Aligning, Genotyping, and Annotating Sequence 13 Leads to Identification of Nonsynonymous Variant in Significant Gene to Parkinson's Disease

Alice Saparov
BIOS 26120

Introduction

Personal genomics analyzes the information in a person's genome. Information is used from across the whole genome, and because of this the information can have both clinical or personal value, or be extraneous [1]. Personal genomics may be the future of health care; as the amount of information available about individual genes and disease risk increases, it is possible to make more informed decisions about preventing and delaying the onset of certain diseases. This can include knowing the information that a set of parents are carriers of a detrimental disease, and investigating other options for child bearing to prevent passing on the heritable disease [2].

Personal genomics can be performed using next-generation sequencing (NGS). NGS has the capacity to sequence DNA extremely fast and therefore allows for the rise of personal genomics [3]. NGS determines the order of nucleotides in a DNA sequence using parallel sequencing technology that enables high-throughput, scalable, and rapid sequencing to occur [4]. It does so by mapping many shorter reads of an individual's genome to a known reference sequence, and determining variations that occur. Ultimately, it has allowed genome sequencing to become faster, cheaper, and more efficient, enabling personal genomes to be sequenced. The increased amount of sequencing data available because of this has further led to the identification of variations in the genome and how these variations correlate to disease [3]. Now, through NGS, it is much more cost-friendly for individuals to participate in personal genomics and become educated about their susceptibility for certain diseases or carrier-risk to pass on a disease.

Exome technology is another form of sequencing that allows the protein coding regions (exomes) of DNA to be sequenced. First, fragmented Exome DNA samples are selectively hybridized by biotinylated oligonucleotide probes (capture). Next, the non-targeted regions are washed away and PCR is used to amplify the targeted sample which is then sequenced [5]. Nowadays, exome sequencing is increasingly used to understand various symptoms and diseases in healthcare [6].

In this project I will be aligning, genotyping, and annotating a sequence. This includes aligning raw NGS reads, obtained from an Illumina HiSeq 2000 instrument, of exome-captured DNA to the reference human genome. Next, the aligned sequence is genotyped by variant calling which finds locations in the aligned sequence that differ from the reference sequence. Finally, these results are annotated utilizing information from various databases. This allows me to identify genomic context, coding mutation type, and obtain relevant IDs, functions, or pathologies of the variants. With this information, the sequence can be analyzed and meaningful variants can be identified.

Methods

First, raw NGS reads of exome-captured DNA were aligned to the reference human genome. The raw reads were obtained from an Illumina HiSeq 2000 instrument. I performed the alignment using the Burrows Wheeler alignment (BWA) algorithm by submitting my paired end

FASTQ files to Midway (UChicago Supercomputer). This algorithm then aligned and merged my sequences, converted them to .bam format, and sorted the results. Next, I genotyped the sorted .bam file by variant calling which allowed me to identify single nucleotide polymorphisms (SNPs) and insertion-deletion events (indels). This enables me to find locations in my sequence where it differs from the reference sequence. These results are ultimately filtered by various quality control parameters to make sure variants originate from well-aligned regions (not just sequencing errors). This genotyping was performed on all chromosomes within my sequence, by submitting another job to Midway. First, a raw set of genotype calls is generated using the *samtools mpileup* program, directing this output into a variant call format (.vcf) file. Next, the genotypes are filtered according to read depth, and then by their phred score (greater than 50). Finally, using the ANNOVAR tool I annotated my results utilizing information from multiple databases (refGene, avsnp150, clinvar_20170905, and dbnsfp33a). This allowed me to identify genomic context (exonic, intronic, intergenic, etc.), coding mutation type (synonymous, non-synonymous, frame-shift etc.), RefGene ID of the affected gene, previous variant identification (dbSNP rs ID), variant pathology from genome-wide association studies, and metrics that describe how deleterious/functional the variant may be. I also filtered by exonic regions since we are more interested in the coding regions, and with this information, the interpretation and analysis of results can begin.

