Biases Down to Our Genes: An Investigation of the Biases Driving the Genetic Ignorome

Elizabeth Nemeti[1]
[1] Department of Biomedical Informatics, Emory University, Atlanta, GA USA

E-mail: enemeti@emory.edu

Abstract.

Genomic "big data" has been increasing exponentially as a result of breakthrough technologies in high throughput sequencing. Consequently, a plethora of opportunities have emerged toward making advancements in disease therapies and our mechanistic understanding of biological processes to inform future treatments. To keep up with this new era of genomics, our research systems need to catch up by attending to their limitations concerning bias. The purpose of this narrative review article is to identify bias trends in the current literature that have directly resulted in the gene annotation gap, whereby 10% of genes from the Human Genome make up 90% of research pursuits. 30% of the genes have never been annotated at all, a rejected subset termed the ignorome, and one that impedes novel gene-disease associations from being discovered. The drivers behind this striking discrepancy in gene study have arisen from the confluence of two categories. First, historical biases such as conceptual bias, gene medical relevance and social, economic, and academic factors like publication bias and funding limitations. Second, through technical biases via experimental design, model organisms, and equipment limitations. We present recommended solutions outlined by the recent literature that primarily address calls for deliberate funding initiatives for lesser-studied genes, propose reframing experimental designs, and push to support risk-averse research in a system designed to reward investigating only heavily cited-genes for recognition.

1. Introduction

Our ability to sequence genomes has increased exponentially during the past 20 years. Whereas sequencing a singular gene might be the focus of a single lab in the pregenomics era, new high-throughput and genome sequencing technologies have made it possible to identify thousands of genes [1]. Genomic "big data" is increasing quickly, opening a wealth of possibilities for biological advancements and treatments [2]. A significant amount of funds has been invested in genome-wide association studies over the past ten years (GWAS). One of the main objectives of GWAS is to find novel genes linked with complicated human diseases and encourage study into their characterization [3]. Understanding human gene functions will help us better understand disease and enable the creation of novel therapeutic strategies [1]. The sequencing of the human genome was anticipated to be a key breakthrough because it would allow researchers to investigate previously unstudied genes by identifying all human genes [4]. Funders, governments, business, and researchers made bold claims about how genome-based discoveries would revolutionize our understanding of human biology and disease when a draft of the human genome was announced in 2000 [5]. Therefore, the expectation was set that functional annotation would be distributed across the entire human genome [1]. Today's research into human genes, however, rarely focuses on genes that have not already been carefully studied in the past [6], as in Figure 1. This revelation suggests the Human Genome Project has not fulfilled its objective. Even with a low threshold for determining function, where a protein sequence is considered functionally annotated if exhibiting similarity to other genes with given functions or having a conserved domain, most genes are still only partially annotated [2]. While relevant genes are the focus of biomedical research, an abnormal portion of research effort is directed toward already well researched genes. This positive feedback loop, also known as reinforcement or the "Matthew effect", has emerged as a strong bias in genomic literature. Less than 10% of genes have been the subject of more than 90% of research articles. More than 30% of genes have not been studied at all [6]. This unstudied subset is referred to as an ignorome, due to our ignorance of their functions [7]. Since the release of the human genome sequence, annotation inequality has increased and nearly doubled. Biomedical advancements are impeded by annotation inequality since the mechanistic studies of gene-disease connections are solely concentrated on familiar genes [4]. The discovery of these gene-disease associations cannot be further implemented in development of therapies on the foundation of incomplete biological knowledge of human gene functions [6][8]. Preclinical research is additionally biased towards the pedestaled 10% of well-studied genes. There exists a trade-off between exploration and exploitation, whereby concentrating research on extensively researched genes offers benefits to researchers, such as, the applicability of current research tools [1]. For genes of the ignorome to be researched and the historical bias gap to be overcome, the intersection of pressures and influences from economic factors, medical relevance, technological limitations, and the structure of academic systems must be addressed and adjusted.

