate over multiple years, providing a more sensitive and robust readout of long-term impact. **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. For more recent years, where articles have had less time to accumulate citations, insufficient signal may cause correlation to converge toward zero.

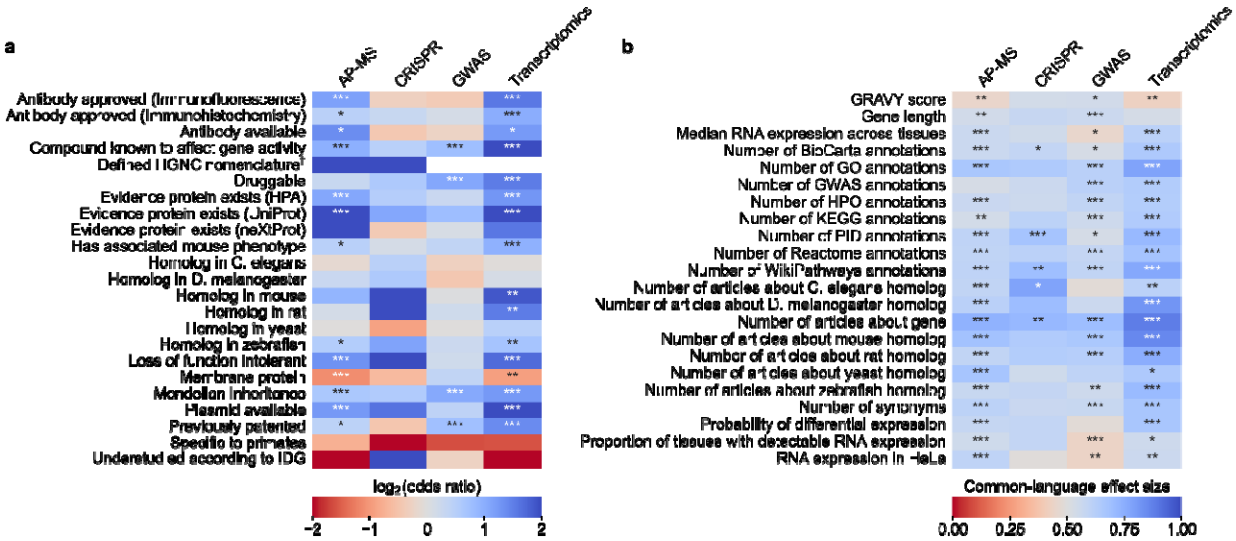
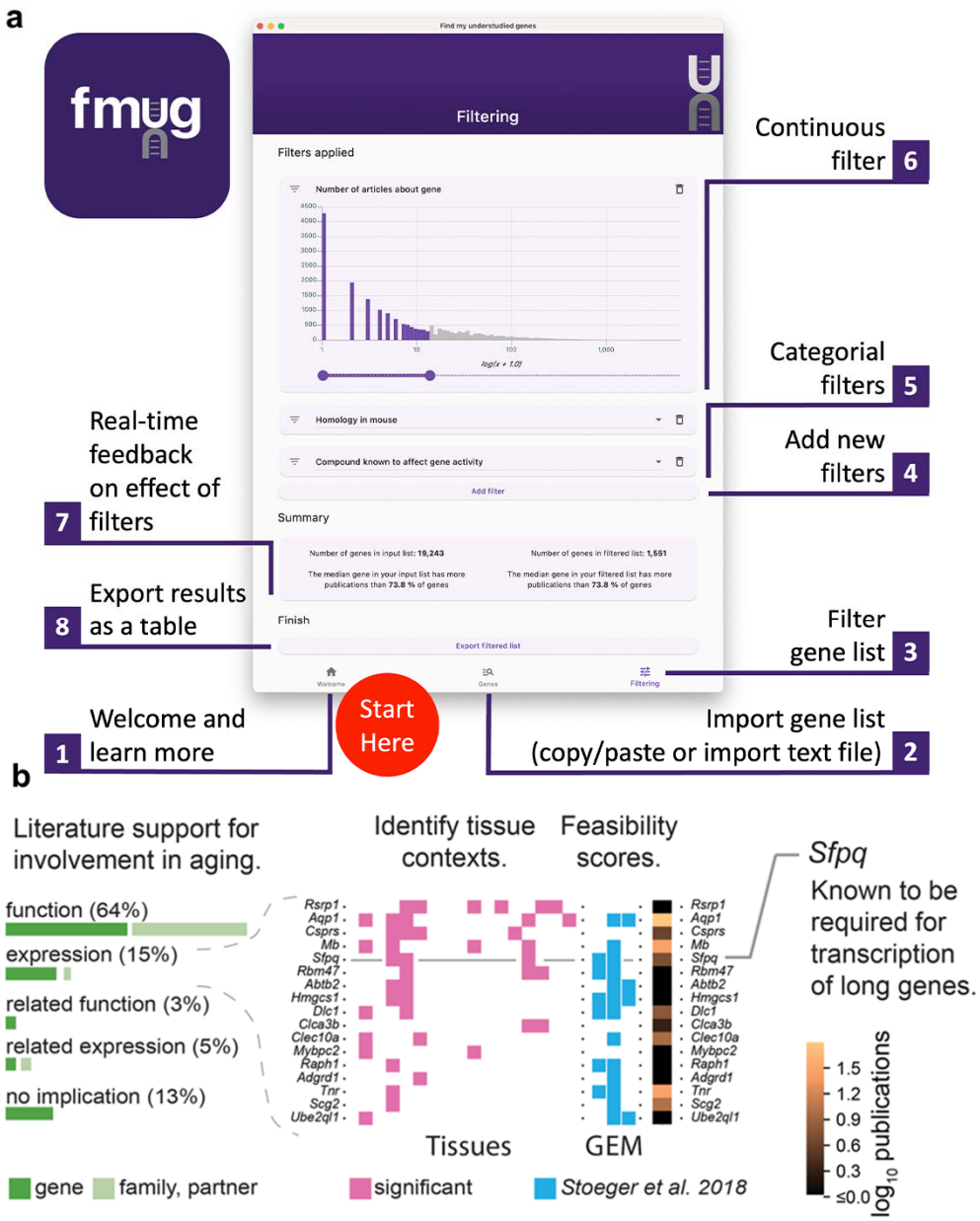


