

META-RESEARCH

COVID-19 research risks ignoring important host genes due to pre-established research patterns

Abstract It is known that research into human genes is heavily skewed towards genes that have been widely studied for decades, including many genes that were being studied before the productive phase of the Human Genome Project. This means that the genes most frequently investigated by the research community tend to be only marginally more important to human physiology and disease than a random selection of genes. Based on an analysis of 10,395 research publications about SARS-CoV-2 that mention at least one human gene, we report here that the COVID-19 literature up to mid-October 2020 follows a similar pattern. This means that a large number of host genes that have been implicated in SARS-CoV-2 infection by four genome-wide studies remain unstudied. While quantifying the consequences of this neglect is not possible, they could be significant.

THOMAS STOEGER* AND LUÍS A NUNES AMARAL*

Introduction

Shortly after SARS-CoV-2, the coronavirus that causes COVID-19, had emerged as a global threat to human health in January 2020, researchers had identified the host proteins required for viral entry into cells (Hoffmann et al., 2020; Monteil et al., 2020; Wrapp et al., 2020), repurposed drugs for treating COVID-19 patients (Grein et al., 2020; Recovery Collaborative Group, 2020), and initiated vaccine development (Folegatti et al., 2020; Jackson et al., 2020). A common feature of these advances was that they drew upon previous lines of research. A major question, however, is whether research into COVID-19 is pursuing all important host genes implicated in COVID-19.

To answer this question we used LitCOVID, a literature hub curated by the National Library of Medicine that tracks publications on COVID-19 (Chen et al., 2020). LitCOVID tags genes within the publicly accessible text of individual publications through PubTator (Wei et al., 2019), which first applies an ensemble of automated

approaches to tag genes, and then allows for a revision of these tags through biocurators. We consider genes tagged within the title, abstract or results sections of individual publications, and use MEDLINE to exclude reviews and other non-research publications (see Methods). This yields 10,395 research publications featuring 3733 human protein-coding genes that have been tagged at least once. This enables us to ask whether the choices by scientists to investigate these genes can be understood in terms of current biological knowledge on COVID-19.

Results

The most prominently tagged genes up to this point are: *Angiotensin-converting enzyme 2*, which serves as receptor for SARS-CoV-2 to enter cells (Hoffmann et al., 2020); *C-reactive protein*, a serum marker for inflammation (Sproston and Ashworth, 2018); and *Interleukin 6*, a mediator of systemic inflammatory responses (Kang et al., 2019). They account for 10.8%, 9.7%, and 4.5% of the total research on

*For correspondence: thomas.stoeger@northwestern.edu (TS); amaral@northwestern.edu (LANA)

Competing interests: The authors declare that no competing interests exist.

Funding: See page 9

Reviewing editor: Peter Rodgers, eLife, United Kingdom

© Copyright Stoeger and Nunes Amaral. 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.

human protein-coding genes within the COVID-19 literature, respectively (see Methods). Gene Ontology Enrichment analysis of the human protein-coding genes tagged in the COVID-19 literature finds them enriched for annotations on immune response (false-discovery rate $<10^{-66}$), inflammatory response (false-discovery rate $<10^{-65}$), and defense response to virus (false-discovery rate $<10^{-31}$) (**Supplementary file 1**). These two observations would thus suggest that the choice of host genes tagged in the COVID-19 literature is biologically grounded and in accord with current knowledge about respiratory viruses.

Most host genes identified by genome-wide studies have not been pursued

Genome-wide datasets provide another window on SARS-CoV-2 infection. As genome-wide approaches circumvent research patterns that may have been pre-established within the scientific literature (Haynes et al., 2018; Nelson et al., 2015; Stoeger et al., 2018), they might identify additional genes implicated in COVID-19. RNA-sequencing (RNA-seq) was used recently to identify 1726 host genes that change the expression of their transcripts in the lungs of COVID-19 patients at an adjusted p -value <0.05 (Blanco-Melo et al., 2020). Affinity-purification mass spectrometry (Aff-MS) was used to identify 293 host proteins following the pulldown of exogenously expressed SARS-CoV-2 proteins (Gordon et al., 2020). Using genome-wide association studies (GWAS), the Host Genetics Initiative identified 52 genes through their association at a P -value of 10^{-5} or lower in one of three comparisons: COVID-19 vs lab or self-reported negative; hospitalized COVID-19 patients vs population; or very severe respiratory COVID-19 vs population (COVID-19 Host Genetics Initiative, 2020a; Ellinghaus et al., 2020). 15 genes were identified in two comparisons and one gene, *Leucine zipper transcription factor-like protein 1* (LZTFL1), was identified in all three comparisons (**Supplementary file 2**). Using a pooled CRISPR screen to affect SARS-CoV-2 induced cell death in African green monkey cells, Wei et al. identified 41 genes, which we mapped to their human homologs using BioMart (Wei et al., 2020; see Methods). 48 genes are identified in two of the four different genome-wide datasets

(**Supplementary file 3**), but no gene is identified in more than two.

However, an analysis of the COVID-19 literature reveals that most (56%–71%) of the genes identified in these four datasets have not yet been tagged in the COVID-19 literature (**Figure 1A**). Thus, the genes identified by the four genome-wide datasets are 10–25% more likely to have been tagged than a randomly chosen gene because we also observe that 19% of all human protein-coding genes have been tagged at least once in the COVID-19 literature. Similarly, the fraction of tagged genes only increases by 0–7% if we include preprints (**Figure 1—figure supplement 1**). We conclude that many genes identified by genome-wide datasets on COVID-19 have not been investigated yet in more detail in the context of COVID-19.

At the same time, we observe that genes, which have been identified by multiple of the four distinct genome-wide datasets (**Figure 1B**), or multiple GWAS comparisons (**Figure 1C**), are more likely to have been tagged in the COVID-19 literature. This, reassuringly, demonstrates that research into COVID-19 host genes enriches for host genes identified by multiple different lines of support – particularly if there exists support from human genome-wide association studies.

Yet, overall, genes identified by multiple genome-wide datasets remain only a minority of all identified genes (2%), and many of them are still ignored in the COVID-19 literature (52%) (**Supplementary file 2**), suggesting that research into SARS-CoV-2 host genes might be missing important pieces of the puzzle.