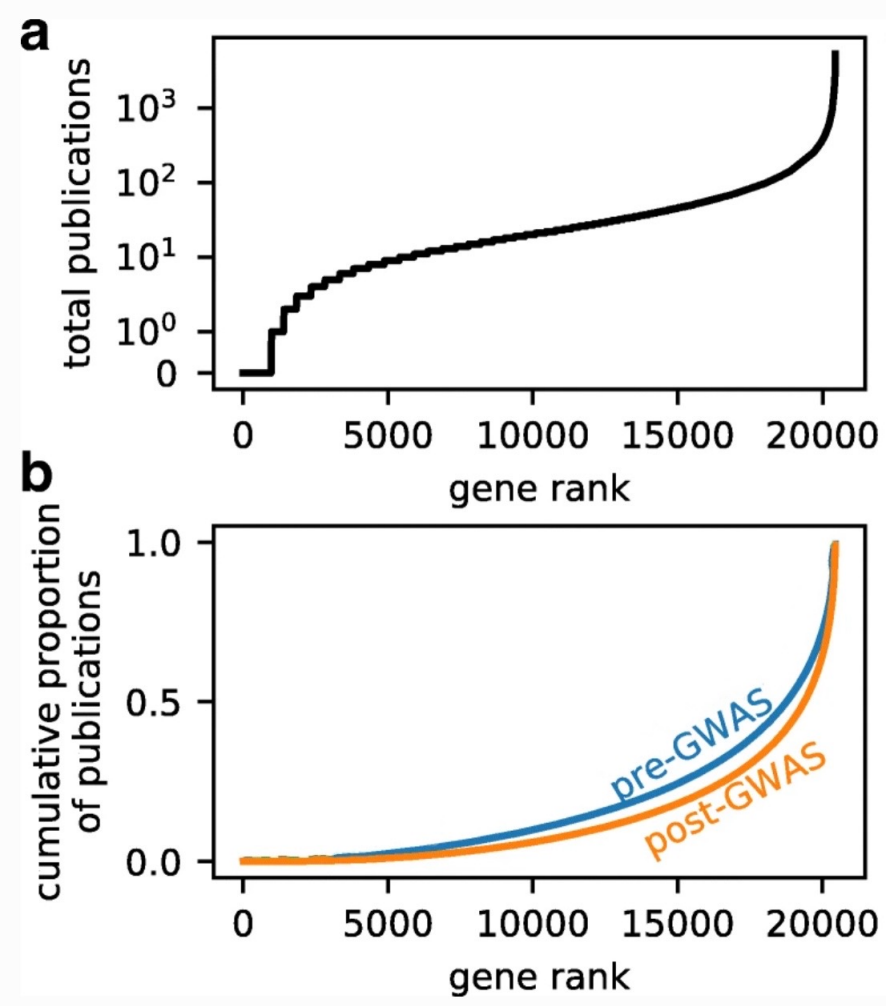

Figure 1: A) The distribution of publications among all human genes is highly uneven and skewed toward well-known genes, where the gene with fewest publications is plotted as rank 1, and the gene with the most publications as rank 20,422. B) Following the event of GWAS, research on novel genes are more uneven [3].

2. Inclusion and Exclusion Criteria

The publications selected for this narrative review were collected using the database Google Scholar advanced search option. The parameters chosen were open-access publications in a time range from 1990 (the begin of the Human Genome project)-2022. Searches were led using various key words and phrases: 'racial bias' + 'genomics', 'annotation inequality' + 'genomics', 'bias' + 'genomic data', 'technical bias' + 'genomic data". From these surveyed results, the most-highly cited publications were selected. From those papers, further publications were gathered by back-rolling from their reference lists. News and commentary articles were omitted.

3. Medical Relevance

3.1. Gene-Disease Association

Gene association with human disease via GWAS is correlated with an increase in subsequent publications [6]. Genes that have been discussed most frequently in publications were three to five times more likely to have been found in GWAS studies [1]. However, the excitement around new associations from GWAS are short-lived, as their chances of remaining "hot" or popularly studied genes is declining [3]. Genes that are investigated by many researchers are often regarded to be functionally more relevant, although this is not supported by GWAS data, hence a conceptual bias lingers heavily in the literature [4].