Results

The sequence alignment resulted in 184,854 variants identified, 74,486 of which were high quality. Of the variants, 14,543 were exonic. 69,621 SNPs were found and 4,920 indels. There were 7,325 synonymous variants identified and 6,214 nonsynonymous. There were 32 frameshift mutations, with 19 deletions and 13 insertions. Lastly, there were 46 premature stop variants. In *Table 1* there are thirty variants described, all of which had a Phred score greater than 50. Some of these more interesting variants included a gene change from T to C (Rs12021720 [10]) on chromosome 1, position 100206504. This was a nonsynonymous SNV in the DBT gene that results in intermediate maple syrup urine disease type 2. This disease clinically presents with mental and physical retardation, feeding problems, and a maple syrup urine odor [7]. Next, on chromosome 5 position 74685445, there was a nonsynonymous SNV in the gene HEXB that contained a change from T to C (Rs820878 [11]). The implications of this substitution is infantile Sandhoff disease. This is a rare inherited lipid storage disorder that destroys nerve cells progressively in the brain and spinal cord and this causes seizures, vision and hearing loss, intellectual disability, and paralysis [8]. Also, on chromosome 12 in position 40225280, there was a nonsynonymous SNV in the LRRK2 gene with a change from G to A (Rs2256408 [12]). This results in autosomal dominant Parkinson's disease which is a progressive nervous system disorder. Clinically, the symptoms include tremors, slow movement or inability to move, and imparied balance and coordination. It can also affect emotions and cognition [9].

Investigating the data further, it appears the number of variants generally decreases with each chromosome. Chromosome 1 contains the most variants, and the Y chromosome contains

the fewest number of variants (*Fig. 1*). Also, most mutations are intronic and intergenic (*Fig. 2*). Now looking deeper into the types of variants, the majority of the variants are synonymous (over 7,000) and nonsynonymous (over 6,000) (*Fig. 3*). Next, excluding these variants, many of the other variants are classified as unknown and the next highest type is a premature stop codon mutation (*Fig. 4*). Comparing chromosome sizes to the number of mutations on each, there is a positive correlation. Thus, as the size of the chromosome increases, so does the number of mutations. There is some variation with chromosomes of about 60 to 150 million base pairs as the number of mutations appear to be more scattered (*Fig. 5*). However, when comparing the number of mutations on each chromosome relative to the number of genes encoded, there is a stronger positive linear correlation. Hence, as the number of genes per chromosome increases, so does the number of mutations per chromosome (*Fig. 6*).

Chr	Position	Type of Variant	Implications
	100206504	nonsynonymous	Intermediate maple syrup urine disease type 2
	158654738	nonsynonymous	Spherocytosis type 3 autosomal recessive
	161629903	nonsynonymous	Neutrophil-specific antigens NA1/NA2
	171114092	stopgain	Trimethylaminuria
Chr1	196690107	nonsynonymous	Age-related macular degeneration 4; Basal laminar drusen; Mesangiocapillary glomerulonephritis, type II; Atypical hemolytic uremic syndrome
Chr2	214986579	synonymous	Congenital ichthyosiform erythroderma
	121788384	nonsynonymous	Nephronophthisis - Renal dysplasia and retinal aplasia
Chr3	165046931	synonymous	Sucrase-isomaltase deficiency
Chr4	186236880	nonsynonymous	Prekallikrein deficiency
	35860966	nonsynonymous	Severe combine immunodeficiency, autosomal recessive, T-cell-negative, B-cell-positive, NK-cell positive; Severe Combined Immune Deficiency
	35871088	nonsynonymous	Severe combine immunodeficiency, autosomal recessive, T-cell-negative, B-cell-positive, NK-cell positive; Severe Combined Immune Deficiency
Chr5	74685445	nonsynonymous	Sandhoff disease, infantile type
Chr6	26090951	nonsynonymous	Hemochromatosis type 1; Microvascular complications of diabetes 7
Chr7	150999023	nonsynonymous	Coronary artery spasm 1; Alzheimer disease late-onset; Hypertension pregnancy-induced; Ischemic heart disease; Ischemic stroke
	6532368	synonymous	Not specified
Chr8	86667075	nonsynonymous	Achromatopsia; Stargardt Disease Recessive
	23193706	nonsynonymous	Permanent neonatal diabetes mellitus (PNDM)
	54195850	nonsynonymous	Nonsyndromic Hearing Loss Recessive; Retinitis pigmentosa-deafness syndrome
Chr10	68885620	nonsynonymous	Preeclampsia, eclampsia 4
	18269312	nonsynonymous	Serum amyloid A variant
Chr11	66560624	stopgain	Actinin alpha-3 polymorphism (ACTN3 deficiency), Sprinting performance
	40225280	nonsynonymous	Parkinson disease 8, autosomal dominant
	120999579	nonsynonymous	Maturity-onset diabetes of the young, type 3
Chr12	121857429	nonsynonymous	4-Alpha-hydroxyphenylpyruvate hydroxylase deficiency
	56514589	nonsynonymous	Bardet-biedl syndrome 2/6, digenic
Chr16	69711242	nonsynonymous	Benzene toxicity; Leukemia post-chemotherapy; Breast cancer, post-chemotherapy poor survival; Lung cancer
	13011692	nonsynonymous	Prostate cancer, hereditary, 2
Chr17	64496464	nonsynonymous	Progressive External Ophthalmoplegia with Mitochondrial DNA Deletions
Chr18	2705702	synonymous	Not specified
Chr21	45537880	nonsynonymous	Gastrointestinal stromal tumor

