**Genetic Insights from a Genetically Isolated Community**

Alyssa Wheeler

Juniata College

Huntingdon, Pennsylvania

**Abstract**

The Old Order Amish and Mennonite communities of Pennsylvania, known as the Plain population, represent a unique and valuable resource for studying rare genetic variants due to their genetic isolation. This study generated and analyzed a genetic variant database derived from the exome sequences of 210 Kish Valley Amish and Mennonite individuals, with a focus on identifying potentially harmful mutations, both genome-wide and within specific genes of clinical interest. Exome data were initially stored as individual VCF files, which were merged and annotated using wANNOVAR. Subsequent analysis was conducted using PLINK and RStudio, incorporating the tidyverse and infer libraries. The dataset was filtered and explored for variant pathogenicity predictions and for mutations in genes of interest to Dr. Morton, a physician at the Clinic for Special Children, where the data were collected. Tables and visualizations were generated to summarize the frequency and distribution of harmful variants. Chromosome 1 was found to harbor the greatest number of predicted harmful mutations across both the full gene set and the targeted genes. While many variants were classified as tolerated by multiple predictor tools, a substantial subset were predicted to be disease-causing. The resulting database offers a valuable resource for clinicians and researchers, enabling improved understanding of gene function, disease risk, and variant frequency in a genetically distinct population. Moreover, these findings may have broader implications, as shared variants between the Plain community and the general population could inform preventative care strategies and contribute to cost-effective genetic screening beyond this population.

**Introduction**

The Old Order Amish and Mennonite populations of Pennsylvania, also known as the Plain populations, are a community that, through their interesting history, offer a unique avenue for gene variant discovery (Strauss et al., 2012). The Plain people are descended from a small group of Swiss immigrants who have remained genetically isolated for the past 12-14 generations (Puffenberger et al., 2012; Strauss et al., 2012). This isolation has resulted in higher rates of specific autosomal recessive disorders in their population (Crowgey et al., 2019; Puffenberger et al., 2012). They have also experienced many genetic bottlenecks resulting in a pronounced founder effect (Payne et al., 2011a). This effect means there is a reduced amount of genetic variation compared to the rest of the population because their genes are derived from the same founding communities (Crowgey et al., 2019; Payne et al., 2011a). This results in many specific genetic disorders being more prevalent in this population, making them very suitable for genetic study (Payne et al., 2011a).

The Plain people’s settlements are socially constructed and not exclusively composed of one type of religion. One settlement could have multiple Anabaptist affiliations sharing the space and resources (Anderson & Bacon, 2024). With trusted medical practitioners, the Plain people tend to be cooperative towards new studies, and many of the studies conducted have allowed doctors, geneticists, and other health professionals to collaborate in better understanding the issues at hand (Nolt, 2020). Moreover, many groups in the Plain community have extensive genealogical records maintained by local ministers that can help trace the population back to its founding ancestors (Payne et al., 2011a). The Plain people also tend to have large families with low rates of nonpaternity, and the parents of Plain children tend to be more altruistic (Payne et al., 2011a; Nolt, 2020). In particular, the parents tend to understand that although new studies may not help their children directly, they could help future children (Nolt, 2020). Culturally, the Plain people tend not to assimilate into the rest of the population, resulting in low genetic inflow (Payne et al., 2011a; Miller et al., 2019). In part to this, the diseases they contract tend to be place-specific because they tend to marry within their area or somewhere very close to their area (Anderson & Potts, 2022). This causes the same geographic area to have a specific pattern of genetic disorders that differ from other communities around it (Ehrenberg et al., 2021).

Because of the Amish and Mennonite’s closed gene pool, they have a shared culture that shapes how they live their lives (Anderson & Potts, 2022). They concentrate in existing communities and migrate to new places, causing an outside impact on the local culture and infrastructure (Anderson & Potts, 2022). There have been many health studies done on the Plain people to study how their lifestyle may affect their risk of certain diseases (Anderson & Potts, 2020; Anderson & Potts, 2021). They base many of their healthcare decisions on their religious or cultural beliefs and information they have gathered from their community or trusted healthcare practitioners (Anderson & Potts, 2020). Compared to the rest of the Pennsylvania population, the Plain people are more likely to drink well water, eat fruits and vegetables, and drink raw milk. Conversely, they are also more likely to be exposed to agricultural chemicals than the general Pennsylvania population. Overall, despite being genetically isolated and receiving fewer screening exams than the general public, they are in good physical and mental health compared to the rest of the population (Miller et al., 2019). The Plain people will use modern medicine, but they prefer to use alternative medicine like supplements, prayer, massage therapy, and other non-traditional methods. They avoid modern medicine when they can because they have a distrust of the medical industry and feel culturally disoriented in modern medical environments (Anderson & Potts, 2020).

One of the main contributors to studying the Old Order Amish and Mennonite population is the Clinic for Special Children. Founded in 1989 for uninsured Plain children, the clinic strives to provide accessible and affordable healthcare to this population. The clinic lowered health costs for the community by bringing together biochemical and genetic knowledge. Initially, they had only an analytical laboratory for research, but in 1998, the clinic expanded its laboratory by bringing in a geneticist. They focused on known disease and gene combinations and then subsequently added on genetic mapping studies and whole-exome sequencing (Crowgey et al., 2019). Whole exome sequencing is a second-generation sequencing technique that can identify variants in the coding regions of the human genome (Daud et al., 2015). It can be costly but is highly practical as it can integrate the 1.5% of the genome that contains 95% of pathogenic variants and can identify up to 20,000 variants (Puffenberger et al., 2012; Daud et al., 2015). As Next-Generation Sequencing became available, the clinic used it to advance their knowledge of genetic variations. Next-Generation Sequencing allows the ability to simultaneously sequence millions of DNA fragments, compared to Sanger sequencing, which can only sequence one DNA fragment at a time. Along with improvements in sequencing, they developed screening assays that allowed the opportunity to give presymptomatic treatment. This screening assay is a custom gene panel consisting of 202 known pathogenic variants with a capability of determining the carrier status for 168 conditions (Crowgey et al., 2019). For perspective, about 25% of the variability associated with death in this population is attributed to genetic factors (Anderson & Potts, 2022). Therefore, the potential impact of these assays on patient care is high because couples that are at-risk can be informed of their carrier status before their affected children are born (Crowgey et al., 2019).  Early diagnosis and treatment of diseases can prevent severe disability and cut down on future medical costs for patients (Ehrenberg et al., 2021).

The Clinic for Special Children has managed to save local communities countless expenses on medical costs. Integrating molecular technologies into primary care can not only improve diagnosis, but also reduce costs for laboratories and hospitals. Having population-specific genetic information can strongly help with preventative healthcare (Strauss et al., 2012). Moreover, discovering rare and highly phenotypically showing alleles in a small group can prove more useful than large genome-wide association studies (Puffenberger et al., 2012). Scaling studies to small populations is a better alternative than large studies as it may help solve more difficult human disease problems (Strauss et al., 2012). This can reveal basic genetic foundations for complex diseases because the alleles can be seen against the population-specific genetic variation (Puffenberger et al., 2012). In the Plain community, about 10-20% of each generation of children have decided to leave their communities, contributing to the spread of rare alleles beyond the Plain community (Payne et al., 2011a). Additionally, the Plain people overall are beginning to decline in their geographical isolation and increasingly connect more to society (Conlin, 2021). These people are similar to the general public. For instance, many of the mutations causing disease can be found in Europe where the founding populations originated (Payne et al., 2011a). Furthermore, the Plain people population is growing exponentially, relocating in rural North America (Anderson & Potts, 2021). They are expected to double in size within the next twenty years and double again by 2040 (Conlin, 2021). As the population increases, the Amish increasingly provide an opportunity to better understand disease by comparing them genetically with non-Amish people (Anderson & Potts, 2020).

              One way to find genetic comparisons in a population is through a genetic database. Creating a database of searchable genetic mutations and their impact can be useful to healthcare providers performing genetic testing for patients. Databases can help with finding a better understanding of gene function and can estimate the prevalence of genes in a population. They are helpful for searching for associations between genetics and health to improve medical intervention. Many pharmaceutical companies use databases for gene-related variabilities in medication responsiveness and metabolism. They can tailor medication to a particular genetic makeup and screen a patient for genetic suitability to a medication. These databases have also helped in toxicological investigations in being able to sort out the cause of an adverse drug reaction, characterize genetic pathology for some cancers, and reveal gene expression in different stages of life (Lowrance, 2001).

Furthermore, Plain population-specific databases can be useful for identifying harmful variants that are not only found in the Plain population but also in the general population. Because each database has its own genomic focus and goal, the data on the Plain population of Pennsylvania specifically can be found in a variety of databases. One database is the Anabaptist Database. This database was created as a way to search through this community’s genealogy, and it allows for an easier mapping of mutated genetic disorders. The database is searchable and keeps patient information within it confidential. Although it started as only an Amish database, it has expanded to include Anabaptists that are not Amish and people living outside Pennsylvania (Agarwala et al., 2003). Another database that is useful is the Amish, Mennonite, and Hutterite Genetic Disorder Database. This database was intended for medical doctors to be better able to treat this population of people. It focuses on single-gene disorders and rare genetic mutations specific to the Plain people (Payne et al., 2011b). Lastly, the Amish Complex Genetic Disease Database is a good source for studying the Plain people of Lancaster County, Pennsylvania. This database is researched and led by the University of Maryland School of Medicine. Data has been collected for this database starting in 1993 and includes over 7,000 adults with a wide variety of genetic mutations (University of Maryland Baltimore, 1995). Overall, creating a different database can add to the knowledge in other databases and possibly reveal new insights into the Plain people.

