Decoding Sandhoff Disease: Unveiling the Impact of the HEXB Gene Mutation

Arnav Brahmasandra

March 5, 2023

1 Introduction

Genomics has revolutionized the field of medicine and specifically personalized healthcare. Personal genomics involves the sequencing and analysis of the individual's genome to understand the genetic basis of inherited and acquired diseases. Personal genomics has also allowed for advances in the field of precision medicine, a medical approach that takes into account the genetic variation, environment, and lifestyle of each person. It involves tailoring the prevention, diagnosis, and treatment of diseases to a specific patient's characteristics, and can lead to more effective treatments and lower healthcare costs [15]. Precision medicine is particularly important for complex diseases, such as cancer and cardiovascular disease, which have several types and require different treatment approaches for each patient.

Personal genomics has been able to flourish both because of novel next-generation sequencing (NGS) techniques, as well as advances in high performance computing power. NGS techniques have made genotyping and sequencing more affordable, efficient, and accurate, by sequencing many reads in parallel.

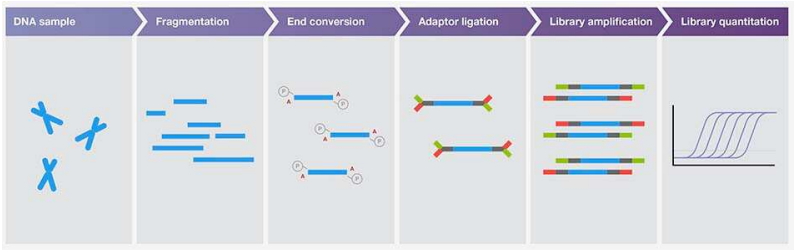


Figure 1: NGS Sequencing Process Flow

In NGS, genomic DNA is fragmented into many reads by sonication and specific adapters are added to each fragment. These fragments make up the NGS library. The fragments are then amplified using polymerase chain reaction (PCR). Finally, the sequencing machine reads the DNA sequence through a variety of methods: one being pyrosequencing. In pyrosequencing, nucleotides are added one at a time to each cluster of DNA reads. The pyrophosphate released by the addition of a complementary nucleotide can release light through a sequence of chemical reactions, allowing for the sequencing of the reads [15].

Once all of the reads have been sequenced, the reads are aligned to the reference human genome and assembled to form the full DNA sequence. The assembly of the reads has been massively sped up by the completion of the Human Genome Project in the early 2000s, which sequenced a reference human genome. In addition, massive improvements in high performance computing, specifically in parallel computing and storage, has allowed for quick analysis on massive genomic sequences.

One NGS-based technology that has enabled the identification of rare genetic variants is exome sequencing, a cost-effective approach that selectively captures

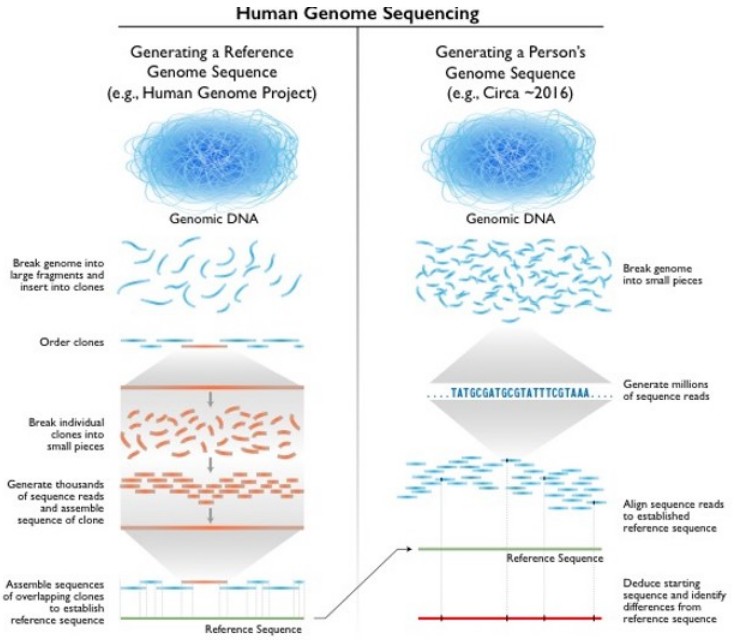


Figure 2: Image illustrating whole genome sequencing before and after the completion of the Human Genome Project. Once the reference genome was completed, the alignment of all the individual reads became much easier.

and sequences only the protein-coding regions of the genome [12]. There are two main methods of exome capture: hybrid capture and PCR capture. In hybrid capture, a set of biotinylated probes that specifically target the exome regions of interest are hybridized to the fragmented DNA [10]. In PCR capture, a set of PCR primers that target the exome regions of interest are used to amplify the DNA fragments, and then they are sequenced using NGS [2].

