

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

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2

March 12, 2024 (this version)

Reviewed preprint version 1



December 15, 2023

Posted to preprint server

October 31, 2023

Sent for peer review

October 18, 2023

Reese AK Richardson, Heliodoro Tejedor Navarro, Luis A Nunes Amaral , Thomas Stoeger 

Interdisciplinary Biological Sciences, Northwestern University • Department of Chemical and Biological Engineering, Northwestern University • Northwestern Institute on Complex Systems, Northwestern University • Department of Physics and Astronomy, Northwestern University • Department of Molecular Biosciences, Northwestern University • The Potocsnak Longevity Institute, Northwestern University • Simpson Querrey Lung Institute for Translational Science, Northwestern University

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

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.

eLife assessment

This study investigated the factors related to understudied genes in biomedical research. It showed that understudied genes are largely abandoned at the writing stage, and it identified a number of biological and experimental factors that influence which genes are selected for investigation. The study is an **important** contribution to this branch of meta-research, and the evidence in support of the findings is **solid**.

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^{1,2,3,4,5}. 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^{8,9,10,11,12,13}. 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 disease¹³. Furthermore, when comparing gene-disease pairs, there is no correlation between the ranks of support by transcriptomics and occurrence in annotation databases⁹.

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^{14,15,16,17}, this discrepancy has persisted^{9,18,19,20,21,22} even as the popularity of -omics technologies has risen^{5,23,24,25}. 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 catalog²⁵), 296 studies using affinity purification–mass spectrometry (AP-MS, indexed by BioGRID²⁶), 148 transcriptomic studies (indexed by the EBI Gene Expression Atlas, EBI-GXA²⁷), and 15 genome-wide screens using CRISPR (indexed by BioGRID Open Repository of CRISPR Screens, BioGRID ORCS²⁶) (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 gene2pubmed²⁸ and PubTator²⁹. We used NIH iCite v32 for citations³⁰. 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 studies³².

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 impact¹⁷, as observed around the turn of the millenium³³. 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

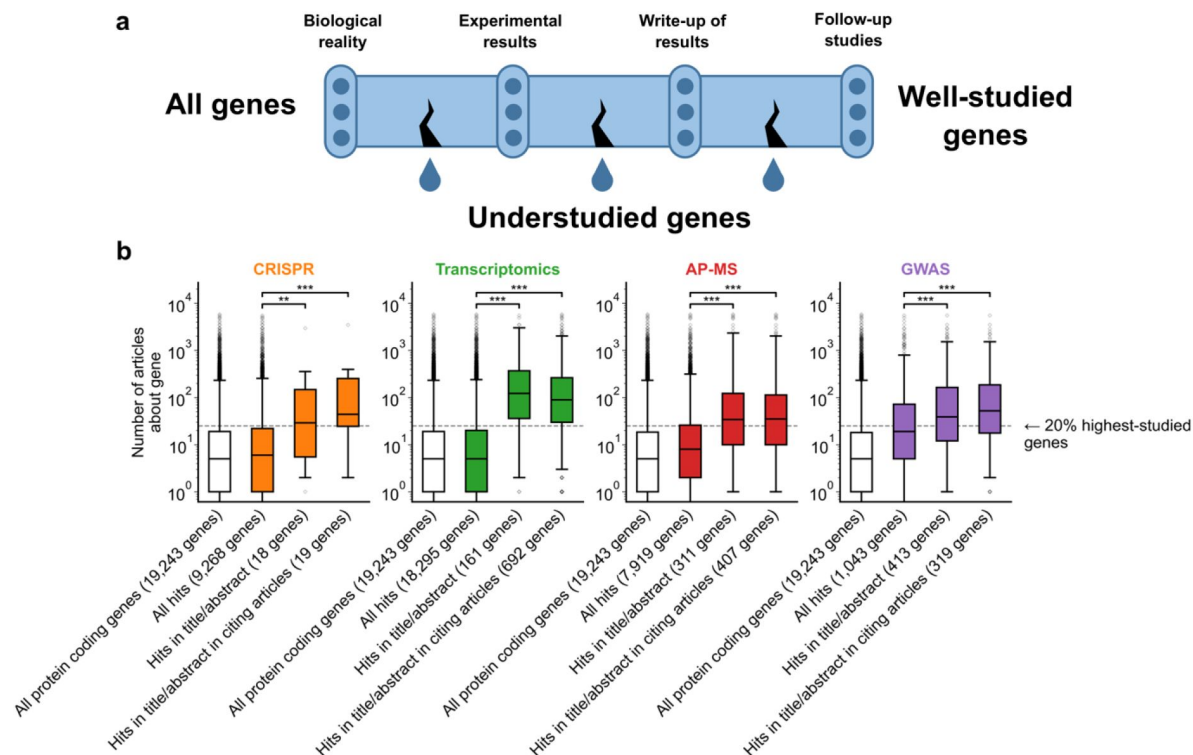


Figure 1

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.

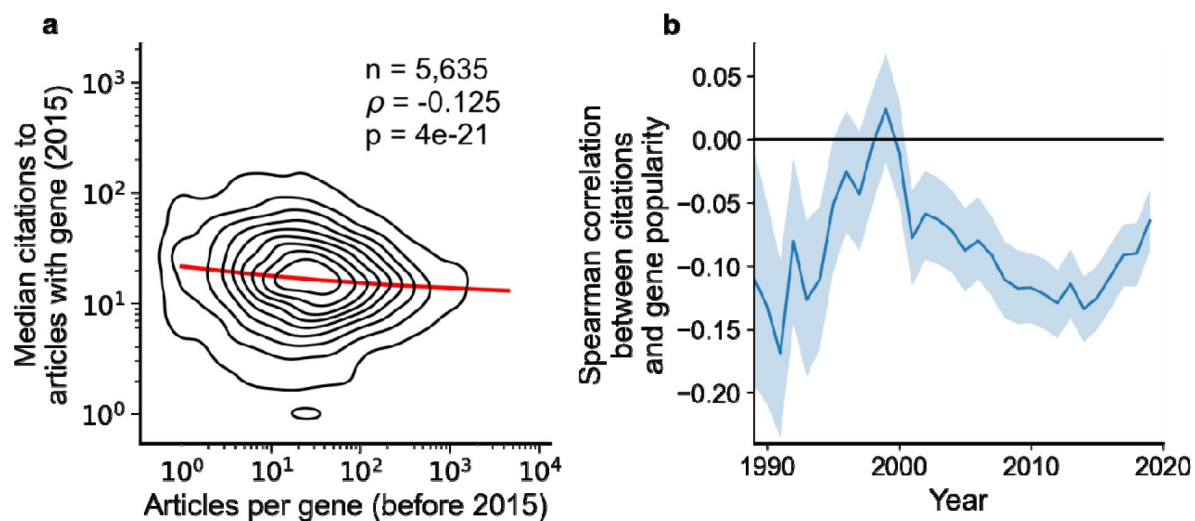


Figure 2

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. 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.

(**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 sections of the publication most accessible to other scientists. While, overall, understudied genes (and high-throughput assays themselves⁵) 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 1B**) and measured how much these factors align with the selective highlighting of hit genes in the title or abstract of GWAS, AP-MS, transcriptomics, and CRISPR studies.

We identified 45 factors that relate to genes and found 33 (12 out of 23 binary factors and 21 out of 22 continuous factors) associated with selection in at least one assay type at Benjamini-Hochberg FDR < 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. We cautiously hypothesize that this might reflect on many different research groups producing reagents surrounding the genes that they actively study. The most informative continuous factor is the number of research articles about a gene (**Figure 1B**).

To better understand how all 45 factors are related, we performed a cluster analysis of the collected factors (**Figure S11** and **Figure S12**). This clustering suggests that many 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 polymorphisms³⁷.