Once a database is created, there are many options for analyzing the information obtained from it. Data analysis has a common pattern regardless of the type of analysis being done. After the data is collected, it is quality checked and cleaned (Akalin, 2020). This means that missing values are handled by either removing them or replacing them with a similar or average value (Akalin, 2020; Wickham et al., 2023). Any data quality issues are addressed and cleaned from the data. The data is then processed into a format suitable for exploratory analysis and modeling (Akalin, 2020). Exploratory data analysis includes applying machine learning or other statistical methods to look at the relationship between the variables in the data. The variables of interest are modeled with statistics such as linear regression or hypothesis testing (Akalin, 2020; Wickham et al., 2023). Once modeling is done, the data is visualized through different figures and tables to help describe the outcome of the analysis. After all of this is done, the findings in the data can be reported on (Akalin, 2020). In this case, the genetic database data can be visualized in R to gain a better insight into the discovered gene variants.

              This study created a database from the exomes of 210 Kish Valley Amish and Mennonite patients. For this project, the Plain population data used was provided by Dr. Morton, a medical doctor practicing in the Clinic for Special Children. This data allowed for the creation of a searchable database that can be used by healthcare workers or patients to better understand Kish Valley genetic mutations. Because the Plain populations have remained genetically isolated for generations, they have developed a higher concentration of certain inherited disorders. This makes them a valuable group for studying rare genetic variants and creating resources tailored to their unique genetic makeup. The study focused on identifying harmful mutations found in the database after doing exploratory data analysis. These mutations could be used for further studies or preventative care. Once the data was explored more, new insights emerged that could also be helpful for comparisons between similar mutations found in the general population.

**Methods**

***Database Creation***

Exomes of 210 Kish Valley Amish and Mennonite patients had been sequenced and stored as VCF files. All VCF file analysis was done on the Juniata College cluster accessed through PuTTY (McMullen, n.d.).

The individual VCF files, each representing one of the 210 patients, were merged into a single file called “merged-vars.vcf.gz.” This merged file already included a header; however, an additional ID column was added, containing the chromosome number, position, reference allele, and first alternative allele for each variant (McMullen, n.d.).

Next, the merged file with the new ID column was annotated using wANNOVAR to provide additional information on each variant, such as pathogenicity and allele frequency statistics. To obtain more detailed allele frequency statistics not included in the annotation, the annotated VCF file was further analyzed using PLINK, run through PuTTY. PLINK generated a detailed genotype count report and calculated the minor allele frequency for each variant. Included in this is the number of individuals homozygous for the variant and number of individuals heterozygous for the variant (McMullen, n.d.; “Whole Genome Association Analysis Toolset”, 2025).

After running PLINK, the two files created (merged-ID.frq and merged-ID.frqx) were downloaded to the local computer. The results from these files were added to the annotated Excel file, “Table1\_database”, in a new tab labeled “PLINK” (McMullen, n.d.).

To improve readability, the columns in the annotated Excel sheet were color-coded and the sheet was formatted to consolidate all results into one sheet. A column containing direct links to the Broad Institute’s gnomAD browser was added for easy access to detailed information about each variant. Finally, relevant PLINK results were transferred from the “PLINK” tab into the main annotated sheet (McMullen, n.d.).

***Exploratory Data Analysis***

Data analysis was continued in RStudio. The tidyverse and infer libraries were loaded into the R Markdown (RMD) file. The infer library enables statistical inference within the tidyverse framework, while tidyverse supports data manipulation and visualization using ggplot. The database, created from the exomes of the 210 Kish Valley Amish and Mennonite patients, was imported into RStudio as a CSV file (Wickham et al., 2023).

To begin, the data was cleaned and manipulated for easier analysis. This process included renaming variables to remove spaces or parenthesis, ensuring each variable had the correct data type, and verifying that variable values were accurate and consistent. For the variables ExonicFunc, SIFT\_pred, LRT\_pred, MutationTaster\_pred, MutationAssessor\_pred, FATHMM\_pred, all entries containing "." or missing values were converted to NA. Similarly, for the quantitative variables AF, AF\_nfe, REVEL\_score, and GERP++\_NR, "." values were changed to NA. After this, all NA values were removed from the dataframe.

Following data cleaning, bar charts were generated to better understand individual variables. These plots included distributions for Chromosome, Function, Exonic Function, and each of the predictor variables (SIFT, LRT, Mutation Taster, Mutation Assessor, FATHMM). Covariation among variables was also explored through bar charts comparing Gene Function vs Exonic Function, Chromosome Number vs Function, Chromosome Number vs Exonic Function, and Chromosome Number vs each of the predictor variables.

Next, the analysis focused on genes of interest identified by Dr. Morton, a physician at the Clinic for Special Children. The dataset was filtered to include only these genes: APOB, PCCB, PAH, ZDHHC16, RYR2, BCKDHA, GCDH, PKLR, ACADM, SLC12A3, FLVCR1, MTHFR. Columns unrelated to the analysis (such as gnomad, GeneDetail, and rsNum) were excluded. This filtered dataset was saved as GOI\_data. Using this dataset, a bar chart was made for the Genes of Interest vs Chromosome Number and the Genes of Interest vs each predictor variable. Then, bar charts were made for the Genes of Interest vs Chromosome Number filled with color for each predictor variable. Subsequently, the GOI\_data was filtered for genes classified as harmful by each predictor tool, and the result was stored as a new dataframe called “GOI\_Harmful.” The count of the Gene variable was shown for GOI\_Harmful to see out of the genes of interest which genes were marked harmful by every predictor tool.

In addition, the original dataset was filtered for all harmful genes variants and stored as a new dataframe called “Harmful\_data.” The count of all the Gene variable was shown for this dataframe. The data was then further filtered for homozygous for allele 1, with counts of Gene, Chr, MAF\_percent, REVEL\_score, CADD\_raw, and GERP++\_NR variables displayed in a table and sorted in descending order by MAF\_percent. Similar filtering was performed for homozygous for allele 2, with gene counts displayed, and for heterozygous variants, where counts for Gene were again shown in descending order by MAF\_percent.

**Results**

*Table 1. (Attached separately) This is a searchable database of gene variants identified in the Kish Valley Amish and Mennonite community. It includes variant location, frequency, and the predicted impact of each variant using various bioinformatics tools. This resource can support future research and clinical assessments by identifying potentially disease-causing mutations.*

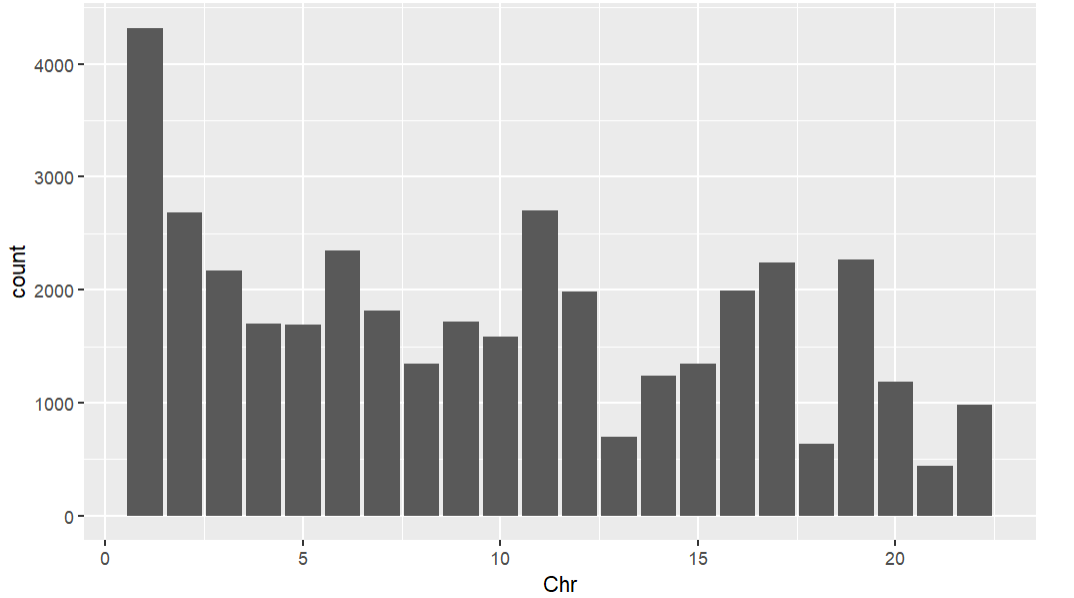


Figure 1. The bar plot illustrates the distribution of variant counts across chromosomes 1 to 23. Chromosome 1 exhibits the highest variant count, over 4000, greater than all other chromosomes. Chromosomes 2, 11, 16, 17, and 19 show variant counts ranging between 2000 and 2700. In contrast, chromosomes 13, 18, and 22 display the lowest variant counts, with chromosome 22 having the smallest value under 1000. The remaining chromosomes have variant counts more in the middle ranging from 1300 to 2300.