3.2. Intrinsic Gene Properties

While it is certain that the majority of protein-coding genes have biological significance because they are present in the human genome, certain genes will be favourably studied based on their intrinsic gene properties [9]. For example, some genes may have obvious importance, such as the delta and beta-globins, which rank among the earliest genomic clones of humans and code for the hemoglobin subunits. Most of the physiological significance of other genes might not become apparent until after their fundamental characterization outside of medical contexts [1].

3.3. Global Health Implications

In research focused on genes linked to human disease, less-studied genes could provide researchers and policymakers novel opportunities to integrate their work with societal goals. In the months that followed COVID-19's emergence as a worldwide health concern, it was discovered that more than half of the human host genes associated with COVID-19 lie outside of established research paradigms, as seen in as in Figure 2. Just as we see the stark discrepancy in the annotation inequality for the human genome, the 20% top-tagged human protein-coding genes are currently responsible for 90% of the COVID-19 literature, which is becoming increasingly dominated by a small number of genes [10]. Unless previously studied, genes that are identified by genome-wide datasets and are therefore likely to have biological importance in the context of COVID-19 have up until now gone unnoticed. This case study relating to COVID-19 demonstrates the impeding and dangerous qualities that our current research systems encourage, that have real-life consequences on global health.

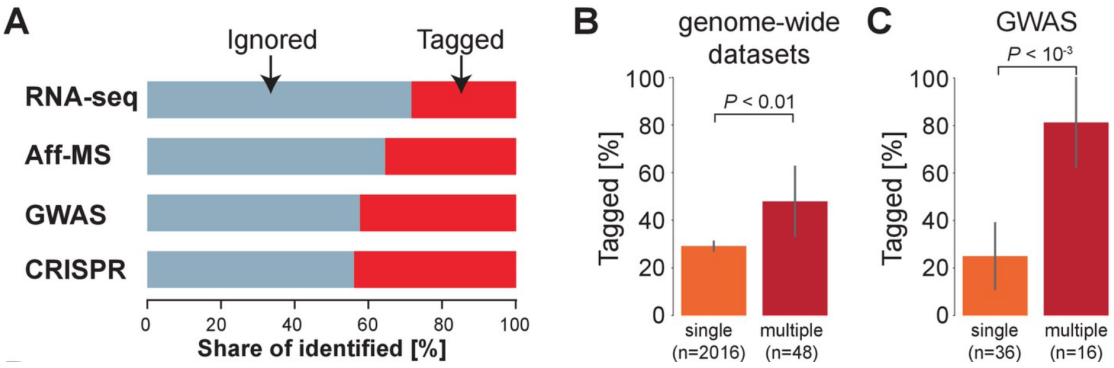

Figure 2: A) The ratio of identified genes that are ignored (never tagged, blue) or tagged (at least once, red) within the COVID-19 literature. B) The ratio of tagged genes identified by a single (orange) or multiple (maroon) genome-wide datasets, where n is the number of genes, and P-values have been calculated with Fisher's exact test. C)The ratio of tagged genes identified by a single (orange) or multiple (maroon) GWAS comparisons [10].

.

4. Social, Economic, and Academic Factors

4.1. Risk Aversion

Working in a wide research area improves the probability of being cited, which increases the likelihood of publishing in high-impact journals, which is necessary for academic success [4]. Researcher independence is 50% less likely for postdocs and PhD candidates who concentrate on weakly classified genes [6]. Original research into a novel biological function is important, but it takes time and money. A protein must have minimum annotation for researchers to invest their time in such an endeavor. Without it, there is no basis for hypotheses probing a protein's function [4]. Moreover, researchers may wind up selecting "the lowest-hanging fruit" within their domain because of the intense competition in academia [1]. The natural inclination of scientists is to delve further into their fields of specialization [5], and because funding and peer-review mechanisms are risk-averse, scientists frequently choose to investigate a topic they are already working on in greater detail, as in Figure 3.