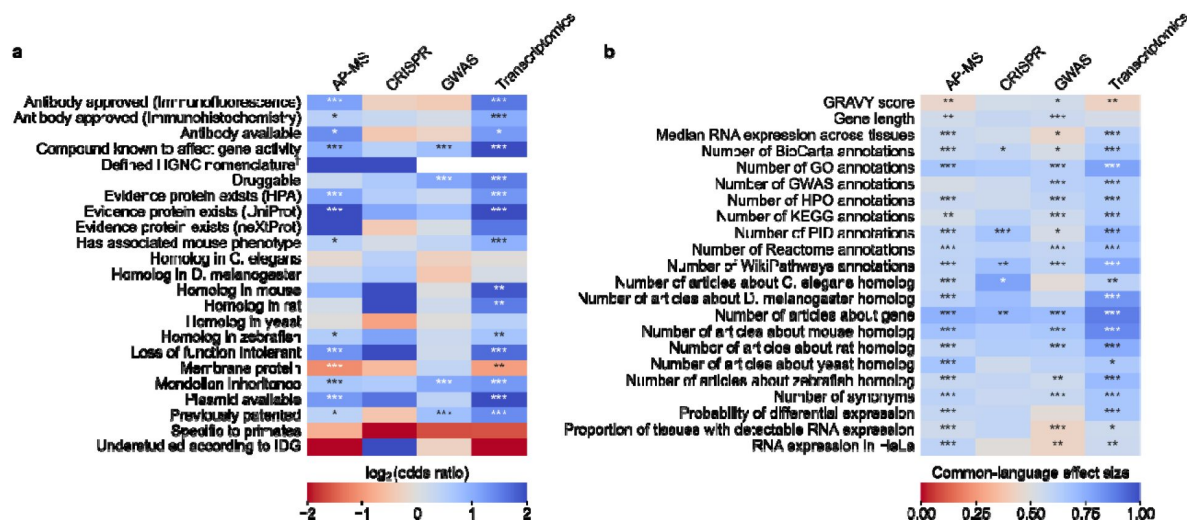


Figure 3

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**.

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

To promote the investigation of understudied genes, 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 genes³⁸⁻⁴⁵ or by providing information about hit genes^{7,46-50}. 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 S11** and **Figure S12**), some factors 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 S11**) 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 4A** 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 genes⁵¹. This led us to hypothesize that during vertebrate aging, the transcripts of long genes become downregulated in most tissues

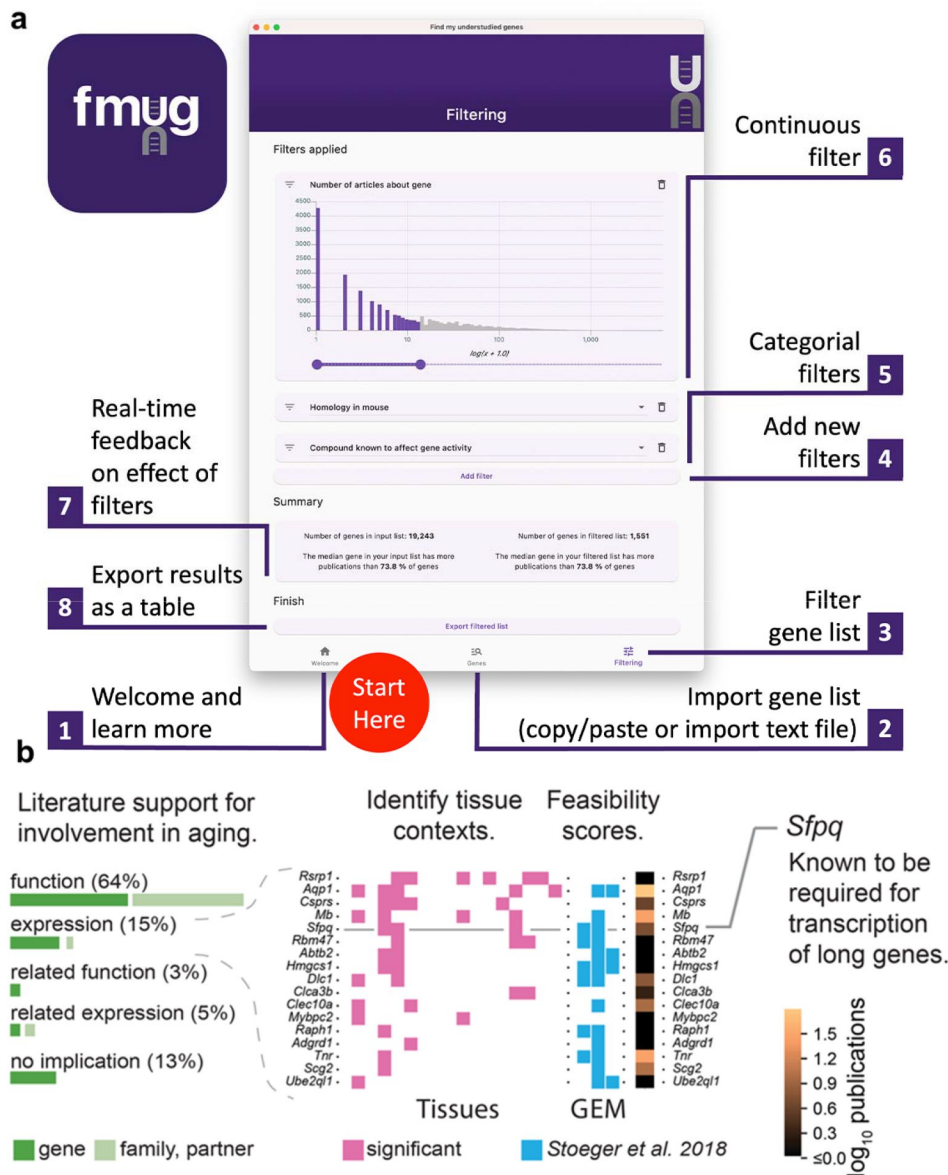


Figure 4

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 domain close to aging. For genes reported by others to change expression with age, we identified tissues in which transcripts of the genes change during aging. We computed 'feasibility scores' scientific strategies (GEM: G: strong genetic support, E: and experimental potential, M: homolog in invertebrate model organism) as described by Stoeger et al.^{7,10} and total number of publications in MEDLINE. Splicing factor, proline- and glutamine-rich (*Sfpq*) had previously been demonstrated by Takeuchi et al. to be required for the transcriptional elongation of long genes^{51,52}. When performing a data-driven analysis of factors that could possibly explain age-dependent changes of the entire transcriptome, we thus included gene and transcript lengths, and subsequently found them to be more informative than transcription factors or microRNAs^{52,53}.

(**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 2023⁵⁴ [↗](#).

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 expertise³ [↗](#), a perceived lack of accuracy of -omics studies¹⁷ [↗](#),⁵⁵ [↗](#), -omics serving research purposes beyond target gene identification²² [↗](#), the absence of good protocols for mass spectrometry⁵⁶ [↗](#), the electronic distribution and reading of research articles⁵⁷ [↗](#), rates of reproducibility¹⁷ [↗](#), career prospects of investigators⁷ [↗](#),⁵⁸ [↗](#), authors beginning manuscripts with something familiar before introducing something new⁵⁹ [↗](#), and the human tendency to fall back to simplifying heuristics when making decisions under conditions with uncertainty⁶⁰ [↗](#). 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 genes¹⁶ [↗](#),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 unknowomics 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 even if not investigated further by the authors. This intervention would need to be carefully designed since abstracts are limited in their size. Further, the observed discrepancy between the popularity of hits highlighted by GWAS versus other technologies suggests that some -omics technologies may be more powerful than others for characterizing understudied genes. This possibility merits further research and researchers participating in unknowomics should consider the relative strengths of each technology towards providing tractable results for follow-up.

We believe that enabling scientists to consciously engage with bias in research target selection will enable more biomedical researchers to participate in unknowomics, to the potential benefit of their own research impact and towards the advancement of our collective understanding of the entire human genome.

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 “-”.

Genes in title/abstract of primary research articles

Homo sapiens gene information was downloaded from NCBI Gene on Aug 16, 2022 [ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens/gene_info.gz]. gene2pubmed was download from NCBI Gene on Aug 16, 2022 [<ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz>]²⁸. PubTator gene annotations were downloaded from NIH-NLM on July 12, 2022 [<https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/>]^{28,67}. PubMed was downloaded on Dec 17, 2021 [<https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/>].

