COMMENT OPEN

Characterizing the pathogenicity of genetic variants: the consequences of context

Timothy H. Ciesielski[1,2,3] ✉, Giorgio Sirugo[1,4,5], Sudha K. Iyengar[1,6,7] and Scott M. Williams[1,6,7]

Beyond initial discovery of a pathogenic variant, establishing that a variant is recurrently associated with disease is important for understanding clinical impact and disease etiology. Disappointingly, our ability to characterize pathogenicity under varied circumstances is limited. Here we discuss the role of genetic and environmental background and how it affects variant penetrance and outcomes. Specifically, genetic and environmental settings determine penetrance, and we should expect lower penetrance where contexts are diverse. For example, when over 5000 ClinVar pathogenic and loss-of-function variants were assessed in two large biobanks, UK Biobank and BioMe, the mean penetrance was only 7%. This indicates that the participants in the family-based, clinical, and case-control studies that identified these variants were more homogenous and enriched for etiologic co-factors, and the winner's curse was at play. We also emphasize that the outcome of interest can vary across conditions. The variant that causes hemoglobin S can increase the risk of death from sickling, lower the risk of death from malaria, or increase the risk of kidney disease, depending on the presence of other variants, the endemicity of malaria, and a suite of other factors. Overall, annotation on a single continuum from benign to pathogenic attempts to shoehorn a complex phenomenon into an overly simplistic framework. Variant effects often vary by context, and thus it is critical to assess potential pathogenicity in different settings. There is no panacea or easy fix, but we offer two recommendations for consideration. First, we need to routinely evaluate contexts such as sex and genetic ancestry by conducting stratified analyses and developing methods that can detect heterogenous effects (e.g. female-to-male allele proportion ratios). Second, we need to consistently document what we know about effect modifiers in our annotation databases. These are not the only possible approaches, but they begin to provide means to create robust annotations of pathogenicity.

npj Genomic Medicine (2024)9:3 ; https://doi.org/10.1038/s41525-023-00386-5

When we talk about the pathogenicity of genetic variants, what exactly are we talking about? Although this question on its surface may appear to be a trivial or simply philosophical question, it is not. It shapes the foundational logic of human genetics research and determines the utility of our work with respect to disease risk and clinical intervention. In brief, our standard definitions of pathogenicity refer to variants that are deleterious, harmful, or increase the probability of disease[1]. This sounds simple, but it is too simple, as this definition often leads us to ignore a key principle: Genes evolve and function in the contexts created by their environment, including other genetic variants. These contexts can determine penetrance and thus the ability of a variant to cause disease.

VARIANT PATHOGENICITY OFTEN DEPENDS ON CONTEXT

A simple but informative example of the heterogeneity of pathogenicity is the beta globin variant that causes hemoglobin S (HbS). The HbS allele in an individual who is homozygous for this variant has sickle cell disease, thereby increasing risk of death at a young age[2,3]. However, this same allele in the context of a second allele that encodes HbA will reduce the risk of risk of death at a young age in malaria endemic regions[4,5]. This decreased risk of death is the reason that the HbS allele is common in malaria endemic regions, and has not been culled by evolution[6,7]. Furthermore, in regions without malaria, being heterozygous for the HbS allele may not affect risk of death at a young age, unless there exists another precipitating variant in that individual's genome, or the carrier experiences hypoxia when exercising at high altitude[8–10]. In these distinct, yet malaria-free, contexts, the HbS variant may again increase death risk at a young age. To add to this complexity, data now indicate that older heterozygotes may have an increased risk of subclinical kidney pathology, and increased rates of acute renal failure when exposed to Sars-CoV-2[11]. Finally, variants that decrease the expression of alpha globin subunits (HBA1 and HBA2—alpha thalassemia)[12,13] or allow for the persistent expression of gamma globin subunits into adulthood (HBG1 and HBG2 – persistence of fetal hemoglobin)[14] can greatly mitigate the risk of death due to HbS homozygosity. Thus, the pathogenicity of the HbS variant depends heavily on other alleles, the environment, and the health outcome being evaluated. HbS can be considered a "simple" case, but even in this situation, pathogenic potential is strongly shaped by multiple contextual factors (Fig. 1).

This example clarifies that the process of making a universal pathogenicity assessment, uses an oversimplistic framework to describe an inherently complex phenomenon. Even when a variant can cause disease, it often does not, and knowing the modifying factors is critical to evaluating pathogenicity. Thus assuming that genetic variants have a single unidirectional effect

[1]The Department of Population and Quantitative Health Sciences at Case Western Reserve University School of Medicine, Cleveland, OH, USA. [2]Mary Ann Swetland Center for Environmental Health at Case Western Reserve University School of Medicine, Cleveland, OH, USA. [3]Ronin Institute, Montclair, NJ, USA. [4]Institute of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5]Division of Translational Medicine and Human Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]The Department of Genetics and Genome Sciences at Case Western Reserve University School of Medicine, Cleveland, OH, USA. [7]Cleveland Institute for Computational Biology, Cleveland, OH, USA. ✉email: thc23@case.edu

