# 1. Introduction

Since the introduction of Sanger sequencing, almost 50 years ago, the field of DNA sequencing has made enormous technical advances. Massive parallel DNA sequencing techniques such as Illumina can sequence a human genome within a few days, and more recent methods for long-read sequencing such as PacBio or Oxford Nanopore sequencing allow easier *de novo* assembly while increasing availability and portability of sequencing techniques (Shendure et al. 2017). The relative technical ease of sequencing humane genomes stands opposed to the continuous challenge of interpreting genetic variation. Analyses of the 1000 Genomes Project revealed that an individual has - depending on their ancestry - between 4.1 and 5 million Variants compared to the reference genome. Most of these Variants were Single Nucleotide Variants (SNVs) or short insertions and deletions (indels), they tend to be located in the non-coding genome, and up to 200 000 Variants per individual had allele frequencies below 0.5% (Auton et al. 2015). All of this illustrates that accurately distinguishing neutral (or even beneficial) variants from malignant variants that cause or increase risks for diseases is not a trivial task, which currently still limits the clinical utility of whole genome sequencing technologies (Lappalainen and MacArthur 2021; Shendure et al. 2017). Several methods to approach the challenge of predicting the effects of genetic variants have been developed. However, the use of such tools is often constrained by their limited scope on for example the coding genome (Brandes et al. 2023; Cheng et al. 2023), splice Sites (Jaganathan et al. 2019), or specific regulatory sites (Sample et al. 2019). An alternative approach to the problem of estimating the effect of genetic variants is the Combined Annotation-Dependent Depletion (CADD) Framework, introduced by Kircher et al. (2014) . Instead of trying to predict direct molecular effects of genetic variants, CADD uses the predicted deleteriousness, I.e. an evolutionary measure of reduced fitness, of a variant as an indicator for pathogenicity. The method is based on the assumption, that alleles that can be observed at high frequencies (>95%) in humans but don’t exist in the human-chimpanzee ancestral genome are neutral or benign while simulated variants contain a significant number of deleterious variants that in reality would be removed by purifying selection. At the core of the CADD framework is a linear regression model, that was trained to use a large set of more than 100 annotations to distinguish the observed (proxy-neutral) from the simulated (proxy-deleterious) variants (Kircher et al. 2014). Based on the annotations, the model computes a score (raw CADD score) that indicates how likely a variant belongs to the proxy-neutral or the proxy-deleterious class. After training, the model was applied to compute raw CADD scores for all theoretically possible SNVs in the human genome. To improve comparability of the results, the raw CADD score has then been scaled onto a Phred (-10 log10) scale from 0 to 99 where a score of 10 indicates a variant being amongst the top 10% with the highest predicted deleteriousness, while a score of 20 indicates the top 1% etc. (Kircher et al. 2014; Rentzsch et al. 2019). Since its introduction, CADD has frequently been updated to incorporate the updated reference genome (GRCh38) and to improve its predictive performance (Rentzsch et al. 2019, 2021; Schubach et al. 2024). The model has been used in various research projects and has been found to be especially useful in prioritizing variants of unknown significance (Bongaerts et al. 2022; Tollefson et al. 2023).

Single Nucleotide Variants can have different molecular consequences. Depending on their location and the kind of substitution they can for example affect regulatory sites and splices sites or they can lead to changes in protein coding regions where they might result in loss or gain of stop codons or changes in the amino acid sequence. Kircher et al. (2014) found, that variants with a lower Phred score tend to be located in the non-coding regions of the genome, while variants with a higher score are more often located in protein-coding genes. At a Phred above 40, virtually all variants were stop-gains (nonsense mutations). Since stop-gain mutations often lead to a loss of function of a gene by nonsense-mediated mRNA decay or by synthesis of truncated proteins, this prediction appears to be biologically plausible. However, the SNVs that CADD predicts to be highly deleterious have not been further investigated to date. Thus, it is unknown which of these variants have been observed before and if they are known to be associated with phenotypes or diseases. This project aimed to address the questions of which of the SNVs with a Phred score of at least 40 have previously been observed and what distinguishes the observed variants from those that were only simulated. For this, the Genome Aggregation Database (GnomAD) (Karczewski et al. 2020) has been used to identify observed SNVs with a PHRED above 40. Identified variants with the highest allele frequencies have been further investigated using additional databases as well as literature research. Furthermore, gene ontology and phenotype enrichment analyses have been performed to compare genes that are affected by observed stop gains with those that are affected by the simulated stop gains. Lastly, the relative positions of stop-gains on their respective genes have been compared. The analysis showed that approximately 10% of the SNVs with Phred above 40 were found in GnomAD. The observed variants with high allele frequencies seemed to be neutral and simulated but not observed variants were more strongly associated with pathogenic effects than those that were observed. The results highlight that while CADD’s use of deleteriousness as an indicator of pathogenicity is a good approach for prioritizing large sets of candidate variants, individual SNVs might be misclassified when using a defined threshold for binary classification of variants.