Tagged host genes follow pre-established research patterns

A possible explanation for the relative lack of interest in the additional genes implicated in SARS-CoV-2 infection by these genome-wide datasets is that research on COVID-19 is constrained by pre-established research patterns (Chu and Evans, 2018). Briefly, we know that knowledge on human genes is heavily skewed toward a subset of genes (Gans et al., 2008; Gillis and Pavlidis, 2013; Hoffmann and Valencia, 2003; Oprea et al., 2018; Su and Hogenesch, 2007) that were being investigated prior to the Human Genome Project (Edwards et al., 2011; Grueneberg et al., 2008; Stoeger et al., 2018). As a result, if assessing their importance

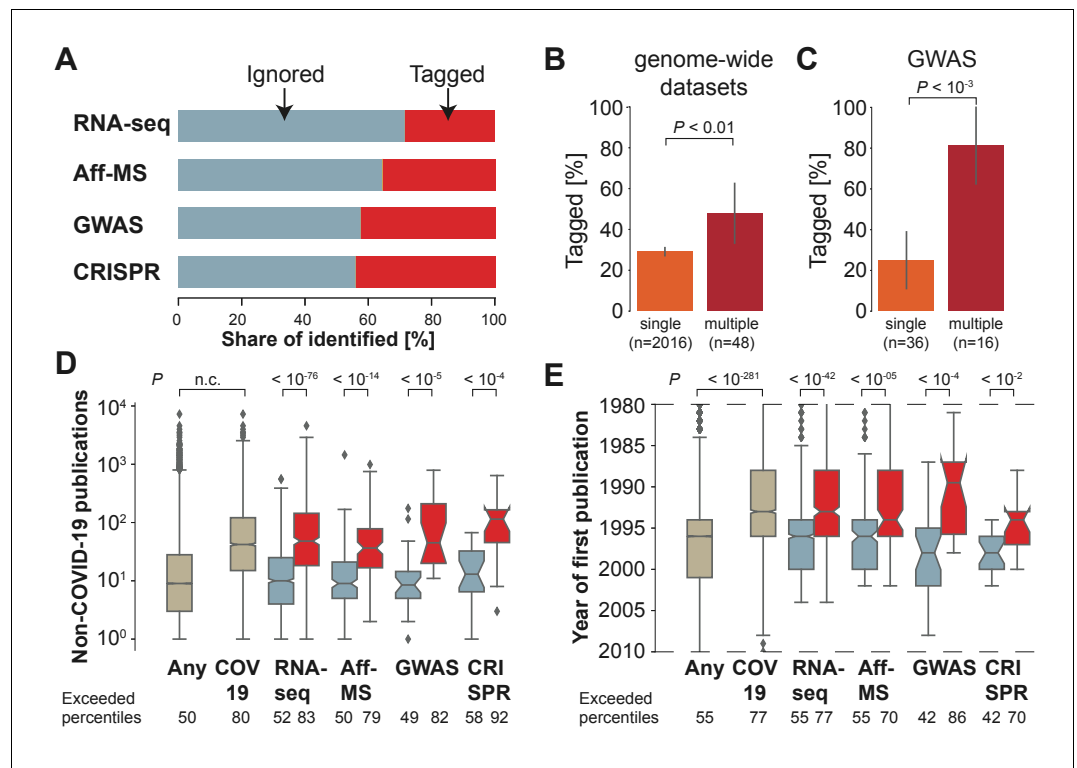


Figure 1. Most host genes implicated in COVID-19 identified by genome-wide approaches are not being investigated. (A) Share of identified genes, which are ignored (never tagged, blue) or tagged (at least once) within the COVID-19 literature. (B) Share of tagged genes identified by a single (orange) or multiple (maroon) genome-wide datasets. P -values are calculated via Fisher's exact test. n is the number of genes. (C) Share of tagged genes identified by a single (orange) or multiple (maroon) GWAS comparisons. P -values are calculated via Fisher's exact test. n is the number of genes. (D) Non-COVID-19 publications measured for any human protein-coding gene (ocher, any) and those occurring in the COVID-19 literature (ocher, COV19) and genes identified in A (colors as in A). Notches indicate 95% confidence interval of the median. P -values are calculated via Mann-Whitney U test. Exceeded percentiles indicates percentiles of all genes exceeded by the median gene of the genes in an individual boxplot. n.c. marks non-computable P -values that approximate 0. (E) As D, but for year of initial publication on the gene. Dashed lines indicate limit of visualized values. Some genes had their first publication before or afterwards.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Share of identified genes that are ignored or tagged.

through genetic loss-of-function intolerance or findings of GWAS (Haynes et al., 2018; Stoeger et al., 2018), the most frequently investigated protein-coding genes tend to be only marginally more important to human physiology and disease than a random selection of genes.

To test the hypothesis that COVID-19 research is constrained by patterns similar to those seen in non-COVID-19 research, we take advantage of the ability of gene2pubmed (a service provided by the National Center for Biotechnology Information) to link human protein-coding genes to individual publications, and compare 465,770 non-COVID-19 papers published until December 2015 with 10,395 COVID-19 research publications indexed by LitCOVID

until October 16th, 2020. For the non-COVID-19 research we exclude publications that contain any viral gene (irrespective of whether the virus in question is a coronavirus) and publications tagging 100 or more genes.

We find that genes that are tagged in the COVID-19 literature are also frequently investigated in the non-COVID-19 literature. To assess how frequently individual genes have been investigated in the non-COVID-19 literature relative to other genes, we rank all genes according to the number of publications in the non-COVID-19 literature. The median rank of genes tagged in the COVID-19 literature exceeds the rank of 80% of human protein-coding genes (Figure 1D). This demonstrates that the majority