Only using PMIDs annotated as primary research articles, a human gene 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.

CRISPR articles

BioGRID ORCS²⁶ v1.1.6 was downloaded on April 25, 2022 [<https://downloads.thebiogrid.org/BioGRID-ORCS/Release-Archive/BIOGRID-ORCS-1.1.6/>]. Any genome-wide CRISPR knockout screens in human with an associated PubMed ID in which hit genes were mentioned in the title or abstract was considered (n = 15). 9,268 unique genes were found as hits. Of these, 18 (0.19%) were highlighted in titles/abstracts in the reporting articles and 19 (0.21%) were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**.

Transcriptomics articles

EBI-GXA²⁷ release 36 was downloaded on Sep 15, 2020 [<https://web.archive.org/web/20201022184159/https://www.ebi.ac.uk/gxa/download>]. This is the most recent release of EBI-GXA available as a bulk download. Any transcriptomics comparisons with an associated PubMed ID in which hit genes were mentioned in the title or abstract was considered (n = 148). Analysis was restricted to protein-coding genes (some screens featured non-protein-coding genes, but this was not common to all analyses). DE was called at Benjamini-Hochberg FDR $q < 0.05$. 18,295 unique genes were found as hits. Of these, 161 (0.88%) were highlighted in titles/abstracts in the reporting articles and 692 (3.78%) were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**.

Affinity purification–mass spectrometry articles

BioGRID²⁶ v3.5.186 was downloaded on April 25, 2022 [<https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.186/>]. Any interactions involving a human gene as the prey protein with an experimental evidence code of ‘Affinity Capture-MS’ labeled as ‘High-Throughput’ that had an associated PubMed ID in which hit genes were mentioned in the title or abstract was considered (n = 296). Prey proteins in these interactions were considered hits. 7,919 unique genes were found as hits. Of these, 311 (3.93%) were highlighted in titles/abstracts in reporting articles and 407 (5.14%) were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**.

GWAS articles

The NHGRI-EBI GWAS catalog²⁵ (associations and studies) was download on Aug 17, 2022 [<https://www.ebi.ac.uk/gwas/docs/file-downloads>]. Any GWAS screens with an associated PubMed ID in which hit genes were mentioned in the title or abstract was considered (n= 450). Only SNPs occurring within a gene were considered hits. 1,043 unique genes were found as hits. Of these, 413 (39.6%) were highlighted in titles/abstracts in reporting articles and 319 (30.6%) were highlighted in titles/abstracts in citing articles. A full list of PubMed IDs is available in **Table S6**

Citing articles

NIH iCite v32 was downloaded on Aug 25, 2022³⁰ [https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collection_/4586573/32].

Functional annotations

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\[Subtree\]](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 the most recent release of EBI-GXA available as a bulk download. For probability of DE, only RNA-seq comparisons were considered and DE was called at Benjamini-Hochberg $q < 0.05$.

Global RNA expression

RNA consensus tissue gene data from HPA release 21.1 was downloaded on Sep 20, 2022 [<https://www.proteinatlas.org/about/download>]. Global RNA expression was estimated by taking the median expression (nTPM) across tissues for each gene and the proportion of tissues with detectable (≥ 1 nTPM) expression for each gene.

Expression in HeLa cells

RNA cell line gene data from HPA release 21.1 was downloaded on Sep 20, 2022 [<https://www.proteinatlas.org/about/download>]. Expression is in nTPM.

Previous patent activity

Genes with patent activity were defined from Table S1 of Rosenfeld and Mason, 2013⁶⁸. Genes were mapped with their HGNC symbol. This analysis aligned sequences in patents to the human genome to estimate patent coverage of human coding sequences. Although this does not necessarily reflect whether the mapped genes were claimed directly by the patent holder, as noted by others⁶⁹, 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., 2017⁷⁰. 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., 2018⁷.

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 BioPython⁷¹.

Loss of function intolerance

Data was obtained from Karczewski et al.⁷², 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.⁷³.

Number of GWAS hits

EBI GWAS catalog²⁵ (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/?_id=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.

Data Availability

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

Acknowledgements

We thank Xiaojing Sui for testing FMUG and Northwestern Information Technology for technical assistance. RAKR was supported in part by the National Institutes of Health Training Grant (T32GM008449) through Northwestern University's Biotechnology Training Program. RAKR also acknowledges support from the Dr. John N. Nicholson fellowship from Northwestern University and Moderna Inc., "Identifying bias and improving reproducibility in RNA-seq computational pipelines". LANA was supported by NSF 1956338, NIH U19AI135964 and Simons Foundation DMS-1764421. TS was supported by NIH K99AG068544. We thank Alexander Misharin, Richard Morimoto, and Scott Budinger for feedback on an early prototype of FMUG which we used as part of our shared research into the biology of aging.

References

- 1 Hoffmann R., Valencia A. (2003) **Life cycles of successful genes** *Trends Genet* **19**:79–81 [https://doi.org/10.1016/S0168-9525\(02\)00014-8](https://doi.org/10.1016/S0168-9525(02)00014-8)
- 2 Su A. I., Hogenesch J. B. (2007) **Power-law-like distributions in biomedical publications and research funding** *Genome Biol* **8** <https://doi.org/10.1186/gb-2007-8-4-404>
- 3 Edwards A. M., et al. (2011) **Too many roads not taken** *Nature* **470**:163–165 <https://doi.org/10.1038/470163a>
- 4 Gillis J., Pavlidis P. (2013) **Assessing identity, redundancy and confounds in Gene Ontology annotations over time** *Bioinformatics* **29**:476–482 <https://doi.org/10.1093/bioinformatics/bts727>
- 5 Stoeger T., Nunes Amaral L. A. (2022) **The characteristics of early-stage research into human genes are substantially different from subsequent research** *PLoS Biol* **20** <https://doi.org/10.1371/journal.pbio.3001520>
- 6 Grueneberg D. A., et al. (2008) **Kinase requirements in human cells: I. Comparing kinase requirements across various cell types** *Proc. Natl. Acad. Sci. U. S. A* **105**:16472–16477 <https://doi.org/10.1073/pnas.0808019105>
- 7 Stoeger T., Gerlach M., Morimoto R. I., Nunes Amaral L. A. (2018) **Large-scale investigation of the reasons why potentially important genes are ignored** *PLoS Biol* **16** <https://doi.org/10.1371/journal.pbio.2006643>
- 8 Riba M., et al. (2016) **Revealing the acute asthma ignorome: characterization and validation of uninvestigated gene networks** *Sci Rep* **6** <https://doi.org/10.1038/srep24647>
- 9 Haynes W. A., Tomczak A., Khatri P. (2018) **Gene annotation bias impedes biomedical research** *Sci Rep* **8** <https://doi.org/10.1038/s41598-018-19333-x>
- 10 Border R., et al. (2019) **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
- 11 Stoeger T., Nunes Amaral L. A. (2020) **COVID-19 research risks ignoring important host genes due to pre-established research patterns** *Elife* **9** <https://doi.org/10.7554/eLife.61981>
- 12 Zhang D., et al. (2020) **Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders** *Science Advances* **6**
- 13 Byrne J. A., et al. (2022) **Protection of the human gene research literature from contract cheating organizations known as research paper mills** *Nucleic Acids Research* **50**:12058–12070 <https://doi.org/10.1093/nar/gkac1139>
- 14 Collins F. S., Green E. D., Guttmacher A. E., Guyer M. S. (2003) **Institute, U. S. N. H. G. R. A vision for the future of genomics research** *Nature* **422**:835–847 <https://doi.org/10.1038/nature01626>