# 2. Methods

## 2.1 Identification of observed SNVs

To identify SNVs with a PHRED above 40 that have been observed in the GnomAD database, CADD v1.6 pre-scored Variants [available here: <https://cadd.bihealth.org/download>] have been filtered for a Phred score of greater or equal than 40. GnomAD release 4.0 Exome data have been retrieved from <https://gnomad.broadinstitute.org/downloads>. Variants that did not pass die GnomAD quality control filters were removed and only SNVs were retained.

For each variant, a unique identifier consisting of chromosome, position, reference allele, and alternative allele was generated. This identifier was then used to identify observed variables by computing the overlap between the two datasets so that all CADD variants that were also found in the GnomAD dataset were retained.

Filtering of the CADD and GnomAD datasets and generation of the overlap has been performed using the BASH command line interpreter. Subsequent data manipulation and analysis were performed using R version 4.2.2 (2022) and the packages data.table version 1.14.8 (Dowle and Srinivasan 2023) as well as dplyr version 1.1.4, tidyr version 1.3.0, stringr version 1.5.1 and ggplot2 version 3.4.4 provided by the tidyverse metapackage (Wickham et al. 2019).

Variants can be annotated using different gene models leading to duplicated variants that are differentially annotated in the CADD dataset. However, the CADD output contains only the highest score for each variant without indicating which annotation the score is based on. To account for this, a three-step filtering strategy was used to ensure that only unique variants were retained. Firstly, for each variant, the annotation with the highest ConsScore was prioritized. The ConScore is a score from 0 to 8 which is based on the assumption that some types of consequences (such as stop-gain, non-synonymous base exchange, variants in splice sites) tend to be more deleterious than others (such as synonymous or intergenic exchanges) [assignment of ConScore: [https://github.com/kircherlab/CADD-scripts/blob/master/src/scripts/lib/Annotations.py#L175](https://github.com/kircherlab/CADD-scripts/blob/master/src/scripts/lib/Annotations.py" \l "L175)]. The score is returned as one of the CADD annotations, it is however not used in the model itself. Secondly, variants that occur earlier in a gene (have a smaller relative cDNA position) have been prioritized over those that appear later. Lastly, duplicates that could not be resolved by the aforementioned methods were resolved by random selection.

## 2.2 Analysis of Variants with the highest allele frequencies

From Variants that were observed in GnomAD, the five variants with the highest allele frequencies were selected for further literature research. Relevant literature that might explain the consequences of the variants was identified using the LitVar2 text mining tool (Allot et al. 2023) Furthermore, the ClinVar Database (Landrum et al. 2020) was used to search for the clinical significance of the variance, and OMIM (Hamosh et al. 2005) was used to investigate phenotypes/diseases that are associated with genes that are affected by the variants.

## 2.3 Enrichment Analysis

Gene ontology and phenotype enrichment analysis have been performed with the enrichr gene set analysis tool (Xie et al. 2021) using the enrichR package (Jawaid 2023). For the gene ontology (GO) enrichment analysis, two gene lists have been created based on the variants with a PHRED of at least 40. One list contained genes that were observed more than once in GnomAD while the other one contained genes derived from simulated variants not found in GnomAD. For the phenotype enrichment analysis, a third gene list containing genes where singleton variants were found on. The used databases were GO\_Biological\_Process\_2023 and MGI\_Mammalian\_Phenotype\_Level\_4\_202 for gene ontology and phenotype enrichment analysis respectively. Terms with a Benjamini-Hochberg corrected P-value of 0.05 or lower were considered significantly enriched. Redundant GO terms were reduced based on semantic similarities using the rrvgo package (Sayols 2020) to first compute a similarity matrix of the GO terms and then to reduce the terms in the matrix using -log10 transformed adjusted p-value as weights and a threshold of 0.7.

## 2.4 Analysis of relative cDNA positions

To identify where on the genes deleterious stop gains are most often located, the distribution of the relative cDNA position of observed and simulated stop-gains with Phred scores below and above 40 was plotted using the geom\_density function provided by ggplot2 using default parameters.