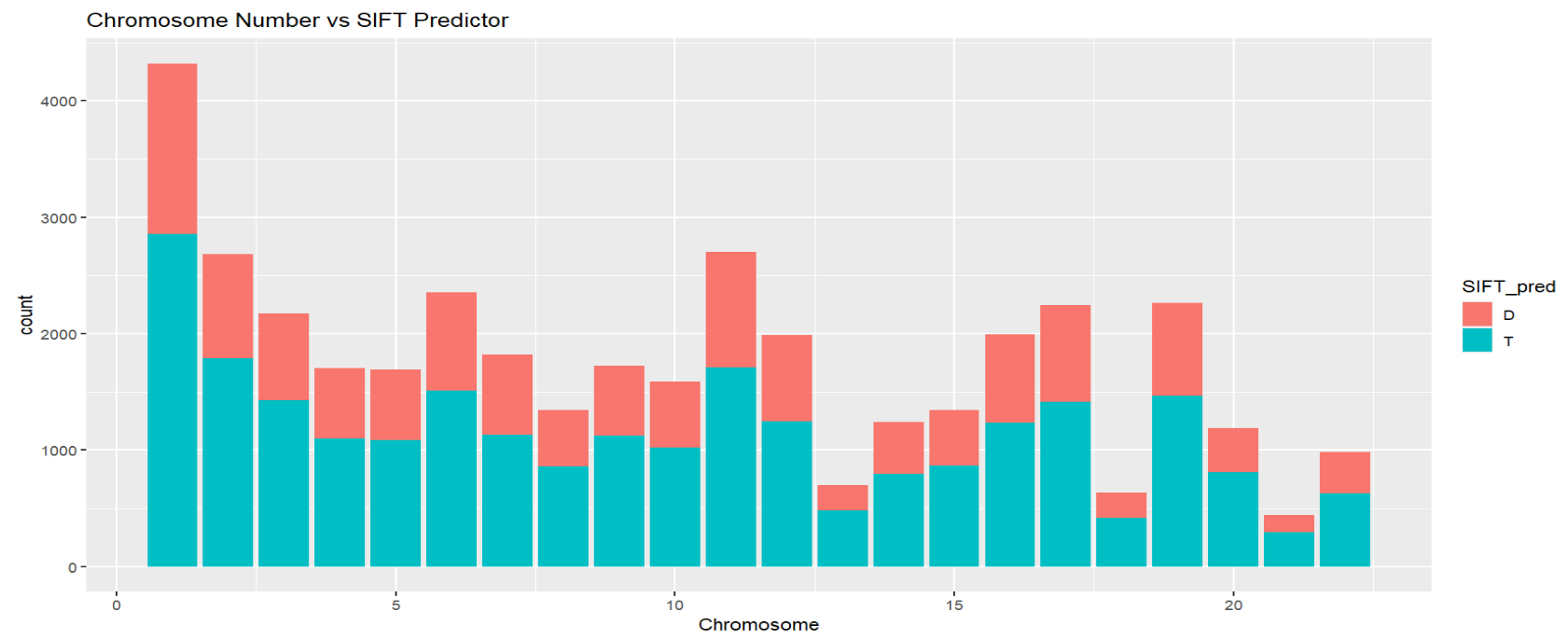


Figure 2. The bar graph shows the gene variant counts on each chromosome colored by the SIFT predictor tool results. Most variants were classified as tolerated (T), with the highest and lowest counts on chromosomes 1 and 21, respectively.

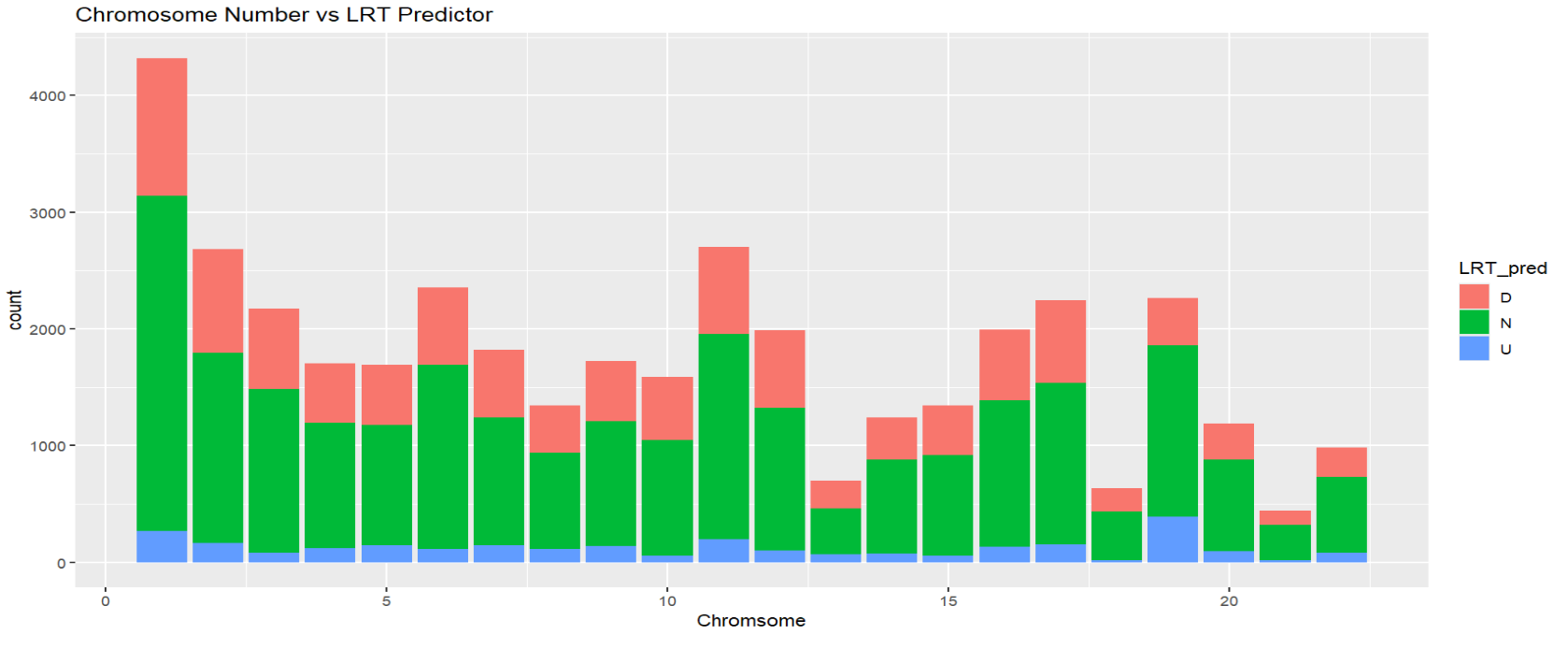


Figure 3. The bar graph shows the gene variant counts on each chromosome colored by the LRT predictor tool results. The majority of variants were classified as neutral (N). Chromosome 1 had the highest damaging (D) and neutral counts, while chromosomes 18 and 21 had the lowest unknown (U) predictions.

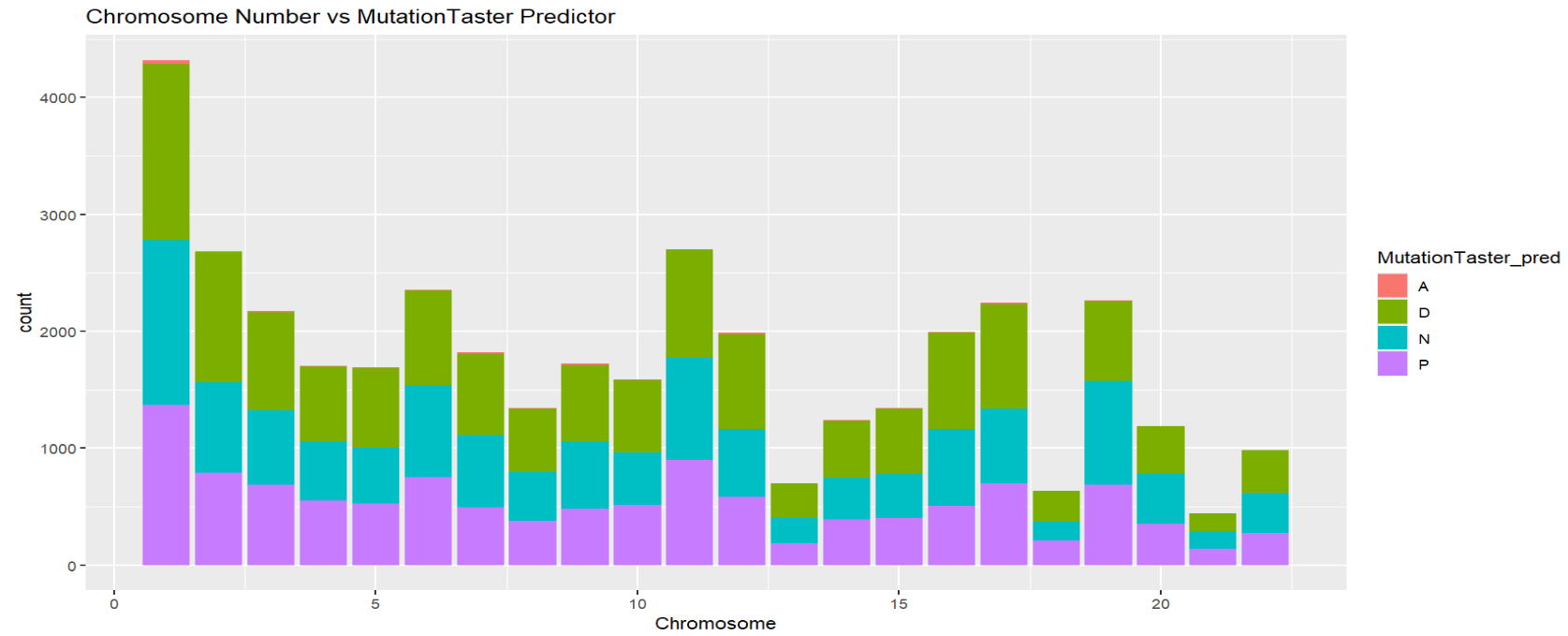


Figure 4. The bar graph shows the gene variant counts on each chromosome colored by the Mutation Taster predictor tool results. Most variants were classified as disease-causing (D or A) on chromosome 1. Chromosomes 18 and 21 had fewer neutral (N) variants.