- 15 Shendure J., Findlay G. M., Snyder M. W. (2019) **Genomic Medicine-Progress, Pitfalls, and Promise** *Cell* **177**:45–57 <https://doi.org/10.1016/j.cell.2019.02.003>
- 16 Lloyd K. C. K., et al. (2020) **The Deep Genome Project** *Genome Biol* **21** <https://doi.org/10.1186/s13059-020-1931-9>
- 17 Kustatscher G., et al. (2022) **Understudied proteins: opportunities and challenges for functional proteomics** *Nat Methods* **19**:774–779 <https://doi.org/10.1038/s41592-022-01454-x>
- 18 Rodriguez-Esteban R., Jiang X. (2017) **Differential gene expression in disease: a comparison between high-throughput studies and the literature** *BMC Med Genomics* **10** <https://doi.org/10.1186/s12920-017-0293-y>
- 19 Oprea T. I., et al. (2018) **Unexplored therapeutic opportunities in the human genome** *Nat Rev Drug Discov* **17**:317–332 <https://doi.org/10.1038/nrd.2018.14>
- 20 Sinha S., Eisenhaber B., Jensen L. J., Kalbuajji B., Eisenhaber F. (2018) **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** <https://doi.org/10.1002/pmic.201800093>
- 21 Wood V., et al. (2019) **Hidden in plain sight: what remains to be discovered in the eukaryotic proteome?** *Open Biol* **9** <https://doi.org/10.1098/rsob.180241>
- 22 Donohue C., Love A. **Perspectives on the Human Genome Project and genomics**
- 23 Pena-Castillo L., Hughes T. R. (2007) **Why are there still over 1000 uncharacterized yeast genes?** *Genetics* **176**:7–14 <https://doi.org/10.1534/genetics.107.074468>
- 24 Ellens K. W., et al. (2017) **Confronting the catalytic dark matter encoded by sequenced genomes** *Nucleic Acids Res* **45**:11495–11514 <https://doi.org/10.1093/nar/gkx937>
- 25 Buniello A., et al. (2019) **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019** *Nucleic Acids Res* **47**:D1005–D1012 <https://doi.org/10.1093/nar/gky1120>
- 26 Oughtred R., et al. (2021) **The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions** *Protein Science* **30**:187–200
- 27 Papatheodorou I., et al. (2018) **Expression Atlas: gene and protein expression across multiple studies and organisms** *Nucleic acids research* **46**:D246–D251
- 28 Maglott D., Ostell J., Pruitt K. D., Tatusova T. (2007) **Entrez Gene: gene-centered information at NCBI** *Nucleic acids research* **35**:D26–D31
- 29 Wei C.-H., Allot A., Leaman R., Lu Z. (2019) **PubTator central: automated concept annotation for biomedical full text articles** *Nucleic acids research* **47**:W587–W593
- 30 Hutchins B. I., Santangelo George. iCite (2019) **Hutchins, B. I., Santangelo George. iCite, <10.35092/yhjc.c.4586573> (2019).** <https://doi.org/10.35092/yhjc.c.4586573>
- 31 Stoeger T., Nunes Amaral L. A. (2020) **COVID-19 research risks ignoring important host genes due to pre-established research patterns** *Elife* **9**

- 32 Stoecker T., Gerlach M., Morimoto R. I., Nunes Amaral L. A. (2018) **Large-scale investigation of the reasons why potentially important genes are ignored** *PLoS biology* **16**
- 33 Pfeiffer T., Hoffmann R. (2007) **Temporal patterns of genes in scientific publications** *Proc. Natl. Acad. Sci. U. S. A* **104**:12052–12056 <https://doi.org/10.1073/pnas.0701315104>
- 34 Knorr Cetina K. (1999) **Epistemic Cultures**
- 35 Alberts B. M. (1985) **Limits to growth: In biology, small science is good science** *Cell* :337–338
- 36 Richardson S. S. a. S. (2015) **Hallam**
- 37 Lopes I., Altab G., Raina P., De Magalhães J. P. (2021) **Gene size matters: an analysis of gene length in the human genome** *Frontiers in Genetics* **12**
- 38 Duek P., Gateau A., Bairoch A., Lane L. (2018) **Exploring the Uncharacterized Human Proteome Using neXtProt** *J Proteome Res* **17**:4211–4226 <https://doi.org/10.1021/acs.jproteome.8b00537>
- 39 Crow M., Lim N., Ballouz S., Pavlidis P., Gillis J. (2019) **Predictability of human differential gene expression** *Proc Natl Acad Sci U S A* **116**:6491–6500 <https://doi.org/10.1073/pnas.1802973116>
- 40 Perdigao N., Rosa A. (2019) **Dark Proteome Database: Studies on Dark Proteins** *High Throughput* **8** <https://doi.org/10.3390/ht8020008>
- 41 Essegian D., Khurana R., Stathias V., Schurer S. C. (2020) **The Clinical Kinase Index: A Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer** *Cell Rep Med* **1** <https://doi.org/10.1016/j.xcrm.2020.100128>
- 42 Sheils T. K., et al. (2021) **TCRD and Pharos 2021: mining the human proteome for disease biology** *Nucleic Acids Res* **49**:D1334–D1346 <https://doi.org/10.1093/nar/gkaa993>
- 43 Rocha J., et al. (2022) **Functional unknowns: closing the knowledge gap to accelerate biomedical research** *bioRxiv*
- 44 Higgins D. P., Weisman C. M., Lui D. S., D'Agostino F. A., Walker A. K. (2022) **Defining characteristics and conservation of poorly annotated genes in *Caenorhabditis elegans* using WormCat 2.0** *Genetics* **221** <https://doi.org/10.1093/genetics/iyac085>
- 45 Wainberg M., et al. (2021) **A genome-wide atlas of co-essential modules assigns function to uncharacterized genes** *Nat Genet* **53**:638–649 <https://doi.org/10.1038/s41588-021-00840-z>
- 46 Rebhan M., Chalifa-Caspi V., Prilusky J., Lancet D. (1998) **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support** *Bioinformatics* **14**:656–664 <https://doi.org/10.1093/bioinformatics/14.8.656>
- 47 Tan J., et al. (2017) **ADAGE signature analysis: differential expression analysis with data-defined gene sets** *Bmc Bioinformatics* **18** <https://doi.org/10.1186/s12859-017-1905-4>
- 48 Kustatscher G., et al. (2019) **Co-regulation map of the human proteome enables identification of protein functions** *Nat Biotechnol* **37**:1361–1371 <https://doi.org/10.1038/s41587-019-0298-5>

- 49 Wu T., et al. (2021) **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data** *Innovation (Camb)* **2** <https://doi.org/10.1016/j.xinn.2021.100141>
- 50 Jiang J., et al. (2022) **Systematic illumination of druggable genes in cancer genomes** *Cell Rep* **38** <https://doi.org/10.1016/j.celrep.2022.110400>
- 51 Takeuchi A., et al. (2018) **Loss of Sfpq Causes Long-Genes Transcriptopathy in the Brain** *Cell Rep* **23**:1326–1341 <https://doi.org/10.1016/j.celrep.2018.03.141>
- 52 Stoeger T., et al. (2022) **Aging is associated with a systemic length-associated transcriptome imbalance** *Nature Aging* **2**:1191–1206 <https://doi.org/10.1038/s43587-022-00317-6>
- 53 Gyenis A., et al. (2023) **Genome-wide RNA polymerase stalling shapes the transcriptome during aging** *Nat Genet* **55**:268–279 <https://doi.org/10.1038/s41588-022-01279-6>
- 54 Ibañez-Solé O., Barrio I., Izeta A. (2023) **Age or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression** *iScience* <https://doi.org/10.1016/j.isci.2023.106368>
- 55 Brown L. A., Peirson S. N. (2018) **Improving Reproducibility and Candidate Selection in Transcriptomics Using Meta-analysis** *Journal of Experimental Neuroscience* **12** <https://doi.org/10.1177/1179069518756296>
- 56 Cesar-Razquin A., et al. (2015) **A Call for Systematic Research on Solute Carriers** *Cell* **162**:478–487 <https://doi.org/10.1016/j.cell.2015.07.022>
- 57 Evans J. A. (2008) **Electronic publication and the narrowing of science and scholarship** *Science* **321**:395–399 <https://doi.org/10.1126/science.1150473>
- 58 Alberts B., Kirschner M. W., Tilghman S., Varmus H. (2014) **Rescuing US biomedical research from its systemic flaws** *Proc Natl Acad Sci U S A* **111**:5773–5777 <https://doi.org/10.1073/pnas.1404402111>
- 59 Uzzi B., Mukherjee S., Stringer M., Jones B. (2013) **Atypical combinations and scientific impact** *Science* **342**:468–472 <https://doi.org/10.1126/science.1240474>
- 60 Gilovich T., Griffin D. W., Kahneman D. (2002) **Heuristics and biases** *the psychology of intuitive judgment*
- 61 Gerlt J. A., et al. (2011) **The Enzyme Function Initiative** *Biochemistry* **50**:9950–9962 <https://doi.org/10.1021/bi201312u>
- 62 Carter A. J., et al. (2019) **Target 2035: probing the human proteome** *Drug Discov Today* **24**:2111–2115 <https://doi.org/10.1016/j.drudis.2019.06.020>
- 63 Rodgers G., et al. (2018) **Glimmers in illuminating the druggable genome** *Nat Rev Drug Discov* **17**:301–302 <https://doi.org/10.1038/nrd.2017.252>
- 64 Kustatscher G., et al. (2022) **An open invitation to the Understudied Proteins Initiative** *Nat Biotechnol* **40**:815–817 <https://doi.org/10.1038/s41587-022-01316z>
- 65 EUBOPEN **EUBOPEN**, <<https://www.eubopen.org/>>(<