Figure 3: We evaluated which gene-related factors are associated with elevation to the title/abstract of an article featuring a high-throughput experiment. a) Association between factors with binary (True/False) identities and highlighting hits in title/abstract of reporting

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 4.262 times as likely to be mentioned in the title/abstract in an article using transcriptomics, corresponding to an odds ratio of 4.331). Collected articles are described in **Figure 1B** and **Figures S5**, **S6** and **S7.** 95% confidence interval of odds ratio is shown in parentheses. * = Benjamini-Hochberg FDR < 0.05, ** = FDR < 0.01, and *** = FDR < 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**, **S6** and **S7.** * = Benjamini-Hochberg FDR < 0.05, ** = FDR < 0.01, and *** = FDR < 0.001 by two-sided Mann-Whitney U test. Results are shown numerically in **Table S3**.



575

576 **Figure 4: We created FMUG to help researchers identify understudied genes among their**
577 **genes of interest and characterize their tractability for future research. a,** Diagram describing
578 use of FMUG. **b,** An early prototype of FMUG led us to the hypothesis that transcript length
579 negatively correlates with up-regulation during aging. First, we identified genes that strongly
580 associate with age-dependent transcriptional change across multiple cohorts. We then
581 performed a literature review for each of these genes to identify the most direct way the genes
582 (or evolutionally closely related genes or functionally closely related partner proteins) had been
583 studied in aging. 64% had been functionally investigated in aging, 15% shown to change a
584 measure of gene expression, 3% functionally investigated in a biological domain close to aging
585 (such as senescence), and 5% shown to change a measure of gene expression in a biological

