Insights into Disease-Causing Variants: A Genome-Wide Exploration through Personal Genomics, Next-Generation Sequencing, and Exome Analysis

Bernard Shen
University of Chicago

Introduction

Imagine having a book that holds the most intimate and vital information about yourself—the story of your genetic makeup, ancestry, and potential health pathways. This is the essence of personal genomics, a field delving into the comprehensive examination and interpretation of an individual's genetic blueprint, extending beyond conventional clinical genetic testing. It involves understanding the contents of that precious book without the need for a comprehensive knowledge of family history, or before official clinical diagnoses or tests [11]. The advent of Next-Generation Sequencing (NGS) technologies represents a monumental leap forward in analyzing genomes and uncovering an individual's genetic blueprint. These cutting-edge sequencing methodologies, such as Roche 454, Illumina, and SOLiD, represent a massive step forward in sequencing capabilities. NGS operates on a massively parallelized platform, generating millions to billions of sequencing reads, each comprising hundreds to thousands of base pairs. The resulting efficiency gain from parallelization (as compared to older techniques, like Sanger sequencing) reduces the time required for a run to a timeframe of hours or days; hence the cost of a whole-genome sequence has reduced by several orders of magnitude (from millions to thousands of dollars). The relevance of such cost reductions are the expanded applications in the everyday world, where this technology is now accessible to more individuals than ever before [12].

Among the specialized techniques within genomic analysis is exome sequencing. This method concentrates on the protein-coding regions (exons) of the genome, employing a selective targeting and capturing (or enriching) process for sequencing; the capture process involves biotinylated oligonucleotides that are complementary to exonic regions, which selectively bind to those regions, capturing only the exons as opposed to the entire sequence. Despite examining only exons, which comprise just 1-2% of the genome, this focused approach yields a high success rate in identifying disease-causing mutations. Historically, exome sequencing has been preferred over whole-genome sequencing due to its cost-effectiveness, but recent advancements in technology have blurred the cost disparity between the two approaches [12].

In this project, the primary objective is to dissect two sequence reads—forward and reverse—representing an individual's entire genome. The aim is to convert raw sequence data of exome-captured DNA into a comprehensive understanding of genetic variations, particularly disease-causing genetic variants present within the sequence. The analytical steps of this project include alignment, genotyping, and annotation. Alignment serves as the initial step, involving the mapping of individual sequence reads onto the reference human genome. This process unveils the origin within the

genome for each sequencing read and highlights variations from the reference genome. To accomplish this, the Burrows Wheeler alignment (BWA) algorithm is employed for its consistency and speed in aligning short sequence reads to the vast human genome [4]. Subsequently, genotyping or variant calling comes into play, aiming to identify single nucleotide polymorphisms (SNPs) and insertion-deletion events (indels) by comparing an individual's genomic sequence reads against the reference human genome. This step leverages repeated reads to determine the individual's genotype at specific positions. I use this process to discern variations between the individual's genomic sequence and the reference, identifying potential genetic differences. Finally, the automated annotation process enriches the genetic data by adding additional information about each variant. This step aids in categorizing different types of genetic variants and discerning their potential implications in diseases. Employing ANNOVAR enables the layering of existing biological information onto the variants, including genomic context (exonic or intronic), mutation types, implications in pathologies from genome-wide association studies, and more [19].

These steps pave the way for a detailed exploration into the genetic landscape, with the aim to unravel disease-causing variants and their potential implications in the context of the individual's genomic profile. This endeavor seeks to bridge the gap between raw genetic data and real biological insights, providing an understanding of the genetic makeup's relevance to potential health outcomes.

Methods

To streamline the process, all steps involved in converting sequence reads into an annotated list of mutations were conducted on the Midway3 supercomputer. Each phase entailed the submission of a shell script file specifying job names, requested time (twenty-four hours per task), and the relevant file and script paths. Beginning with alignment, the BWA program mapped short sequencing reads to a reference genome, optimizing matches for the best alignment [4]. The script included paths to two "fastq" files, representing the sequences, and the BWA alignment program. Following alignment, genotyping involved comparing the individual's genomic sequence reads to the reference using the "mpileup" command from the SAMtools package [5]. Script parameters included addresses for the sorted bam file from alignment and the genotyping command.