In this paper, exome sequencing data from an individual with rare genetic disorders is analyzed, using alignment, genotyping, and annotation of the variants. Alignment involves mapping the sequenced genomic data to the reference genome. Genotyping involves the identification of variants in the sequenced exomes. Finally, annotation involves the interpretation of the identified variants outputted from genotyping.

This work aims to identify and understand the genetic variants that lead to rare disorders. This analysis provides insight into how these genetic variations lead to changes in protein structure and loss of function, potentially aiding in the development of novel therapeutic approaches.

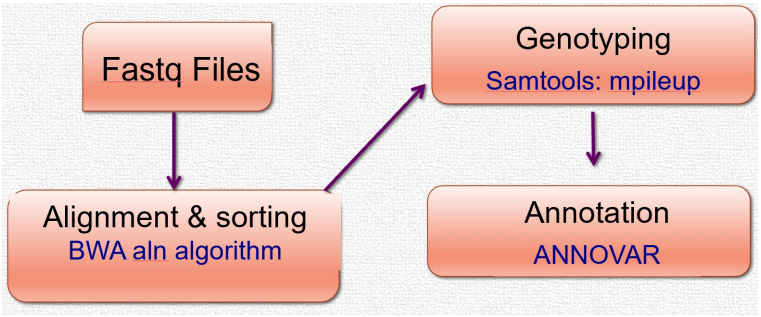


Figure 3: Analysis Pipeline Schematic

2 Methods

The data was comprised of two fastq files, due to the paired end approach of NGS sequencing, where each read is sequenced from the left and the right. First, the sequences were aligned to the reference genome using BWA alignment [13]. The -q parameter was set as 5 in the bwa aln step to specify the quality threshold for filtering out low-quality reads, meaning that any base with a quality score under 5 would not be used in the alignment process. In addition to -q, the -t parameter was set to 28 to specify the number of CPU cores to use for the alignment.

Then, bwa sampe was used to create the BAM file, and the -P flag was passed to indicate that the input files contained data from paired-end sequencing reads. Then, samtools sort was used to sort and index the BAM file. The -@ and -m parameters were used to specify the number of threads and maximum amount of memory samtools could use (-@ 28 -m 1500M) [13].

The second step of the analysis process was genotyping using samtools mpileup [14]. First, samtools index was used to index the BAM file. Next, mpileup was used to create the raw vcf file. The -t parameter is set to SP to specify the format of the output columns for the pileup as a space separated list of fields containing the nucleotide base, read depth, and mapping quality for each position in the reference sequence. Both the -u and -v flags are set to tell samtools to output the results in an uncompressed BAM format, as well as output positions with zero coverage. Next, the raw vcf file is filtered using Samtools vcfutils varFilter and the output is written to another file. The -D parameter is set to 100 to specify the maximum read depth allowed for a variant call. Finally, the variants with a Phred quality score below 50 are removed.

The third and final step of the analysis pipeline was annotation using annovar [17]. The vcf files from genotyping were converted to annovar format using convert2annovar. The variants were then annotated using information from the databases GWAS, dbSNP, refgene, as well as various functional scores.

The result of this pipeline was a large file containing information about many different genetic mutations: Single-nucleotide polymorphisms (SNPs) and

Insertion-Deletions (Indels). To select variants of interest I searched for mutations that were in exonic regions and nonsynonymous, meaning the mutation changes the amino acid sequence of the coded protein. In addition, I looked for mutations that were listed as pathogenic on dbSNP, or that were nonsense mutations, a mutation resulting in a premature stop codon.

3 Results

198,000 variants were identified through the genomic analysis pipeline. 123,984 of these variants were deemed high quality, as defined with a Phred quality score cutoff of 50, meaning there is less than a 1 in 100,000 chance of an incorrect base call. 13,297 of the high quality variants were identified as exonic, meaning they actually code for proteins.