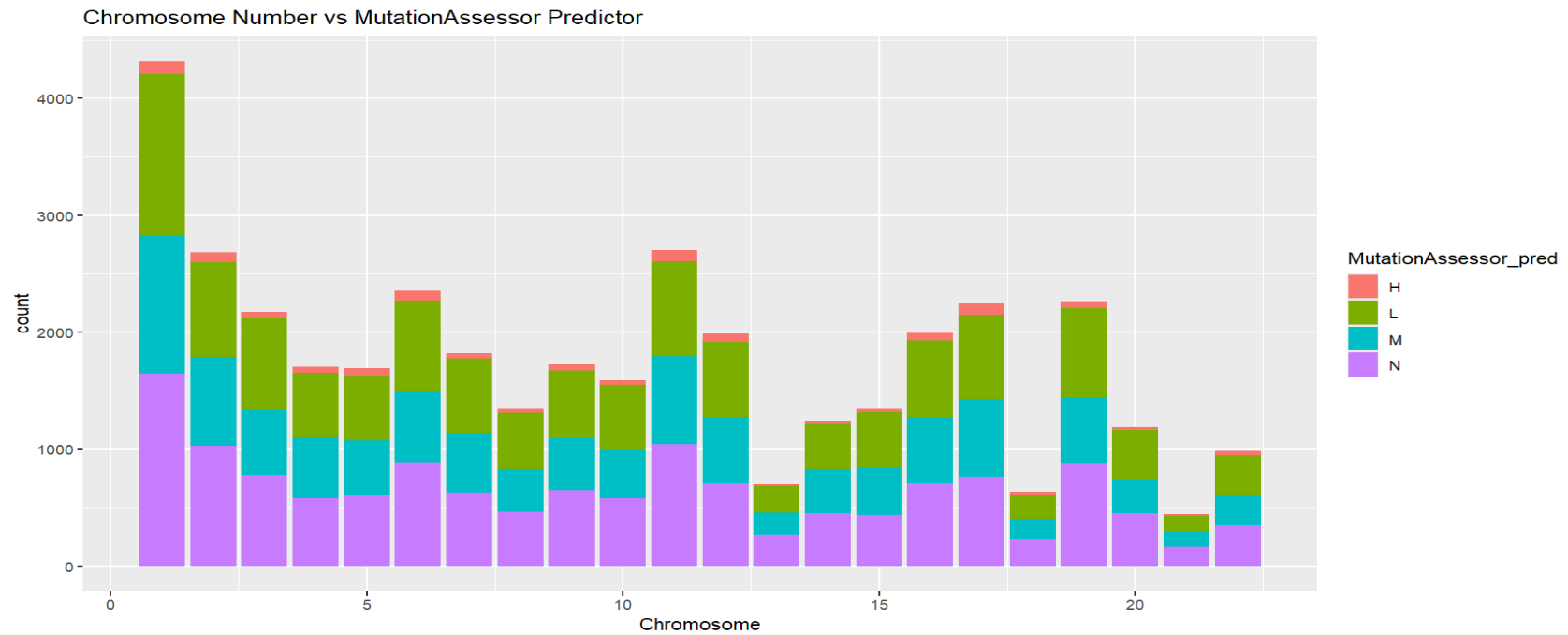


Figure 5. The bar graph shows the gene variant counts on each chromosome colored by the Mutation Assessor predictor tool results. High (H) impact variants were rare, while medium (M) impact variants were more common, especially on chromosome 1. Chromosome 21 consistently had the lowest counts.

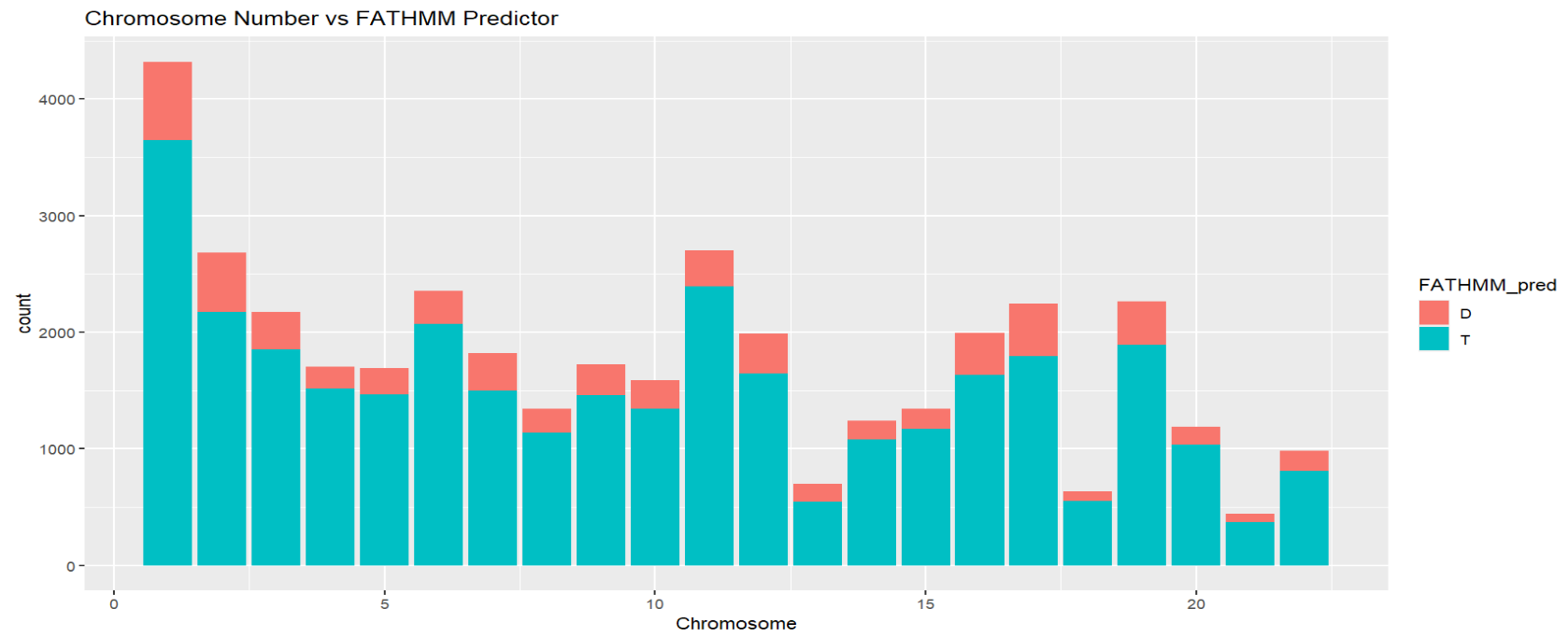


Figure 6. The bar graph shows the gene variant counts on each chromosome colored by the FATHMM predictor tool results. Most variants were tolerated (T), with chromosome 1 showing the highest number of potentially damaging (D) variants. Chromosomes 18 and 21 had the fewest damaging predictions.

