Shriya Bansal
Multiscale Modeling for Biological Sciences

Understanding the Effects of SNPs on Gene Expression through NGS

Sequence 4 Analysis

Introduction

The field of personal genomics is critical for the advancement of medical technology in the modern era. Specialized medicine allows doctors to prescribe more customized drug selections and better predict the status of inherited diseases. Personal genomics sequencing can also be employed to understand genetic links and identify the likelihood of contracting common diseases. Scientists are more easily retrieving genetic information through Next Generation Sequencing (NGS) for data analysis in order to improve the methods with which they personalize medicine.

NGS has become popular method for retrieving an individual's genome sequence and is based on a highly parallelized process that allows for efficient sequencing of nucleotide bases [4]. This method utilizes array-based techniques with short read lengths to reduce sequencing costs and time constraints in comparison to the previously used Sanger Sequencing method. Precision medicine has been greatly enhanced by NGS, as researchers can now more effectively determine where genetic variants reside and can better formulate medications to target these regions.

NGS includes three steps consisting of library preparation, DNA amplification, and genome sequencing. The genetic library is first created by enzymatically separating DNA into short fragments and ligating these strands to synthetic adapters. A polymerase chain reaction (PCR) composed of the denaturing, annealing, and extending phases is then performed on the DNA in order to amplify it. Capillary electrophoresis is then used to separate the amplified fragments, and pyrosequencing techniques are applied to "synthesize the complementary DNA strands in the presence of a polymerase enzyme" [5]. In this process, fluorescent tags are utilized to label each nucleotide. The amount of light released is measured to determine the how much of a given nitrogenous base exists in the sample. This allows for the strand to be sequenced by synthesis, and occurs simultaneously with the detection of the bases. Paired end sequencing is also used in order to refine the signal-to-noise ratio between the shorter strand reads [6].

Exome sequencing is more commonly applied when performing genomic analysis because exonic regions contain the most critical protein encoding components of DNA. With exome capture technology, "probes are bound to a high-density microarray" [1] which allows nucleotides to bind to their respective complementary strands of the exonic regions. Non-coding regions such as introns are then washed away through multiple cycles of this process. The most common

form of such exome capture technology is Illumina NGS.

In contrast with common, high frequency variants, rare variants are considered to be more difficult to identify due to their extreme allelic heterogeneity [3]. Furthermore, these rare variants often are functionally critical in producing complex traits. NGS makes it more plausible to sequence small portions of entire genomes to determine the locations and frequencies of these variants. In addition, these variants have a tendency to be population specific and may need to be aggregated to determine their influence.

Throughout this project, I will perform sequence alignment to the reference human genome using the Burrows Wheeler Alignment to identify exonic variants and discern insertion-deletion (indels) locations from substitutions and single nucleotide polymorphisms (SNPs). I will then apply the ANNOVAR tool to annotate the genetic variants by comparing against the NCBI dbSNP and GWAS databases through the process known as variant calling. The pathogenic diseases that this individual is more prone to contract, along with caveats of genetic sequencing will then be discussed.

Methods

The sequence alignment was performed using the raw reads from the Illumina Technology NGS machine. Every individual read was compared against the reference human genome to understand which portion of the genome a given read mapped to. After aligning the genome, I implemented the Burrows Wheeler Alignment (BWA) algorithm to determine where a given nucleotide substitution or indel event occurred in the individual's genetic sequence. The first step involved the *bwa aln* program which aligned the paired-end reads and produced separate files indexed in an efficient manner. The parameters for the number of parallel threats "-t" and trimming "-q" were applied to properly complete this portion of the alignment. This was then followed by usage of the *bwa sampe* and *samtools* programs. In these, the most accurate alignment was extracted from the paired-end reads with reference to the alignment locations from each. The parameter "-p" was applied to define a path at the beginning of filename outputs. A binary alignment map was then produced with the most ideal alignment.

The alignment of the genomic sequence was followed by its genotyping in which SNPs and indel events were visualized. These were identified by correlating the individual's genetic code to the reference human genome through variant calling. Numerous reads were performed over each nucleotide in order to understand the state of the individual's genotype at a given location. The probability of a mutation occurring at that position was then extracted through these multiple reads and compared against the frequencies of specific alleles. The *samtools mpileup* program contains the SAMtools algorithm. This was used with the flag "-f" to specify the fasta format of the file to be read by the program, the flag "-t" to indicate additional reference data, and the parameters "-uv" and "-mv."

Annotations were then performed on these variant call format (.vcf) file. The ANNOVAR tool was applied to retain biological information about the identified genetic variants. This produced data on the genomic context of the given variant, the mutation type, and the RefGene ID that can be further investigated through public databases. It also provides researchers with information about whether the specific variant has been studied by others given that it has a dbSNP or rsID, if it is associated with any pathological disease, and a measure of how functional it is within the organism. The "--buildver," "--protocol," and "--operation" parameters were implemented in the script for this program. The protocol flag specifically indicated the databases from which the annotations were conceived. These included refGene, avsnp150 from the NCBI dbSNP database, clinvar_20170905 to convey if the variant has been associated with a pathology from GWAS, and dbnsfp33a, which provided annotations from dbNSFP about the given SNP's functionality.