586 domain close to aging. For genes reported by others to change expression with age, we
587 identified tissues in which transcripts of the genes change during aging. We computed
588 'feasibility scores' scientific strategies (GEM: G: strong genetic support, E: and experimental
589 potential, M: homolog in invertebrate model organism) as described by Stoeger et al.[7] and total
590 number of publications in MEDLINE. Splicing factor, proline- and glutamine-rich (*Sfpq*) had
591 previously been demonstrated by Takeuchi et al. to be required for the transcriptional
592 elongation of long genes[51]. When performing a data-driven analysis of factors that could
593 possibly explain age-dependent changes of the entire transcriptome, we thus included gene
594 and transcript lengths, and subsequently found them to be more informative than transcription
595 factors or microRNAs[52].
596
597 **Data Availability**
598
599 All underlying data for figures are available at github.com/amarallab/fmug_analysis.
600
601 **Acknowledgements**
602
603 We thank Xiaojing Sui for testing FMUG and Northwestern Information Technology for technical
604 assistance. RAKR was supported in part by the National Institutes of Health Training Grant
605 (T32GM008449) through Northwestern University's Biotechnology Training Program. RAKR also
606 acknowledges support from the Dr. John N. Nicholson fellowship from Northwestern University
607 and Moderna Inc., "Identifying bias and improving reproducibility in RNA-seq
608 computational pipelines". LANA was supported by NSF 1956338, NIH U19AI135964 and Simons
609 Foundation DMS-1764421. TS was supported by NIH K99AG068544. We thank Alexander
610 Misharin, Richard Morimoto, and Scott Budinger for feedback on an early prototype of FMUG
611 which we used as part of our shared research into the biology of aging.
612
613 **References**
614
615 1 Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79-81
616 (2003). https://doi.org:10.1016/S0168-9525(02)00014-8
617 2 Su, A. I. & Hogenesch, J. B. Power-law-like distributions in biomedical publications and
618 research funding. *Genome Biol.* **8**, 404 (2007). https://doi.org:10.1186/gb-2007-8-4-404
619 3 Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163-165 (2011).
620 https://doi.org:10.1038/470163a
621 4 Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology
622 annotations over time. *Bioinformatics* **29**, 476-482 (2013).
623 https://doi.org:10.1093/bioinformatics/bts727
624 5 Stoeger, T. & Nunes Amaral, L. A. The characteristics of early-stage research into human
625 genes are substantially different from subsequent research. *PLoS Biol* **20**, e3001520
626 (2022). https://doi.org:10.1371/journal.pbio.3001520
627 6 Grueneberg, D. A. *et al.* Kinase requirements in human cells: I. Comparing kinase
628 requirements across various cell types. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16472-16477
629 (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 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** (2020). https://doi.org:10.7554/eLife.61981

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

673 23 Pena-Castillo, L. & Hughes, T. R. Why are there still over 1000 uncharacterized yeast
674 genes? *Genetics* **176**, 7-14 (2007). https://doi.org:10.1534/genetics.107.074468
675 24 Ellens, K. W. *et al.* Confronting the catalytic dark matter encoded by sequenced
676 genomes. *Nucleic Acids Res* **45**, 11495-11514 (2017).
677 https://doi.org:10.1093/nar/gkx937
678 25 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
679 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-
680 D1012 (2019). https://doi.org:10.1093/nar/gky1120
681 26 Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of
682 curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187-200 (2021).
683 27 Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple
684 studies and organisms. *Nucleic acids research* **46**, D246-D251 (2018).
685 28 Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered
686 information at NCBI. *Nucleic acids research* **35**, D26-D31 (2007).
687 29 Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept
688 annotation for biomedical full text articles. *Nucleic acids research* **47**, W587-W593
689 (2019).
690 30 Hutchins, B. I., Santangelo, George. *iCite*, <https://doi.org/10.35092/yhjc.c.4586573 >
691 (2019).
692 31 Stoeger, T. & Nunes Amaral, L. A. COVID-19 research risks ignoring important host genes
693 due to pre-established research patterns. *Elife* **9**, e61981 (2020).
694 32 Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation
695 of the reasons why potentially important genes are ignored. *PLoS biology* **16**, e2006643
696 (2018).
697 33 Pfeiffer, T. & Hoffmann, R. Temporal patterns of genes in scientific publications. *Proc.*
698 *Natl. Acad. Sci. U. S. A.* **104**, 12052-12056 (2007).
699 https://doi.org:10.1073/pnas.0701315104
700 34 Knorr Cetina, K. *Epistemic Cultures*. (Harvard University Press, 1999).
701 35 Alberts, B. M. Limits to growth: In biology, small science is good science. *Cell*
702 *(Cambridge)* **41**, 337-338 (1985).
703 36 Richardson, S. S. a. S., Hallam. (Duke University Press, 2015).
704 37 Lopes, I., Altab, G., Raina, P. & De Magalhães, J. P. Gene size matters: an analysis of gene
705 length in the human genome. *Frontiers in Genetics* **12**, 559998 (2021).
706 38 Duek, P., Gateau, A., Bairoch, A. & Lane, L. Exploring the Uncharacterized Human
707 Proteome Using neXtProt. *J Proteome Res* **17**, 4211-4226 (2018).
708 https://doi.org:10.1021/acs.jproteome.8b00537
709 39 Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential
710 gene expression. *Proc Natl Acad Sci U S A* **116**, 6491-6500 (2019).
711 https://doi.org:10.1073/pnas.1802973116
712 40 Perdigao, N. & Rosa, A. Dark Proteome Database: Studies on Dark Proteins. *High*
713 *Throughput* **8** (2019). https://doi.org:10.3390/ht8020008
714 41 Essegian, D., Khurana, R., Stathias, V. & Schurer, S. C. The Clinical Kinase Index: A
715 Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer.
716 *Cell Rep Med* **1**, 100128 (2020). https://doi.org:10.1016/j.xcrm.2020.100128