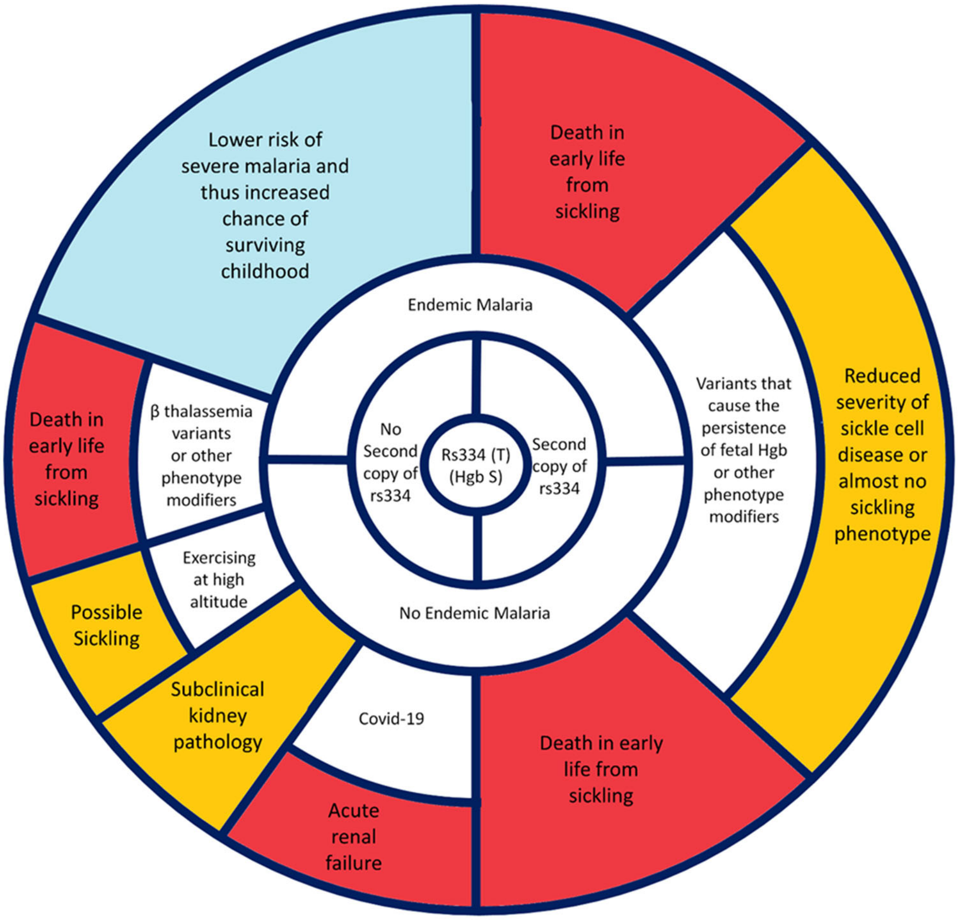
Lower risk of severe malaria and thus increased chance of surviving childhood
Death in early life from sickling
Reduced severity of sickle cell disease or almost no sickling phenotype
Death in early life from sickling
Acute renal failure
Subclinical kidney pathology
Possible Sickling
Death in early life from sickling
β thalassemia variants or other phenotype modifiers
Exercising at high altitude
Covid-19
Variants that cause the persistence of fetal Hgb or other phenotype modifiers
Endemic Malaria
No Endemic Malaria
No Second copy of rs334
Second copy of rs334
Rs334 (T) (Hgb S)

Fig. 1 The complex determinants of phenotype and pathogenicity: example of a relatively simple case—Hemoglobin S (rs334). Starting at the center of this schematic and moving out radially in any direction, different relevant contexts are encountered. These contexts determine the type and severity of the observed phenotypes. This schematic is based on our current understanding and is not intended to be an exhaustive description of all relevant and all possible phenotypes linked to rs334. Some of the modifying contexts and relevant phenotypes may yet to be discovered. Finally, although it cannot be comprehensively depicted on this figure, phenotypes may serve as competing risks for one another, and this becomes more complex with age. As an example, a person cannot develop chronic kidney disease in older age if they die of sickling complications at a younger age.

on one outcome, obscures the complex genetic architecture of disease[15]. Regulatory processes, genetic buffering, environmental interactions, and epistasis can all play roles in determining the impact of a given variant[16–19], and these contexts cannot be ignored if we want to understand variant pathogenicity[15].

DEFINING PATHOGENICITY IS ESPECIALLY HARD FOR VARIANTS WITH LOW PENETRANCE AND VARIABLE EXPRESSIVITY

Nonetheless, attempts are still made to produce "universal pathogenicity" assessments[20]. These assessments may make sense in the context of highly penetrant variants that cause Mendelian disease, but what about low penetrance variants with variable expressivity? Allelic expression levels, epigenetic changes, cis variants, trans variants, environmental exposures, and other factors, including lifestyle, collectively shape variant impact[21,22] and low penetrance variants make up a very large proportion of our annotations. When over 5000 pathogenic and loss-of-function variants were assessed in the UK Biobank and BioMe, the mean penetrance was unexpectedly low (6.9%, 95% CI: 6.0–7.8%)[23]. While some of this pattern can be partly explained by the factors that drive the winner's curse (i.e. inflated magnitude of initial associations due to low power, publication bias, model overfitting, etc.)[24,25], it must be added that smaller associations should be expected when the study participants are more diverse. Family-based, clinical, and case-control studies have more homogenous participants and because study entry is partly conditioned on disease status, these study groups are enriched for etiologic co-factors. This means lower penetrance and smaller effect sizes will often be observed in large population-based cohorts[22,26,27], even when there are subgroups where penetrance is high. When a variant has a smaller effect size and reduced penetrance in a heterogenous, population-based sample, it is important to examine that variant in multiple contexts. This can identify potentially sensitive subgroups, such as an ancestries, environments, or multiplexed families with higher penetrance and pathogenicity. Overall, assessment of variants in multiple contexts[28,29] is critical to understanding differences in the causal mechanisms of disease in distinct groups.

DOWNPLAYING THIS HETEROGENEITY IMPAIRS CLINICAL COMMUNICATION AND PRACTICE

Regardless of the reason for low penetrance, it creates a problem for pathogenicity assessments and clinical genetic practice. When these annotations are used as screening tests for disease risk, there is a systematic problem with test specificity (i.e., the ability of a test to identify true negatives and avoid false positives[30]). Since penetrance among many pathogenic variants is often low, most people with these variants will not develop disease. Thus, when applied clinically this can result in a very large number of false positives and subsequent unnecessary actions. While a strong argument can be made for tolerating false positives (type 1 error) in the early stages of genetic discovery research[31,32], false positives in clinical settings can lead to patient anxiety, needless expense, and harm[33].

One way to vet putative pathogenicity is to perform experiments that biologically validate the effects of genetic variants. However, it should be noted that such experiments are limited in their generalizability, and they are restricted by the conditions under which the experiments are performed. In vitro experiments and animal models can clearly demonstrate causal and mechanistic evidence of pathogenicity, but they cannot test or create all relevant contexts. For example, the experimental temperature, day night cycle, diet, air quality, or hormonal milieu may not reflect those of the humans that carry a potentially pathogenic variant. Geneticists are aware of these dynamics, known as *reaction norms*, and they have been taught in genetics classes for decades[34,35]. However some physicians and the general public may not be as familiar with how this fundamental principle of genetic variation can affect our annotations.