- 66 Rocha J. J., et al. (2023) **Functional unknowns: Systematic screening of conserved genes of unknown function** *PLoS biology* **21**
- 67 Wei C. H., Allot A., Leaman R., Lu Z. (2019) **PubTator central: automated concept annotation for biomedical full text articles** *Nucleic Acids Res* **47**:W587–W593 <https://doi.org/10.1093/nar/gkz389>
- 68 Rosenfeld J. A., Mason C. E. (2013) **Pervasive sequence patents cover the entire human genome** *Genome Med* **5** <https://doi.org/10.1186/gm431>
- 69 Tu S., et al. (2014) **Response to ‘pervasive sequence patents cover the entire human genome’** *Genome medicine* **6**:1–3
- 70 Finan C., et al. (2017) **The druggable genome and support for target identification and validation in drug development** *Science translational medicine* **9**
- 71 Cock P. J. A., et al. (2009) **Biopython: freely available Python tools for computational molecular biology and bioinformatics** *Bioinformatics* **25**:1422–1423 <https://doi.org/10.1093/bioinformatics/btp163>
- 72 Karczewski K. J., et al. (2020) **The mutational constraint spectrum quantified from variation in 141,456 humans** *Nature* **581**:434–443 <https://doi.org/10.1038/s41586-020-2308-7>
- 73 Lek M., et al. (2016) **Analysis of protein-coding genetic variation in 60,706 humans** *Nature* **536**:285–291 <https://doi.org/10.1038/nature19057>
- 74 Kendall M. G., Stuart A. (1973) **Kendall, M. G. & Stuart, A. Inference and Relationship The Advanced Theory of Statistics 2 (1973).** *Inference and Relationship The Advanced Theory of Statistics 2*
- 75 Ward J. H. (1963) **Hierarchical grouping to optimize an objective function** *Journal of the American statistical association* **58**:236–244

Article and author information

Reese AK Richardson

Interdisciplinary Biological Sciences, Northwestern University, Department of Chemical and Biological Engineering, Northwestern University
ORCID iD: [0000-0002-6058-5886](https://orcid.org/0000-0002-6058-5886)

Heliodoro Tejedor Navarro

Department of Chemical and Biological Engineering, Northwestern University, Northwestern Institute on Complex Systems, Northwestern University
ORCID iD: [0000-0001-5441-8101](https://orcid.org/0000-0001-5441-8101)

Luis A Nunes Amaral

Department of Chemical and Biological Engineering, Northwestern University, Northwestern Institute on Complex Systems, Northwestern University, Department of Physics and Astronomy, Northwestern University, Department of Molecular Biosciences, Northwestern University

For correspondence: amaral@northwestern.edu

ORCID iD: [0000-0002-3762-789X](https://orcid.org/0000-0002-3762-789X)

Thomas Stoeger

Department of Chemical and Biological Engineering, Northwestern University, The Potocsnak Longevity Institute, Northwestern University, Simpson Querrey Lung Institute for Translational Science, Northwestern University

For correspondence: thomas.stoeger@northwestern.edu

ORCID iD: [0000-0002-5540-4278](https://orcid.org/0000-0002-5540-4278)

Copyright

© 2023, Richardson et al.

This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor

Peter Rodgers

eLife, Cambridge, United Kingdom

Senior Editor

Peter Rodgers

eLife, Cambridge, United Kingdom

Reviewer #1 (Public Review):

The authors have addressed most of the concerns I had about the original version in this revised version.

<https://doi.org/10.7554/eLife.93429.2.sa2>

Reviewer #2 (Public Review):

The authors have successfully addressed all of the concerns I had about the original version.

<https://doi.org/10.7554/eLife.93429.2.sa1>

Reviewer #3 (Public Review):

The message conveyed by figure 1b is now clearer, but could still be improved. The authors explained the meaning of this figure well in their response to the reviewers: "For example, the results for CRISPR were obtained from 15 focus studies (original research) and 18 subsequent studies (papers citing focus articles). Those 15 studies identified 9,268 genes where loss-of-function changed phenotypes but, in their titles and abstracts, mentioned only 18 of those 9,268 genes. While the 9,268 hit genes have received similar research attention to the entirety of protein-coding genes, the 18 hit genes mentioned in the title or abstract are significantly more well studied. The articles citing the focus articles also only mentioned in their titles and abstracts 19 highly studied hit genes".

The new Figure S8 is good.

<https://doi.org/10.7554/eLife.93429.2.sa0>

Author Response

The following is the authors' response to the original reviews.

eLife Assessment

This study investigated the factors related to understudied genes in biomedical research. It showed that understudied genes are largely abandoned at the writing stage, and it identified a number of biological and experimental factors that influence which genes are selected for investigation. The study is a valuable contribution to this branch of meta-research, and while the evidence in support of the findings is solid, the interpretation and presentation of the results (especially the figures) needs to be improved.

We thank the editor and reviewers for their detailed and thoughtful assessment of our work. Below, we present detailed responses to reviewers' comments and suggestions. We are also submitting a version edited for clarity of presentation and precision of interpretation.

Following the eLife assessment, we also tried to identify further statements where results could be presented in a more precise way.

First, in the section Subsequent reception by other scientists does not penalize studies on understudied genes, we now state "This result again opposes the hypothesis that less-investigated genes will yield articles with lower impact."

Second, in section Identification of biological and experimental factors associated with selection of highlighted genes, we now state:

"We cautiously hypothesize that this might reflect on many different research groups producing reagents surrounding the genes that they actively study. The most informative continuous factor is the number of research articles about a gene (Figure 1B).", removing claims of causality.

Finally, for improved readability, we have moved all supplemental tables into separate .xlsx files.

Reviewer #1 (Public Review):

Summary and strengths

The authors tried to address why only a subset of genes are highlighted in many publications. Is it because these highlighted genes are more important than others? Or is it because there are non-genetic reasons? This is a critical question because in the effort to discover new genes for drug targets and clinical benefit, we need to expand a pool of genes for deep analyses. So I appreciate the authors' efforts in this study, as it is timely and important. They also provided a framework called FMUG (short for Find My Understudied Gene) to evaluate genes for a number of features for subsequent analyses.

We thank the reviewer for their insightful comments and are pleased that the reviewer shares our appreciation for the gravity of these questions. As the reviewer emphasizes, it is critical to understand whether the choice of genes reflects their importance or non-genetic reasons. Previously we and others demonstrated that this choice does not reflect biological importance, when the latter is assessed through unbiased genome-wide data (e.g.: Haynes et al., 2018; Stoeger et al. 2018). Now we contribute to this critical question by systematically evaluating individual non-genetic reasons. We address the reviewer's comments below.