Before the final computational task, annotation, I segregated high-quality variants from the variant call format file, based on Phred quality scores exceeding fifty. I utilized the ANNOVAR tool [19] for the annotation itself, and employed the specific databases refGene, avsnp150, "clinvar," and functional scores from "dbnsfp 42c" for a more detailed annotation. Both the high-quality and standard file were annotated, and the former was later copied to my local machine for additional analysis using Excel. The selection of variants for further investigation relied on the amount of accessible information on those variants, both in the annotated spreadsheet and online. I began by filtering the “Clinical Significance” column to eliminate blank and “benign” entries; and looked for rows whose SIFT, LRT, and Mutation Taster columns contained values. After identifying candidates rows who met these criteria, I researched the corresponding rs IDs on the NCBI database and dbSNP to learn what genetic diseases were caused by the mutation and gauge the general amount of scientific literature written about them. Those with the largest online footprints were chosen for further analysis. The resulting variants of interest were primarily nonsynonymous mutations found within the exonic regions of the genome.

Results

A total of 115,625 variants were identified by the genotyping program. However, adhering to different Phred quality score (Q) thresholds for the variants produced smaller datasets for analysis. A total of 103,728 variants met the criteria at a standard of $Q > 30$.[1] Adhering to the higher threshold of $Q \geq 50$ revealed a subset of 86,021 high-quality variants.[2] Additionally, the count of exonic variants amounted to 16,326; 105,865 SNP genotypes and 10,011 indels/substitutions were identified, and the total count for each mutation type consisted of 7,639 synonymous variants, 6,555 non-synonymous variants, 57 frameshift mutations, and 49 premature stop codons.[3]

Table 1. Some High Quality Variants and Their Implications

Location of Variant	Type of Variant	Implications
Chromosome 1, Position 976215	Nonsynonymous SNV	Renal tubular epithelial cell apoptosis
Chromosome 1, Position 11796321 (624, 1653 37)	Nonsynonymous SNV	Thrombophilia and Hypertension
Chromosome 1, Position 203186754	Unspecified	Risk of asthma
Chromosome 4, Position 38798089	Nonsynonymous SNV	Associated with helicobacter pylori infection and tuberculosis
Chromosome 5, Position 132660272	Nonsynonymous SNV	Allergic rhinitis, susceptibility to asthma
Chromosome 6, Position 154093438	Synonymous SNV	Heroine/opioid addiction
Chromosome 7, Position 101128436	Nonsynonymous SNV	Hereditary angioedema; C1-inhibitor deficiency
Chromosome 10, Position 114045297	Nonsynonymous SNV	Associated with unbalanced cardiac sympathetic modulation, left ventricular hypertrophy, and sudden cardiac death

[1] This is the threshold outlined in Lab 8.
[2] This is the threshold outlined in the final project documentation.
[3] All of these metrics include low-quality variants, as the documentation did not specify whether to answer these questions for all variants or just the high quality ones.

Chromosome 11, Position 75172532	Nonsynonymous SNV	Lower rates of prostate cancer progression
Chromosome 12, Position 47879112 (435, 834, 15)	Start-loss	Affects vitamin D levels; vitamin D dependent rickets Type II
Chromosome 13, Position 109782884	Nonsynonymous SNV	Predisposition to Type 2 Diabetes; increased nonalcoholic fatty liver disease susceptibility
Chromosome 16, Position 27344882	Nonsynonymous SNV	Immunodeficiency syndrome; promotes a predisposition to the development of bullous pemphigoid
Chromosome 16, Position 69711242	Nonsynonymous SNV	Associated with increased cancer risk, susceptibility to pneumonitis and esophagitis
Chromosome 19, Position 43551574	Nonsynonymous SNV	Reduced risk of (thyroid, breast) cancer; associated with rheumatoid arthritis
Chromosome 22, Position 43928847 (576, 1729, 18)	Nonsynonymous SNV	Affects metabolic dysfunction-associated steatotic liver disease (elevated ferritin); predictor for (in combination with other mutations) tardive dyskinesia

All of the mutations featured in the table above have rs IDs and are documented in at least one scientific journal, indicating that they have all been discovered previously. A few mutations were particularly well documented in the National Library of Medicine; of these I chose three to examine further. On chromosome 1 at position 11796321 I identified a non-synonymous SNP in gene MTHFR that changed the amino acid alanine to valine (rs1801133). Similarly, I discovered another non-synonymous SNP at position 43928847 of Chromosome 22; the mutation occurred in gene PNPLA3, changing isoleucine to methionine (rs738409). Moreover, located on chromosome 12 at position 47879112 is a start-loss mutation in gene VDR, where the start codon for methionine is changed to threonine (rs2228570). Each of these mutations, along with their respective genes, has associations with specific diseases and symptoms, as uncovered through research on databases like OMIM, the National Library of Medicine, and Clinvar.