Universal pathogenicity assessments also create a systematic problem with sensitivity (i.e., the ability of a test to identify true positives and avoid false negatives[30]). This is partly because our annotation guidelines[36], even when thoughtfully refined[37] have traditionally considered the “absence of evidence” to be “evidence of absence”. In other words, when a variant is observed in a high number of healthy people (e.g., minor allele frequency [MAF] >5%) and it has not been yet linked to disease, then it can be labeled benign. Unfortunately, this approach fails to account for the determinants of penetrance. If a key determinant of penetrance was not present among the observations, then a conditionally pathogenic variant can be labeled a *Variant of Unknown Significance* or even *Benign*. This creates many issues but it seems particularly troublesome in the clinic when sequencing patients to identify the cause of rare syndromes[38]. Imagine trying to annotate the phenylalanine hydroxylase gene variants that cause phenylketonuria[39] in a population with almost no access to foods that contain phenylalanine. Phenylalanine hydroxylase variants would appear benign in this context. Hence, in most cases when variant pathogenicity is assessed, the process identifies what *can* cause disease, but importantly, it does not identify what *will* cause disease in a given person at a given time[40,41]. This context agnostic approach has utility, but its limitations must be acknowledged and accounted for.

EXISTING GENOMIC METHODS IMPROVE WHEN CONTEXT IS CONSIDERED

Despite the drawbacks of often defining pathogenicity as a binary and immutable feature of variants, genetic researchers have created many techniques of great utility. For example, molecular algorithms have been developed that can predict loss of protein function and these have high value in many settings[42–44]. We also now have protocols for molecular and clinical validation with laboratory-based functional assays[45], and the longitudinal tracking of sequenced individuals in electronic health records[46]. Furthermore, several key papers have improved our thinking about the necessity of using diverse convergent evidence for causal reasoning in genomics[31,47,48]. Perhaps the most impressive advance in this area, is the scoring system developed by ClinGen that assembles and interprets empirical evidence for pathogenicity[49]. However, these approaches can only do so much when context is not explicitly considered. For example, even if we could develop a prediction algorithm that perfectly determined loss-of-function in any protein, we would still not know if loss-of-function was good or bad for any individual (given the remainder of their genome, and their environment, and the phenotype in question)[50–56]. Take for instance a protein that can convert procarcinogenic compounds to carcinogens. Loss-of-function of this protein may be beneficial in the context of high procarcinogen exposure[57]. Hence, the context, in this case the environment, can change a variant from beneficial to pathogenic and vice versa.

Therefore, even if we are using the best methods, we can observe conflicting evidence of pathogenicity when we do not explicitly consider context. This is particularly relevant for common variants. If a given variant is detrimental in all contexts, then this variant will usually be observed as a rare or de novo variant. In other words, variants are persistently culled by evolution when they reduce reproductive fitness in all contexts, but they can be maintained in the contexts where they do not reduce reproductive fitness. This may be especially evident when we consider pleiotropy, because antagonistic pleiotropy appears to play a major role in the persistence of several human disease variants[58,59]. For example, the strongest genetic determinant of Alzheimer’s Disease, APOE4[60,61], also prevents death from diarrhea in childhood[62,63]. Our ancestors probably needed infection protection for their reproductive fitness and one of the variants that met this early life requirement, also increased the risk of a late life disease, Alzheimer’s Disease[62–66]. Thus, it makes very little sense to talk about the universal pathogenicity of any common variant. However, from a practical perspective, it is hard to do anything else.

CONTEXT IS COMPLEX—HOW CAN WE SPECIFY IT?

Context is easy to invoke as a concept, but the relevant context or determinants of penetrance, can differ for virtually every variant. Thus, when operationalizing research questions: What contexts do we measure? What contexts do we analyze? What phenotype do we examine? Even in the simplest research case with a single SNP, the potentially relevant context can be a cryptic and computationally impractical search space. Unfortunately, this explodes into intractability when considering Genome Wide Association or Next Generation Sequencing data (millions of SNPs and potentially thousands of environmental exposome variables). So, how can this problem be addressed? How can contexts that need attention be identified? It may be most practical to start with common and easily measured “contexts” that are known to have strong biological functions. This will help to optimize precision, statistical power, and the likelihood of documenting context-dependent pathogenicity.

With these features in mind, biological sex is among the easiest contexts to evaluate. It is easily measurable, it divides all human populations approximately in half, and there are many anatomic, physiologic, and pathophysiologic distinctions that align with it. Thus we can, and probably should, run sex-stratified sensitivity analyses in most genetic research studies[67–69] especially when a trait is sexually dimorphic[70]. Failure to do this can obscure important biological patterns. Another step would be to encourage new methods for probing the X-chromosome, a chromosome that is often-ignored in association analyses. We have already started this strategy by analyzing the female-to-male allele frequency ratio as tool for the discovery of pathogenic variants (Equation 1)[71]. The reasoning is as follows: females have 2 copies of all Non-Pseudoautosomal X-chromosome loci and males only have one. Thus, females can be biologically more resilient to the presence of harmful variants at these sites. The exception is variants with dominant effects, in which case ratios will not be useful for detecting these variants. In any dataset of adult humans, when a Non-Pseudoautosomal X-chromosome variant exists at a higher proportion in females, this pattern can serve as evidence that the variant may increase the probability of premature death.

Following this simple logic, we used gnomAD data[72] to characterize this phenomenon. Our methods are fully described in[71], but in short, we obtained exome data from the

X-Chromosomes of 76,702 males and 64,754 females. Then, we calculated female-to-male allele frequency ratios for the 44,606 variants that had an allele count of at least 5. None of the pseudoautosomal variants had a ratio above 11, but 319 of the non-pseudoautosomal variants had ratios above this empiric threshold.

Only 25 of these high-ratio variants were annotated in ClinVAR and had a rs number. Most of these variants had high sex-averaged MAFs and no known associations with disease, and they were listed as *benign* or *likely benign* (Table 1). As an example, one of the 25 variants had a sex-averaged MAF of 0.13, no known disease associations, and was listed as *likely benign*. This site had been genotyped 38,527 times in males (one locus each) and 104,056 times in females (2 loci each), so there was no shortage of data. Overall, the variant was observed a total of 18,736 times, but not one of these observations came from a male or a homozygous female. It was only found in heterozygous females. Thus, it is likely that this variant is almost 100% lethal (perhaps even embryonic lethal) in males and homozygous females, but is without large effect in heterozygous females. When we considered the other 24 variants, we found similar patterns, although the comparisons were less extreme.