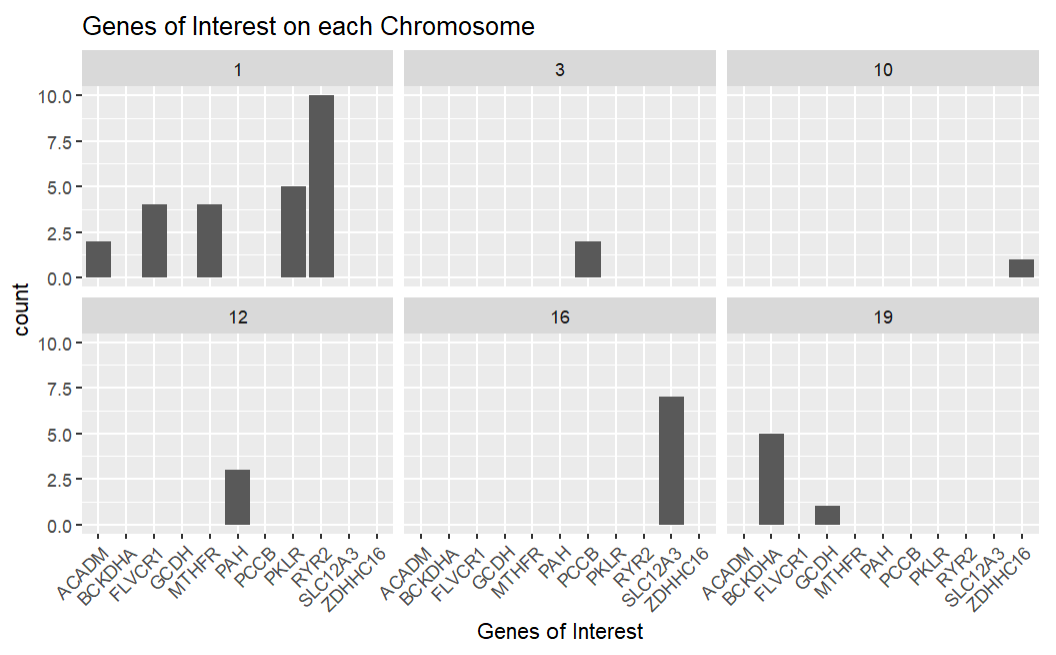


Figure 7. These bar graphs show the number of variants from each of the genes of interest organized by chromosome. The RYR2 gene on chromosome 1 had the most variants, while GCDH on chromosome 19 and ZDHHC16 on chromosome 10 had the fewest.

Table 2. This table shows the results of filtering for genes of interest and for only the harmful indicators in the predictor tools. The results are placed in descending order of the minor allele frequency (MAF). The PCCB gene had the highest MAF and REVEL score. The PAH gene had three distinct variants flagged as harmful across all prediction tools.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNPid | Gene | Chr | MAF\_percent | REVEL\_score | CADD\_raw | GERP++\_NR |
| 3:136048854:A:G | PCCB | 3 | 3.571429 | 0.919 | 5.072618 | 5.3 |
| 19:41930487:T:A | BCKDHA | 19 | 1.904762 | 0.953 | 5.097043 | 5.45 |
| 16:56919275:C:G | SLC12A3 | 16 | 1.428571 | 0.511 | 6.313977 | 5.4 |
| 12:103246653:C:T | PAH | 12 | 0.714286 | 0.985 | 6.13065 | 5.72 |
| 1:155261636:C:T | PKLR | 1 | 0.47619 | 0.947 | 5.750361 | 4.85 |
| 12:103234252:T:C | PAH | 12 | 0.238095 | 0.982 | 4.988886 | 5.63 |
| 12:103249091:C:T | PAH | 12 | 0.238095 | 0.89 | 5.048905 | 5.73 |

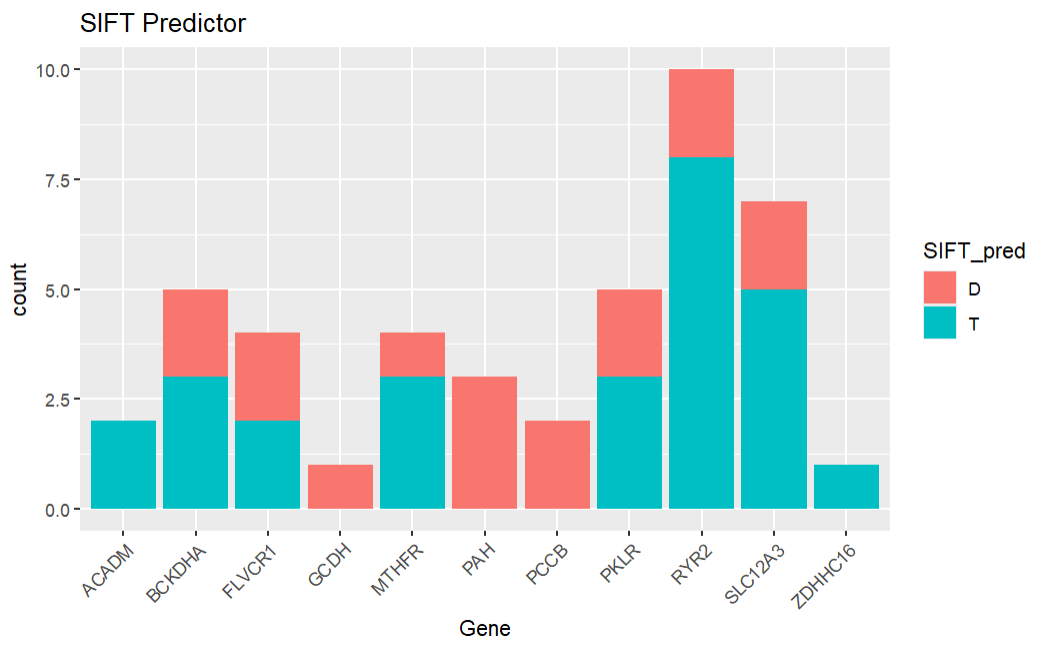


Figure 8. The bar graph shows the gene variant counts of each of the genes of interest colored by the SIFT predictor tool results. All variants in GCDH, PAH, and PCCB were classified as damaging. ACADM and ZDHHC16 variants were all tolerated.

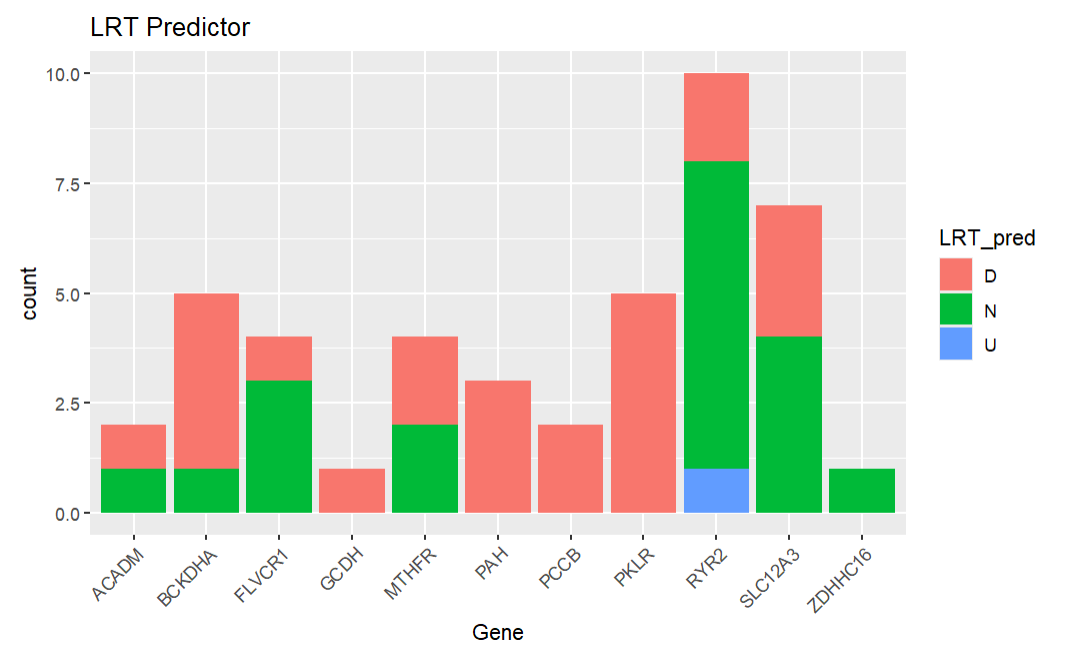


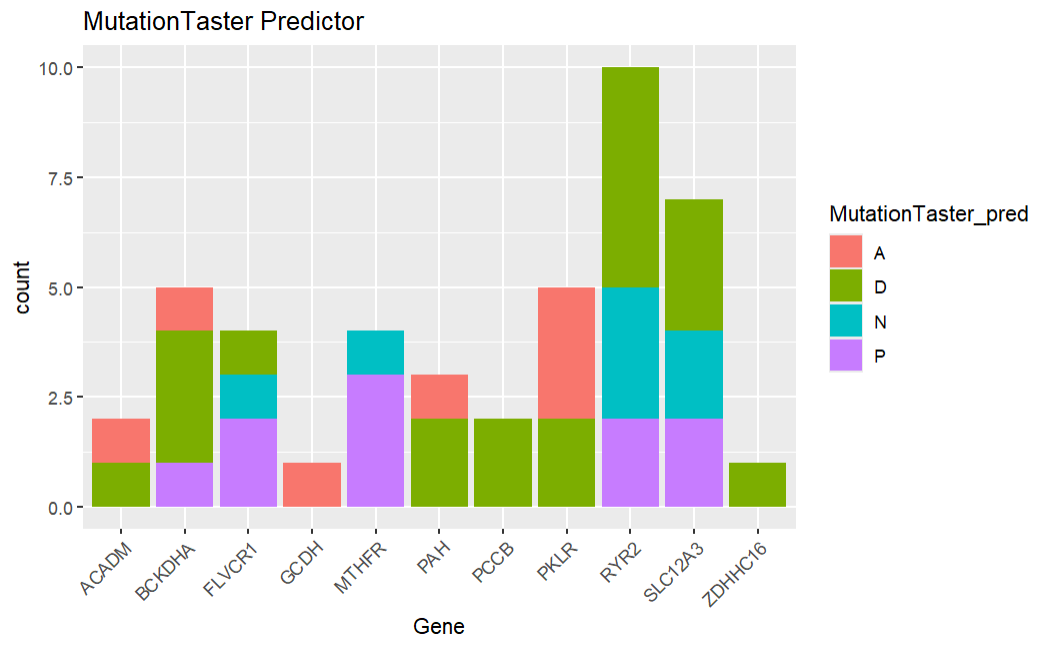
Figure 9. The bar graph shows the gene variant counts of each of the genes of interest colored by the LRT predictor tool results. All variants in PAH, PCCB, and PKLR were classified as damaging. RYR2 had the only unknown variant and the most neutral predictions.