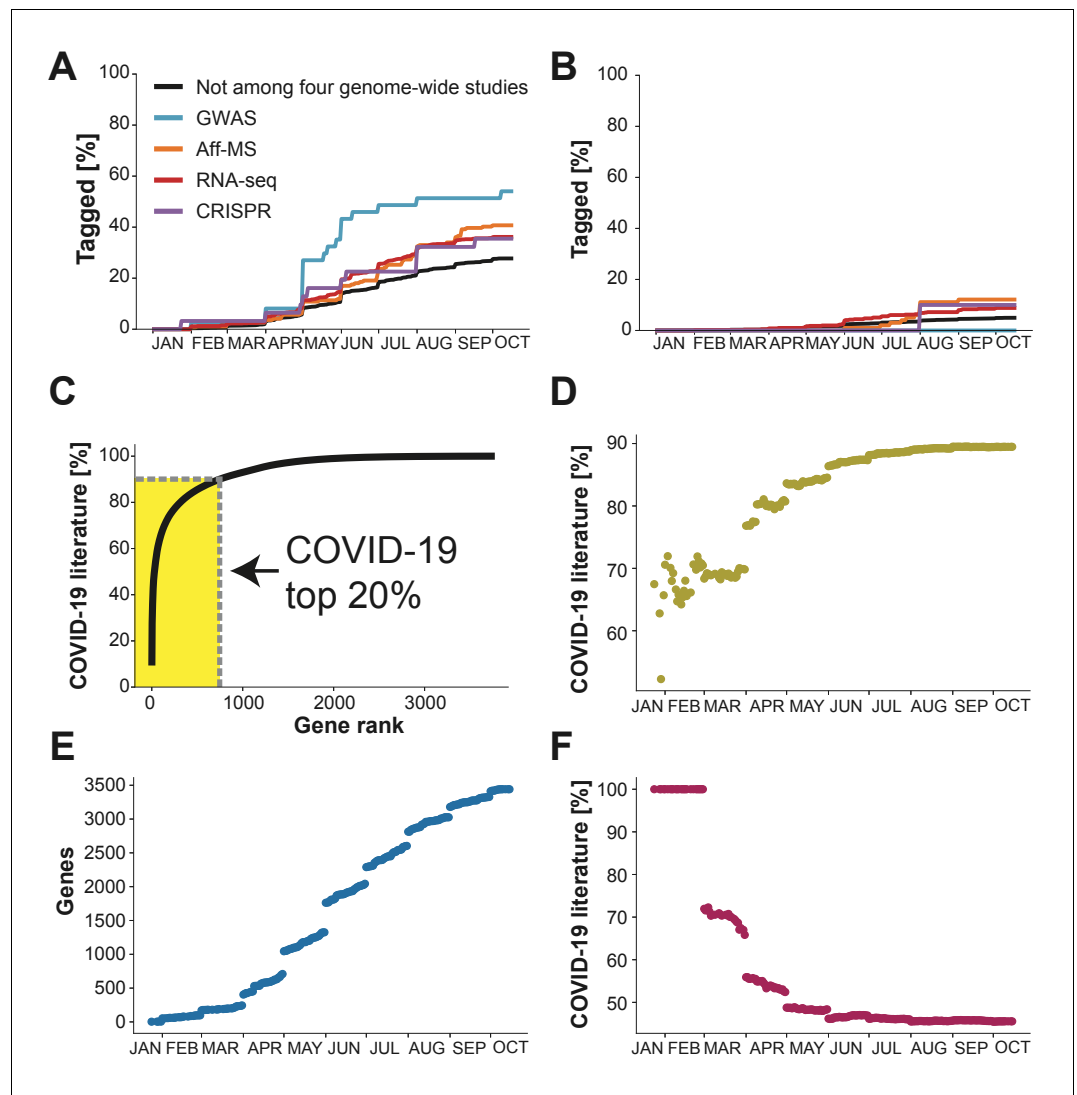


Figure 2. What the future holds? Percentage of genes with indicated levels of support by the four genome-wide studies which have been tagged at least once in the COVID-19 literature. (A) Analysis restricted to the 50% of genes with highest number of publications in non-COVID-19 literature. (B) Analysis restricted to the 50% of genes with the lowest number of publications in the non-COVID-19 literature. (C) Cumulative share of literature on human protein-coding genes tagged in the COVID-19 literature. Top 20% indicates the 20% of genes that occur the most in the non-COVID-19 literature. Gene rank refers to the order of human protein-coding genes. The gene with the most publication equivalents would be have rank 1. Yellow area indicates share of literature accounted for by the top 20% genes. (D) Share of COVID-19 literature accounted for by the 20% of genes that had occurred the most in the COVID-19 literature by a given date. (E) Number of distinct human protein coding genes that have been tagged in the literature by a given date. (F) Share of COVID-19 literature accounted for by first 100 genes to be tagged in the COVID-19 literature by a given date.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Temporal trends in the diversity of COVID-19 research.

of protein-coding human genes tagged in the COVID-19 literature was already heavily investigated in the context of research unrelated to COVID-19.

Next we return to our earlier observation on the majority of the implicated host genes

reported by the four different genome-wide datasets being ignored within the COVID-19 literature. As anticipated, we observe that for each of the four distinct datasets investigated, ignored genes also occur less in this non-COVID-19 literature (Figure 1D). When we