To further characterize these variants, we probed them with a diverse set of web-based bioinformatic resources: dbSNP[73], VarSome[74], OMIM[75], and VENUS[76,77]. These databases provide additional information on evolutionary conservation, gene-phenotype relationships, protein-structure predictions, and other aspects of these variants that need consideration in pathogenicity assessments. We found that:

1. Existing annotation methods can miss sex-specific pathogenicity. We observed that 22 out of 25 (88%) high ratio variants are listed as Benign or Likely Benign in ClinVar (1 is listed as Conflicting [Uncertain Significance and Benign] 2 are listed as Uncertain Significance). These variants are commonly observed in healthy heterozygous females and they achieve high sex-averaged MAFs so they appear benign, but males are rarely observed (i.e., these variants are not often tolerated in males)
2. QC procedures can mislabel evidence of sex-specific pathogenicity as genotyping error. We looked in the second dataset from gnomAD site (the genomes data) and observed that 22 out of the 25 (88%) high ratio variants failed QC filters[74]. Sex differences in MAF were assumed to be error rather than putative evidence of sex-specific pathogenicity. Thus, these QC filters may systematically remove variants with sex-specific pathogenicity before they can even be assessed.
3. Our ratio method identified genes that were already linked to clinical syndromes through other variants. In all, 23 of 25 (92%) genes implicated by the high ratio variants have specific links to clinical syndromes listed in OMIM[75]. The other two genes have tentative links to pathology described in their OMIM entry.
4. Structural predictions are not available or useful for most of these top ratio hits. Michaelangelo-VENUS structural predictions[76,77] were only possible for 6 of the 25 variants (24%). VENUS requires the specification of a specific amino acid substitution at a specific site in the protein. This makes sense for some variants, but 19 of the 25 variants do not have that impact, or their exact impact on amino acid sequence cannot be yet specified (synonymous, intronic, splice donor variants, etc.)
5. Additional heterogeneity exists and some high ratio variants might be better tolerated by males and homozygous females in specific contexts. Some high ratio alleles had frequencies that differed by ancestry group, and this is consistent with the interpretation that these variants may not have sex-specific pathogenicity in all contexts.

Overall, these 5 points indicate that seeking and documenting evidence of sex-specific effects could improve pathogenicity annotations. The existing tools for variant characterization can only do so much if context is not explicitly evaluated. Finally, we note that the many potential mechanisms for sex-specific pathogenicity remain to be characterized, but there is some indication in our initial results that regulatory function may sometimes be involved. RegulomeDB evaluations of the 25 high-ratio variants provide diverse and nuanced information on the likelihood of regulatory function at these loci (Table 2). They reveal that 13 of the 25 high ratio variants (52%) have some indication of regulatory function: a rank less than three or a score greater than 0.5. A rank less than three indicates the presence of at least two strong pieces of experimental evidence that are consistent with regulatory function, and scores greater than 0.5 are in the top half of possible scores from models that predict transcription factor binding.

Sex differences in allele frequency on the X chromosome are a special case, but this pattern may also be found in autosomal variants that affect disease risk differently between males and females. Very large and very small allele proportion ratios in the autosomes may also be indicative of sex-specific effects that deserve further investigation. While this area of genetic research is still in its infancy, and thresholds for discovery and confirmatory findings are not yet established, we have already observed extreme female-to-male allele proportion ratios on autosomes (many standard deviations above or below the mean). Work in progress has already revealed a distribution of ratios on chromosome 21 that demonstrates this point (Table 3). Ratios this high are very unlikely occur by chance. Finally, we note that biological sex is just the first and simplest context to consider. More complex situations such as ancestry and environmental exposures will need increased attention. For example, we already know that failing to assess ancestry-specific associations can generate ancestry-specific misinterpretations of genetic tests that disproportionally harm marginalized groups[78]. We need to collect genetic data on diverse ancestry groups[79] and explicitly consider this context in order to avoid generating health disparities with ancestry-specific medical error[80].

Overall, considering context will not solve all the problems in pathogenicity assessment, but it is a necessary step for addressing key clinical and translational issues in genetics. Sex-stratified GWAS[70], and female-to-male allele proportion ratios[71] can start us on a path that probes multiple determinants of penetrance. A lot of work remains in determining how to best explore contextual frameworks for variant pathogenicity, and other tools will be needed to evaluate additional factors, such as xenobiotic exposures and ancestry. However, biological sex is an ideal context to start with, because it will not require any new data. Information on biological sex is extractable from virtually all existing genomic data, and these data can be easily re-evaluated at low cost. Furthermore, it will not be hard or expensive to better evaluate sex differentials in allele frequency and improve the definition of *benign* in pathogenicity annotations. As an easy first step, ClinVar could present MAFs by sex. Overall, we call on the genetic research community to proactively consider context. While the optimal frameworks for achieving this goal are not fully established, we can to start by routinely evaluating the sexes separately, and documenting what is known about effect modifiers in our annotations. We have proposed a deeper dive into sex as a common effect modifier but other strata should be explored and documented in annotations. Covariates should be collected in our datasets and exploratory sensitivity analyses should be more routine or we will fail to identify many determinants of penetrance that have clinical relevance.

Table 1. Variants identified in gnomAD exome data that have an allele proportion ratio above 11 and a ClinVar entry.