Weaknesses

Many of the figures are hard to comprehend, and the figure legends do not sufficiently explain them.

For example, what was plotted in Fig 1b? The number of articles increased from results -> write-ups -> follow-ups in all four categories with different degrees. But it does not seem to match what the authors meant to deliver.

We apologize for the lack of clarity. We identified two interrelated elements that we have now fixed: i) the prior figure legend provided for each genomics approach n number of articles, such as “GWAS (n=450 articles)””; ii) the prior y-axis was labelled “Number of articles”.

Addressing the first element, we now rephrased the legend for clarity:

“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.”

The number of genes in each box plot is now reported in the x-axis labels for each step. For example, the results for CRISPR were obtained from 15 focus studies (original research) and 18 subsequent studies (papers citing focus articles). Those 15 studies identified 9,268 genes where loss-of-function changed phenotypes but, in their titles and abstracts, mentioned only 18 of those 9,268 genes. While the 9,268 hit genes have received similar research attention to the entirety of protein-coding genes, the 18 hit genes mentioned in the title or abstract are significantly more well studied. The articles citing the focus articles also only mentioned in their titles and abstracts 19 highly studied hit genes.

Addressing the second element, we updated the axis label to “Number of articles about gene”, to distinguish it from number of articles mentioned in the legend, convey that this is the number of articles about each gene that were published independently of the genomics assays we inspect. To further underscore this point we now label the “20% highest-studied genes” that we mention in the main text, and reworded the figure caption to better capture where the critical increase occurs: “A shift in focus towards well-studied genes occurs during the summarization and write-up of results and remains in subsequent studies.”.

Fig 4 is also confusing. It appears that the genes were clustered by many features that the authors developed. But does it have any relationship with genes being under- or over-studied?

We again apologize for the lack of clarity. As is described in the main text, while the results of Figs. 1-2 suggest that gene popularity may be predict the highlighting of a differentially expressed gene in the title or abstract, we want to conduct a systematically analysis of the factors that correlate with such a decision. We thus build a set of 45 factors that have been discussed as factors explaining why some genes receive increased research attention.

The data in Fig. 4 shows that those 45 factors are not independent but that some are highly correlated. Because of those correlations, we are able to select a smaller number as representative of the full set. Those are the default factors shown to users of FMUG. While users can choose all factors that are significantly correlated with the highlighting in title or

abstract, the default of presenting factors representing different clusters of factors enabled us to limit the number of factors that are initially displayed.

Please note that following the suggestion of Reviewer 3, we have now moved this Figure to the supplemental material, as Figure S11.

Reviewer #2 (Public Review)

Summary and strengths

In this manuscript the authors analyse the trajectory of understudied genes (UGs) from experiment to publication and study the reasons for why UGs remain underrepresented in the scientific literature. They show that UGs are not underrepresented in experimental datasets, but in the titles and abstracts of the manuscripts reporting experimental data as well as subsequent studies referring to those large-scale studies. They also develop an app that allows researchers to find UGs and their annotation state. Overall, this is a timely article that makes an important contribution to the field. It could help to boost the future investigation of understudied genes, a fundamental challenge in the life sciences. It is concise and overall well-written, and I very much enjoyed reading it. However, there are a few points that I think the authors should address.

We thank the reviewer for their kind assessment.

Weaknesses

The authors conclude that many UGs "are lost" from genome-wide assay at the manuscript writing stage. If I understand correctly, this is based on gene names not being reported in the title or abstract of these manuscripts. However, for genome-wide experiments, it would be quite difficult for authors to mention large numbers of understudied genes in the abstract. In contrast, one might highlight the expected behaviour of a well-studied protein simply to highlight that the genome-wide study provides credible results.

We agree that it is not reasonable to expect a title or abstract to highlight hundreds or even thousands of differentially expressed genes. We've now extended our Study Limitations section to address this:

"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."

We also agree that highlighting the expected behavior of a well-studied protein may provide credibility to a study and increase confidence on other results. The soundness of such a strategy was quantitatively studied in a study by Uzzi et al. (Science 2013), which we now include in the section on study limitations as:

"authors beginning manuscripts with something familiar before introducing something new".

To convey the practical limitation of abstracts needing to be concise, we added the following sentence to our discussion section, when suggesting controlled trials that add genes to abstracts:

"This intervention would need to be carefully designed since abstracts are limited in their size."

To avoid over-interpretation we have in the discussion also extended the sentence on “lost in a leaky pipeline” to “lost to titles and abstracts of research articles in a leaky pipeline”.

Our focus on titles and abstracts has been equally motivated by their availability (full text still is often behind paywalls and/or not accessible for bulk-download and text-mining) and by abstracts being the most visible and most read parts of research articles (e.g.: bioRxiv estimates that for the preprint for the present manuscript, the abstract was read ~10 times more frequently than full-text HTML and 4 times more frequently than the pdf).

Could this bias the authors' conclusions and, if so, how could this be addressed? For example, would it be worth to normalise studies based on the total number of genes they cover?

We previously described that – in line with the reviewer’s expectations – unstudied genes are preferentially added to the title or abstract of articles that feature more genes in the title or abstract (Stoeger et al., Plos Biology, 2022; Fig. 2B). Normalizing by the total number of genes should thus preserve the pronounced division between well-studied genes and unstudied genes show in Figure 1B. In line with these predictions, we randomly select one gene per title/abstract and find that the effect remains (see new Figure S7).

Author response image 1.

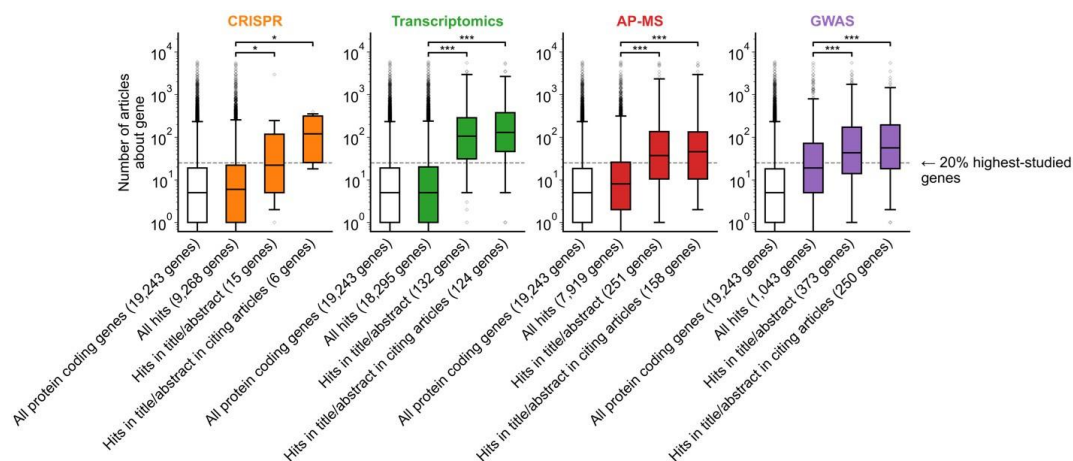
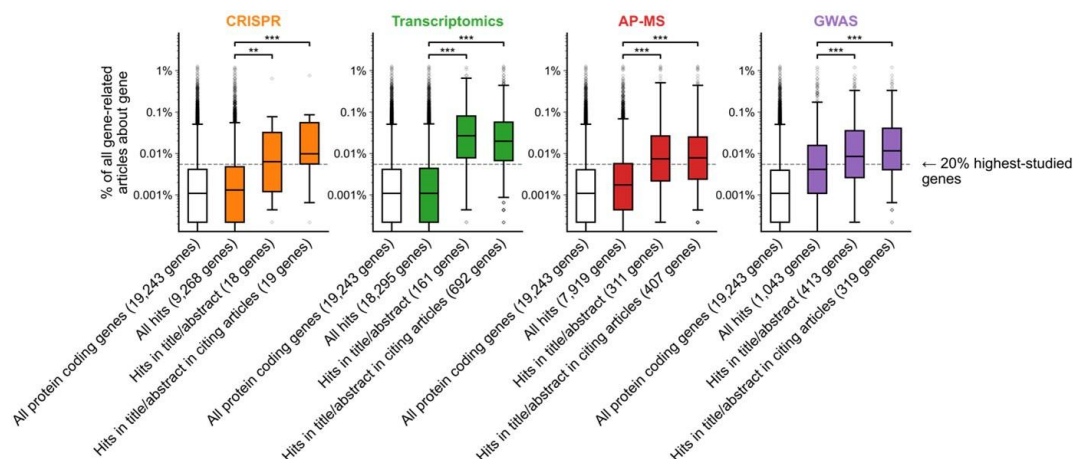