The MTHFR gene provides instructions for producing an enzyme that is involved in folate metabolism. The alanine to valine mutation identified in this gene is associated with a reduction of this enzyme's activity and high levels of homocysteine in the blood, which can increase the risk of vascular disease. Different populations show varying frequencies of this mutation. Homozygous individuals tend to have higher homocysteine levels, which can be normalized with low-dose folic acid supplementation [15]. Common symptoms of this disease include chest pain, leg cramping during physical activity, numbness and weakness. This mutation has higher allele frequencies in Latin American, Asian, and European populations, and occurs at lower frequencies in African and African American populations [17].

The PNPLA3 gene is responsible for coding a protein involved in lipid metabolism, particularly in the liver; this protein breaks down fats, storing or mobilizing them as needed for energy [7]. Clinvar identifies this mutation of isoleucine to methionine as a risk factor for non-alcoholic fatty liver disease. Symptoms of this disease include fatigue, abdominal pain, and weight loss. This mutation has higher allele frequencies in East Asian, Asian, and Latin American populations, and occurs at lower frequencies in African and African American populations [16].

The VTD gene provides instructions for producing the Vitamin D receptor protein, which acts as a transcription factor, binding to specific DNA sequences and regulating the expression of various genes in response to the vitamin's active form. It plays a crucial role in calcium and phosphate absorption, and bone health [9]. Ponasenko et al. identify a relationship between this mutation and coronary artery disease severity, with individuals carrying the A/A-A/G genotypes found to have a significantly decreased serum levels of vitamin D in high-risk patients. This suggests a potential link between this VTD gene variant and the severity of coronary artery disease [13]. Another study done by Zhong et al. found that lower vitamin D levels are linked to an increased risk of cardio cerebrovascular diseases in prediabetic individuals, and that this mutation interacted significantly with vitamin D levels in blood serum [21]. Common symptoms of coronary artery disease are chest pains, shortness of breath, and heart attacks. This mutation has higher allele frequencies in African and African American populations, and occurs at lower frequencies in Latin American, Asian, and East Asian populations [18].

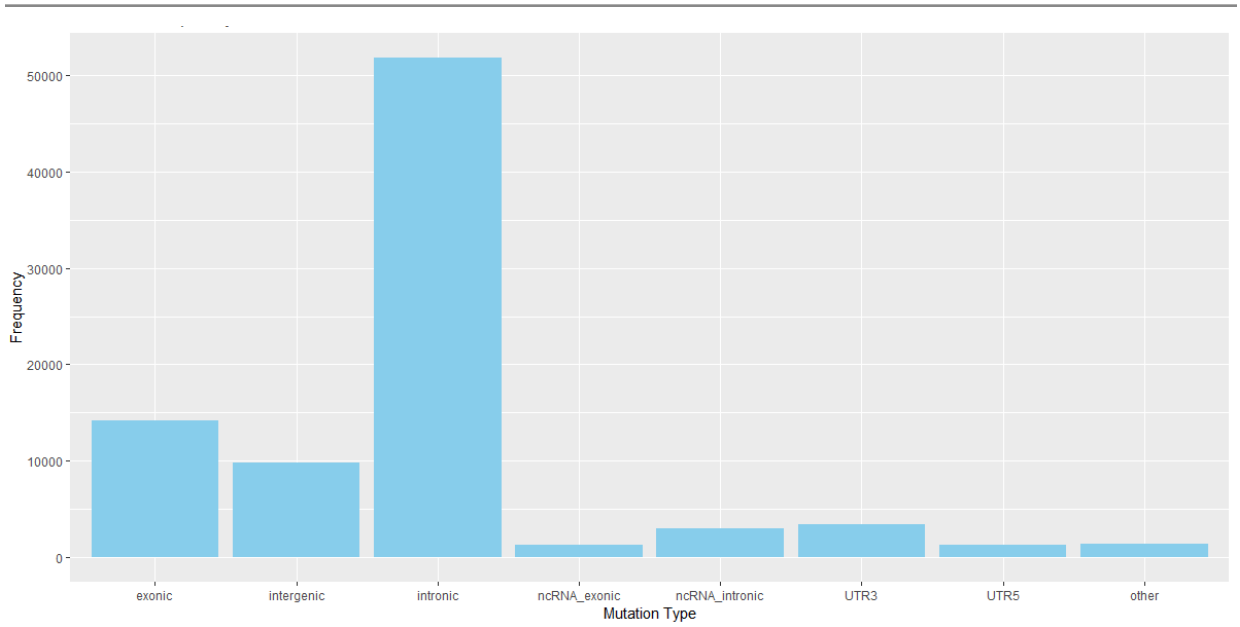

Figure 1. Mutation Frequency [20]