rs number and gene from dbSNP[73]	ClinVar entry[20,81]	Female to male allele proportion ratio[a71]	Listed as failing QC in the gnomAD Genome data[74]	Total no. of observations in the gnomAD Exome data[72,74]	Varsome pathogenicity scores summary[74]	PhyloP 100way Conservation score listed in varsome (higher = more conserved)[74]	Complications in interpretation noted in varsome[74]	Unique feature observed in varsome[74]	Gene- phenotype association from OMIM[75]	Predicted effect of variant on protein structure from Michelangelo – VENUS[76,77]
rs201580891 FMR1	Likely benign	6937.5	yes	18,736	1 Pathogenic 8 Uncertain 6 Benign	2.471	NA	Only observed in heterozygous females	Missense variant in FMR1—the gene linked to Fragile X Syndrome, Fragile X tremor/ataxia syndrome, Premature ovarian failure 1	Structurally neutral —K119N (a variant may be structurally neutral, but phenotypically deleterious)
rs1315062158 IQSEC	Benign/likely benign	1809.3	yes	3617	no data	0.985	NA	Only observed in heterozygous females	Synonymous variant in IQSEC – a gene linked X-linked intellectual developmental disorder-1	N/A Synonymous Variant
rs782666190 SMC1A	Benign	929.4	yes	4418	no data	0.100	NA	Males are rare, and homozygous females are absent	Intronic Variant in SMC1A – a gene tied to Cornelia de Lange syndrome-2, and developmental and epileptic encephalopathy-85	N/A Intronic Variant
rs1432363549 SMC1A	Likely benign	487.7	yes	968	no data	−0.691	NA	Only observed in heterozygous females	Intronic Variant in SMC1A – a gene tied to Cornelia de Lange syndrome-2, and developmental and epileptic encephalopathy-85	N/A Intronic Variant
rs782705493 HDAC8	Benign	379.3	yes	1657	no data	0.131	NA	Males and homozygous females are rare	Intronic Variant in HDAC8—a gene tied to Cornelia de Lange syndrome-5	N/A Intronic Variant
rs777010333 COL4A6	Likely benign	279.1	yes	563	no data	−0.995	NA	Only observed in heterozygous females	Intronic Variant in COL4A6—tied to X-linked deafness-6	N/A Intronic Variant
rs782664878 SMC1A	Benign	263.8	yes	8386	no data	−0.522	NA	Males and homozygous females are rare	Intronic Variant in SMC1A – a gene tied to Cornelia de Lange syndrome-2, and developmental and epileptic encephalopathy-85	N/A Intronic Variant
rs372580592 SLC9A6	Likely benign	223.1	yes	493	6 Benign	−0.104	NA	Only observed in heterozygous females	Intronic Variant in SLC9A6 – a gene tied to Christianson type of X-linked syndromic intellectual developmental disorder	N/A Intronic Variant

Table 1 continued

rs number and gene from dbSNP[73]	ClinVar entry[20,81]	Female to male allele proportion ratio[a71]	Listed as failing QC in the gnomAD Genome data[74]	Total no. of observations in the gnomAD Exome data[72,74]	Varsome pathogenicity scores summary[74]	PhyloP 100way Conservation score listed in varsome (higher = more conserved)[74]	Complications in interpretation noted in varsome[74]	Unique feature observed in varsome[74]	Gene- phenotype association from OMIM[75]	Predicted effect of variant on protein structure from Michelangelo – VENUS[76,77]
rs782792601 NDUFB11	Benign	201.2	yes	624	4 Pathogenic 1 Uncertain 2 Benign	1.656	NA	Males are absent and homozygous females are rare	Splice Donor Variant in NDUFB11—a gene tied to Linear skin defects with multiple congenital anomalies 3	N/A Splice Donor Variant (VENUS requires the specification of a single AA change at a single site)
rs782032695 EBP	Likely benign	194.5	yes	398	no data	0.342	NA	Only observed in heterozygous females	5′ UTR Variant in EBP—a gene tied to X-linked dominant chondrodysplasia punctata-2 and MEND syndrome	N/A 5′ UTR Variant
rs781824575 HDAC8	Benign	131.5	yes	6367	no data	−1.565	NA	Hemizygous Males and homozygous females are rare	Intronic Variant in HDAC8—a gene tied to Cornelia de Lange syndrome-5	N/A Intronic Variant
rs745354475 USP9X	Benign	129.4	yes	432	6 Benign	2.378	In a segmental duplication region	Only observed in heterozygous females	Intronic variant in USP9X – a gene tied to X-linked intellectual developmental disorder-99 and Female-restricted X-linked syndromic intellectual developmental disorder-99	N/A Intronic Variant
rs199626569 GJB1	Uncertain significance	89.3	NA	176	10 Pathogenic 8 Uncertain 4 Benign	8.015	NA	Only observed in heterozygous females	Missense variant in GJB1 – a gene tied to X-linked dominant Charcot-Marie-Tooth neuropathy 1	Structurally neutral —V166G (a variant may be structurally neutral, but phenotypically deleterious)
rs782072345 NDUFB11	Benign	87.6	no	1340	1 Pathogenic 1 Uncertain	1.254	NA	Hemizygous Males and homozygous females are rare	Intronic Variant in NDUFB11—a gene tied to Linear skin defects with multiple congenital anomalies 3	N/A Intronic Variant
rs745338783 POLA1	Likely benign	87.1	yes	201	10 Benign	−0.413	In a low complexity region	Only observed in heterozygous females	Intronic Variant in POLA1 – a gene tied to X-linked reticulate pigmentary disorder (PDR) with systemic manifestations and Van Esch-O'Driscoll syndrome	N/A Intronic Variant

Table 1 continued

rs number and gene from dbSNP[73]	ClinVar entry[20,81]	Female to male allele proportion ratio[a71]	Listed as failing QC in the gnomAD Genome data[74]	Total no. of observations in the gnomAD Exome data[72,74]	Varsome pathogenicity scores summary[74]	PhyloP 100way Conservation score listed in varsome (higher = more conserved)[74]	Complications in interpretation noted in varsome[74]	Unique feature observed in varsome[74]	Gene- phenotype association from OMIM[75]	Predicted effect of variant on protein structure from Michelangelo – VENUS[76,77]
rs751314374 RPGR	Conflicting (Uncertain significance and Benign)	70.7	yes	308	3 Uncertain 18 Benign	−1.198	In a low complexity region	Only observed in heterozygous females (more common in East and South Asians)	Missense Variant in RPGR—a gene tied to X-linked cone-rod dystrophy-1, X-linked atrophic macular degeneration, retinitis pigmentosa-3, and X-linked retinitis pigmentosa and sinorespiratory infections, with or without deafness	Structurally neutral - E934G (a variant may be structurally neutral, but phenotypically deleterious) This was run on the only isoform in VENUS that has at least 934 residues (Q92834)
rs1250133030 RPGR	Likely benign	64.9	yes	318	2 Uncertain 24 Benign	−0.145	In a low complexity region	Only observed in heterozygous females	Missense Variant in RPGR—a gene tied to X-linked cone-rod dystrophy-1, X-linked atrophic macular degeneration, retinitis pigmentosa-3, and X-linked retinitis pigmentosa and sinorespiratory infections, with or without deafness	Unclear - ClinVar lists this variant as creating a K857E amino acid change - only one isoform listed in VENUS has this many residues, and this isoform has an E at position 857
rs72609545 VCX3A	Benign	58.1	yes	1,218	2 Uncertain 21 Benign	−4.362	In a segmental duplication region	Hemizygous Males and homozygous females are rare	Missense Variant in VCX3A – a gene putatively tied to X-linked Ichthyosis (it is in a region that is implicated)	Structurally neutral - V140M (a variant may be structurally neutral, but phenotypically deleterious)
rs12849277 MED12	Benign/Likely benign	55.2	yes	83	No data	−1.295	In a low complexity region	Only observed in heterozygous females	Intronic variant in MED12 – a gene tied to Hardikar syndrome, Lujan-Fryns syndrome, X-linked Ohdo syndrome, and Opitz-Kaveggia syndrome	N/A Intronic Variant
rs781379769 USP9X	Benign/Likely Benign	53.3	yes	6602	2 Benign	−0.241	In a low complexity region	Hemizygous Males and homozygous females are rare	Intronic variant in USP9X – a gene tied to X-linked intellectual developmental disorder-99 and Female-restricted X-linked syndromic intellectual developmental disorder-99	N/A Intronic Variant