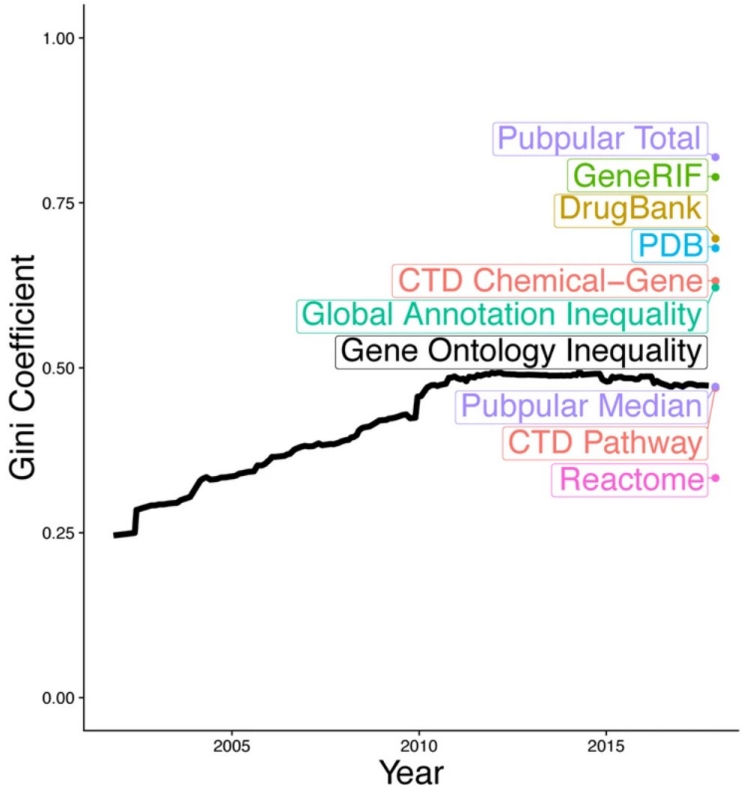

Figure 3: Inequality observed in gene annotations by measurement of the Gini coefficient across a variety of gene annotation resources [11]

4.2. Journals and Publication

The gap in the annotation of the human genome may be caused by the undervaluation of many genes or by the problematic and biased design of the academic research system [6][12]. Because there is a strong association between the frequency of a gene in scientific publications and the journal impact of the publications, researchers primarily publish on genes that have already appeared in many publications [13]. Research publications on recently discovered gene targets may be better received in certain research fields than others. Journals dedicated to neurobiology, for example, can be more enriching towards recently highlighted genes over newly highlighted genes [1]. Since researchers conduct recursive literature searches and copy references from other publications, they frequently cite works that have been cited by other works [13]. This procedure results in a power-law among the frequency distribution of genes in academic articles. This power law that governs gene annotation has researchers actively favouring continuous study on a small number of genes [14]. Consequently, beyond the relevance of a paper's subject, social processes are factors in research articles' popularity [13].

4.3. Funding and Financial Support

Many genes that have a strong link to human disease are still not comprehensively studied. Instead, funding mechanisms continue to support the current state of science's emphasis on past areas of research [6]. The power-law concerning publication is deeply present in research funding as well [14]. Only 8 of the 112 funding mechanisms utilized between 2010 and 2018 were found to encourage publications that considerably enrich early-stage research, according to a study that drew on grant databases and biomedical literature. Research on poorly understood genes is less likely to receive funding and favorable reviews, due to a greater difficulty in justifying reasoning for the study [5]. In light of the lack of financial support, researchers are not incentivized to pursue work concerning lesser-known genes.

5. Technical Bias

5.1. Experimental Design

Annotation inequality may additionally stem from error in experimental design [11]. Rather than question-driven research, the bias in gene annotations may reflect hypothesis-driven research [14]. This form of experimental design can lead to the streetlight effect, a bias whereby researchers look for results where the "light is better" that tare more easily accessible or supportive, over accuracy and truth [11]. This causes researchers to focus on genes with extensive annotations rather than those with the strongest molecular data in biomedical research, despite the fact that it has been repeatedly demonstrated that research conducted 'outside the streetlight' can uncover new gene-disease associations [11]. However, developing theories about the mechanistic molecular function of an uncharacterized gene might be challenging [5]. Yet, by having an unproven or unfounded hypotheses will result in the deduction of equally unfounded findings [15]. Researchers should reduce these biases by pursuing data-driven hypotheses [16]. The importance of the data-driven approach lies in making inferences from genomic data to develop a hypotheses and unbiased understanding of the material, without attempting to fit a pre-theorized hypothesis onto the data instead [17].