Table 1. Includes thirty high quality exonic variants with a Phred score greater than 50.

Barplots

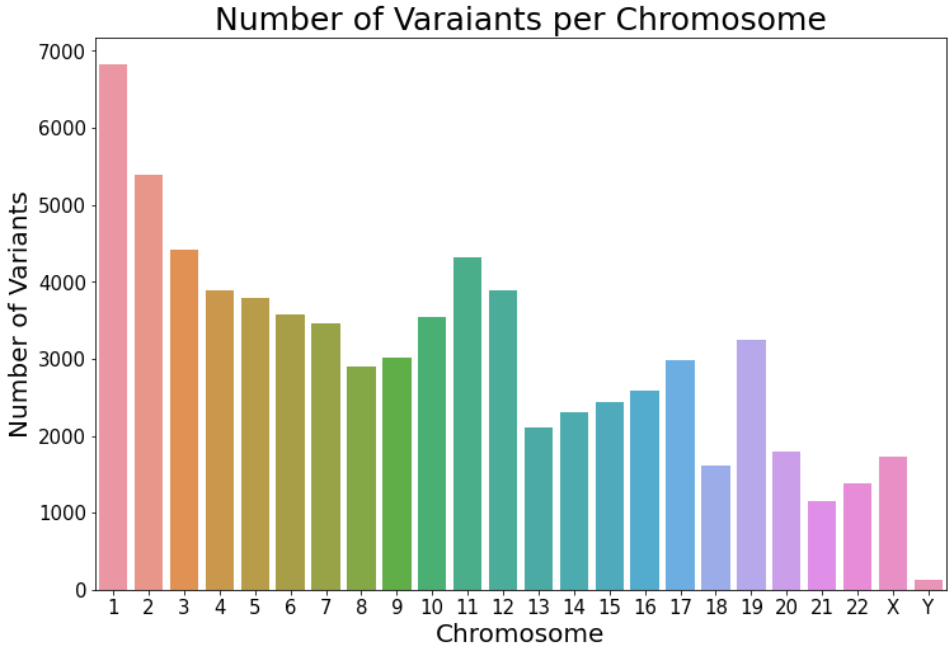

Figure 1. A barplot of the number of variants per chromosome.