Figure 10. The bar graph shows the gene variant counts of each of the genes of interest colored by the Mutation Taster predictor tool results. PCCB and ZDHHC16 had only disease-causing predictions. RYR2 had the highest number of disease-causing variants, while PKLR had the most known disease-causing variants (A).

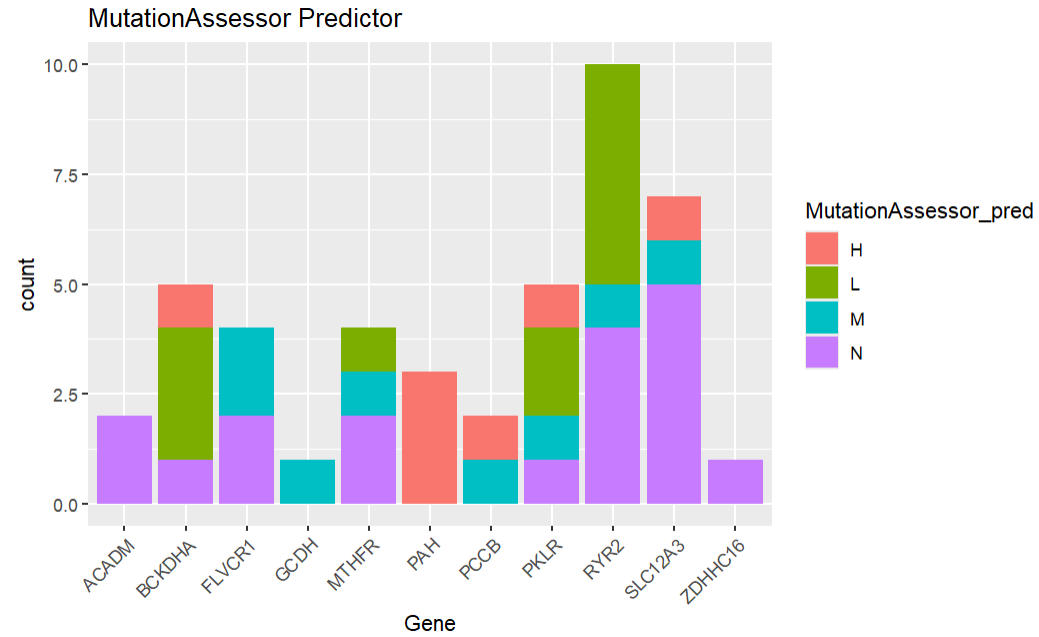


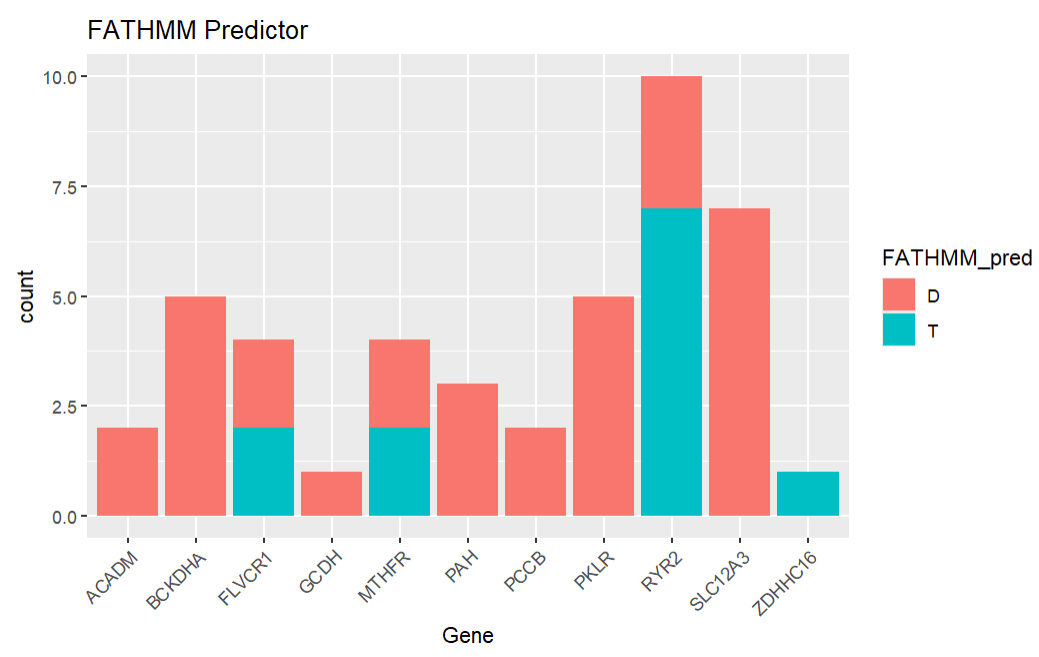
Figure 11. The bar graph shows the gene variant counts of each of the genes of interest colored by the Mutation Assessor predictor tool results. All PAH variants were classified as high impact. ACADM and ZDHHC16 were neutral, and GCDH was mostly medium impact.

Figure 12. The bar graph shows the gene variant counts of each of the genes of interest colored by the FATHMM predictor tool results. Most genes had damaging variants. Only ZDHHC16 and some variants of FLVCR1, MTHFR, and RYR2 were tolerated.

*Table 3. (Attached separately) The table shows the resulting gene variants when filtered for each harmful predictor. It is also sorted in descending order of the minor allele frequency. The results show 251 gene variants that meet the criteria after filtering.*

Table 4. The table represents the gene variants that were filtered for only the harmful predictor results and were homozygous. Of the 23 identified, four had minor allele frequencies above 5%, suggesting higher prevalence and potential for recessive disease inheritance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNPid | Gene | Chr | MAF\_percent | REVEL\_score | CADD\_raw | GERP++\_NR |
| 14:94847262:T:A | SERPINA1 | 14 | 7.142857 | 0.692 | 5.001297 | 5.18 |
| 6:110763935:T:C | SLC22A16 | 6 | 5.952381 | 0.873 | 4.104905 | 4.74 |
| 19:34868776:T:C | GPI | 19 | 5 | 0.744 | 4.180961 | 5.58 |
| 2:178565913:T:C | PDE11A | 2 | 5 | 0.841 | 5.393832 | 5.7 |
| 11:67809268:C:T | TCIRG1 | 11 | 4.52381 | 0.67 | 6.122438 | 3.72 |
| 17:39525750:C:T | KRT33B | 17 | 4.285714 | 0.769 | 5.269234 | 4.41 |
| 9:140357962:C:T | PNPLA7 | 9 | 4.285714 | 0.834 | 3.939536 | 4.48 |
| 22:32628900:C:T | SLC5A4 | 22 | 3.809524 | 0.783 | 5.495204 | 4.74 |
| 22:35478529:G:A | ISX | 22 | 3.809524 | 0.691 | 5.576396 | 4.94 |
| 3:136048854:A:G | PCCB | 3 | 3.571429 | 0.919 | 5.072618 | 5.3 |
| 15:76578762:G:A | ETFA | 15 | 3.095238 | 0.635 | 5.553098 | 5.65 |
| 15:89444912:G:A | MFGE8 | 15 | 3.095238 | 0.692 | 4.260405 | 5.36 |
| 1:160141491:C:T | ATP1A4 | 1 | 2.857143 | 0.889 | 5.076141 | 4.19 |
| 9:131193551:G:C | CERCAM | 9 | 2.142857 | 0.771 | 5.831571 | 5.02 |
| 19:41930487:T:A | BCKDHA | 19 | 1.904762 | 0.953 | 5.097043 | 5.45 |
| 2:224866427:A:G | SERPINE2 | 2 | 1.904762 | 0.848 | 4.618518 | 5.67 |
| 19:33702165:C:T | SLC7A10 | 19 | 1.666667 | 0.958 | 5.925265 | 5.45 |
| 6:155597134:G:A | CLDN20 | 6 | 1.428571 | 0.724 | 5.226399 | 5.38 |
| 15:78921602:G:A | CHRNB4 | 15 | 0.952381 | 0.774 | 4.792431 | 5.13 |
| 20:13514760:A:G | TASP1 | 20 | 0.952381 | 0.896 | 4.999185 | 5.79 |
| 12:103246653:C:T | PAH | 12 | 0.714286 | 0.985 | 6.13065 | 5.72 |
| 2:167162345:G:A | SCN9A | 2 | 0.238095 | 0.832 | 6.036371 | 5.96 |
| 10:71008478:G:A | HKDC1 | 10 | 0 | 0.929 | 5.465443 | 4.85 |

Table 5. (Attached separately) The table represents the gene variants that were filtered for only the harmful predictor results and were heterozygous. The results show 251 variants deemed harmful, representing carrier status. While these individuals may not be affected, their offspring could inherit pathogenic mutations.