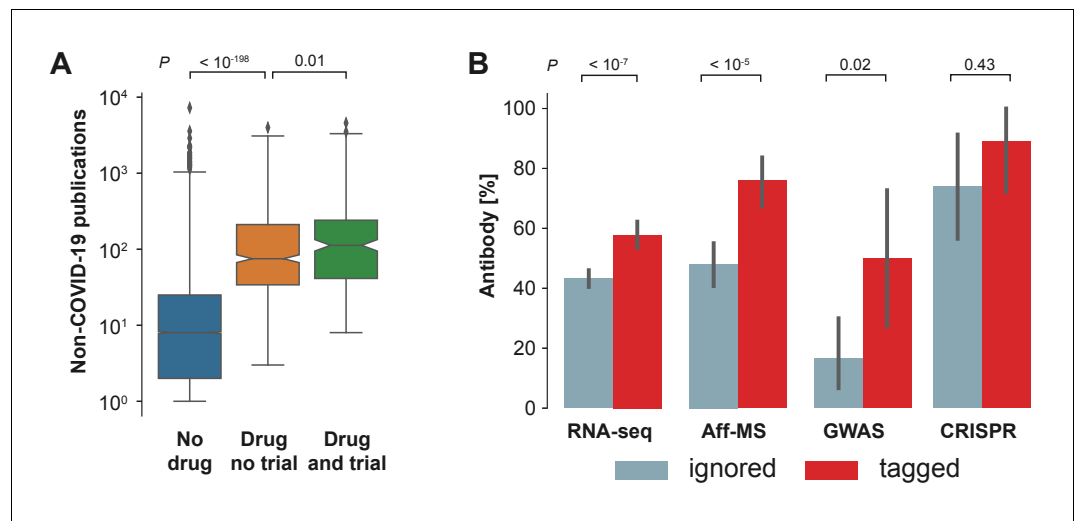


Figure 3. Availability of reagents. (A) Drugs studied in COVID-19 related clinical trials are frequently studied within the non-COVID-19 literature. We compare non-COVID-19 publications measured for human protein-coding genes that are not listed as pharmaceutical targets in DrugBank (other, No drug), against those that are listed as pharmaceutical targets but have not occurred as an intervention in a clinical trial on COVID-19 (orange, Drug no trial), and against those that are listed as pharmaceutical targets and have occurred as an intervention in a clinical trial on COVID-19 (green, Drug and trial). Notches indicate 95% confidence interval of the median. *P*-values are calculated via *Mann-Whitney U* test. (B) Fraction of genes with reported usage of an antibody to detect the encoded protein as a prey in BioGRID. Bars are genes identified by the four different genome-wide studies that have either been tagged in the COVID-19 literature (red) or ignored (blue). Error bars indicate 95% confidence interval. *P*-values are calculated via Fisher's exact test.

compare the number of publications on implicated but ignored host genes to the number of publications on any protein-coding gene encoded in the human genome, this difference is modest, and only reaches statistical significance for RNA-seq (RNA-Seq: $p < 10^{-2}$; Interactomics: $p = 0.20$; GWAS: $p = 0.93$; CRISPR: $p = 0.31$), where ignored genes had occurred slightly more in the non-COVID-19 literature (median percentile: 52). In contrast, implicated and tagged host genes have occurred significantly more frequently in the non-COVID-19 literature (RNA-Seq: $p < 10^{-98}$; Interactomics: $p < 10^{-18}$; GWAS: $p < 10^{-6}$; CRISPR: $p < 10^{-6}$). We conclude that implicated host genes that are ignored in the COVID-19 literature have in the past been studied as much as randomly chosen human protein-coding genes, whereas implicated host genes that are tagged in the COVID-19 literature have in the past already been investigated much more frequently than randomly chosen human protein-coding genes.

Before the COVID-19 pandemic it had been shown that the literature is skewed toward a subset of genes that were being investigated prior to the productive phase of the Human Genome Project. These features include the fraction of organs with detectable transcript

expression, the length of the genes, the hydrophobicity of the coded proteins, their loss-of-function insensitivity, and studies on orthologous genes in model organisms (Stoeger et al., 2018). We decided to explore if the genes tagged in the COVID-19 literature had been studied before the pandemic, and found that they had occurred earlier (Figure 1E), with many also first being studied before the productive phase of the Human Genome Project (NHGRI, 2003). Similarly, the host genes identified by the four genome-wide datasets that are ignored in the COVID-19 literature first appeared in the non-COVID-19 literature after the host genes that are tagged in the COVID-19 literature (Figure 1E).

Trends over time

The COVID-19 pandemic has ravaged for less than a year, which is a short period of time compared to most research projects. Thus, we might not yet be observing research addressing poorly-studied implicated host genes because not sufficient time has passed for research to catch up to the new information.

To anticipate the near future, we follow the occurrence of genes in the COVID-19 literature

over time. Based on our insight that ignored host genes have not been studied more than other genes in the non-COVID-19 literature (**Figure 1D**), we separate genes into two classes: genes that are among the 50% top-studied human protein-coding genes in the non-COVID-19 literature, and genes that are among the 50% least-studied human protein-coding genes. The second class holds 35% of the genes identified by RNA-seq, 33% of the genes identified by Aff-MS, 29% of the genes identified by GWAS, and 24% of the genes identified by CRISPR. If research is catching up to the new knowledge, we would expect to see the fraction of the COVID-19 literature addressing the 50% least-studied human protein-coding genes to increase over time.

When focusing on the genes that are among the 50% top-studied human protein-coding genes, we observe their occurrence in the COVID-19 literature to increase steadily. Extrapolating from the observed trends, we anticipate that it will take around one year till nearly all genes of this class will have been tagged at least once within the COVID-19 literature (**Figure 2A**). When focusing on the genes that are among the 50% least-studied human protein-coding genes, we too observe their occurrence in the COVID-19 literature to increase steadily over time (**Figure 2B**). As for each of the four genome-wide datasets the increase is, however, slower than for the 50% most studied protein-coding genes, we project that multiple years could pass until each gene of the 50% least-studied human protein-coding genes will have been tagged at least once within the COVID-19 literature.

Pursuing this observation further, we turn to the entire COVID-19 literature. Notably, 83% of all human protein-coding genes tagged in the COVID-19 literature have not been identified by any of the four genome-wide datasets. Further, the different genome-wide datasets together only account for 26% of the COVID-19 literature (RNA-seq: 11.7%, Aff-MS: 2.4%, GWAS: 0.5%, CRISPR: 11.1%) (see Methods).

We ask whether the COVID-19 literature might become dominated by a few genes that are tagged more commonly than other genes that are also tagged in the COVID-19 literature. If we consider the current literature, we do indeed observe support for our hypothesis that the COVID-19 literature is becoming dominated by a few genes as currently the 20% top-tagged human protein-coding genes (747 of 3,733) in the COVID-19 literature account for 90% of the literature (**Figure 2C**). This share exceeds the

80% anticipated for scientific processes subjected to anthropogenic biases (*Jia et al., 2019*). We conclude that a surprisingly small fraction of genes dominates the COVID-19 literature.

Finally, we inspect whether the extent to which the COVID-19 literature tags each tagged gene is becoming more or less expansive over time. We observe that the COVID-19 has become less expansive, whether we quantify expansiveness through the share of the literature that is accounted for by the 20% top-tagged genes or the Gini coefficient over the share of the COVID-19 literature attributable to individual genes (*Gini, 1912; Figure 2D, Figure 2—figure supplement 1A,B*). However, if assessing expansiveness by the total number of genes that have been tagged at least once, then this literature did become more expansive after the first months (**Figure 2E**).

Interestingly, we also observe that the share of the COVID-19 literature, which is accounted for by the 100 genes that were tagged first within the COVID-19 literature has been decreasing (**Figure 2F, Figure 2—figure supplement 1C**) – though stabilizing at an astonishingly high share of roughly 45% since June (**Figure 2F, Figure 2—figure supplement 1C**). We conclude that, overall, the literature on COVID-19 became less expansive during the first months of the pandemic and has since stayed focused on a restricted subset of genes.