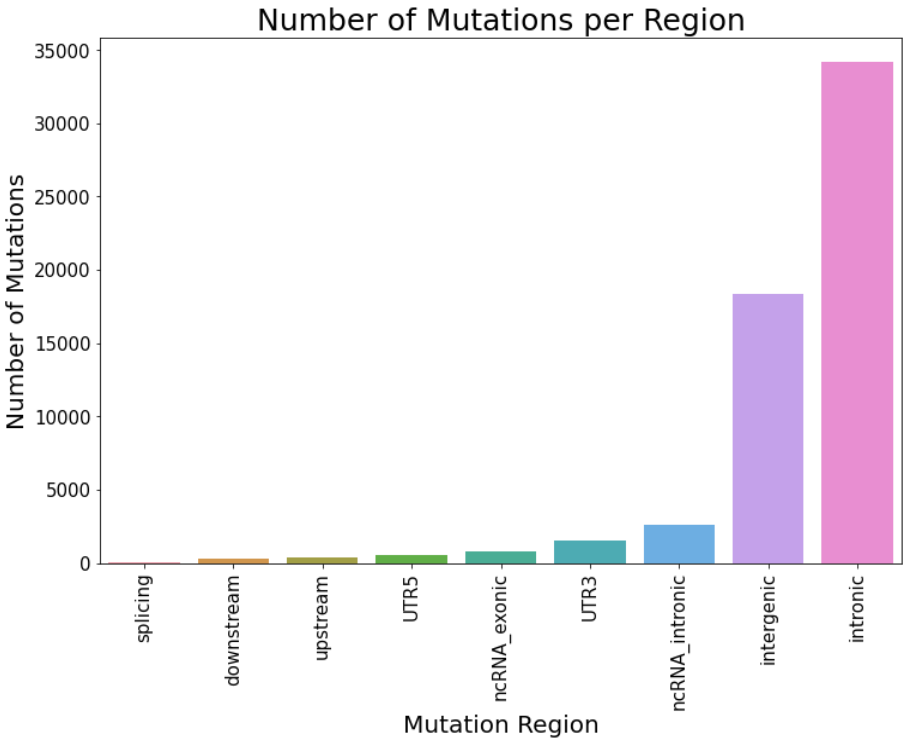

Figure 2. The figure shows a barplot of the number of mutations in each mutation region.

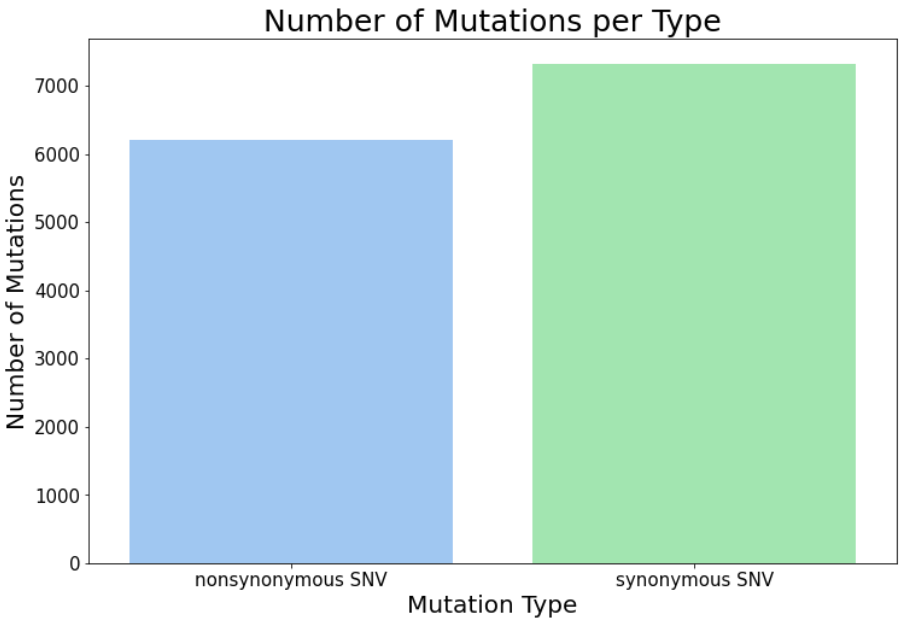

Figure 3. The barchart above contains the totals of the top two mutation types (nonsynonymous and synonymous).