5.2. Model Organisms

Research on human genes is driven by studies on model organisms. Through the use of model organisms, we can better understand biological processes [18]. These models offer a fundamental understanding of biological mechanisms that can be applied to human biology. No model organism, however, can fully replicate the human phenotype. There is an overrepresentation of publications that cite research on nonhuman genes in reports of new human genes [6]. Although the models can help in understanding fundamental biological concepts, they lack relevance in modeling human disease. As the human population grows exponentially, there has been a radical proliferation of new alleles

unique to humans and not present in model organisms. A substantial portion of them are believed to have an impact on how diseases manifest phenotypically. However, various incentives and biases encourage scientists to keep concentrating on well-researched genes in their favored model organism [2]. To target unfamiliar genes, labs must go beyond the common laboratory setting and the limited genetic backgrounds attributed with model organisms [4]. Moreover, the surge in genomic sequence availability cannot be fully addressed by researching single genes in a model organism due to their slow-paced nature.

5.3. Equipment

The ease with which research questions involving a gene can be created and addressed depends on the available tools. The availability of experimental instruments serves as a powerful incentive to study well-researched genes, hence reinforcing the annotation inequality [4]. Where there has been a change in research activity, it was frequently prompted by the development of tools to investigate a specific gene rather than by a change in the gene's perceived importance. To close any systemic gaps left by current methodologies, it will be required to create new tools and processes. To examine genes' functioning acutely, for instance, instruments might not be sensitive enough to detect their biochemical characteristics. Due to these sensor technologies constraining dynamic ranges, mistakes in genome annotation may occur [4]. To encourage study into the understudied regions of the human genome, improving genomic tools must be a priority as new genes are discovered [5]. Without tools available, researchers will be unable to design methodologies, let alone create data-driven hypotheses to investigate novel genes. Additionally, only a limited number of highly qualified researchers tend to use genomic computational annotation tools. Experimental biologists my not be able to access the code or the analysis results because of their difference in skillset. Similar to other statistical techniques, it is frequently challenging for experimental biologists to determine which resources are the most dependable and suitable for their application, and they may select inappropriate tools for analysis [2].

6. Racial Bias

The ability to identify molecular variants in both individuals and communities has changed how society regards genetics [19]. In many cases, factors other than genetics—such as differences in culture, diet, stress, availability of medical care, education, environment, and socioeconomic status—will be more important contributors to health disparities than genetics. However, it is false to claim that personal genetics and differences in population genetics never contribute to health disparities. An example, would be the asymmetrical distribution of disease-associated alleles for the recessive disease sickle cell anemia [10]. While some scientists argue that race is a reliable indicator of ancestry, others contend that race misrepresents how genetic diversity is

distributed [20] and argue for exclusion of racial and ethnic classifications in biomedical research. However, within racial or ethnic subpopulations, population genetics research has found significant genetic variation [21]. It is critical to identify genetic distinctions between races and ethnic groups, whether they pertain to discovering novel genetic markers, pathogenic genes, or variations in treatment response [22]. However, the identification of these genes is more challenging because less is known about the pathogenicity of variations from predominantly non-European ancestry populations [23]. The ClinVar and Human Gene Mutation databases, two of the top genomic datasets, show a discernible bias in favor of genetic information based on European heritage over that based on African ancestry [24]. According to a 2009 study, 96% of GWAS participants were of European ancestry. When the same analysis was conducted again in 2016, the new results showed that the percentage of people included in the GWAS who are not of European heritage had climbed to over 20%. Nonetheless, the proportion of people with Hispanic, African, Latin American, and indigenous heritage has little changed. The GWAS is funded by the US National Institutes of Health, that mandated over 20 years ago the inclusion of diverse participants in the biomedical research that it finances. Yet, it continues to miss a significant fraction of global genetic diversity [25]. This bias makes it more difficult and expensive to apply genomic medicines to minority populations [24]. Therefore, it is not only a matter of the possible advantages to be obtained by diversifying genomic resources, but the possible dangers that not including diversity entails for patients at risk [26]. Genetically, racial, and ethnic groupings should not be considered equal in disease risk or drug responses[27]. Neither will "race-neutral" or "color-blind" approaches be equitable and reduce disparities in disease risk or treatment effectiveness between populations [22]. The number of variations discovered in a person's genome and the diseases they are susceptible to contracting may depend on the percentage ratios of their heritages. At least 15 million genetic polymorphisms are thought to exist. The significance of these variants is highlighted by the fact that several well-known inherited diseases and disorders can be brought on by a change of just one base pair, such as Alzheimer's or cystic fibrosis[23][8]. To develop more accurate medical genetic diagnoses, we must therefore broaden these databases to encompass a wider variety of ancestries [28]. The search for novel gene-disease associations across racial and ethnic groups needs to see greater support to achieve the medical goal of defining genetic global human variation [20][22]. With the continuum of racial bias in genomic databases left virtually unrecognized, patients remain at risk, and the challenges concerning application and cost-effectiveness for treating minorities will not be resolved [23].