From the exonic variants, 13,120 were SNPs and 177 were indels (80 insertions and 97 deletions). 6,325 variants were synonymous, also known as silent mutations, where the genetic mutation does not result in a change to the amino acid sequence of the coded protein. 5,170 variants were nonsynonymous, meaning that the mutation did effect the primary structure of the protein. 172 of the indels resulted in frameshift mutations, meaning that the indel caused a shift in the reading frame of the codons. Frameshift mutations are especially dangerous because they can cause a change in all amino acids in the protein downstream from the mutation site, which can lead to premature stop codons and truncation or completely non-functional proteins. 41 of the variants led to nonsense mutations, or stop gain mutations, leading to a premature stop codon in the genetic code. Lastly, 8 variants led to stop loss mutations, leading to the loss of a stop codon in the reading frame of the genetic code. Stop gain and loss mutations can be extremely dangerous as they result in the production of non-functional or truncated proteins, potentially disrupting important biological processes.

3.1 Mutation Distribution Analysis

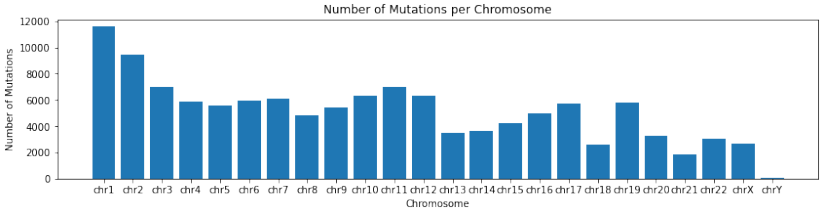


Figure 4: Chromosomal Distribution

The distribution of variants can be visualized in many different ways. In Figure 4, we can see the distribution of the mutations per chromosome. Chromosome 1 clearly has the greatest number of mutations, with 11,569. Chromosome 2 has the second greatest number of mutations with 9,465. We can also

visualize the fraction of mutations within each gene type (intergenic, intronic, exonic, etc.) with a pie chart. In Figure 5, we can see that most mutations occur in intronic genes, with intergenic genes second, and exonic genes third. Very few mutations occur in the UTR3 and UTR5 genetic regions, most likely because of their extremely short length when compared to intronic, exonic, and intergenic portions of the DNA.

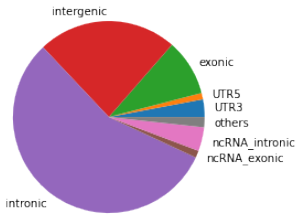


Figure 5: Gene Type Distribution

In Figure 6, we can see the correlation between the number of mutations on a given chromosome and its size, measured in the total amount of base pairs. The smallest chromosome is comprised of approximately 20Mb, while the largest chromosome has around 250Mb (Mb - million base pairs). The maximum number of mutations on a given chromosome is nearly 12,000 mutations, while the smallest number of mutations on a chromosome is nearly 0. Overall, there is a clear positive linear correlation between the number of mutations on a chromosome and its size. This correlation can be quantified using the Pearson correlation coefficient (r-value), which can range from -1 to 1. The closer $|r|$ is to 1, the more the best-fit line describes the variation in the data. The sign of the r-value indicates the direction of correlation (negative sign implies negative correlation and positive sign indicates positive correlation). This correlation has an r-value of 0.79, which indicates a strong positive correlation.

In Figure 7, we can see the correlation between the number of mutations on a given chromosome and the number of genes, exons, and introns. The number of genes per chromosome was obtained from the NCBI Human Genome Database [1], and the both the numbers of introns and exons were obtained from Table 1 in *Distribution of introns and exons in the human genome* [16].

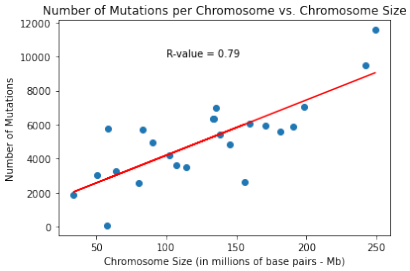
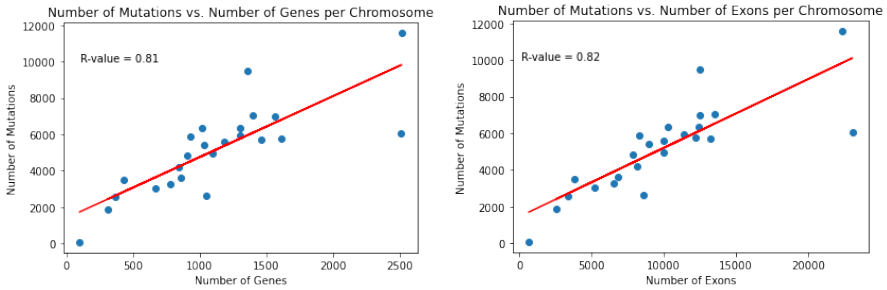
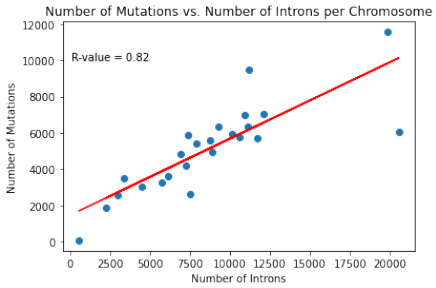