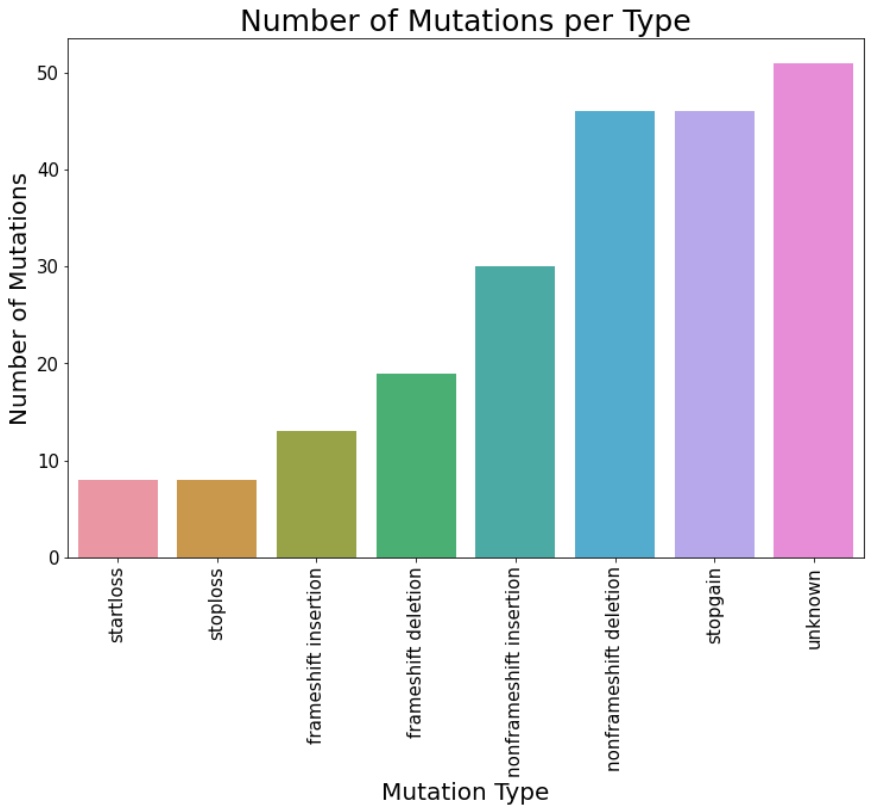

Figure 4. The barplot above shows exonic mutations only, and contains the number of mutations for each mutation type.

Scatterplots

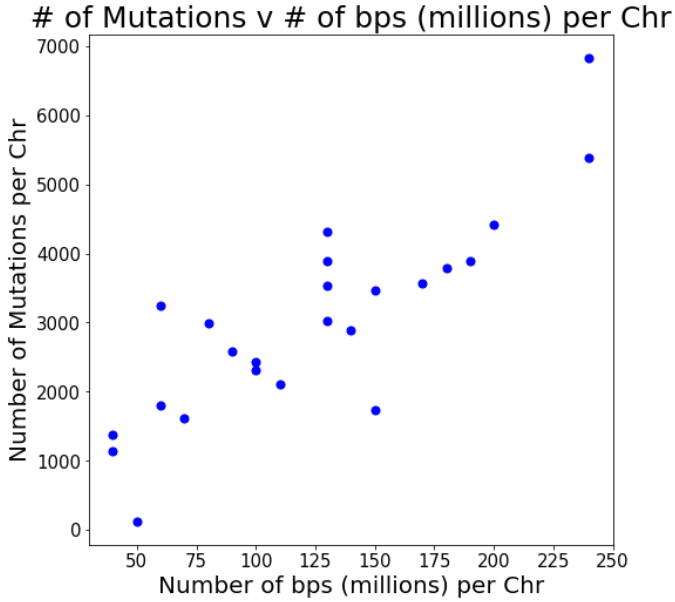

Figure 5. Scatterplot of the number of mutations per chromosome vs the number of base pairs (bps) in millions per chromosome.

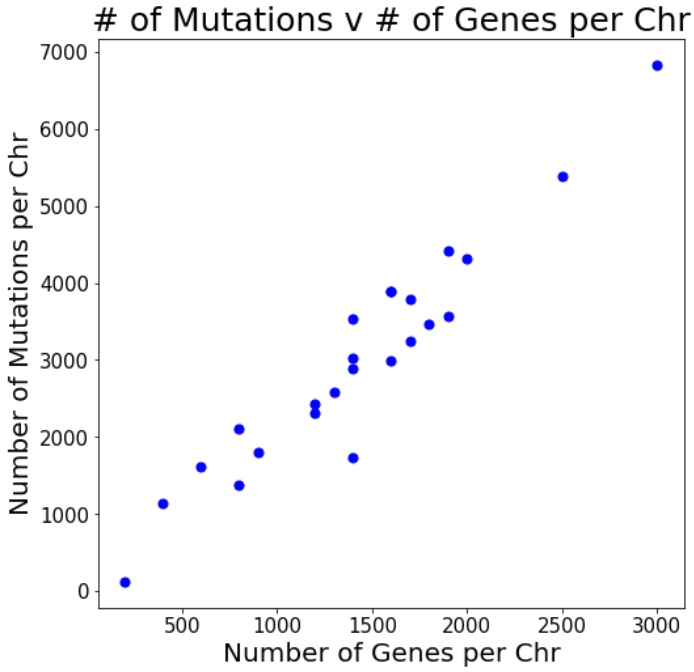

Figure 6. Scatterplot of the number of mutations per chromosome vs the number of genes per chromosome.