# 3. Results

A total of 1 223 918 variants on the autosomes had a PHRED score of at least 40. Of those variants, 127871 (10.4%) were found in the GnomAD database.

Almost all variants (127400) were stop-gains, while only 445 non-synonymous, 18, intronic, 5 splice site, 2 regulatory, and 1 one Variant in the 3’ UTR were identified. The majority of the observed SNVs (59.8%) are singletons (I.e. they have only been observed once). Only 38 variants had an allele frequency (AF) of at least 1% (i.e. are not rare variants). Interestingly, three variants with an AF greater than 50% were found in the observed data. Table 1 shows an overview of the identified SNVs with the highest allele frequencies.

| **Chrom** | **Pos** | **Ref** | **Alt** | **PHRED** | **Gene** | **AF** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 171208951 | C | T | 40 | *FMO2* | 0.995225 |
| 5 | 75669297 | G | A | 49 | *ANKDD1B* | 0.635824 |
| 1 | 54716627 | T | A | 45 | *MROH7-TTC4* | 0.544624 |
| 7 | 21543345 | G | T | 42 | *DNAH11* | 0.484693 |
| 19 | 48703417 | G | A | 46 | *FUT2* | 0.459996 |

## 3.1 Literature research of variants with high allele frequencies

### 3.1.1 1-171208951-C>T (rs6661174)

The observed SNV with the highest allele frequency was rs6661174. It leads to a premature stop codon on exon 9 of *FMO2*, a gene coding for Flavin Containing Dimethylaniline Monooxygenase 2. The alternative allele (T-allele) has an overall AF of 1. It has allele frequencies of or close to 0 (100%) throughout all GnomAD ancestry groups except African / African American where the frequency is only 0.88. The T allele leads to *FMO2* loss of function while the ancestral C allele leads to a functional allele. A study by Veeramah et al. (2008) found high frequencies of the C allele in the population of sub-Saharan Africa where approximately one-third was at least heterozygote, but could not find evidence for selective pressure on either allele. In contrast, Mekonnen and Bekele (2017) propose that in specific African populations, the persistence of the functional allele can be explained by its association with reduced Tuberculosis disease activity.

**For Discussion: C allele does not segregate together with known TB risk alleles**

* Problem - do the risk alleles have a molecular effect?
* If so –> might they be actually responsible for the observed reduced TB?
  + Relevant section marked in paper

### 3.1.2 5-75669297-G>A (rs34358)

This variant is a stop-gain in *ANKDD1B* , coding for Ankyrin Repeat And Death Domain Containing 1B. The alternative A allele is the major allele in all GnomAD ancestry groups except South Asian and East Asian where it has AFs of 0.44 and 0.35 respectively. The variant has been described as a risk allele for type 2 diabetes and dyslipidemia (Vujkovic et al. 2020) and as a protective allele for migraine without aura (Zhang et al. 2022), however no studies investigating the evolutionary history of the allele and no evidence for a deleterious effect (like a pathogenic effect with early onset and high penetrance) could be found.

### 3.1.3 1-54716627-T>A (rs1147990)

This variant has an overall frequency of 0.55. It is most frequent in the East Asian (0.98) and South Asian (0.72) ancestry groups. Although being annotated as stop gain in *MROH7-TTC4* by CADD v1.6, the variant is located in a read-trough transcript between *MROH7* and *TTC4,* two neighboring genes. Multiple variants of the transcript can arise from alternative splicing, however, they are likely degraded by nonsense-mediated decay and do not encode any protein. In CADD v1.7 which uses updated Ensemble annotations, the transcript is no longer annotated as a coding gene. With that change, the variant’s PHRED score was markedly reduced to 19.97.

### 3.1.4 7-21543345-G>T (rs2285943)

Rs2285943 is predicted to result in a stop gain on exon 1/82 of *DNAH11.* The variant is common (AFs 0.29 - 0.55) throughout all GnomAD ancestry groups. It is known that DNAH11 loss of function leads to Primary Ciliary Dyskineasia (Bartoloni et al. 2002; Lai et al. 2016; Schwabe et al. 2008), however, multiple ClinVar submissions indicated that rs2285943 is not linked to that condition.