Figure 6: Chromosomal Size Distribution



(a) Number of Genes

(b) Number of Exons



(c) Number of Introns

Figure 7: Distribution of the Variants

Visually, there is clearly a positive correlation between the number of genes, exons, and introns in a chromosome and the number of mutations. This is quantified by the r-value of the correlations which are 0.81, 0.82, and 0.82 respectively. These large, positive r-values indicate a strong positive, linear correlation between the variables.

3.2 Specific Variant Analysis

Next, I wanted to examine the specific variants even closer. Out of the exonic variants, I looked for variants that were both non-synonymous and pathogenic. In Table 1, I listed 15 specific mutations throughout the genome that fit these criteria. In addition to listing the chromosome and location of the mutation on the chromosome, I have also noted their phenotypic implications.

One interesting mutation is a non-synonymous variant on chromosome 1 (rs35948326) at position 158654738. The mutation was in the SPTA1 gene, which changed the amino acid Alanine to Aspartic Acid. The implications of this change were either Type 3 autosomal recessive, or just recessive Spherocytosis. Spherocytosis is a disease that affects red blood cells, and people with the disease typically experience anemia, jaundic, and an enlarged spleen [8]. Globally, this mutation occurs in around 4.5% of the population, however this mutation has higher allele frequencies in the European population (at 4.9%) [3].

Location of Variant	Implications
Chromosome 1, Position 100206504	Intermediate maple syrup urine disease type 2
Chromosome 1, Position 158654738	Spherocytosis autosomal recessive type 3, Spherocytosis recessive
Chromosome 3, Position 10289773	Predisposition to Obesity, Susceptibility to Metabolic Syndrome
Chromosome 3, Position 45772602	Hyperglycinuria, Digenic Iminoglycinuria
Chromosome 5, Position 74685445	Infantile type Sandhoff Disease
Chromosome 5, Position 177093242	Cancer Progression and Tumor Cell Motility
Chromosome 7, Position 150999023	Susceptibility to Coronary Artery Spasm, Late Onset Alzheimer's Disease, Hypertension, Ischemic Heart Disease, Ischemic Stroke
Chromosome 9, Position 133436862	Upshaw-Schulman syndrome
Chromosome 10, Position 68885620	Preeclampsia
Chromosome 12, Position 120737875	Deficiency of butyryl-CoA dehydrogenase
Chromosome 12, Position 120999579	Maturity onset diabetes type 3
Chromosome 12, Position 121857429	4-alpha-hydroxyphenylpyruvate hydroxylase deficiency
Chromosome 16, Position 27344882	Resistance to Atopy, Acquired immunodeficiency syndrome
Chromosome 16, Position 56514589	Barder-biedl syndrome
Chromosome 16, Position 69711242	Susceptibility to Benzene Toxicity, Leukemia post chemotherapy, Lung Cancer, etc.

Table 1: Table of Pathogenic Mutations

Another interesting mutation is a non-synonymous variant on chromosome 3 (rs696217) at position 10289773. The mutation was in the GHRL gene, which changed the amino acid Leucine to Methionine. The implications of this change were a genetic predisposition to obesity, or susceptibility to metabolic syndrome. Ghrelin, the protein coded for by the GHRL gene, is involved with regulating growth hormone release, hence the mutation is correlated with obesity [7]. Globally, this mutation occurs in around 7.8% of the population, however this mutation has higher allele frequencies in the Asian population (at $18-20\%$) [4].

The third interesting variant I found is a non-synonymous variant on Chro-

mosome 5 (rs820878) at position 74685445. The mutation was in the HEXB gene, which changed the amino acid Leucine to Serine. The implication of this change can potentially be infantile type Sandhoff disease, a rare, inherited disease that destroys nerve cells in the brain and spinal cord [6]. Globally, this mutation occurs in around 3.8% of the population, however this mutation has higher allele frequencies in the European and Latin American populations (at $4-5\%$) [5].