Discussion

Of the variants identified, infantile Sandhoff disease and Parkinson's disease were the most interesting. Sandhoff disease results from a mutation in the HEXB gene. This gene encodes the beta subunit of hexosaminidase which participates in the breakdown of gangliosides [14]. The mutation causes a deficiency in the degradation of gangliosides, and this results in the accumulation of lipids in the brain and spinal cord. This ultimately gives rise to a multitude of neurological issues such as seizures and intellectual disabilities [8]. Parkinson's diseases can occur from a mutation in the LRRK2 gene which codes for dardarian. The protein has multiple enzymatic domains, and participates in a wide range of cellular functions and signalling pathways. This includes mitochondrial function, vesicle trafficking and endocytosis, retromer complex modulation and autophagy. Its mechanistic role in Parkinson's disease is still being researched, and many of its physiological and neurotoxic properties are yet to be completely understood [15]. While the mechanisms have yet to be fully determined, multiple studies have found that LRRK2 may have great potential for targeted gene therapy for Parkinson's. Pathogenic mutations in the LRRK2 gene have been found to increase LRRK2 kinase activity, and small-molecule LRRK2 kinase inhibitors can be neuroprotective in preclinical models of Parkinson's [13]. Furthermore, it was found that LRRK2 interacted with the microRNA pathway to regulate protein synthesis. Ultimately, the study suggested that novel miRNA-based therapeutic strategies have potential for targeting Parkinson's disease [16]. A structure for part of the protein domain is shown (*Fig. 7*).

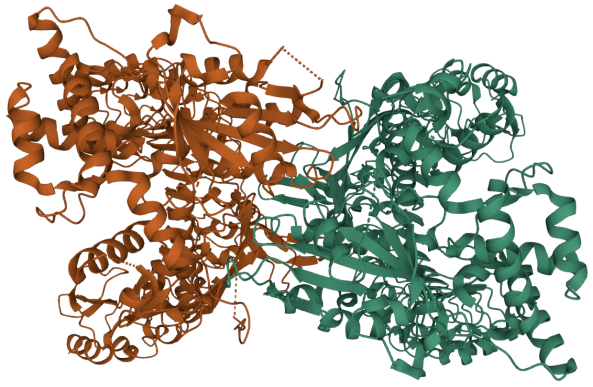

Figure 7. The LRR-Roc-COR domain of the LRRK2 gene which corresponds to Parkinson's disease.

As previously stated, research on Parkinson's disease and LRRK2 is still ongoing. Many papers describe that while significant work has been published to describe the structure, function, and biochemical properties encoded by the gene, there are seemingly now even more unanswered questions. This is in part due to the sheer size of the LRRK2 gene which encodes five functional domains and consists of approximately 2,500 amino acids [16].

Overall, more improvements can be made in exome sequencing and personal genomics. Exome sequencing is a crucial tool in identifying variants within an individual's genome. However, the analytical and validation process is complex and the results require thorough application and interpretation [17]. Furthermore, NGS and exome sequencing have allowed personal genomics to become more wide-spread and accessible, however there is not one standardized way to perform the alignments introducing variability to the results. The scientific community should proceed with caution though as there is much policy to be created regarding the ethics and privacy of sequencing. Also, for personal genomics to become available to the public, there would have to be a standard for sequencing machines since the accuracy of genotyping is very important. Lastly, while personal sequencing can provide a significant amount of information that has the potential to greatly benefit an individual, it should all be taken and interpreted in the context of environmental and life-history [18]. Ultimately, identifying these variants in a clinical setting can greatly benefit individuals in attempting to provide clarity or further support to current diagnoses made. As well as potentially illuminating future risk for certain diseases, allowing individuals to have advanced knowledge and the chance to mitigate their risk (if at all possible).