One possible reason for why some genes are tagged more than others in the COVID-19 literature could be that compared to other genes they are more important in the context of COVID-19. To probe this hypothesis, we consider groups of genes and the four different genome-wide datasets. When contrasting the 100 initially tagged genes against the other genes tagged in the COVID-19 literature, we reassuringly find that the 100 initially tagged genes are 29% more likely to have been identified by one of these four datasets (**Supplementary file 4**). However, they are on average tagged 2993% more (**Supplementary file 4**). If we contrast the 20% top-tagged genes against the other tagged genes, we find them to be 3% less likely to have been identified by one of the four genome-wide datasets, while they on average are tagged 3512% more (**Supplementary file 4**). Cumulatively, this suggests that the present focus of the COVID-19 literature on a restricted subset of genes cannot be explained by those genes having been identified by genome-wide datasets

reporting on transcriptomic changes, protein interactions, genetic associations and loss-of-function perturbations.

Study limitations

Our study has several important limitations. First, we cannot say whether a gene tagged in the COVID-19 literature is truly investigated for its potential role in COVID-19. Second, we cannot yet assess how important individual genes are in COVID-19. Third, and despite our projections, it remains formally unclear, whether the findings reported in this manuscript will hold in the upcoming months as more genome-wide datasets will become available and researchers will have had sufficient time for follow-up studies. For instance, there might already be research initiatives specifically targeted toward the ignored COVID-19 host genes.

Nonetheless, in the past, genome-wide experiments have rarely guided subsequent studies in the non-COVID-19 literature (Haynes et al., 2018; Stoeger et al., 2018). Thus, there is a significant risk that the COVID-19 literature will continue to ignore host genes that have not already been extensively studied independently of COVID-19.

Discussion

Our study reiterates prior observations that research into human protein-coding genes is disproportionately skewed towards a comparably small set of genes (Haynes et al., 2018; Nelson et al., 2015; Stoeger et al., 2018). Likewise, our current analysis on COVID-19 already allows us to conclude that genes that are identified by genome-wide datasets, and hence are likely to have biological significance in the context of COVID-19, have hitherto remained ignored if they had not already been investigated more than other genes prior COVID-19.

We realize that there is an exploration-exploitation trade-off at play and that focusing research on genes that have already been heavily investigated yields significant advantages to investigators: applicability of existing research tools, the ability to place findings in a broader context, and the identification of drugs and other reagents that could be repurposed. Supporting a focus on exploitation, we find that: interventions in clinical trials on COVID-19 are biased toward pharmaceutical targets that occurred frequently in the non-COVID-19 literature (Figure 3A); and that antibodies – a class of reagent that cannot be produced for arbitrary

genes within a few days – are less available for those genes identified by RNA-seq or Aff-MS or GWAS which have been ignored in the COVID-19 literature (Figure 3B).

Further, additional factors might affect the exploration of studies on ignored host genes. First, the number of laboratories working on ignored genes was quite small prior COVID-19 (Supplementary file 5), and plausibly only a small fraction of the laboratories studied host responses toward respiratory viruses. Second, the risk of being outcompeted by other laboratories might discourage individual laboratories from pursuing publicly acknowledged research targets (Bergstrom et al., 2016). Third, scientists rarely switch topics (Zeng et al., 2019). Likewise, laboratories already working on COVID-19 might have little incentive to move toward distinct host genes as it is possible to contribute to the COVID-19 literature irrespective of whether the genes had been identified by genome-wide datasets (Figure 2A,B and Supplementary file 4). Moreover, it might be beneficial overall if research into COVID-19 is mainly driven by researchers with a background on pathogens (Kwon, 2020). Lastly, concerns have been expressed about the possibility that fraudulent gene knockdown studies that target under-studied human genes may be corrupting the literature and impeding research into biomarkers (Byrne et al., 2019).

We believe that a more complete understanding of host biology could open novel directions for interventions against SARS-CoV-2 and other viruses. However, the challenge remains of how to promote research on ignored host genes. For example, we cannot speculate whether researchers that turn their attention toward ignored genes in the context of COVID-19, will face a similar disadvantage to their career as did those that studied less studied genes prior to COVID-19 (Stoeger et al., 2018).

In the hopes of prompting greater investigation into implicated host genes, we list genes occurring in multiple of the four datasets described earlier in the supplemental material of this manuscript (Supplementary file 3). Most of the genes identified by multiple datasets appear multiple times because of the large volume of genes identified by the RNA-seq dataset. For this reason, we highlight four genes that were identified by multiple smaller datasets: (1) *Mitochondrial import inner membrane translocase subunit Timm10* (TIMM10) has been identified through Aff-MS and CRISPR, and (2) *FYVE And Coiled-Coil Domain Autophagy Adaptor 1*

(FYCO1) and (3) *Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2* (PLOD2) and (4) Ras GTPase-activating protein-binding protein 2 (G3BP2) have been identified through Aff-MS and GWAS. Of these four genes only G3BP2 has been tagged in the COVID-19 literature; and TIMM10 and FYCO1 have both occurred in nine publications in the non-COVID-19 literature, matching the expectation for a randomly selected gene. Of additional interest in the context of COVID-19, FYCO1 is associated with the levels of the monocyte chemoattractant protein-1 (Ahola-Olli et al., 2017; Buniello et al., 2019), which contributes to COVID-19 through hyperinflammation (Mehta et al., 2020).

Methods

COVID-19 literature

We downloaded LitCOVID from <https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/> on 2020-10-16 and parsed the contained json file for the presence of concepts annotated as genes. For studies annotated with proteins, we used their PubMed identifiers, to query MEDLINE on 2020-10-16 via their efetch API. Subsequently we parsed the MEDLINE entries via `pubmed_parser 2.2` (https://github.com/titipata/pubmed_parser; Achakulvisut et al., 2020). We then excluded publications carrying at least one of the following publication types: Review, Comment, Editorial, Meta-Analysis, Systematic Review, News, Published Erratum, Historical Article, Interview, Retracted Publication, Retraction of Publication, Webcast, Expression of Concern or Portrait. Further we excluded publications whose abstract contain either the phrase ‘this review’ or the phrase ‘this perspective’. We considered genes tagged within LitCOVID in the annotated TITLE, INTRO, ABSTRACT or RESULTS sections.