\*\*For discussion\*\*

* Genome Browswer shows alternative promotor just downstream of variant
* alt promotor: <https://genome.ucsc.edu/cgi-bin/hgc?hgsid=1885897182_19ARA0SsykGacZ7HT0dy80zu9yoK&db=hg38&c=chr7&l=21543005&r=21547379&o=21543345&t=21543495&g=knownAlt&i=altPromoter>
* Range: [chr7:21543346-21543495](https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=1885897182_19ARA0SsykGacZ7HT0dy80zu9yoK&db=hg38&position=chr7%3A21543346-21543495)

### 3.1.5 19-48703417-G>A (rs601338)

The fifth of the selected variants leads to a loss of function in *FUT2.* With allele frequencies between 0.30 in Admixed Americans and 0.52 in Middle Easterns, it is common in most ancestry groups. However, the variant is rare in East Asians (AF 0.002). The variant affects the function of the alpha(1,2)fucosyltransferase which is responsible for secretion of ABO antigens. Carriers of a functional *FUT2* allele secrete these antigens in their bodily fluids while homozygous or compound heterozygous carriers of non-functional FUT2 (for example caused by rs601338) do not secret the antigens. The secretor phenotype (secretor polymorphism) seems to be associated as both, a risk or a protective factor, with numerous conditions and diseases (Krog et al. 2023; Mottram et al. 2017).

## 3.2 Enrichment analysis

Gene ontology enrichment analysis revealed 103 significantly enriched biological processes for the genes that contained observed SNVs, resulting in 18 clusters with semantically distinct functions. Analysis of the gene list derived simulated but not observed SNVs yielded 115 enriched terms in 26 distinct clusters. Figure 1 shows the enriched clusters for both groups.

|  |
| --- |
| **Figure 1:** Treemaps of the biological processes that are enriched in gene lists derived from observed (left) and simulated (right) variants. The treemaps reflect the hierarchical nature of gene ontology terms by grouping terms that were found to be semantically similar under a parent term. The size of the rectangles indicates the negative log10 adjusted p-value. |

Genotype enrichment analysis showed that genes with observed SNVs with a Phred over 40 tend to be involved in processes that affect cell and tissue morphology (extracellular structure organization and intermediate filament organisation, cell-matrix adhesion), cation transport, and metabolic processes (fatty acid metabolic process, proteolysis, glycerophospholipid biosynthetic processes). The genelist from the simulated SNVs on the other hand was mainly enriched for processes that regulate gene expression, cell proliferation, and differentiation or are involved in cellular signalling.

The phenotype enrichment analysis revealed 347, 513, and 83 significantly enriched terms for the gene lists derived from simulated, singleton and non-singleton SNVs respectively. The 10 most significantly enriched terms (lowest adjusted P-value) are shown in Figure 2.

|  |
| --- |
| **Figure 2:** The ten most significantly enriched phenotypes for genes on which simulated, singleton, and non-singleton SNVs with Phred scores above 40 are found. The colour of the bars represents the statistical significance as negative log10 adjusted P-value and the bars show the fraction of genes found in the respective gene lists compared to all genes that map to the phenotype. |

The phenotype enrichment analysis showed, that in the simulated and singleton groups most of the most significantly enriched phenotypes lead to early (embryonic, pre- and postnatal or preweaning) lethality while the phenotypes in the non-singleton group appear less likely to lead to this early lethality. Furthermore, genes from simulated variants tend to have higher overlaps than those from singleton or non-singleton SNVs.

## 3.3 Relative positions of simulated and observed SNVs

To see where deleterious SNVs are located on the individual genes, the relative cDNA positions of observed and simulated stop-gains have been plotted.

|  |
| --- |
| **Figure 3:** **(a)** Distribution of the stop-gains in different Phred groups. **(b)** Relative cDNA positions of stop-gains with Phred scores above 40 that have been observed (red) or simulated (blue). |

Figure 3a shows that stop gains with a lower Phred score tend to have a lower relative cDNA position, meaning they are located closer to the start of a gene. In figure 3b it can be seen that the distribution of the observed variants roughly follows the distribution of the simulated ones. However, while the simulated variants appear to be normally distributed around their mean of 0.47 while the observed variants have a slightly lower mean (0.45) and a distribution with a plateau around 0.25.

# 4. References

Allot, Alexis, Chih-Hsuan Wei, Lon Phan, Timothy Hefferon, Melissa Landrum, Heidi L. Rehm, and Zhiyong Lu. 2023. “Tracking Genetic Variants in the Biomedical Literature Using LitVar 2.0.” *Nature Genetics* 55 (6): 901–3. <https://doi.org/10.1038/s41588-023-01414-x>.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. <https://doi.org/10.1038/nature15393>.