Table 1 continued

rs number and gene from dbSNP[73]	ClinVar entry[20,81]	Female to male allele proportion ratio[a71]	Listed as failing QC in the gnomAD Genome data[74]	Total no. of observations in the gnomAD Exome data[72,74]	Varsome pathogenicity scores summary[74]	PhyloP 100way Conservation score listed in varsome (higher = more conserved)[74]	Complications in interpretation noted in varsome[74]	Unique feature observed in varsome[74]	Gene- phenotype association from OMIM[75]	Predicted effect of variant on protein structure from Michelangelo – VENUS[76,77]
rs145404090 SAGE1	Uncertain significance	44.8	yes	79	3 Uncertain 23 Benign	0.252	In a segmental duplication region	Only observed in heterozygous females 92% were European	Missense variant in SAGE1 – a geneputatively tied to cancer Part of a set of genes that are only expressed in tumors, spermatogenic and placental cells	Structurally neutral —T203A (a variant may be structurally neutral, but phenotypically deleterious)
rs148934011 RBMX	Uncertain significance	33.6	yes	56	5 Pathogenic 8 Uncertain 2 Benign	6.024	In a segmental duplication region	Only observed in heterozygous females 95% were African	Missense variant in RBMX – a gene tied to X-linked syndromic intellectual developmental disorder-11, Sashi type	Stabilizing - D333Y (a variant may be stabilizing, but phenotypically deleterious)
rs782233695 RHOXF2	Likely benign	22.7	yes	39	4 Uncertain 24 Benign	−4.628	In a segmental duplication region	Only observed in heterozygous females 97% were South Asian	Missense variant in RHOXF2 – a gene with no phenotypes noted in OMIM	Unclear - ClinVar notes that the change is H139Y, but VENUS notes there is an N at position 139
rs1446705794 RPGR	Likely benign	20.9	yes	91	no data	−1.556	In a low complexity region	Only observed in heterozygous females	Synonymous Variant in RPGR—a gene tied to X-linked cone-rod dystrophy-1, X-linked atrophic macular degeneration, retinitis pigmentosa-3, and X-linked retinitis pigmentosa and sinorespiratory infections, with or without deafness	N/A Synonymous Variant
rs201558029 DMD	Benign	15.7	no	26	no data	1.541	NA	Only observed in heterozygous African or Latino Females	Intronic Variant in DMD – a gene linked to Becker muscular dystrophy, Duchenne muscular dystrophy, and X-linked dilated cardiomyopathy-3B	N/A Intronic Variant

[a] The female to male allele proportion ratio: $\boldsymbol{R} = \frac{(V_f+1)/(A_f+1)}{(V_m+1)/(A_m+1)}$

R allele proportion ratio, V_f the minor allele count in females, A_f the total allele count in females, V_m the minor allele count in males, A_m the total allele count in males. Note that 1 is added to the numerators and denominators to avoid dividing by 0. This allows a female-to-male allele proportion ratio to be calculated when no male carriers are observed with a given variant.

Table 2. Evidence of regulatory function among the high ratio variants.

rs number from dbSNP[73]	RegulomeDB—rank[82,83] [1] Integrative metric based on existing evidence 1 = strong evidence 7 = no evidence	RegulomeDB—score[82,83] prediction based on transcription factor binding models 0 = lowest probability of regulatory function 1 = highest probability of regulatory function
rs201580891	7	0.18412
rs1315062158	5	0.38000
rs782666190	5	0.00454
rs1432363549	5	0.00000
rs782705493	2b	0.73553
rs777010333	5	0.01895
rs782664878	5	0.00000
rs372580592	5	0.09659
rs782792601	4	0.70497
rs782032695	2b	0.48000
rs781824575	2b	0.79371
rs745354475	5	0.00000
rs199626569	4	0.60906
rs782072345	4	0.70497
rs745338783	5	0.00125
rs751314374	5	0.58955
rs1250133030	5	0.58955
rs72609545	5	0.58955
rs12849277	5	0.58955
rs781379769	5	0.00000
rs145404090	5	0.23589
rs148934011	5	0.58955
rs782233695	4	0.60906
rs1446705794	5	0.98500
rs201558029	7	0.18412

[1] Possible scores in the RegulomeDB ranking system.
1a eQTL/caQTL + TF binding + matched TF motif + matched Footprint + chromatin accessibility peak.
1b eQTL/caQTL + TF binding + any motif + Footprint + chromatin accessibility peak.
1c eQTL/caQTL + TF binding + matched TF motif + chromatin accessibility peak.
1d eQTL/caQTL + TF binding + any motif + chromatin accessibility peak.
1e eQTL/caQTL + TF binding + matched TF motif.
1f eQTL/caQTL + TF binding / chromatin accessibility peak.
2a TF binding + matched TF motif + matched Footprint + chromatin accessibility peak.
2b TF binding + any motif + Footprint + chromatin accessibility peak.
2c TF binding + matched TF motif + chromatin accessibility peak.
3a TF binding + any motif + chromatin accessibility peak.
3b TF binding + matched TF motif.
4 TF binding + chromatin accessibility peak.
5 TF binding or chromatin accessibility peak.
6 Motif hit.
7 Other.

Table 3. Summary statistics for the 21493 female-to-male allele proportion ratios calculated on chromosome 21 in the GnomAD exomes data.