Intronic mutations were much more common than exonic and intergenic mutations, which in turn were more common than mRNA and UTR mutations.

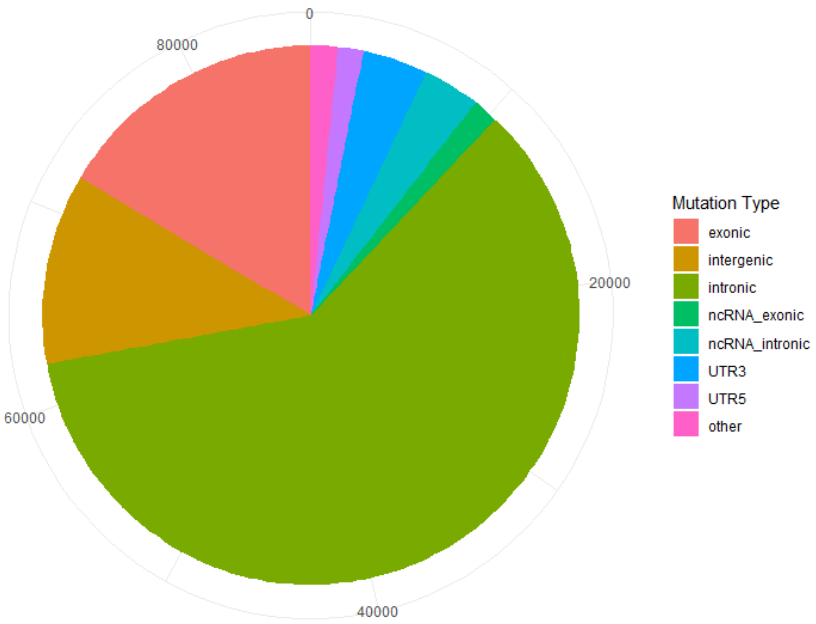

Figure 2. Mutation Frequency [20]