Bartoloni, Lucia, Jean-Louis Blouin, Yanzhen Pan, Corinne Gehrig, Amit K. Maiti, Nathalie Scamuffa, Colette Rossier, et al. 2002. “Mutations in the DNAH11 (Axonemal Heavy Chain Dynein Type 11) Gene Cause One Form of Situs Inversus Totalis and Most Likely Primary Ciliary Dyskinesia.” *Proceedings of the National Academy of Sciences* 99 (16): 10282–86. <https://doi.org/10.1073/pnas.152337699>.

Bongaerts, Michiel, Ramon Bonte, Serwet Demirdas, Hidde H. Huidekoper, Janneke Langendonk, Martina Wilke, Walter de Valk, Henk J. Blom, Marcel J. T. Reinders, and George J. G. Ruijter. 2022. “Integration of Metabolomics with Genomics: Metabolic Gene Prioritization Using Metabolomics Data and Genomic Variant (CADD) Scores.” *Molecular Genetics and Metabolism* 136 (3): 199–218. <https://doi.org/10.1016/j.ymgme.2022.05.002>.

Brandes, Nadav, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. 2023. “Genome-Wide Prediction of Disease Variant Effects with a Deep Protein Language Model.” *Nature Genetics* 55 (9): 1512–22. <https://doi.org/10.1038/s41588-023-01465-0>.

Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, et al. 2023. “Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense.” *Science* 381 (6664): eadg7492. <https://doi.org/10.1126/science.adg7492>.

Dowle, Matt, and Arun Srinivasan. 2023. “Data.table: Extension of ‘Data.frame‘.”

Hamosh, Ada, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. 2005. “Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders.” *Nucleic Acids Research* 33 (Database Issue): D514–17. <https://doi.org/10.1093/nar/gki033>.

Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, et al. 2019. “Predicting Splicing from Primary Sequence with Deep Learning.” *Cell* 176 (3): 535–548.e24. <https://doi.org/10.1016/j.cell.2018.12.015>.

Jawaid, Wajid. 2023. “enrichR: Provides an r Interface to ’Enrichr’.”

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43. <https://doi.org/10.1038/s41586-020-2308-7>.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O’Roak, Gregory M Cooper, and Jay Shendure. 2014. “A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants.” *Nature Genetics* 46 (3): 310–15. <https://doi.org/10.1038/ng.2892>.

Krog, Grethe Risum, Henriette Lorenzen, Frederik Banch Clausen, and Morten Hanefeld Dziegiel. 2023. “Secretor Status of Blood Group O Mothers Is Associated with Development of ABO Haemolytic Disease in the Newborn.” *Vox Sanguinis* 118 (5): 402–6. <https://doi.org/10.1111/vox.13420>.

Lai, Michele, Massimo Pifferi, Andrew Bush, Martina Piras, Angela Michelucci, Maria Di Cicco, Ambra del Grosso, et al. 2016. “Gene Editing of DNAH11 Restores Normal Cilia Motility in Primary Ciliary Dyskinesia.” *Journal of Medical Genetics* 53 (4): 242–49. <https://doi.org/10.1136/jmedgenet-2015-103539>.

Landrum, Melissa J., Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, et al. 2020. “ClinVar: improvements to accessing data.” *Nucleic Acids Research* 48 (D1): D835–44. <https://doi.org/10.1093/nar/gkz972>.

Lappalainen, Tuuli, and Daniel G. MacArthur. 2021. “From Variant to Function in Human Disease Genetics.” *Science* 373 (6562): 1464–68. <https://doi.org/10.1126/science.abi8207>.

Mekonnen, Ephrem, and Endashaw Bekele. 2017. “An Ancestral Human Genetic Variant Linked to an Ancient Disease: A Novel Association of FMO2 Polymorphisms with Tuberculosis (TB) in Ethiopian Populations Provides New Insight into the Differential Ethno-Geographic Distribution of FMO2\*1.” *PLOS ONE* 12 (10): e0184931. <https://doi.org/10.1371/journal.pone.0184931>.

Mottram, Lynda, Gudrun Wiklund, Göran Larson, Firdausi Qadri, and Ann-Mari Svennerholm. 2017. “FUT2 Non-Secretor Status Is Associated with Altered Susceptibility to Symptomatic Enterotoxigenic Escherichia Coli Infection in Bangladeshis.” *Scientific Reports* 7 (September): 10649. <https://doi.org/10.1038/s41598-017-10854-5>.