Table 1 showed the database generated from the exomes of individuals from the Kish Valley Amish and Mennonite community. This database included the minor allele frequency of each gene variant and indicated how many individuals were homozygous or heterozygous for a given mutation, along with other relevant genetic details. Figure 1 further explored this data by displaying variant counts across chromosomes. Notably, chromosome 1 had the highest number of variant counts, followed by chromosomes 2 and 11.

To assess the functional impact of these gene variants, several predictive tools were employed. One such tool was SIFT, which evaluates whether an amino acid substitution is likely to affect protein function. It operates on the assumption that functionally important amino acids are evolutionarily conserved; therefore, changes in conserved regions are more likely to be deleterious. To make the prediction, it considers the position where the change occurs and the specific type of amino acid change. The tool categorizes substitutions as damaging (D) or tolerated (T) (Ng, 2003). According to the SIFT predictions shown in Figure 2, most gene mutations were predicted to be tolerable and not damaging to protein function. However, predictions varied across tools, so additional analyses were conducted.

The LRT (Likelihood Ratio Test) evaluates whether a gene is under selective constraint compared to a model of neutral evolution. For LRT, “D” designates a mutation has a harmful functional impact, “N” means there is no significant impact, and “U” means there is not enough information to determine impact (“Population Genetics”, 2023). As seen in Figure 3, most gene mutations were considered neutral. However, there were mutations across all chromosomes where insufficient data prevented a definitive impact assessment.

Another layer of analysis was added using MutationTaster, which uses a classifier trained on known disease-causing and tolerable mutations to predict the likelihood of a mutation being pathogenic. For this tool, “A” means it is known to be disease-causing, “D” represents potentially disease-causing, “N” means probably harmless, and “P” means automatically classified as harmless based on the 1000 Genomes dataset (“MutationTaster: Documentation”, n.d.). Figure 4 indicated that the most disease-causing and potentially disease-causing gene variants were found on chromosome 1, though there were similarly classified variants present on other chromosomes.

Similarly, the MutationAssessor tool evaluates the functional impact of amino acid substitutions based on evolutionary conservation. Predicted from the functional impact on protein, the tool assigns scores of H (high), M (medium), L (low), or N (neutral) (NIF, 2025). As shown in Figure 5, few variants were classified as having a high functional impact, though many had medium impact.

FATHMM predicts the functional consequences of both coding and non-coding variants through functional analysis using hidden Markov models. Its classifications are D (potentially disease-causing) and T (tolerated) (Shihab, n.d.). According to Figure 6, most variants were tolerated, with the most potentially disease-causing mutations found on chromosome 1 followed by chromosomes 2 and 11.

Looking particularly at specific genes of interest to Dr. Morton, Figure 7 highlighted their mutation counts by chromosome. Most of these genes were found on chromosome 1, but there were genes with a larger number of variants overall such as genes RYR2 and SLC12A3. Additional figures provided insight into the functional consequences of these genes’ mutations. For example, Figure 8 revealed that GCDH, PAH, and PCCB carried only damaging substitutions according to SIFT. Conversely, ACADM and ZDHHC16 showed only tolerable substitutions. LRT results in Figure 9 aligned with these findings, showing all variants of GCDH, PAH, PCCB, and PKLR as potentially harmful. Substantial numbers of harmful variants were also found in BCKDHA and SLC12A3.

Figures 10 further evaluated protein-level impact using the MutationTaster tool. According to this figure, genes of interest that were most likely disease-causing were ACADM, BCKDHA, GCDH, PAH, PCCB, PKLR, RYR2, SLC12A3, and ZDHHC16. In contrast, the genes FLVCR1 and MTHFR had mostly neutral or potentially harmless mutations. Figure 11 showed the genes of interest with high potential impact on protein function to be BCKDHA, PAH, PCCB, PKLR, and SLC12A3. There were a few genes with some medium potential impact, but gene FLVCR1 had the most.

Furthermore, Figure 12 demonstrated that nearly all genes of interest exhibited functional consequences, except ZDHHC16 and some variants of FLVCR1, MTHFR, and RYR2. Table 2 summarized filtered results showing only harmful predictions from each tool for Dr. Morton’s genes of interest. Among these, the PCCB gene variant displayed the highest minor allele frequency, suggesting a larger proportion of people in this population carried this mutation. Notably, the PAH gene contained three different mutations that were considered harmful by all predictor tools.

Table 2 also showed REVEL, CADD, and GERP++ scores that help with determining pathogenicity. REVEL (rare exome variant ensemble learner) scores are based on a model that is trained on multiple predictive tools and are applied to missense variants to assess pathogenicity (Ioannidis et al., 2016). Among the genes of interest, all but one had a REVEL score above 0.8, indicating a high likelihood of disease causation, as REVEL scores range from 0 to 1. CADD (Combined Annotation Dependent Depletion) scores are also used to predict the deleteriousness of gene variants as a higher score indicates that a variant is more likely to be damaging. CADD compares observed variants with simulated mutations and generates scores that correlate with both allelic diversity and disease association (University of Washington, n.d.). Within the genes of interest, the CADD scores centered around 5-6, indicating a lower likelihood of pathogenicity compared to more deleterious variants. GERP++ (Genomic Evolutionary Rate Profiling) scores identify constrained genomic regions by comparing observed substitutions to those expected under neutral evolution (Davydov et al., 2010). For the whole database, the highest GERP++ scores were about 9, while scores for the genes of interest were generally around 6. Since scores above 2 suggest a region is under selective pressure, these variants are considered deleterious and less likely to occur by random chance.

Lastly, Table 3 identified a total of 251 gene variants considered harmful by the predictive tools. These were then filtered by being either homozygous or heterozygous in Tables 4 and 5. Table 4 revealed 23 homozygous variants, while Table 5 showed 251 heterozygous variants.

**Discussion**

This study developed a searchable database of exomes from 210 Kish Valley Amish and Mennonite patients. Once the database was constructed, exploratory analysis was conducted to identify meaningful insights into the gene variants it contained. A particular emphasis was placed on the use of predictive tools to reveal patterns in variant distribution, pathogenicity, and gene-specific impacts.

The resulting database includes key genetic metrics such as minor allele frequency, homozygosity, heterozygosity, and more. It serves as a valuable resource for both healthcare professionals and patients seeking a better understanding of gene function. The database can be used to estimate the prevalence of specific variants in the population to better understand their potential impact (Lowrance, 2001). Additionally, the database can be compared to variant data from outside populations to better understand disease risk and the frequency of specific mutations. There are characteristics of the outside population that might benefit from the discovery of disease-causing variants (Fatumo et al., 2022). There are also many examples of clinically important variants that were only discovered in underrepresented populations, such as the Plain population (Fatumo et al., 2022).

Furthermore, analysis revealed a high concentration of potentially pathogenic variants on chromosome 1 across all predictive tools. Considering the large size of chromosome 1 compared to other chromosomes, it is not surprising to observe more gene variants on this chromosome. Chromosome 1 accounts for about 8% of the entire genome and is six times longer than chromosome 21, the smallest in the genome (Pearson, 2006). This suggests this chromosome may need particular attention in future research. This concentration may be a result of the founder effect or reduced genetic diversity within this closed population.

Across the genome, the SIFT predictor indicated that most variants were tolerated and not damaging to protein function. However, there were still many variants determined by this predictor to be damaging. Additionally, some variants from the LRT predictor tool were classified as “unknown,” meaning there was not enough information for the tool to classify them. These variants could be prioritized for further study to better understand their pathogenicity.

Figure 5, which used the MutationAssessor tool, showed few variants with high-impact on amino acid substitutions, though a significant number were classified as having medium impact. These results align similarly with Figure 2, where the medium-impact variants were also predicted by SIFT to be damaging, and with Figure 4, where MutationTaster similarly identified many of these variants as disease-causing. Together, these results indicate these variants found to be damaging through each of these predictors should be prioritized when considering gene variant research and clinical care.

In examining the genes of interest prioritized by Dr. Morton, chromosome 1 again contained the most genetic variants. Among these, the genes RYR2 and SLC12A3 exhibited the most variants, emphasizing them for further genetic investigation to better understand their susceptibility to mutation. In Figure 8, the genes GCDH, PAH, and PCCB were all predicted to be damaging. Among the genes of interest, these genes should be prioritized for their harm to the patient’s health. In particular for genetic screening assays similar to the ones done in the Clinic for Special Children, these are genes that should be screened for in the Plain population because of their larger potential harm to the patient (Crowgey et al., 2019). On the other hand, the genes ACADM and ZDHHC16 were found to be tolerated. However, if a patient had these gene mutations and was still suffering symptoms, this could be an indicator of a complex gene interaction or a misclassification by the predictor tools. The results of the LRT predictions seen in Figure 9 showed the same genes along with the gene PKLR as potentially harmful and many other as having at least one variant that was harmful, such as BCKDHA and SLC12A3. These results can help with prioritizing certain gene variants over others for initial research or treatment. Figures 10 and 11 further indicated that most genes of interest had variants that were potentially disease-causing or carried medium to high functional impacts. Similarly, the FATHMM predictor (Figure 12) found that most genes, except ZDHHC16, contained disease-causing variants. The lack of distinguishing results from this predictor tool made these results unhelpful for identifying differences between the gene variants.