4 Discussion

The most interesting variant I observed was within the HEXB gene on Chromosome 5, which has the potential to lead to infantile-type Sandhoff disease. The HEXB gene encodes the beta subunit of the enzyme hexosaminidase, which is involved in the breakdown of fats, particularly gangliosides. In the mutation, a Leucine residue is replaced with Serine. This amino acid change disrupts the structure and function of the hexosaminidase enzyme, impairing its ability to break down gangliosides (fats) in the nervous system, and leading to an accumulation of these fats in the brain and other tissues. This accumulation leads to progressive damage to nerve cells in the brain and spinal cord. Weakness typically will begin in the first 6 months of life. Infants typically have early blindness, progressive mental and motor deterioration, seizures, etc. Most children with the disease die within 3 years [18].

Figure 8: Human Beta-hexosaminidase protein B [11]

There is existing research on mutations in HEXB and their connections to Sandhoff disease. For example, Zhang et. al. describe this specific mutation in detail in their 1995 paper studying a patient with infantile Sandhoff disease [19]. The most common mutation found with this disease is a 16kb deletion in the HEXB gene. In this case, a different large deletion involving the HEXB

gene was discovered and the patient was also found to be homozygous for a $C \rightarrow T$ mutation in exon 1. In addition, the authors suggest that the father of the patient has two distinct haplotypes, A and B, and the mother has a third haplotype, C, and a null allele. They used several experimental techniques involving Southern blot and pulsed-field electrophoresis to understand the genetic deletions and SNP [19]. There are still limitations of this study, especially since they only analyzed this mutation in one patient. They also observed a deletion in addition to the SNP, so it is difficult to isolate which mutation is most correlated with the disease. The researchers even concede that there could have been a much larger deletion in the genome, but due to experimental limitations, they were unable to observe it.

Although the effect of this specific mutation is not conclusively understood, mutations in the HEXB gene are correlated to Sandhoff disease. HEXB gene codes for the beta subunit of the beta-hexosaminidase protein. The protein functions as a glycosyl hydrolase, breaking down fats in the lysosome. The mutation above results in the replacement of a nonpolar Leucine residue with a polar Serine residue. The increased hydrophilicity of the residue may alter protein folding in order to place the Serine towards the surface of the protein. Additionally, the Leucine residue may have participated in hydrophobic interactions with other distant residues, stabilizing the protein structure. Serine, on the other hand, has a hydroxyl functional group, and can therefore form hydrogen bonds, stabilizing the tertiary structure in different ways than before the mutation.

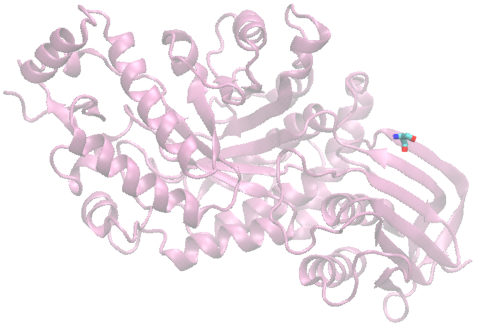


Figure 9: Chain A of Protein with Highlighted Mutated Residue [11]

In Figure 9, we can clearly see the highlighted residue on Chain A of the beta-hexosaminidase B protein, which undergoes a mutation. The mutation occurs in the beta sheets domain of the protein, which is important for stabilizing the structure of the protein through secondary structure. The mutation may affect the stability of the beta sheets domain, and alter the overall structure and function of the protein. The exact role of the mutation in infantile Sandhoff disease, however, is not fully understood.

4.1 Limitations

There are several potential sources of error that may have impacted the accuracy of the variant analysis, including imperfect sample quality, sequencing errors, and imperfect variant annotation. If the sample was contaminated or degraded, it could lead to false positive or negative results. Sequencing errors

could lead to an incorrect mapping of reads in the alignment step. Inaccurate or incomplete annotations can result in the misclassification of variants or incorrect interpretation of their phenotypic implications.

In addition, this work focused on utilizing exome sequencing and analysis, but this method has several pitfalls. Coverage bias, for example, is a large problem. Exome sequencing only captures a subset of the genome, and therefore there may be regions of the genome that are not well-covered. In addition, there may be regions that are difficult to sequence due to high GC content, leading to uneven coverage throughout the exome. Exome sequencing can also heighten the potential for errors, especially introduced during sample preparation. These variants can lead to false positive results, where a variant is reported as pathogenic, but is actually benign. Burdick et al. elucidated the limitations of exome sequencing (ES) in their study analyzing its ability to detect rare and undiagnosed diseases. They found that 33% of diagnoses were not solved exclusively by ES, and several other methods were needed to detect the variants that were missed [9].