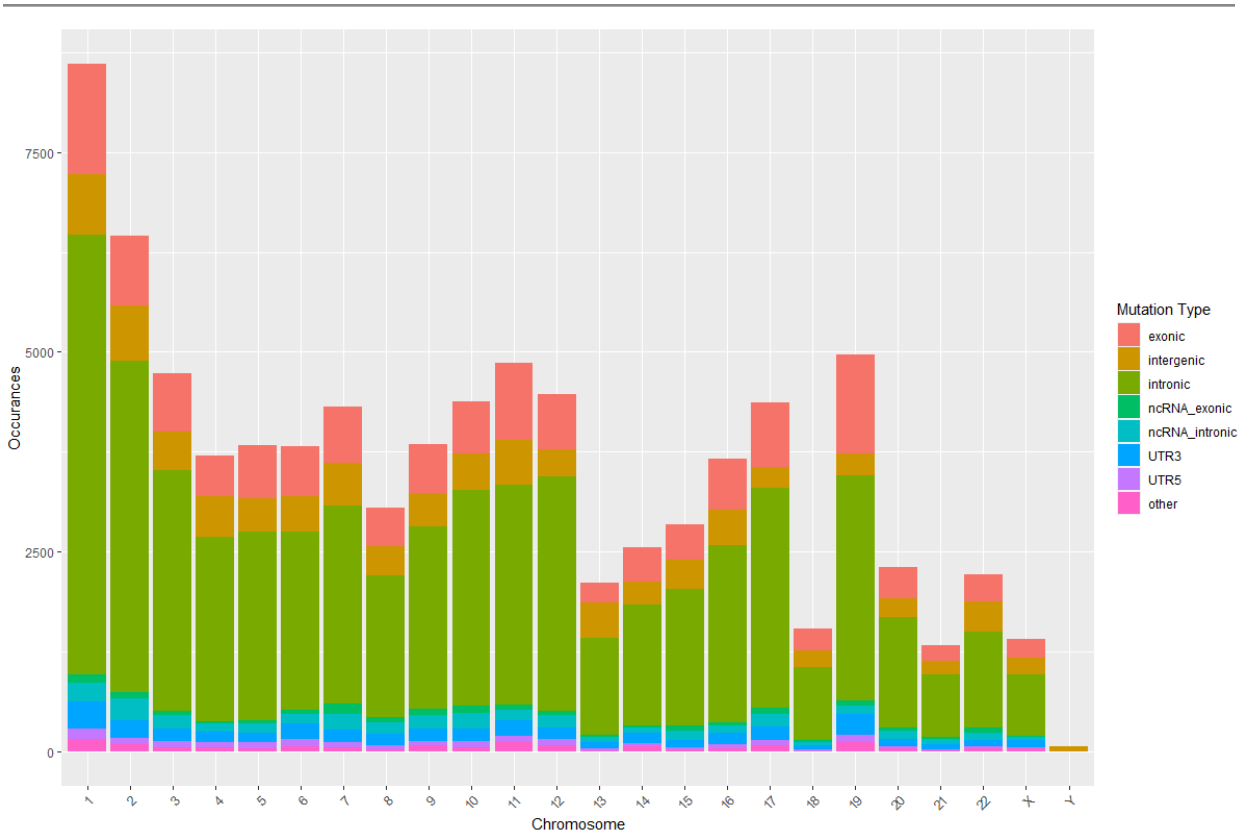

Figure 3. Number of Mutations per Chromosome [20]

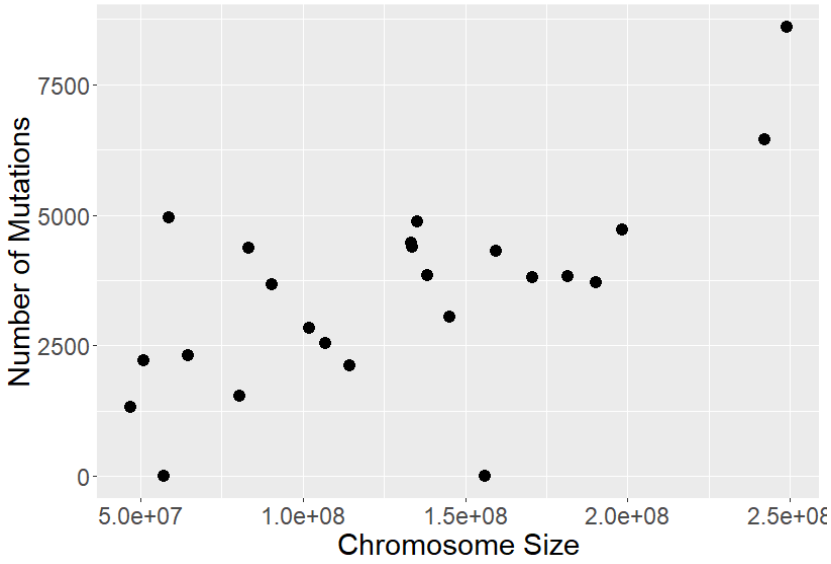

Figure 4. Chromosome Size vs. Number of Mutations [20]

The statistical correlation between chromosome size and the number of mutations in a chromosome is 0.64, suggesting a moderate positive linear relationship between the explanatory and response variable; larger chromosomes will generally have more mutations.