The identification of variants with high REVEL scores (mostly above 0.8) further supports their potential to cause disease. This is especially true when the GERP++ scores are around 6, indicating these variants are under strong evolutionary constraint. However, the CADD scores were low for these genes (5-6). Meaning these gene variants may not be among the most deleterious genome-wide, but they still carry harmful consequences. The gene PCCB stands out among the genes of interest due to its high minor allele frequency in the database. This indicates it is relatively common in this population despite being predicted as harmful. This may reflect a form of genetic drift in this isolated group.

Tables 4 and 5 showed that only 23 of these harmful variants are homozygous, indicating a higher likelihood of these variants being passed to offspring. All of the gene variants had patients that were heterozygous for the variant. This means there is a chance the variant is passed on to offspring or that the patient is a carrier for the disease. The higher proportion of heterozygous compared to homozygous alleles shows the expected autosomal recessive inheritance. In other words, while many individuals carry the potentially disease-causing mutation, fewer individuals are at risk of expressing recessive conditions. This might also reflect the potential that the survivors of the homozygous alleles may be limited because of the damaging mutations.

Overall, the consistency across multiple prediction tools, allele frequency data, and conservation scores provides strong evidence for prioritizing the genes found to be harmful in this database for future clinical and functional studies. Further research would be necessary to establish the clinical significance of the variants. Importantly, findings from this closed population could have broader impacts, as the genetic background of the Plain community could overlap with parts of the general population. This could also be relevant given that some individuals from these communities have decided to leave their community and integrate into the broader society (Payne et al., 2011a). As demonstrated in the study by Strauss et al., the Clinic for Special Children significantly reduced medical costs in their community through the integration of genetic research and patient care. This further supports the potential value of incorporating this database and its findings into preventative care strategies and cost reduction efforts for not only the Plain population, but also other populations.

Despite these findings, this study had several limitations, the most significant being its small sample size of only 210 individuals. While this may be substantial for a closed or specific population, it limits the applicability of the findings to the broader population. As a result, the database may not be as applicable to the general public. Nevertheless, it still includes known disease-causing variants that may be of relevance to wider groups. Another limitation lies in the tools used to predict variant pathogenicity, which are not perfect and are prone to errors. For instance, the LRT tool was unable to classify certain variants and marked them as “unknown.” These unclassified variants could still pose a risk and merit further investigation. Future research should examine how this database compares to those from other populations and how its findings could inform the prioritization of specific genetic variants in patient care.

**Acknowledgements**

Thank you to Dr. Regina Lamendella, Dr. Holmes Morton, and Brittney McMullen for their help and guidance throughout this project.

**Literature Cited**

Agarwala, R., Biesecker, L. G., & Schäffer, A. A. (2003). Anabaptist genealogy database. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *121C*(1), 32–37. https://doi.org/10.1002/ajmg.c.20004

Akalin, A. (2020, September 30). Computational genomics with R. https://compgenomr.github.io/book/

Anderson, C., & Bacon, R. (2024). Profiles of major Amish settlements in North America: A guide for researchers and service providers. *Journal of Amish and Plain Anabaptist Studies*, 47–79. https://doi.org/10.62192/japas.v12i1-2n05

Anderson, C., & Potts, L. (2020). The Amish health culture and culturally sensitive health services: An exhaustive narrative review. *Social Science &amp; Medicine*, *265*, 113466. https://doi.org/10.1016/j.socscimed.2020.113466

Anderson, C., & Potts, L. (2021). Research Trends in Amish Population Health, a Growing Literature about a Growing Rural Population. *Journal of rural social sciences*, *36*(1), 6.

Anderson, C., & Potts, L. (2022). Physical health conditions of the Amish and intervening social mechanisms: an exhaustive narrative review. *Ethnicity & health*, *27*(8), 1952–1978. https://doi.org/10.1080/13557858.2021.1968351

Conlin W. E. (2021). Ethical Considerations for Treating the Old Order Amish. *Ethics & behavior*, *31*(6), 419–432. https://doi.org/10.1080/10508422.2020.1805614

Crowgey, E. L., Washburn, M. C., Kolb, E. A., & Puffenberger, E. G. (2019). Development of a novel next-generation sequencing assay for carrier screening in Old Order Amish and Mennonite populations of Pennsylvania. *The Journal of Molecular Diagnostics*, *21*(4), 687–694. https://doi.org/10.1016/j.jmoldx.2019.03.004

Daud, D., Griffin, H., Douroudis, K., Kleinle, S., Eglon, G., Pyle, A., Chinnery, P. F., & Horvath, R. (2015). Whole exome sequencing and the clinician: We need clinical skills and functional validation in variant filtering. *Journal of Neurology*, *262*(7), 1673–1677. https://doi.org/10.1007/s00415-015-7755-y

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12). https://doi.org/10.1371/journal.pcbi.1001025

Ehrenberg, S., Walsh Vockley, C., Nelson, E., Baker, J., Arcieri, M., Lindenberger, J., & Ghaloul-Gonzalez, L. (2021). Under-referral of Plain community members for genetic services despite being qualified for genetic evaluation. *Journal of genetic counseling*, *30*(4), 1084–1090. https://doi.org/10.1002/jgc4.1395

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., & Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. Nature medicine, 28(2), 243–250. https://doi.org/10.1038/s41591-021-01672-4

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W. (2016). Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

McMullen, Brittney (n.d.). “Exome Database Project (EDP) – Schedule and Instructions.”

Lowrance, W. W. (2001). The promise of human genetic databases. *BMJ*, *322*(7293), 1009–1010. https://doi.org/10.1136/bmj.322.7293.1009

Miller, K., Yost, B., Abbott, C., Thompson Buckland, S., Dlugi, E., Adams, Z., Rajagopalan, V., Schulman, M., Hilfrank, K., & Cohen, M. A. (2019). Health Needs Assessment of five pennsylvania plain populations. *International Journal of Environmental Research and Public Health*, *16*(13), 2378. https://doi.org/10.3390/ijerph16132378

*MutationTaster: Documentation*. MutationTaster. (n.d.). https://www.mutationtaster.org/MutationTaster69/info/documentation.html

Ng, P. C. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

NIF. (2025, July 17). Mutationassessor. Resource Summary Report. https://neuinfo.org/data/record/nlx\_144509-1/RRID:SCR\_005762/resolver/pdf&i=rrid:scr\_005762

Nolt, S. M. (2020). The emergence of Amish Genetic Studies: A brief history of collaboration and reciprocity. The Journal of Plain Anabaptist Communities, 1(1), 38–51. https://doi.org/10.18061/jpac.v1i1.7659

Payne, M., Rupar, C. A., Siu, G. M., & Siu, V. M. (2011a). Amish, Mennonite, and Hutterite Genetic Disorder Database. Pediatrics &amp; Child Health, 16(3). https://doi.org/10.1093/pch/16.3.e23

Payne, M., Rupar, C. A., Siu, G. M., & Siu, V. M. (2011b). Amish, Mennonite, and Hutterite Genetic Disorder Database. *Paediatrics &amp; Child Health*, *16*(3). https://doi.org/10.1093/pch/16.3.e23

Pearson, H. (2006, May 17). *Human genome completed (again)*. Nature News. https://www.nature.com/news/2006/060515/full/news060515-12.html#:~:text=Chromosome%201%20is%20the%20largest,in%20their%20eight%2Dyear%20quest.

*Population Genetics and statistics for forensic analysts: Likelihood ratio*. National Institute of Justice. (2023, July 17). https://nij.ojp.gov/nij-hosted-online-training-courses/population-genetics-and-statistics-forensic-analysts/statistics/logical-approach/likelihood-ratio

Puffenberger, E. G., Jinks, R. N., Sougnez, C., Cibulskis, K., Willert, R. A., Achilly, N. P., Cassidy, R. P., Fiorentini, C. J., Heiken, K. F., Lawrence, J. J., Mahoney, M. H., Miller, C. J., Nair, D. T., Politi, K. A., Worcester, K. N., Setton, R. A., DiPiazza, R., Sherman, E. A., Eastman, J. T., … Strauss, K. A. (2012). Genetic mapping and exome sequencing identify variants associated with five novel diseases. PLoS ONE, 7(1). https://doi.org/10.1371/journal.pone.0028936

Shihab, H. A. (n.d.). Fathmm. fathmm. https://fathmm.biocompute.org.uk/

Strauss, K. A., Puffenberger, E. G., & Morton, D. H. (2012). One community’s effort to control genetic disease. *American Journal of Public Health*, *102*(7), 1300–1306. https://doi.org/10.2105/ajph.2011.300569

U.S. National Library of Medicine. (n.d.-a). *SNP attributes*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs\_attributes.html

University of Maryland Baltimore. (1995). Amish Complex Genetic Disease Database. https://datacatalog.hshsl.umaryland.edu/dataset/33

University of Washington. (n.d.). *Combined annotation dependent depletion*. CADD- Combined Annotation Dependent Depletion. https://cadd.gs.washington.edu/

*Whole genome association analysis toolset*. PLINK: Whole genome data analysis toolset. (2025, January 17). https://zzz.bwh.harvard.edu/plink/

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science: Import, Tidy, transform, visualize, and model data*. O’Reilly.