Figure 1B is confusing in its present form. I think the plot and/or the legend need revising. For example, what "numbers to the right of each box plot" are the authors referring to? Also, I assume that the filled boxes are understudied genes and the empty/white box is "all genes", but that's not explained in the legend. In the main text, the figure is referred to with the sentence "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 ". I cannot follow how the figure shows this. My interpretation is that the y-axis is not showing the number of articles, but represents the percentage of articles mentioning a gene in the title/abstract, displayed on a log scale. If so, perhaps a better axis labels and legend text could be sufficient. But then one would also need to somehow connect this to the statement in the main text about the 20% highest-studied genes (a dashed line?). Alternatively, the authors could consider other ways of plotting these data, e.g. simply plotting the "% of publication in which a gene appears" from 0-100% or so.

Reviewer 1 raised a similar point on overall figure clarity. We identified two interrelated elements that contribute to overall confusion and have now fixed them (see response to Reviewer 1 beginning on page 2 of this document).

We attempted an alternative plotting of Fig 1B according to the reviewer's suggestion. In the version below, the y-axis instead shows the percent of gene-related articles that are about each gene. We chose to keep the original y-axis (showing number of articles about each gene) as it additionally conveys the absolute scale of scholarship on individual genes.

Author response image 2.



Reviewer #3 (Public Review):

Summary and strengths

The manuscript investigated the factors related to understudied genes in biomedical research. It showed that understudied are largely abandoned at the writing stage and identified biological and experimental factors associated with selection of highlighted genes.

It is very important for the research community to recognize the systematic bias in research of human genes and take precautions when designing experiments and interpreting results. The authors have tried to profile this issue comprehensively and promoted more awareness and investigation of understudied genes.

We thank the reviewer for their kind assessment of our work.

Weaknesses

Regarding result section 1 "Understudied genes are abandoned at synthesis/writing stage", the figures are not clear and do not convey the messages written in the main text. For example, in Figure 1B, figure S5 and S6,

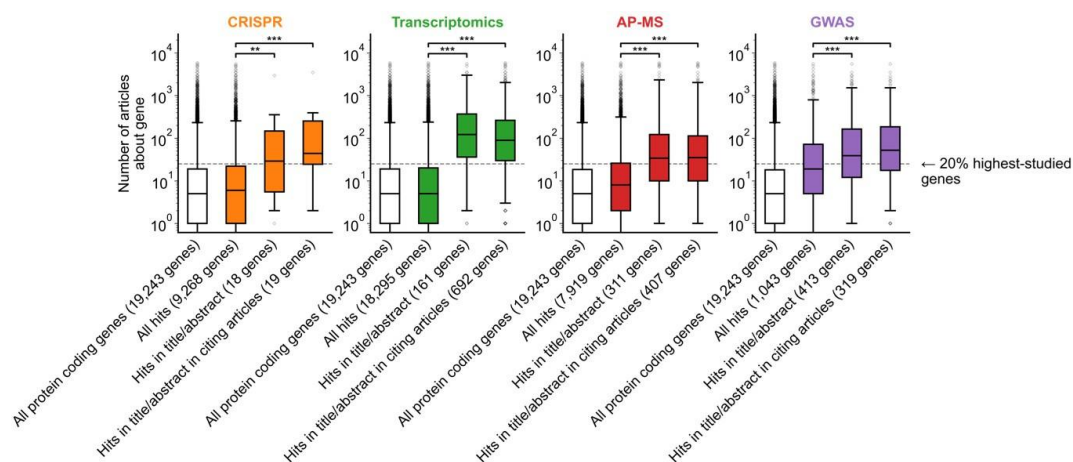
- *There is no "numbers to the right of each box plot".*

The "numbers to the right" statement in the caption was an erroneous inclusion from an earlier version of the figure. We apologize for our error and have now removed this statement.

- Do these box plots only show understudied genes? How many genes are there in each box plot? The definition and numbers of understudied genes are not clear.

The x-axis describes genes featured in each stage of the publication process (from all protein-coding genes to genes found as hits in genome-wide screen to genes found in the title/abstract to genes found in the title/abstract of citing articles) and the y-axis describes the number of articles annotated to those genes. We have also now added the number of genes in each box plot to the figure. This information is also in Materials and Methods under each technology's heading (see also response to Reviewer 1 beginning on page 2 of this document).

Author response image 3.



- "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)". This is not clear from the figure.

We have revised Figure 1B and its caption to better communicate the main point of the figure: that genes which make it to the title/abstract of the reporting article tend to be more popular than genes which are hits in genome-wide experiments from those articles. We have added a horizontal line that shows the cutoff for the top 20% most popular genes.

Regarding result section 2 "Subsequent reception by other scientists does not penalize studies on understudied genes", the authors showed in figure 2 that there is a negative correlation between articles per gene before 2015 and median citations to articles published in 2015. Another explanation could be that for popular genes, there are more low-quality articles that didn't get citations, not necessarily that less popular genes attract more citations.

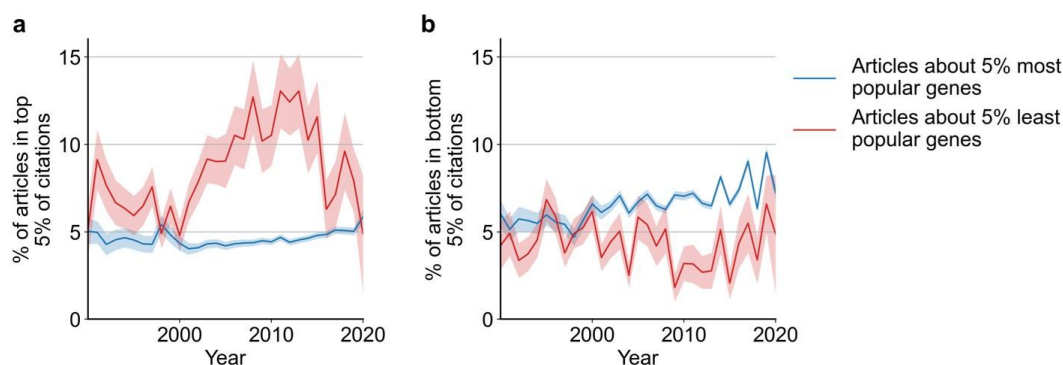
We believe that both explanations for the observed phenomenon are not mutually exclusive. Previously, we focused on the median of citations to articles about a gene to capture the typical effect. In a new analysis, we also find support for the possibility outlined by the reviewer and believe that adding this to our manuscript complements and balances our analysis of citations. Specifically, in the new Figure S8B we find that most popular genes are slightly more likely to be among least cited papers (and in Figure S8A that the least studied genes have been much more likely to be among the most cited papers). In-text, we state:

“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)”.

We thank the reviewer for their suggestion, which strengthens our manuscript. The figure caption reads:

“Figure S8: Likelihoods of being highly cited (top 5% of citations among all articles about genes, panel a) or lowly cited (bottom 5% of citations among all articles about genes, panel b) for articles about the most popular genes (top 5% accumulated articles) versus articles about the least popular genes (bottom 5% accumulated articles) by year of publication. Only articles with a single gene in the title/abstract are considered. Shaded regions show ± 1 standard error of the proportion.”

Author response image 4.



Regarding result section 3 "Identification of biological and experimental factors associated with selection of highlighted genes", in Figure 3 and table s2, the author stated that "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". The number 5.144 comes out of nowhere both in the figure and the table. In addition, figure 4 is not informative enough to be included as a main figure.