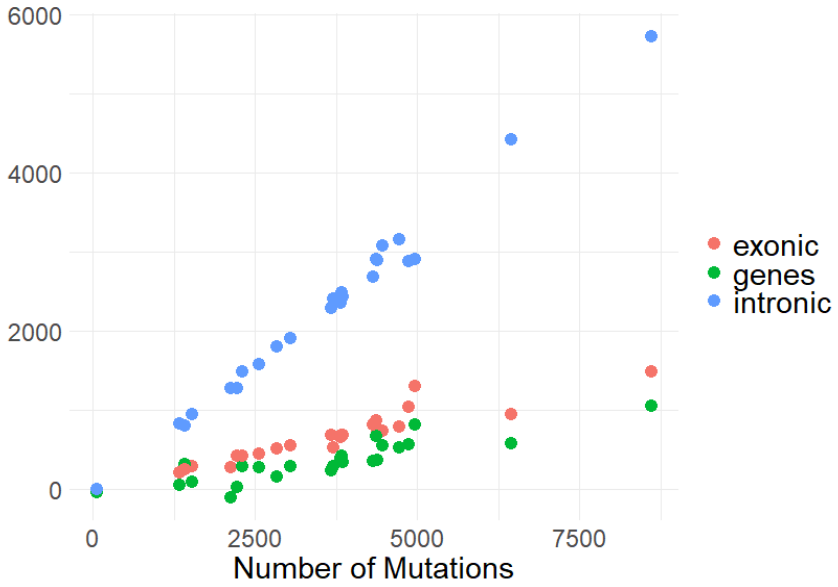

Figure 5. Genes, Exons, and Introns vs. Mutation Count on Each Chromosome [20]

There is a strong linear correlation between the total number of mutations in each chromosome and its associated genes, and exonic and intronic mutations. A simple visual analysis indicates that the rate of increase for exonic variants and genes per thousand mutations is significantly lower than the rate of increase for intronic mutations. Quantifying this relationship with a least-squares regression line, I find the slope associated with the exonic mutations to be 0.1791, and the slope associated with the intronic mutations to be 0.6790. A possible explanation for this difference might be the generally longer intronic regions of chromosomes with more mutations; the chromosomes with more mutations are generally those that are larger and possess more genetic material, which most directly increases the likelihood of mutation occurrence. Moreover, intronic mutations are generally considered to be less damaging than exonic mutations, as while they could impact gene expression, they are not as directly linked to amino acid changes in a protein sequence. Hence, intronic mutations can accumulate, particularly in larger chromosomes where the damage is often unrealized in the form of altered proteins.

Discussion

Of the variants I have discussed so far, the alanine to valine mutation in gene MTHFR is most prominently represented in scientific research. This gene provides instructions for producing an enzyme, methylenetetrahydrofolate reductase, that is involved in folate metabolism, a process by which five 10-methylenetetrahydrofolate are converted to 5-methyltetrahydrofolate. This conversion is essential for the body's processes of converting the amino acid homocysteine to methionine, whose product allows for the production of proteins and other important compounds [7]. Moreover, elevated levels of homocysteine can damage the inner lining of blood vessels, cause a build-up of plaque in the arteries, and increase blood clot formation [6]. The role of mutations in this gene causing vascular disease are tied entirely to the enzymes it codes for; in reducing the activity of the MTHFR enzyme and increasing its thermolability, the body experiences significantly higher levels of plasma homocysteine (particularly for homozygous individuals), a compound linked to various cardiovascular issues and vascular disease [15].

Frosst et al. established the initial connection between this mutation, reduced enzyme activity, and higher homocysteine levels.. They note that the amino acid is highly-conserved, yet this substitution occurs at a frequency of 38% in unselected chromosomes. Their findings form the basis for further research on the relationship between this mutation and vascular disease in different population groups, and in combination with other factors [2]. For instance, Chiu et al. discuss this mutation as it relates to susceptibility for hypertension in Taiwanese adults, finding that the presence of the TT genotype might heighten the risk of hypertension compared to those without. In their study, they also investigate how methylation, a chemical modification that affects gene expression, affects the risk of hypertension as it relates to the MTHFR gene, but find no significant relationship between methylation levels alone and hypertension. The combination of specific genotypes at this position with MTHFR promoter methylation was discovered to jointly impact an individual’s susceptibility to hypertension [1].

The broader impacts of this mutation from valine to alanine can be explained at the molecular level, where the chemical structure of the enzyme is altered, reducing its activity and increasing its thermolability. The difference between these two non-polar amino acids can be found in their R group, where valine consists of a benzine, a methylene, and a methyl radical, and alanine only contains the methyl radical. I was able to locate one scientific discussion of this enzyme’s structure, and how this mutation affects the protein chemically; Frosst et al. claim that the enzyme is stabilized in the presence of folate, to which MTHFR typically binds [2]. In a more general sense, amino acid substitutions are known to compromise the structural integrity of proteins and impact their ability to fold, which directly impacts their functions. With regard to enzymes specifically, the degree to which binding sites and active sites for their substrates are preserved is particularly indicative of how well the mutated enzyme will function. In this instance, since a change to the binding site is suggested, the protein would be rendered less stable and hence less effective.