Research intensity within COVID-19 literature

We measured the research intensity directed toward individual implicated host genes in units of publication equivalents. Each gene tagged within a publication accrues the publication equivalent of number of tags to the gene in that publication divided by total number of tags to any gene in that publication. For example, if a study tags two different genes, and the first gene is tagged three times, whereas the second gene is only tagged once, the first gene would accrue 0.75 publication equivalents, and the second gene would accrue 0.25 publication

equivalents. We expressed the share of literature covered by an individual human protein-coding gene as the sum of its publication equivalents over the sum of all publication equivalents of human protein-coding genes. We excluded the studies of Blanco-Melo et al., 2020 and Gordon et al., 2020 which report the RNA-seq and Aff-MS datasets, respectively.

Gene ontology enrichment analysis

We used the Database for Annotation, Visualization and Integrated Discovery, version 6.8 (Huang et al., 2009).

Data processing and filtering

For CRISPR we considered the African green monkey genes reported in Figure 1D of Jin Wei et al., 2020. To map African green monkey to human genes, we used BioMart’s (Haider et al., 2009) April 2020 release. We used the genetic polymorphisms reported in the 2020-09-30 release of the Host Genetics Initiative (COVID-19 Host Genetics Initiative, 2020b) and mapped them to human genes through the Ensembl Variant Effect Predictor (McLaren et al., 2016), using the Ensembl release 101. For RNA-seq we only considered comparisons flagged with ‘ok’ by the authors (Blanco-Melo et al., 2020). For Aff-MS we used the data as provided by BioGRID (Chattr-Aryamontri et al., 2017), version 3.5.186 (<https://downloads.thebiogrid.org/BioGRID>).

We obtained the list of human protein-coding genes from https://ftp.ncbi.nlm.nih.gov/gene/ATA/gene_info.gz in June 2020.

Occurrence of genes in preprints

We obtained manuscripts abstracts from dimension.ai’s collection of COVID-19 related publications, release 34 (https://dimensions.figshare.com/articles/dataset/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063/34; dimension.ai, 2020), and subsequently select manuscripts listing medRxiv or bioRxiv or arXiv as their source. Next, we matched each word against the gene symbols as downloaded from https://ftp.ncbi.nlm.nih.gov/gene/ATA/gene_info.gz in June 2020.

We excluded the following gene symbols as within the abstracts they would match abbreviations that did not refer to genes: AFM, AIR, AN, APC, APP, AR, ARC, ATM, BCR, BED, BID, CCNC, CFD, CHM, COPD, COPE, CP, CPE, CS, DBI, DCT, ENG, GAN, GC, HP, HPA, HPD, HPO,

HR, IDS, IMPACT, IV, KIT, MCC, MET, MICE, MMD, MS, MS2, NHS, NM, NPS, NSF, NTS, PIP, POLL, REST, SEA, SET, SHE, SI, SPR, STS, TAT, TRAP, WAS.

Non-COVID-19 literature

We downloaded gene2pubmed from <https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz> in early 2017. MEDLINE, containing publication dates and publication types was downloaded from https://www.nlm.nih.gov/databases/download/pubmed_medline.html, and maintained in a local copy of their database in early 2017. We restricted the analysis to research publications published prior 2016.

Temporal profiles

We obtained publication dates from dimension.ai’s collection of COVID-19 publications, release 34 (https://dimensions.figshare.com/articles/dataset/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063/34). We excluded publications dating to January 1st, 2020 – the day linked to the most publications. Manual inspection revealed that the date of January 1st2020 was assigned to publications lacking a concrete 2020 publication date.

Occurrence within clinical trials

We obtained interventions within clinical trials from dimension.ai’s collection of COVID-19 related clinical trials, release 34 (https://dimensions.figshare.com/articles/dataset/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063/34; *dimension.ai*, 2020). We performed a case-insensitive match against drug names and drug synonyms contained within DrugBank, version 5.1.5 (<https://www.drugbank.ca>). Next we used DrugBank’s mapping between drugs and the targets of their pharmaceutical action and used the accompanying gene symbol to identify genes.

Identification of research laboratories

We used disambiguated authorship identifiers from Web of Science and considered the last author of each publication as the laboratory.

Acknowledgements

We thank all members of the Amaral lab for feedback, particularly Jennifer Liu, Kedi Cao, Meagan Bechel, Reese Richardson and Sarah Ben Maamar. We thank Richard Wunderink for feedback and suggesting the analysis of clinical

trials. We also thank Rick Morimoto for feedback.

Thomas Stoeger is in the Successful Clinical Response in Pneumonia Therapy (SCRIPT) Systems Biology Center, the Department of Chemical and Biological Engineering and the Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, United States, and the Center for Genetic Medicine, Northwestern University School of Medicine, Chicago, United States
thomas.stoeger@northwestern.edu
 <https://orcid.org/0000-0002-5540-4278>

Luís A Nunes Amaral is in the Successful Clinical Response in Pneumonia Therapy (SCRIPT) Systems Biology Center, the Northwestern Institute on Complex Systems (NICO), the Department of Molecular Biosciences, and the Department of Physics and Astronomy, Northwestern University, Evanston, United States, and the Department of Medicine, Northwestern University School of Medicine, Chicago, United States
amaral@northwestern.edu
 <https://orcid.org/0000-0002-3762-789X>

Author contributions: Thomas Stoeger, Conceptualization, Resources, Data curation, Software, Formal analysis, Funding acquisition, Investigation, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing; Luís A Nunes Amaral, Conceptualization, Resources, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing

Competing interests: The authors declare that no competing interests exist.

Received 11 August 2020
Accepted 07 November 2020
Published 24 November 2020

Funding

Funder	Grant reference number	Author
National Institute of Allergy and Infectious Diseases	U19AI135964	Luís A Nunes Amaral
National Science Foundation	1956338	Luís A Nunes Amaral
Simons Foundation	DMS-1764421	Luís A Nunes Amaral
Air Force Office of Scientific Research	FA9550-19-1-0354	Luís A Nunes Amaral
National Institute on Aging	K99AG068544	Thomas Stoeger

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.61981.sa1>

Author response <https://doi.org/10.7554/eLife.61981.sa2>

Additional files

Supplementary files

- Source code 1. Source code for curation and analysis of datasets.
- Supplementary file 1. Gene Ontology enrichment analysis for human protein-coding genes tagged in the COVID-19 literature.
- Supplementary file 2. Identification of genes through multiple GWAS comparisons.
- Supplementary file 3. Implicated host genes identified by multiple genome-wide studies.
- Supplementary file 4. Extent of tags in COVID-19 literature compared to rate identification in genome-wide datasets. Per-gene average share of COVID-19 literature and per-gene average identification rate in genome-wide datasets (one if identified, 0 if not identified). Shown are the ratios of this share and the rates in individual groups (100 first tagged genes, and 20% top-tagged genes) over the share and the rates of the other genes that have been tagged in the COVID-19 literature.
- Supplementary file 5. Number of laboratories working on individual genes, identified within one of the four genome-wide datasets, between 2006 and 2015.
- Transparent reporting form

Data availability

No data was generated for this study. Data underlying this study can be downloaded from the sources indicated in the methods section and used under the respective licenses. The data can be preprocessed with the source code accompanying this manuscript, and - for literature until 2015 - the public source code provided in a former publication of ours, https://github.com/tstoeger/plos_biology_2018_ignored_genes.