7. Mitigation and Future Course

To conclude, several publications have suggested ways forward to overcoming this collection of historical, technical, and racial biases that define the gene ignorome. First, Stoeger and Amaral, 2020 suggest that we will be better able to enable the establishment

of initiatives and policies that support research into new or different sets of genes if we are aware of the factors that have historically led to early-stage research on novel genes [10]. To follow, Reynolds et al. 2017 proposed a research initiative to encourage better integration of lesser-studied gene data, by collecting rich annotations on biochemical properties and physiological roles for example into a well-maintained database [2]. Beyond organizing the data, Kustatscher et al. 2022 proposed the formation of the Understudied Protein Initiative, a project actively closing the annotation inequality gap by systematically associating uncharacterized proteins with proteins of known cellular processes, for future reference in mechanistic studies that could possibly be applied to genes as well [4]. To guide biomedical researchers toward understudied genes relating to human diseases, Struck et al. 2018 recommend encouraging follow-up studies for GWAS research to follow progress more closely. For designing such studies, they also promote data-driven approaches to reduce protein annotation inequality. Edwards et al. 2011, reiterate the critical goal of developing more genomic tools during the next decade to accurately conduct these studies and make methodologies more accessible [5]. In terms of funding, Edwards et al. 2011 advocates for two principles. The first concerning a push for granting institutions to reward risks from researchers and foster an improved distribution of funds to create the environment to do this. The second concerning the implementation of mechanisms to avoid researchers' pursuits for redundant and predicable genomic research. Finally, Dunham 2018 points a program that the American National Institutes of Health has already implemented to financially support and promote research into uncharacterized genes through dedicated funding opportunities [1]. Finally, Kessler et al. 2016, encourage the expansion of databases to include a greater variety of ancestries and push for increased research into non-European genetic variants as well. Upon first glance, the field of genomics and gene-disease associations appear to be growing exponentially. Genomic sequencing datasets are in fact, rapidly increasing, however the research system in which they are being studied is heavily littered with biases. Here, we have provided a review of suggestions from the recent literature and have outlined where deliberate care must be taken to avoid these biases, as they have implications for the progress of research, communities, and our global health.

References

[1] Ian Dunham. Human genes: Time to follow the roads less traveled? *PLoS biology*, 16(9):e3000034, 2018.

[2] Kimberly A Reynolds, Eduardo Rosa-Molinar, Robert E Ward, Hongbin Zhang, Breeanna R Urbanowicz, and A Mark Settles. Accelerating biological insight for understudied genes. *Integrative and Comparative Biology*, 61(6):2233–2243, 2021.

[3] Travis J Struck, Brian K Mannakee, and Ryan N Gutenkunst. The impact of genome-wide association studies on biomedical research publications. *Human genomics*, 12(1):1–9, 2018.

[4] Georg Kustatscher, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S Lilley, Emma Lundberg, Edward M Marcotte, Markus Ralser, et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods*, pages 1–6, 2022.

[5] Aled M Edwards, Ruth Isserlin, Gary D Bader, Stephen V Frye, Timothy M Willson, and Frank H Yu. Too many roads not taken. *Nature*, 470(7333):163–165, 2011.