The most pertinent variants were considered to be located in exonic, protein coding regions and of the non-synonymous type, meaning that the SNP resulted in an amino acid mutation and protein change. Analyzing the functional regions of the genome more closely after filtering out variants that had a Phred score below 224 from the original vcf and passing them through GWAS in the ClinVar database allowed me to recognize which mutations were associated with a pathological disease. The NCBI databases also provided information on phenotypic traits associated with a given gene and allowed me to correlate the sequence of the individual with diseases that he or she has a greater likelihood of contracting.

Results

The alignment of this individual's sequence to the reference sequence led to the identification of a total of 212,334 variants, 16,630 of which are high quality with a minimum Phred score set to 224. This Phred score correlates to the certainty with which a variant is considered accurate and illustrates the probability of that variant within the genome. 4,345 of these high quality variants come from the exonic, protein-coding regions of the genome, while 15,853 of these mutations are SNPs and 776 are indels. Within the exonic region, there exist 2,126 synonymous variants, 1,876 non-synonymous variants, 50 frameshift variants, and 11 premature stop codons.

The pie chart in Figure 1 illustrates the proportion of genetic variants that reside on each chromosome in the individual's genome. Chromosome 1 contains 411 variants, while 335 SNVs come from Chromosome 11 and 327 SNVs are encompassed by Chromosome 19. Thus, Chromosome 1 has the greatest amount of variants.

As depicted through the table in Figure 2, a gene that resides on Chromosome 12, HNF1A, is responsible for acute insulin response, coronary heart disease, and maturity onset diabetes of the young, as determined through GWAS. A non-synonymous SNP has been identified

for this gene at position 120999579 where Adenine is substituted for Guanine, resulting in an amino acid mutation from Serine to Glycine. The Online Mendelian Inheritance in Man (OMIM) catalog indicates that HNF1A is mainly associated with maturity onset diabetes of the young. It has been previously studied, as the dbSNP query for rsID of 1169305 indicates that it has an allelic frequency of approximately 0.003.

The gene IL7R from Chromosome 5 is also associated with the diseases of asthma, multiple sclerosis, and type 1 diabetes from GWAS. This non-synonymous SNP occurs at position 35871088, the nucleotide mutation occurs from Guanine to Adenine, causing an amino acid change from Alanine to Valine. Multiple sclerosis is associated with optic neuritis and numbness because the immune system is unknowingly attacking the myelin sheath over the cells in the brain and spine [7]. IL7R has been studied previously, given that it has a rsID of 1494555 and higher allelic frequency of nearly 0.36.

A non-synonymous single nucleotide variant (SNV) exists on Chromosome 7 at position 150999023 that correlates with the gene NOS3. Here, Thymine has been substituted for Guanine causing an amino acid mutation from Aspartic Acid to Glutamic Acid. This gene, Nitric Oxide Synthase 3, primarily correlates to multiple forms of coronary heart disease, as well as cases of diastolic blood pressure, and hypertension. Coronary heart disease is associated with symptoms of angina, or chest pain, indigestion, and shortness of breath [8]. NOS3 has also been considered in past sequencing research, as it has a rsID of 1799983 and an allelic frequency of approximately 0.25.

Additionally, on chromosome 20 at position 44186550, because a synonymous SNP for the gene JPH2 had been identified, no change in the amino acid has occurred. Although Cytosine has been replaced by Thymine, the amino acid Tyrosine has remained. JPH2 is also a protein coding gene associated with Clozapine-Induced Agranulocytosis, as determined through GWAS. This correlates with severely reduced white blood cell counts due to treatment for patients with schizophrenia [9]. JPH2 has also been studied previously, as its rsID 1883790 is associated with an allele frequency of 0.15.

Figure 1 Depicts the proportion of variants on each chromosome in this individual's genome.

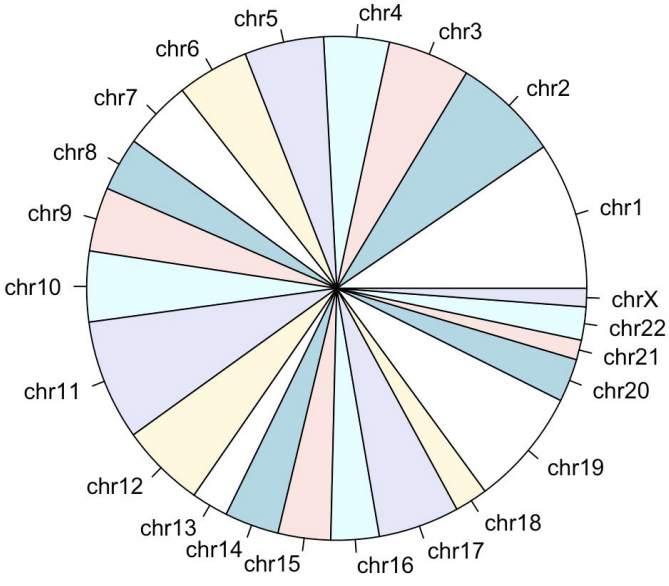

Figure 2 Depicts 15 variants across the genome of this individual, characterizing type of variant and potential diseases associated with exonic mutation.