Whilst the studies I have discussed here shed an important light on the promising capabilities of bioinformatics, many include a list of limitations pertaining to their methodologies. Nearly all reference statistical factors such as potential unrepresentativeness of the population and a lack of existing data or background information. More interestingly, in a paper discussing the relationship between this mutation in the MTHFR gene and serum homocysteine levels in hypertensive patients, Hu et al. note a

different experimental consideration that is a bit more biological in nature; despite them finding a significant correlation between this mutation and serum homocysteine, they note that "the interactions of gene-gene, gene-environment, and environment-environment on serum Hcy levels and the MTHFR SNP remain to be determined" [3]. This observation reflects an inherent limitation of bioinformatics insofar as it cannot account for the impact of an individual's environment on gene expression or their health in general. Even in the context of what can be broken down by bioinformatics, namely genes, the interactions between genes that produce health outcomes can often be challenging to discover and understand. Hence, being aware of potential pitfalls when conducting exome sequencing and analysis are important; here are some considerations [12]:

1. Locating Exons/Genes: discrepancies exist between databases that catalog the location of specific genes, and accurately determining the number of protein-coding and noncoding genes, in addition to the function of some of the proteins they code for, is still an open problem.
2. Complexity and Variability: the process of analyzing exome sequencing reads involves multiple intricate steps, the methods for which are constantly evolving. The impact of using different software tools on results or making different assumptions can be substantial, and the same data may yield very different results depending on how the analysis is conducted.
3. Reference Genome: there is not a single definitive reference genome, and there is not a singular agreed-upon set of variants associated with the human genome; the methodology of exome sequencing weighs heavily on the final description of the genome.

In addition to the gaps in our understanding surrounding the human genome and the inability for personal genomics to account for the influence of environmental and lifestyle factors on an individual's propensity to develop a specific disease, there are also broader concerns related to how this data may impact individuals' lives. In addition to the psychological impact of discovering genetic predispositions in one's genome, which may cause anxiety and unnecessary stress for the individual, there exists notable privacy and discrimination concerns associated with personal genomics that may prevent it from reaching a wide scale implementation in medicine. Nonetheless, the benefits that personal genomics can offer in a clinical setting will only improve as our understanding of the human genome becomes more complete. I see the largest application in preventative care and initial diagnoses; in my discussion of the alanine to valine mutation, for instance, personal genomics would give doctors access to another tool to diagnose vascular disease in a patient, and treat or address it sooner (perhaps by prescribing folate supplements or advising a particular diet to lower the patient's homocysteine levels). According to Frosst et al., "the identification of a candidate genetic risk factor for vascular disease, which may be influenced by nutrient intake, represents a critical step in the design of appropriate therapies for the homocysteinaemic form of arteriosclerosis" [2]. Ultimately, as the field of personal genomics develops and evolves, it should take its place as one of many tools that a physician can use to advise and treat their patients; it is not a replacement for traditional medicine, and the privacy and psychological downsides of its use must be considered alongside the benefits it offers.

Conclusions

The field of personal genomics offers promise for deciphering genetic blueprints and treating the diseases associated with them. Through technologies like NGS and exome sequencing, the once obscure realm of genomics has become more accessible, offering profound insights into an individual's genetic predispositions and ancestral lineage. The analytical process undertaken in this project—from alignment to genotyping and annotation—reflects the bridge between raw genetic data and meaningful biological insights. These steps, while powerful in unraveling disease-causing variants, also highlight the complexities in genomic analysis and the crucial role of methodology in deciphering an individual's genetic landscape.

The focused examination of specific mutations, such as the valine to alanine substitution in the MTHFR gene, underscores how intricate molecular changes can have broader implications within biological systems. This mutation's influence on enzyme activity and thermolability emphasizes the complex interplay between genetic variations and disease pathways, as evidenced by its association with elevated homocysteine levels and increased risk of vascular disease. Yet, amidst the potential capability to diagnose patients based on their genomes, limitations do exist. The limited information regarding how genes interact, in addition to the role that one's lifestyle and environment play in determining health outcomes, is not as easily accounted for through bioinformatics. Moreover, the effects of personal genomics extend beyond the scientific realm, raising ethical, privacy, and psychological concerns. While this field holds immense promise in preemptive healthcare and initial diagnoses, the limitations and concerns associated with it ought to be considered alongside its use. As this field continues to evolve, it should find its place as a complementary tool in the existing healthcare system, enhancing traditional medical practices rather than serving as a standalone solution. Understanding the limitations, ethical concerns, and the evolving nature of genomic knowledge remains crucial as personal genomics continues to shape the future of healthcare.[4]

[4] Note: without graphics (table, graphs) this report is seven pages long.