This is the result of both a typo and imprecise terminology. The number should read 4.262 (the likelihood ratio of being mentioned in the title/abstract between genes with and without a compound), which corresponds to an odds ratio of 4.331. We have clarified this in the table caption, stating:

“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”.

We have removed Figure 4 as a main-text figure and added a version, with revised color scheme along comments of Reviewer 1, as Figure S11. We added to the figure caption “Bold indicates FMUG ‘s default factors, which we selected based on this clustering and based on their strength of association with gene selection (Figure 3, Table S2 and Table S3).”

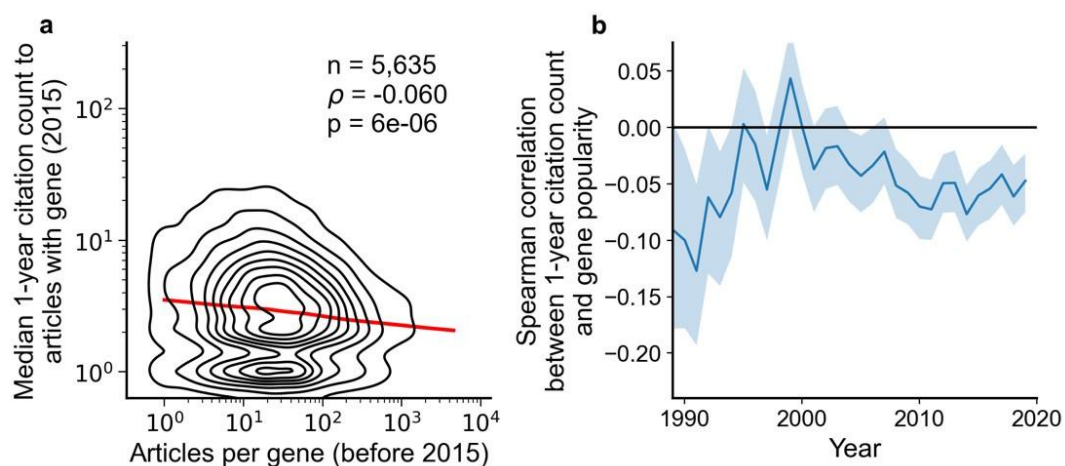
Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

- Fig 2a shows that papers highlighting understudied genes are actually cited more. I wonder why authors only looked at data before 2015. Fig 2b shows an increased correlation since 2015. Please consider redrawing Fig 2a to include data from 2015-2020?

We highlight data from 2015 since, from our used version of iCite (v32, released July 2022, covering citations made through most of 2021), papers published in 2015 have had about 6 years to accumulate citations. With fewer years to accumulate citations, insufficient signal may cause correlation to converge toward zero. Below, we repeat the analysis in Figure 2 but only considering citations made within a year of an article's publication, which substantially reduces correlation (although remaining significant).

Author response image 5.



We added a note to the figure caption:

“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.”

For Figure 2B, we add:

“For more recent years, where articles have had less time to accumulate citations, insufficient signal may cause correlation to converge toward zero.”

- Can FMUG be posted on the web for easy access by researchers with non-computational backgrounds?"

We presently regrettably do not have the resources to create or maintain a web-based version. We hope that the publication of this manuscript will enable us to attract resources to create and maintain a web-based version.

Reviewer #2 (Recommendations for the authors):

- Related to the first weakness in my public review: The observed disparity between CRISPR and GWAS study in terms of which genes they promote to the abstract is interesting. I wonder if this has to do with the application of these techniques. GWAS studies will often highlight that they retrieve known associations between a

gene and a phenotype, to show that a screen is working. I guess often the point is to subsequently identify more genes associated with a particular phenotype, but often it is unclear how to validate/verify newly found associations. In contrast, CRISPR screens might be more focussed on functionally/mechanistically understanding unknown processes, e.g. observing a phenotype that appears/disappears in response to a gene deletion. In such studies, the follow-up of a previously unknown gene could be more straightforward and relevant to the outcome. Does that mean CRISPR screens are better than GWAS studies for addressing the UG problem? Perhaps the authors could briefly discuss this issue.

The number of studies we included featuring CRISPR screens is relatively small ($n = 15$ compared to $n = 450$ for GWAS). Thus, it is not possible to conclude in a statistically sound manner whether authors of CRISPR screens are truly more likely to highlight understudied genes.

However, the reviewer raises compelling reasons for why this might be the case, and we now embed the broader discussion point that some techniques might be more powerful toward understudied genes.

The discussion now includes:

“Further, the observed discrepancy between the popularity of hits highlighted by GWAS versus other technologies suggests that some -omics technologies may be more powerful than others for characterizing understudied genes. This possibility merits further research and researchers participating in unknomics should consider the relative strengths of each technology towards providing tractable results for follow-up.”

- *Affinity capture mass spectrometry (Aff-MS): Perhaps I misunderstood this, but typically this is referred to as affinity purification MS (AP-MS)*

Thank you for the clarification. We have changed ‘Aff-MS’ to ‘AP-MS’ throughout the manuscript.

- *Page 3, line 96. The sentence "The first possibility is that seemingly understudied genes are, in fact, not understudied as they would rarely be identified through experiments.". Would they not still be understudied, just not intentionally?*

We have rephrased this sentence to:

“The first possibility is that some genes are less studied because they are rarely identified as hits in experiments.”

- *Fig 4 is very interesting, but I also found it a bit confusing. First, the choice of colour scheme, where blue shows the absence and white shows the presence of something, seems counterintuitive, especially on a white background. Second, I find it confusing that only some of the experiments are labelled in the heatmap. Could the authors not simply use Fig S9 as Fig 4? Or alternatively, only include the 8 labelled factors in the simplified figure.*

In line with this feedback and that of Review #1 and #3, we have removed Figure 4 as a main-text figure and instead include this figure as Supplementary Figure S11. We have reversed the color scheme so that purple indicates one and white indicates zero. We also now label all factors. Previously we had only listed the default features of FMUG. We also now updated the figure legend to convey how it assisted the choice of default factors in FMUG. It reads:

“Bold indicates FMUG ‘s default factors, which we selected based on this clustering and based on their strength of association with gene selection (Figure 3, Table S2 and Table S3)”.

- *The FMUG app is fantastic and sounds exactly like something that is required to boost the visibility of understudied genes and overcome the understudied gene bias. However, I did not understand the choice of reporting this in the Discussion section.*

We thank the reviewer for their enthusiasm, and have now moved FMUG into the results section.

- To further increase usability of the FMUG app, is there a way it could be deployed online? I appreciate this could require a major amount of coding work, which would not be reasonable to demand. So please consider this a suggestion, potentially for a future implementation.

We presently regretfully do not have the resources to create or maintain a web-based version. We hope that the publication of this manuscript will enable us to attract resources to create and maintain a web-based version.

Reviewer #3 (Recommendations for the authors):

Table s2 and s3: p values are indicated by star signs. However, with so many hypothesis tests, the p values should be corrected for multiple tests.

We have now applied Benjamini-Hochberg multiple hypothesis correction to these tables, correcting p-values within each of the four technologies. We update our significance calling to read:

“We identified 45 factors that relate to genes and found 33 (12 out of 23 binary factors and 21 out of 22 continuous factors) associated with selection in at least one assay type at Benjamini-Hochberg FDR < 0.001.”

Figure S1 - S4

These figures contain too many noninformative boxes. In all the figures, only the last three boxes are informative (reports assessed for eligibility, reports excluded, and studies included in review). The rest boxes convey little information and should be simplified.

We have simplified these diagrams, removing boxes which contained no information.

Figure S6: what does it mean by "prior to the publication of the first article represented in this sample"? What is "this sample"?

“This sample” refers to the collection of 450 GWAS articles, 296 articles using AP-MS, 148 transcriptomics articles, and 15 genome-wide CRISPR screen articles. We have rephrased this sentence to make this clear. It now reads:

“Variant of Figure 1B only considering articles published in 2002 or before, prior to the publication of any of the articles featuring -omics experiments which we considered for this analysis.”