[6] Thomas Stoeger and Luís A Nunes Amaral. The characteristics of early-stage research into human genes are substantially different from subsequent research. *PLoS biology*, 20(1):e3001520, 2022.

[7] Ashutosh K Pandey, Lu Lu, Xusheng Wang, Ramin Homayouni, and Robert W Williams. Functionally enigmatic genes: a case study of the brain ignorome. *PloS one*, 9(2):e88889, 2014.

[8] John Bell. The new genetics in clinical practice. *Bmj*, 316(7131):618–620, 1998.

[9] Elie Dolgin. The most popular genes in the human genome. *Nature*, 551(7681):427–432, 2017.

[10] Thomas Stoeger and Luís A Nunes Amaral. Meta-research: Covid-19 research risks ignoring important host genes due to pre-established research patterns. *Elife*, 9:e61981, 2020.

[11] Winston A Haynes, Aurelie Tomczak, and Purvesh Khatri. Gene annotation bias impedes biomedical research. *Scientific Reports*, 8(1):1–7, 2018.

[12] Marcus R Munafò, Taane G Clark, and Jonathan Flint. Assessing publication bias in genetic association studies: evidence from a recent meta-analysis. *Psychiatry research*, 129(1):39–44, 2004.

[13] Thomas Pfeiffer and Robert Hoffmann. Temporal patterns of genes in scientific publications. *Proceedings of the National Academy of Sciences*, 104(29):12052–12056, 2007.

[14] Andrew I Su and John B Hogenesch. Power-law-like distributions in biomedical publications and research funding. *Genome biology*, 8(4):1–2, 2007.

[15] David J Glass. A critique of the hypothesis, and a defense of the question, as a framework for experimentation. *Clinical Chemistry*, 56(7):1080–1085, 2010.

[16] Teppo Felin, Jan Koenderink, Joachim I Krueger, Denis Noble, and George FR Ellis. Data bias. *Genome biology*, 22(1):1–4, 2021.

[17] Sabina Leonelli. Introduction: Making sense of data-driven research in the biological and biomedical sciences. 2012.

[18] Ali J Marian. Modeling human disease phenotype in model organisms: "it's only a model!". *Circulation research*, 109(4):356–359, 2011.

[19] Francis S Collins. What we do and don't know about'race','ethnicity', genetics and health at the dawn of the genome era. *Nature genetics*, 36(11):S13–S15, 2004.

[20] Richard S Cooper. Race and genomics. *The New England journal of medicine*, 348(12):1166, 2003.

[21] Michael Bamshad, Stephen Wooding, Benjamin A Salisbury, and J Claiborne Stephens. Deconstructing the relationship between genetics and race. *Nature reviews genetics*, 5(8):598–609, 2004.

[22] Esteban González Burchard, Elad Ziv, Natasha Coyle, Scarlett Lin Gomez, Hua Tang, Andrew J Karter, Joanna L Mountain, Eliseo J Pérez-Stable, Dean Sheppard, and Neil Risch. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12):1170–1175, 2003.

[23] Michael D Kessler, Laura Yerges-Armstrong, Margaret A Taub, Amol C Shetty, Kristin Maloney, Linda Jo Bone Jeng, Ingo Ruczinski, Albert M Levin, L Williams, Terri H Beaty, et al. Challenges and disparities in the application of personalized genomic medicine to populations with african ancestry. *Nature communications*, 7(1):1–8, 2016.

[24] Lucia A Hindorff, Vence L Bonham, Lawrence C Brody, Margaret EC Ginoza, Carolyn M Hutter, Teri A Manolio, and Eric D Green. Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, 19(3):175–185, 2018.

[25] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016.

[26] Heidi L Rehm. Evolving health care through personal genomics. *Nature Reviews Genetics*, 18(4):259–267, 2017.

[27] Neil Risch, Esteban Burchard, Elad Ziv, and Hua Tang. Categorization of humans in biomedical research: genes, race and disease. *Genome biology*, 3(7):1–12, 2002.

[28] Morris W Foster and Richard R Sharp. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome research*, 12(6):844–850, 2002.