Personal genomics, at large, is accompanied by numerous caveats, despite the fact that it has led to significant advancements in precision medicine. Personal genomics relies on DNA sequencing technology, which is not 100% accurate. Errors can occur during sequencing, especially if the reads have low quality and/or low coverage, which can lead to false positives or negatives during analysis. While the cost of DNA sequencing has decreased rapidly over the past two decades, personal genomics still can be expensive and not everyone may have access to these services. Lastly, the results of personal genomics are complex and difficult to interpret for anyone without a genetics background. The same genetic variant can also have varying effects on different people, making it difficult to determine its significance.

Still, knowing these exonic variants in a clinical setting could have many benefits. Identifying these genetic variants could aid in diagnosing the conditions and determining treatment options. Many of the variants were associated with an increased risk of certain diseases. Knowing these mutations early can allow for early screening and monitoring, potentially leading to early detection. Additionally, identifying variants that are associated with hereditary diseases can help individuals make informed decisions about family planning and reproductive options.

5 Conclusions

This work aimed to analyze a human genome using a standard bioinformatic pipeline, specifically alignment to the reference genome, genotyping, and annotation. From this pipeline, we were able to discern certain mutations, either SNPs or indels, that were present in the DNA sequence. First, we analyzed some basic trends within the mutations, such as the distribution of mutations throughout the chromosomes or the correlation of the number of mutations with chromosome size. From there, we looked for the most important muta-

tions, those that were disease causing or pathogenic. We picked one pathogenic mutation to narrow in on and research further, specifically a mutation in the HEXB gene that can lead to Sandhoff disease. We found that the mutation causes an amino acid change from Leucine to Serine in Chain A of the beta-hexosaminidase B protein, which alters its structure and function. Knowledge of these genetic mutations can help immensely in a clinical setting for early diagnosis and personalized treatment.

References

[1] Homo sapiens (ID 51) - Genome - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[2] Hybridization Capture vs. PCR Amplification – The Two Enrichment Strategies in NGS - CD Genomics, Mar. 2023. [Online; accessed 4. Mar. 2023].

[3] rs35948326 RefSNP Report - dbSNP - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[4] rs696217 RefSNP Report - dbSNP - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[5] rs820878 RefSNP Report - dbSNP - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[6] VCV000003882.8 - ClinVar - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[7] VCV000005062.6 - ClinVar - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[8] VCV000012846.24 - ClinVar - NCBI, Mar. 2023. [Online; accessed 4. Mar. 2023].

[9] K. J. Burdick, J. D. Cogan, L. C. Rives, A. K. Robertson, M. E. Koziura, E. Brokamp, L. Duncan, V. Hannig, J. Pfotenhauer, R. Vanzo, M. S. Paul, A. Bican, T. Morgan, J. Duis, J. H. Newman, R. Hamid, I. John A. Phillips, and U. D. Network. Limitations of exome sequencing in detecting rare and undiagnosed diseases. *Am. J. Med. Genet. A*, 182(6):1400, June 2020.

[10] M. Gaudin and C. Desnues. Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. *Front. Microbiol.*, 9, Nov. 2018.

[11] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, Feb. 1996.

[12] A. C. Jelin and N. Vora. Whole Exome Sequencing: Applications in Prenatal Genetics. *Obstet. Gynecol. Clin. North Am.*, 45(1):69–81, Mar. 2018.

[13] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[14] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009.

[15] J. Pevsner. *Chapter 9, 20, 21.* John Wiley & Sons, 2015.

[16] M. K. Sakharkar, V. T. K. Chow, and P. Kangueane. Distributions of exons and introns in the human genome. *In Silico Biol.*, 4(4):387–93, Jan. 2004.

[17] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164, Sept. 2010.

[18] C. Xiao, C. Tifft, and C. Toro. Sandhoff Disease. In *GeneReviews® [Internet]*. University of, Seattle, WA, USA, Apr. 2022.

[19] Z. X. Zhang, N. Wakamatsu, B. R. Akerman, E. H. Mules, G. H. Thomas, and R. A. Gravel. A second, large deletion in the HEXB gene in a patient with infantile Sandhoff disease. *Hum. Mol. Genet.*, 4(4):777–780, Apr. 1995.