R Core Team. 2022. “R: A Language and Environment for Statistical Computing.” <https://www.R-project.org/>.

Rentzsch, Philipp, Max Schubach, Jay Shendure, and Martin Kircher. 2021. “CADD-Spliceimproving Genome-Wide Variant Effect Prediction Using Deep Learning-Derived Splice Scores.” *Genome Medicine* 13 (1): 31. <https://doi.org/10.1186/s13073-021-00835-9>.

Rentzsch, Philipp, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. 2019. “CADD: Predicting the Deleteriousness of Variants Throughout the Human Genome.” *Nucleic Acids Research* 47 (D1): D886–94. <https://doi.org/10.1093/nar/gky1016>.

Sample, Paul J., Ban Wang, David W. Reid, Vlad Presnyak, Iain J. McFadyen, David R. Morris, and Georg Seelig. 2019. “Human 5’ UTR Design and Variant Effect Prediction from a Massively Parallel Translation Assay.” *Nature Biotechnology* 37 (7): 803–9. <https://doi.org/10.1038/s41587-019-0164-5>.

Sayols, Sergi. 2020. “Rrvgo: A Bioconductor Package to Reduce and Visualize Gene Ontology Terms.” <https://ssayols.github.io/rrvgo>.

Schubach, Max, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher. 2024. “CADD V1.7: Using Protein Language Models, Regulatory CNNs and Other Nucleotide-Level Scores to Improve Genome-Wide Variant Predictions.” *Nucleic Acids Research* 52 (D1): D1143–54. <https://doi.org/10.1093/nar/gkad989>.

Schwabe, Georg C., Katrin Hoffmann, Niki Tomas Loges, Daniel Birker, Colette Rossier, Margherita M. de Santi, Heike Olbrich, et al. 2008. “Primary Ciliary Dyskinesia Associated with Normal Axoneme Ultrastructure Is Caused by DNAH11 Mutations.” *Human Mutation* 29 (2): 289–98. <https://doi.org/10.1002/humu.20656>.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. “DNA Sequencing at 40: Past, Present and Future.” *Nature* 550 (7676): 345–53. <https://doi.org/10.1038/nature24286>.

Tollefson, Mallory R., Rose A. Gogal, A. Monique Weaver, Amanda M. Schaefer, Robert J. Marini, Hela Azaiez, Diana L. Kolbe, et al. 2023. “Assessing Variants of Uncertain Significance Implicated in Hearing Loss Using a Comprehensive Deafness Proteome.” *Human Genetics* 142 (6): 819–34. <https://doi.org/10.1007/s00439-023-02559-9>.

Veeramah, Krishna R., Mark G. Thomas, Michael E. Weale, David Zeitlyn, Ayele Tarekegn, Endashaw Bekele, Nancy R. Mendell, Elizabeth A. Shephard, Neil Bradman, and Ian R. Phillips. 2008. “The Potentially Deleterious Functional Variant Flavin-Containing Monooxygenase 2\*1 Is at High Frequency Throughout Sub-Saharan Africa.” *Pharmacogenetics and Genomics* 18 (10): 877–86. <https://doi.org/10.1097/FPC.0b013e3283097311>.

Vujkovic, Marijana, Jacob M. Keaton, Julie A. Lynch, Donald R. Miller, Jin Zhou, Catherine Tcheandjieu, Jennifer E. Huffman, et al. 2020. “Discovery of 318 New Risk Loci for Type 2 Diabetes and Related Vascular Outcomes Among 1.4 Million Participants in a Multi-Ancestry Meta-Analysis.” *Nature Genetics* 52 (7): 680–91. <https://doi.org/10.1038/s41588-020-0637-y>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the Tidyverse” 4: 1686. <https://doi.org/10.21105/joss.01686>.

Xie, Zhuorui, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, et al. 2021. “Gene Set Knowledge Discovery with Enrichr.” *Current Protocols* 1 (3): e90. <https://doi.org/10.1002/cpz1.90>.

Zhang, Tianxiao, Hang Wei, Miao Li, Wei Han, Wenjuan Zhang, Xiaojie Zhang, Bo Zhang, Zhao Jiang, and Tao Li. 2022. “Risk of Migraine Contributed by Genetic Polymorphisms of ANKDD1B Gene: A Casecontrol Study Based on Chinese Han Population.” *Neurological Sciences* 43 (4): 2735–43. <https://doi.org/10.1007/s10072-021-05645-w>.