References

1. Chiu MH, Chang CH, Tantoh DM, Hsu TW, Hsiao CH, Zhong JH, Liaw YP. Susceptibility to hypertension based on *MTHFR* rs1801133 single nucleotide polymorphism and *MTHFR* promoter methylation. *Front Cardiovasc Med.* 2023 Oct 2;10:1159764. doi: 10.3389/fcvm.2023.1159764. PMID: 37849939; PMCID: PMC10577234.
2. Frosst, P., Blom, H., Milos, R. et al. *A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase.* Nat Genet 10, 111–113 (1995). https://doi.org/10.1038/ng0595-111
3. Hu XJ, Su MR, Cao BW, Ou FB, Yin RX, Luo AD. *Relationship between the methylenetetrahydrofolate reductase (MTHFR) rs1801133 SNP and serum homocysteine levels of Zhuang hypertensive patients in the central region of Guangxi.* Clin Hypertens. 2023 Oct 1;29(1):26. doi: 10.1186/s40885-023-00250-9. PMID: 37777810; PMCID: PMC10543866.
4. Li, H. (2010). BWA: Burrows-Wheeler Aligner. Retrieved from https://github.com/lh3/bwa
5. Li, H. (2011). SAMtools: Tools for nucleotide sequence analysis. Retrieved from https://www.htslib.org/doc/samtools-mpileup.html
6. National Library of Medicine. (2022, September 28). *Homocysteine Test. MedlinePlus.* https://medlineplus.gov/lab-tests/homocysteine-test/
7. Ibid. *MTHFR gene.* Available from: https://medlineplus.gov/genetics/gene/mthfr/
8. Ibid. *PNPLA3 gene.* Available from: https://medlineplus.gov/genetics/gene/pnpla3/
9. Ibid. *VDR gene*. Available from: https://medlineplus.gov/genetics/gene/vdr/
10. Microsoft. (2016). Excel [Software]. 2309. Redmond, WA: Microsoft.
11. Mountain, J. L., Chapter 6 - Personal Genomics, Editor(s): Geoffrey S. Ginsburg, Huntington F. Willard, *Genomic and Personalized Medicine (Second Edition),* Academic Press, 2013, Pages 74-86, ISBN 9780123822277, https://doi.org/10.1016/B978-0-12-382227-7.00006-9.
12. Pevsner, J. 2015. *Bioinformatics and Functional Genomics.* Vol. Third edition. Chichester, West Sussex, UK: Wiley-Blackwell. https://search-ebscohost-com.proxy.uchicago.edu/login.aspx?direct=true&db=nlebk&AN=1055003&site=eds-live&scope=site.
13. Ponasenko A, Sinitskaya A, Sinitsky M, Khutornaya M, Barbarash O. The Role of Polymorphism in the Endothelial Homeostasis and Vitamin D Metabolism Genes in the Severity of Coronary Artery Disease. *Biomedicines.* 2023 Aug 25;11(9):2382. doi: 10.3390/biomedicines11092382. PMID: 37760823; PMCID: PMC10526004.
14. R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/
15. Tiller G. E. and Kniffen C. L.. “5,10-METHYLENETETRAHYDROFOLATE REDUCTASE; MTHFR”. *Online Mendelian Inheritance in Man* (2013).
16. U.S. National Library of Medicine. (2022, September 21). RS738409 RefSNP report - dbSNP. *National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/snp/rs738409/#clinical_significance
17. Ibid. RS1801133 RefSNP report - dbSNP. Available from: https://www.ncbi.nlm.nih.gov/snp/rs1801133

18. Ibid. RS2228570 RefSNP report - dbSNP. Available from: https://www.ncbi.nlm.nih.gov/snp/rs2228570
19. Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Retrieved from http://annovar.openbioinformatics.org/en/latest/
20. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. *Journal of Statistical Software,* 35(1), 1–65. https://doi.org/10.18637/jss.v035.i01
21. Zhong P, Zhu Z, Wang Y, Huang W, He M, Wang W. Cardiovascular and microvascular outcomes according to vitamin D level and genetic variants among individuals with prediabetes: a prospective study. *J Transl Med.* 2023 Oct 16;21(1):724. doi: 10.1186/s12967-023-04557-x. PMID: 37845735; PMCID: PMC10577927.