	Mean	SD	Min	Max
Ratio	1.5	1.1	0.1	43.6
Log2(Ratio)	0.3	0.8	−2.7	5.4

CONCLUSION

In summary, these strategies will not provide better answers to the old questions; they simply refine the questions so that they are more relevant. The old questions are generally context agnostic, and they have set the basis of our understanding reasonably well, but not well enough. If we want to keep advancing, we must now address the ubiquity of pleiotropy and the contextual determinants of penetrance.

Equation 1. The female-to-male allele proportion ratio[71]

$$\boldsymbol{R} = \frac{(V_f + \mathbf{1})/(A_f + \mathbf{1})}{(V_m + \mathbf{1})/(A_m + \mathbf{1})}$$

R: allele proportion ratio
V_f: the minor allele count in females
A_f: the total allele count in females
V_m: the minor allele count in males
A_m: the total allele count in males

DATA AVAILABILITY

All data are public and available from https://gnomad.broadinstitute.org/.

CODE AVAILABILITY

The data handling and ratio calculations are previously described[71]. We re-evaluated the hits in publicly available web databases, but there is no new code associated with this commentary.

Received: 3 July 2023; Accepted: 15 December 2023;
Published online: 09 January 2024

REFERENCES

1. Pathogenic variant. NCI Dictionary of Genetics Terms, https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/pathogenic-variant.
2. Bender, M. A. & Carlberg, K. Sickle cell disease. In: *GeneReviews(®)* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).
3. Ranque, B. et al. Estimating the risk of child mortality attributable to sickle cell anaemia in sub-Saharan Africa: a retrospective, multicentre, case-control study. *Lancet Haematol.* **9**, e208–e216 (2022).
4. Depetris-Chauvin, E. & Weil, D. N. Malaria and early african development: evidence from the sickle cell trait. *Econ. J. (London)* **128**, 1207–1234 (2018).
5. Gong, L., Parikh, S., Rosenthal, P. J. & Greenhouse, B. Biochemical and immunological mechanisms by which sickle cell trait protects against malaria. *Malar J.* **12**, 317 (2013).
6. ALLISON, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br. Med. J.* **1**, 290–294 (1954).
7. Haldane, J. Disease and evolution. *Ric. Sci.* **19**, 68–76 (1949).
8. Ashorobi, D., Ramsey, A., Yarrarapu, S. N. S. & Bhatt, R. Sickle cell trait. In *StatPearls* (StatPearls Publishing, 2022).
9. Kotila, T. R. Sickle cell trait: a benign state? *Acta Haematol.* **136**, 147–151 (2016).
10. O'Connor, F. G. et al. Summit on exercise collapse associated with sickle cell trait: finding the 'way ahead. *Curr. Sports Med. Rep.* **20**, 47–56 (2021).
11. Verma, A. et al. Association of kidney comorbidities and acute kidney failure with unfavorable outcomes after covid-19 in individuals with the sickle cell trait. *JAMA Intern. Med.* **182**, 796–804 (2022).
12. MedlinePlus. HBA1 gene - hemoglobin subunit alpha 1. https://medlineplus.gov/genetics/gene/hba1/ (2022).
13. MedlinePlus. HBA2 gene - hemoglobin subunit alpha 2. https://medlineplus.gov/genetics/gene/hba2/ (2022).
14. Serjeant, G. R. et al. A plea for the newborn diagnosis of Hb S-hereditary persistence of fetal hemoglobin. *Hemoglobin* **41**, 216–217 (2017).
15. Kumar, S. & Gerstein, M. Unified views on variant impact across many diseases. *Trends Genet.* **39**, 442–450 (2023).