Conclusions

Personal genomics analyzes the information in a person’s genome. It increases the amount of information available about individual genes and disease risk, making it possible to make more informed decisions about preventing and delaying the onset of certain diseases. It can be carried out using NGS or exome sequencing techniques to determine variations that occur. Ultimately, it has allowed genome sequencing to become faster, cheaper, and more efficient, enabling personal genomes to be sequenced. In this project aligned, genotyped, and annotated a sequence. This includes aligning raw NGS reads of exome-captured DNA to the reference human genome, and then variant calling to determine where difference between the sequences occurred. Lastly, I annotated the results using information from various databases. I was able to filter out high quality exonic variants and determine that a majority of the variants were synonymous and nonsynonymous. I chose to focus on three different nonsynonymous variants that resulted in maple syrup urine disease, Sandhoff disease, and Parkinson’s disease. Ultimately, there is a lot of potential in sequencing and personal genomics to positively contribute to diagnostics and disease risk mitigation within a clinical setting. However, more research must be performed and more policies written for ethical, legal, and social grounding before personal genomics becomes commercially available.

References

1. Mountain JL. Chapter 6 - Personal Genomics. In: Ginsburg GS, Willard HF, eds. *Genomic and Personalized Medicine (Second Edition)*. Academic Press; 2013:74-86. doi:10.1016/B978-0-12-382227-7.00006-9
2. Personal genomics: the future of healthcare? yourgenome. Accessed December 7, 2021. https://www.yourgenome.org/stories/personal-genomics-the-future-of-healthcare
3. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics*. 2011;38(3):95-109. doi:10.1016/j.jgg.2011.02.003
4. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*. 2013;98(6):236-238. doi:10.1136/archdischild-2013-304340
5. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda)*. 2015;5(8):1543-1550. doi:10.1534/g3.115.018564
6. Whole Exome Sequencing. Yale Medicine. Accessed December 8, 2021. https://www.yalemedicine.org/conditions/exome-sequencing
7. Flaschker N, Feyen O, Fend S, Simon E, Schadewaldt P, Wendel U. Description of the mutations in 15 subjects with variant forms of maple syrup urine disease. *J Inherit Metab Dis*. 2007;30(6):903-909. doi:10.1007/s10545-007-0579-x
8. Sandhoff Disease Information Page | National Institute of Neurological Disorders and Stroke. Accessed December 9, 2021. https://www.ninds.nih.gov/Disorders/All-Disorders/Sandhoff-Disease-Information-Page
9. Rui Q, Ni H, Li D, Gao R, Chen G. The Role of LRRK2 in Neurodegeneration of Parkinson Disease. *Curr Neuropharmacol*. 2018;16(9):1348-1357. doi:10.2174/1570159X16666180222165418
10. rs12021720 RefSNP Report - dbSNP - NCBI. Accessed December 9, 2021. https://www.ncbi.nlm.nih.gov/snp/rs12021720#frequency_tab
11. rs820878 RefSNP Report - dbSNP - NCBI. Accessed December 9, 2021. https://www.ncbi.nlm.nih.gov/snp/rs820878#frequency_tab
12. rs2256408 RefSNP Report - dbSNP - NCBI. Accessed December 9, 2021. https://www.ncbi.nlm.nih.gov/snp/rs2256408#frequency_tab
13. Jankovic J, Tan EK. Parkinson's disease: etiopathogenesis and treatment. *J Neurol Neurosurg Psychiatry*. 2020;91(8):795-808. doi:10.1136/jnnp-2019-322338
14. Bikker H, van den Berg FM, Wolterman RA, Kleijer WJ, de Vijlder JJM, Bolhuis PA. Distribution and characterization of a Sandhoff disease-associated 50-kb deletion in the gene encoding the human β-hexosaminidase β-chain. *Hum Genet*. 1990;85(3):327-329. doi:10.1007/BF00206756
15. Wallings R, Manzoni C, Bandopadhyay R. Cellular processes associated with LRRK2 function and dysfunction. *The Febs Journal*. 2015;282(15):2806. doi:10.1111/febs.13305
16. Giasson BI, Covy JP, Bonini NM, et al. Biochemical and pathological characterization of Lrrk2. *Ann Neurol*. 2006;59(2):315-322. doi:10.1002/ana.20791
17. Gerhard GS, Bann DV, Broach J, Goldenberg D. Pitfalls of exome sequencing: a case study of the attribution of HABP2 rs7080536 in familial non-medullary thyroid cancer. *npj Genomic Med*. 2017;2(1):1-7. doi:10.1038/s41525-017-0011-x
18. Bolouri H. Computational Challenges of Personal Genomics. *Curr Genomics*. 2008;9(2):80-87. doi:10.2174/138920208784139564