The following datasets were generated:

References

- Achakulvisut T**, Acuna D, Kording K. 2020. Pubmed parser: a Python parser for PubMed Open-Access XML subset and MEDLINE XML dataset XML dataset. *Journal of Open Source Software* **5**:1979. DOI: <https://doi.org/10.21105/joss.01979>
- Ahola-Olli AV**, Würtz P, Havulinna AS, Aalto K, Pitkänen N, Lehtimäki T, Kähönen M, Lyytikäinen LP, Raitoharju E, Seppälä I, Sarin AP, Ripatti S, Palotie A, Perola M, Viikari JS, Jalkanen S, Maksimow M, Salomaa V, Salmi M, Kettunen J, et al. 2017. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *American Journal of Human Genetics* **100**:40–50. DOI: <https://doi.org/10.1016/j.ajhg.2016.11.007>, PMID: 27989323
- Bergstrom C**, Foster J, Song Y. 2016. Why scientists chase big problems: individual strategy and social optimality. *arXiv*. <https://arxiv.org/abs/1605.05822v2>.
- Blanco-Melo D**, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Møller R, Jordan TX, Oishi K, Panis M, Sachs D, Wang TT, Schwartz RE, Lim JK, Albrecht RA, tenOever BR. 2020. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**:1036–1045. DOI: <https://doi.org/10.1016/j.cell.2020.04.026>, PMID: 32416070
- Buniello A**, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousseau O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Jenkins H, Flicek P, et al. 2019. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**:D1005–D1012. DOI: <https://doi.org/10.1093/nar/gky1120>, PMID: 30445434
- Byrne JA**, Grima N, Capes-Davis A, Labbé C. 2019. The possibility of systematic research fraud targeting Under-Studied human genes: causes, consequences, and potential solutions. *Biomarker Insights* **14**:1–12. DOI: <https://doi.org/10.1177/1177271919829162>
- Chatr-Aryamontri A**, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M. 2017. The BioGRID interaction database: 2017 update. *Nucleic Acids Research* **45**:D369–D379. DOI: <https://doi.org/10.1093/nar/gkw1102>, PMID: 27980099
- Chen Q**, Allot A, Lu Z. 2020. Keep up with the latest coronavirus research. *Nature* **579**:193. DOI: <https://doi.org/10.1038/d41586-020-00694-1>, PMID: 32157233
- Chu JSG**, Evans JA. 2018. Too many papers? slowed canonical progress in large fields of science. *SocArXiv*.
- COVID-19 Host Genetics Initiative**. 2020a. The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics* **28**: 715–718. DOI: <https://doi.org/10.1038/s41431-020-0636-6>, PMID: 32404885
- COVID-19 Host Genetics Initiative**. 2020b. COVID-19-hg GWAS meta-analyses round 3. *COVID-19 Hg*. Identifier: round 3.
- dimension.ai**. 2020. Dimensions COVID-19 publications, datasets and clinical trials. *figshare*. DOI: <https://doi.org/10.6084/m9.figshare.11961063.v34>
- Edwards AM**, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. 2011. Too many roads not taken. *Nature* **470**:163–165. DOI: <https://doi.org/10.1038/470163a>
- Ellinghaus D**, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernández J, Prati D, Baselli G, Asselta R, Grimsrud MM, Milani C, Aziz F, Kässens J, May S, Wendorff M, Wienbrandt L, Uellendahl-Werth F, Zheng T, Yi X, et al. 2020. Genome-wide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine* **383**:1522–1534. DOI: <https://doi.org/10.1056/NEJMoa2020283>, PMID: 32558485
- Folegatti PM**, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, Bellamy D, Bibi S, Bittaye M,