16. Castel, S. E. et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
17. Hartman, J. L. 4th, Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291**, 1001–1004 (2001).
18. Domingo, J., Baeza-Centurion, P. & Lehner, B. The causes and consequences of genetic interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).
19. Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. Gene-environment interactions and their impact on human health. *Genes Immun.* **24**, 1–11 (2023).
20. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
21. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
22. Kingdom, R. & Wright, C. F. Incomplete penetrance and variable expressivity: from clinical studies to population cohorts. *Front. Genet.* **13**, 920390 (2022).
23. Forrest, I. S. et al. Population-based penetrance of deleterious clinical variants. *JAMA* **327**, 350–359 (2022).
24. Kraft, P. Curses–winner's and otherwise–in genetic epidemiology. *Epidemiology* **19**, 649–651 (2008).
25. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
26. Xiang, J. et al. Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Sci. Rep.* **10**, 331 (2020).
27. Jackson, L. et al. Influence of family history on penetrance of hereditary cancers in a population setting. *eClinicalMedicine* **64**, 102159 (2023).
28. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
29. Mensah, G. A. et al. Emerging concepts in precision medicine and cardiovascular diseases in racial and ethnic minority populations. *Circ. Res.* **125**, 7–13 (2019).
30. Trevethan, R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front. Public Health* **5**, 307 (2017).
31. Ciesielski, T. H. et al. Diverse convergent evidence in the genetic analysis of complex disease: coordinating omic, informatic, and experimental evidence to better identify and validate risk factors. *BioData Min.* **7**, 10 (2014).
32. Williams, S. M. & Haines, J. L. Correcting away the hidden heritability. *Ann. Hum. Genet.* **75**, 348–350 (2011).
33. Adams, M. C., Evans, J. P., Henderson, G. E. & Berg, J. S. The promise and peril of genomic screening in the general population. *Genet. Med.* **18**, 593–599 (2016).
34. Woltereck, R. Weitere experimentelle Untersuchungen uber Artveranderung, speziell uberdas Wesen quantitativer Artunterschyiede bei Daphniden. *Verh. D. Tsch. Zool. Ges* **1909**, 110–172 (1909).
35. Sultan, S. E. Phenotypic plasticity as an intrinsic property of organisms. *In*: Phenotypic plasticity and evolution: causes, consequences, and controversies 3–24 (CRC Press).
36. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
37. Nykamp, K. et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet. Med.* **19**, 1105–1117 (2017).
38. Sullivan, J. A., Schoch, K., Spillmann, R. C. & Shashi, V. Exome/genome sequencing in undiagnosed syndromes. *Annu. Rev. Med.* **74**, 489–502 (2023).
39. Elhawary, N. A. et al. Genetic etiology and clinical challenges of phenylketonuria. *Hum. Genomics* **16**, 22 (2022).
40. Rothman, K. J. & Greenland, S. Causation and causal inference in epidemiology. *Am. J. Public Health* **95**, S144–S150 (2005).
41. Rothman, K. J. Causes. *Am. J. Epidemiol.* **104**, 587–592 (1976).
42. Gunning, A. C. et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J. Med. Genet.* **58**, 547–555 (2021).
43. Wilcox, E. H. et al. Evaluating the impact of in silico predictors on clinical variant classification. *Genet. Med.* **24**, 924–930 (2022).
44. Pejaver, V. et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).
45. Brnich, S. E. et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).
46. Schiabor Barrett, K. M. et al. Clinical validation of genomic functional screen data: analysis of observed BRCA1 variants in an unselected population cohort. *HGG Adv.* **3**, 100086 (2022).
47. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
48. Geneletti, S., Gallo, V., Porta, M., Khoury, M. J. & Vineis, P. Assessing causal relationships in genomics: from Bradford-Hill criteria to complex gene-environment interactions and directed acyclic graphs. *Emerg. Themes Epidemiol.* **8**, 5 (2011).
49. Strande, N. T. et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *Am. J. Hum. Genet.* **100**, 895–906 (2017).
50. Siddiqui, S. S. et al. The Alzheimer's disease-protective CD33 splice variant mediates adaptive loss of function via diversion to an intracellular pool. *J. Biol. Chem.* **292**, 15312–15320 (2017).
51. Jensen, L. E., Hoess, K., Mitchell, L. E. & Whitehead, A. S. Loss of function polymorphisms in NAT1 protect against spina bifida. *Hum. Genet.* **120**, 52–57 (2006).
52. Orrú, V. et al. A loss-of-function variant of PTPN22 is associated with reduced risk of systemic lupus erythematosus. *Hum. Mol. Genet.* **18**, 569–579 (2009).
53. Mbikay, M. & Chrétien, M. The biological relevance of PCSK9: when less is better…. *Biochem. Cell Biol.* **100**, 189–198 (2022).
54. Mercader, J. M. et al. A loss-of-function splice acceptor variant in IGF2 is protective for type 2 diabetes. *Diabetes* **66**, 2903–2914 (2017).
55. Andersen, M. K. et al. Loss of sucrase-isomaltase function increases acetate levels and improves metabolic health in greenlandic cohorts. *Gastroenterology* **162**, 1171–1182.e3 (2022).
56. Xue, Y. et al. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* **78**, 659–670 (2006).
57. Rifkin, S. B. et al. Wood cookstove use is associated with gastric cancer in Central America and mediated by host genetics. *Sci. Rep.* **13**, 16515 (2023).
58. Byars, S. G. & Voskarides, K. Antagonistic pleiotropy in human disease. *J. Mol. Evol.* **88**, 12–25 (2020).
59. Carter, A. J. R. & Nguyen, A. Q. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med. Genet.* **12**, 160 (2011).
60. Corder, E. H. et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).
61. Raber, J., Huang, Y. & Ashford, J. W. ApoE genotype accounts for the vast majority of AD risk and AD pathology. *Neurobiol. Aging* **25**, 641–650 (2004).
62. Oriá, R. B. et al. ApoE polymorphisms and diarrheal outcomes in Brazilian shanty town children. *Braz. J. Med. Biol. Res.* **43**, 249–256 (2010).
63. Azevedo, O. G. R. et al. Apolipoprotein E plays a key role against cryptosporidial infection in transgenic undernourished mice. *PLoS One* **9**, e89562 (2014).
64. Yassine, H. N. & Finch, C. E. APOE alleles and diet in brain aging and Alzheimer's disease. *Front. Aging Neurosci.* **12**, 150 (2020).
65. Fullerton, S. M. et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).
66. van Exel, E. et al. Effect of APOE ε4 allele on survival and fertility in an adverse environment. *PLoS One* **12**, e0179497 (2017).
67. Powers, M. S., Smith, P. H., McKee, S. A. & Ehringer, M. A. From sexless to sexy: why it is time for human genetics to consider and report analyses of sex. *Biol. Sex Differ.* **8**, 15 (2017).
68. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).
69. Clayton, J. A. Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiol. Behav.* **187**, 2–5 (2018).
70. Ciesielski, T. H. et al. Late-onset neonatal sepsis: genetic differences by sex and involvement of the NOTCH pathway. *Pediatr. Res.* https://doi.org/10.1038/s41390-022-02114-8 (2022).
71. Ciesielski, T. H., Bartlett, J., Iyengar, S. K. & Williams, S. M. Hemizygosity can reveal variant pathogenicity on the X-chromosome. *Hum. Genet.* **142**, 11–19 (2023).
72. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
73. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
74. Kopanos, C. et al. VarSome: the human genomic variant search engine. *Bioinformatics* **35**, 1978–1980 (2019).
75. McKusick-Nathans Institute of Genetic Medicine. OMIM -Online Mendelian Inheritance in Man - An Online Catalog of Human Genes and Genetic Disorders. https://www.omim.org/.
76. Ferla, M. P., Pagnamenta, A. T., Koukouflis, L., Taylor, J. C. & Marsden, B. D. Venus: elucidating the impact of amino acid variants on protein function beyond structure destabilisation. *J. Mol. Biol.* **434**, 167567 (2022).
77. Michelanglo — VENUS Assessing the effect of amino acid variants have on structure [Internet]. [cited 2023 Aug 24]. Available from: https://michelanglo.sgc.ox.ac.uk/venus.
78. Manrai, A. K. et al. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).

79. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
80. Landry, L. G. & Rehm, H. L. Association of racial/ethnic categories with the ability of genetic tests to detect a cause of cardiomyopathy. *JAMA Cardiol.* **3**, 341–345 (2018).
81. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
82. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
83. Dong, S. et al. Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat. Genet.* **55**, 724–726 (2023).

ACKNOWLEDGEMENTS

This work was supported in part by the National Library of Medicine (NLM R01 LM010098) and the National Institutes of Health (NIH U01HG010219).

AUTHOR CONTRIBUTIONS

T.C. drafted the manuscript, and all authors (T.C., S.I., G.S., and S.W.) made substantial intellectual contributions and approved the submitted version.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-023-00386-5.

Correspondence and requests for materials should be addressed to Timothy H. Ciesielski.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.