42 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

43 Rocha, J. *et al.* Functional unknomics: closing the knowledge gap to accelerate biomedical research. *bioRxiv* (2022).

44 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

45 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

46 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

47 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

48 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

49 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

50 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

51 Takeuchi, A. *et al.* Loss of Sfpq Causes Long-Gene Transcriptopathy in the Brain. *Cell Rep* **23**, 1326-1341 (2018). https://doi.org:10.1016/j.celrep.2018.03.141

52 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

53 Gyenis, A. *et al.* Genome-wide RNA polymerase stalling shapes the transcriptome during aging. *Nat Genet* **55**, 268-279 (2023). https://doi.org:10.1038/s41588-022-01279-6

54 Ibañez-Solé, O., Barrio, I. & Izeta, A. Age or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression. *iScience*, 106368 (2023). https://doi.org:https://doi.org/10.1016/j.isci.2023.106368

55 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

56 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

57 Evans, J. A. Electronic publication and the narrowing of science and scholarship. *Science* **321**, 395-399 (2008). https://doi.org:10.1126/science.1150473

58 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

59 Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical combinations and scientific impact. *Science* **342**, 468-472 (2013). https://doi.org:10.1126/science.1240474

60 Gilovich, T., Griffin, D. W. & Kahneman, D. *Heuristics and biases : the psychology of intuitive judgment*. (Cambridge University Press, 2002).

61 Gerlt, J. A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950-9962 (2011). https://doi.org:10.1021/bi201312u

62 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

63 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

64 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

65 *EUbOPEN*, <https://www.eubopen.org/> (

66 Rocha, J. J. *et al.* Functional unknomics: Systematic screening of conserved genes of unknown function. *PLoS biology* **21**, e3002222 (2023).

67 Wei, C. H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* **47**, W587-W593 (2019). https://doi.org:10.1093/nar/gkz389

68 Rosenfeld, J. A. & Mason, C. E. Pervasive sequence patents cover the entire human genome. *Genome Med.* **5**, 27 (2013). https://doi.org:10.1186/gm431

69 Tu, S. *et al.* Response to 'pervasive sequence patents cover the entire human genome'. *Genome medicine* **6**, 1-3 (2014).

70 Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Science translational medicine* **9**, eaag1166 (2017).

71 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009). https://doi.org:10.1093/bioinformatics/btp163

72 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org:10.1038/s41586-020-2308-7

73 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016). https://doi.org:10.1038/nature19057

74 Kendall, M. G. & Stuart, A. Inference and Relationship *The Advanced Theory of Statistics* **2** (1973).

75 Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236-244 (1963).