- Clutterbuck EA, Dold C, Faust SN, Finn A, Flaxman AL, Hallis B, Heath P, Jenkin D, Lazarus R, Makinson R, Minassian AM, et al. 2020. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *The Lancet* **396**:467–478. DOI: [https://doi.org/10.1016/S0140-6736\(20\)31604-4](https://doi.org/10.1016/S0140-6736(20)31604-4), PMID: 32702298
- Gans JS, Murray FE, Stern S. 2008. Patents, papers, pairs and secrets: contracting over the disclosure of scientific knowledge (Preliminary & incomplete). http://fmurray.scripts.mit.edu/docs/Gans.Murray.Stern%20KnowledgeDisclosure_DRAFT_09.30.2008.pdf [Accessed October 23, 2020].
- Gillis J, Pavlidis P. 2013. Assessing identity, redundancy and confounds in gene ontology annotations over time. *Bioinformatics* **29**:476–482. DOI: <https://doi.org/10.1093/bioinformatics/bts727>, PMID: 23297035
- Gini C. 1912. *Variabilità E Mutabilità*. Tipogr. di P. Cuppini.
- Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA, Hüttenhain R, Kaake RM, Richards AL, Tutuncuoglu B, Foussard H, Batra J, Haas K, Modak M, Kim M, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**:459–468. DOI: <https://doi.org/10.1038/s41586-020-2286-9>, PMID: 32353859
- Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, Feldt T, Green G, Green ML, Lescure F-X, Nicastrì E, Oda R, Yo K, Quiros-Roldan E, Studemeister A, Redinski J, Ahmed S, Bernett J, Chelliah D, Chen D, et al. 2020. Compassionate use of remdesivir for patients with severe Covid-19. *New England Journal of Medicine* **382**:2327–2336. DOI: <https://doi.org/10.1056/NEJMoa2007016>
- Grueneberg DA, Degot S, Pearlberg J, Li W, Davies JE, Baldwin A, Endege W, Doench J, Sawyer J, Hu Y, Boyce F, Xian J, Munger K, Harlow E. 2008. Kinase requirements in human cells: I. Comparing kinase requirements across various cell types. *PNAS* **105**:16472–16477. DOI: <https://doi.org/10.1073/pnas.0808019105>, PMID: 18948591
- Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. 2009. BioMart central portal—unified access to biological data. *Nucleic Acids Research* **37**:W23–W27. DOI: <https://doi.org/10.1093/nar/gkp265>, PMID: 19420058
- Haynes WA, Tomczak A, Khatri P. 2018. Gene annotation bias impedes biomedical research. *Scientific Reports* **8**:1362. DOI: <https://doi.org/10.1038/s41598-018-19333-x>, PMID: 29358745
- Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S. 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**:271–280. DOI: <https://doi.org/10.1016/j.cell.2020.02.052>, PMID: 32142651
- Hoffmann R, Valencia A. 2003. Life cycles of successful genes. *Trends in Genetics* **19**:79–81. DOI: [https://doi.org/10.1016/S0168-9525\(02\)00014-8](https://doi.org/10.1016/S0168-9525(02)00014-8)
- Huang daW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**:44–57. DOI: <https://doi.org/10.1038/nprot.2008.211>, PMID: 19131956
- Jackson LA, Anderson EJ, Roupheal NG, Roberts PC, Makhene M, Coler RN, McCullough MP, Chappell JD, Denison MR, Stevens LJ, Puijssers AJ, McDermott A, Flach B, Doria-Rose NA, Corbett KS, Morabito KM, O'Dell S, Schmidt SD, Swanson PA, Padilla M, et al. 2020. An mRNA vaccine against SARS-CoV-2 — Preliminary Report. *New England Journal of Medicine* **383**:1920–1931. DOI: <https://doi.org/10.1056/NEJMoa2022483>
- Jia X, Lynch A, Huang Y, Danielson M, Lang'at I, Milder A, Ruby AE, Wang H, Friedler SA, Norquist AJ, Schrier J. 2019. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**:251–255. DOI: <https://doi.org/10.1038/s41586-019-1540-5>, PMID: 31511682
- Jin Wei MMA, Hanna RE, DeWeirdt PC. 2020. Genome-wide CRISPR screen reveals host genes that regulate SARS-CoV-2 infection. *bioRxiv: The Preprint Server for Biology*.
- Kang S, Tanaka T, Narazaki M, Kishimoto T. 2019. Targeting Interleukin-6 signaling in clinic. *Immunity* **50**:1007–1023. DOI: <https://doi.org/10.1016/j.immuni.2019.03.026>, PMID: 30995492
- Kwon D. 2020. How swamped preprint servers are blocking bad coronavirus research. *Nature* **581**:130–131. DOI: <https://doi.org/10.1038/d41586-020-01394-6>, PMID: 32382120
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome Biology* **17**:122. DOI: <https://doi.org/10.1186/s13059-016-0974-4>, PMID: 27268795
- Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ. 2020. COVID-19: consider cytokine storm syndromes and immunosuppression. *The Lancet* **395**:1033–1034. DOI: [https://doi.org/10.1016/S0140-6736\(20\)30628-0](https://doi.org/10.1016/S0140-6736(20)30628-0)
- Monteil V, Kwon H, Prado P, Hagelkrüys A, Wimmer RA, Stahl M, Leopoldi A, Garreta E, Hurtado Del Pozo C, Prosper F, Romero JP, Wirsberger G, Zhang H, Slutsky AS, Conder R, Montserrat N, Mirazimi A, Penninger JM. 2020. Inhibition of SARS-CoV-2 infections in engineered human tissues using clinical-grade soluble human ACE2. *Cell* **181**:905–913. DOI: <https://doi.org/10.1016/j.cell.2020.04.004>, PMID: 32333836
- Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. 2015. The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**:856–860. DOI: <https://doi.org/10.1038/ng.3314>, PMID: 26121088
- NHGRI. 2003. Press release: International consortium completes human genome project. <https://www.genome.gov/11006929/2003-release-international-consortium-completes-hgp> October 23, 2020].
- Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, Jadhav A, Jensen LJ, Johnson GL, Karlson A, Leach AR, Ma'ayan A, Malovannaya A, Mani S, Mathias SL, McManus MT, et al. 2018. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery* **17**:317–332. DOI: <https://doi.org/10.1038/nrd.2018.14>, PMID: 29472638

Recovery Collaborative Group. 2020.

Dexamethasone in hospitalized patients with Covid-19 - Preliminary report. *New England Journal of Medicine* :NEJMoA2021436. DOI: <https://doi.org/10.1056/NEJMoA2021436>

Sproston NR, Ashworth JJ. 2018. Role of C-reactive protein at sites of inflammation and infection. *Frontiers in Immunology* **9**:754. DOI: <https://doi.org/10.3389/fimmu.2018.00754>, PMID: 29706967

Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. 2018. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology* **16**:e2006643. DOI: <https://doi.org/10.1371/journal.pbio.2006643>, PMID: 30226837

Su AI, Hogenesch JB. 2007. Power-law-like distributions in biomedical publications and research funding. *Genome Biology* **8**:404. DOI: <https://doi.org/10.1186/gb-2007-8-4-404>, PMID: 17472739

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. DOI: <https://doi.org/10.1093/nar/gkz389>

Wei J, Alfajaro MM, Hanna RE, DeWeirdt PC, Strine MS, Lu-Culligan WJ, Zhang SM, Graziano VR, Schmitz CO, Chen JS, Mankowski MC, Filler RB, Gasque V, de Miguel F, Chen H, Oguntuyo K, Abriola L, Surovtseva YV, Orchard RC, Lee B, et al. 2020. Genome-wide CRISPR screen reveals host genes that regulate SARS-CoV-2 infection. *Cell* . DOI: <https://doi.org/10.1016/j.cell.2020.10.028>

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**:1260–1263. DOI: <https://doi.org/10.1126/science.abb2507>, PMID: 32075877

Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley HE, Havlin S. 2019. Increasing trend of scientists to switch between topics. *Nature Communications* **10**: 3439. DOI: <https://doi.org/10.1038/s41467-019-11401-8>, PMID: 31366884