#	Location of Variant	Type of Variant	Implications
1	Chromosome 1, Position 100206504	Non-Synonymous SNV	Maple Syrup Urine Disease
2	Chromosome 2, Position 233799782	Synonymous SNV	Abnormality of Neuronal Migration
3	Chromosome 3, Position 38603929	Non-Synonymous SNV	Long QT Syndrome, Romano-Ward Syndrome, Sick sinus Syndrome
4	Chromosome 4, Position 25252373	Synonymous SNV	Gestational Diabetes mellitus uncontrolled
5	Chromosome 5, Position 35871088	Non-Synonymous SNV	Allergic Disease, Multiple Sclerosis, Type 1 Diabetes
6	Chromosome 7, Position 150999023	Non-Synonymous SNV	Susceptibility to Metabolic syndrome, Coronary Heart Disease
7	Chromosome 12, Position 120999579	Non-Synonymous SNV	Maturity-onset diabetes of the young, Acute Insulin Response
8	Chromosome 13, Position 113147423	Synonymous SNV	Factor X Deficiency
9	Chromosome 14, Position 22812901	Synonymous SNV	Lysinuric protein intolerance
10	Chromosome 15, Position 72345454	Synonymous SNV	Tay-Sachs Disease
11	Chromosome 16, Position 56514589	Non-Synonymous SNV	Bardet-biedl Syndrome 2
12	Chromosome 17, Position 7513777	Frameshift Deletion	Robin Sequence, Intellectual Disability, Bilateral Conductive Hearing Impairment
13	Chromosome 19, Position 53891711	Synonymous SNV	Spinocerebellar Ataxia
14	Chromosome 20, Position 44186550	Synonymous SNV	Familial hypertrophic cardiomyopathy
15	Chromosome 22, Position 17209519	Synonymous SNV	Polyarteritis nodosa, childhood-onset

Discussion

Maturity onset diabetes of the young (MODY) is dependent on a gene of interest, HNF1A. This gene is associated with the protein Hepatocyte nuclear factor 1-alpha. During the development of the embryo, HNF1A is responsible for encoding a critical portion of the auto-regulatory network on the level of transcription within the pancreas [10]. It specifically assists in regulating insulin secretion and glucose transportation within beta cells. HNF1A is primarily expressed within the kidney, pancreas, and liver, where the respective protein acts as a dimer to form a 4-helix bundle [11]. Deletions or interruptions within this transcription factor results in deregulation of the associated molecular mechanism and leads to MODY. The C-terminal containing the transactivational domain is disrupted by this mutation when MODY occurs [12].

A study analyzing the function of the HNF1A gene and its influence on the onset of diabetes by Jesús Miguel Magaña-Cerino and his fellow researchers concludes that mutations in the gene are most commonly indicated by a progressive decrease in insulin secretion within the body as catalyzed by glucose. Such a Mendelian disease based on a single genetic variation occurs most commonly for patients under the age of 25 due to autosomal dominant inheritance. These researchers discovered that the I27L polymorphism, a variant of the original HNF1A gene, is closely linked to an increase likelihood of developing Type 2 Diabetes. This SNP is located within the HNF1A domain and was considered to be an ancestry informative marker within the given population [12]. This served as a clear indicator of decreased transcriptional activity within the genome. This study indicated that greater functional analyses must be completed on the HNF1A gene in order to understand the dependency on glucose metabolism for the stability of this gene and its related protein. The paper also suggests that improvement of sequencing techniques is critical in order to most accurately prescribe sulfonylureas to patients and prevent potential microvascular complications.

The mass-sequencing of complete genomes remains implausible presently due to computationally expensive technology. Exome sequencing is far more efficient, as it provides direct methods to identify diseases associated with SNVs and can be completed in a shorter, more timely manner. Additionally, comprehensive guides to complete genome sequencing and its analyses have not yet been published, making it difficult for others to decipher between the quality of various alignment algorithms and databases to best complete their research. Furthermore, modern sequencing reads are short relative to the length of the genome, and this translates to great difficulty in assembling the entirety of an individual's genetic sequence. Current pyrosequencing machines also struggle to decipher homopolymers in which repeated units of identical nucleotides occur consecutively [4]. This causes incorrect errors to appear within the sequenced

genome that indicate indels and SNPs where there should not be any mutations.

Some other caveats of NGS include the reliance on the availability and functionality of external databases and web interfaces. The data must also be transferred by the user, thus requiring large data to be uploaded to various remote locations. They also provide issues with legality, as confidentiality is not always guaranteed within such studies.

The field of personal genomics is reliant on accurate pathogenic variant specification. Better functional assays must be performed to recognize deviations that occur in rare pathogenic regions. Additionally, so-called hidden mutations may reside outside of exonic protein coding regions and still produce phenotypic morphisms. Reflection of hereditary information as a stimulus for disease may also lead to incorrect prescriptions being given, as "8.5% of families provisionally diagnosed with autosomal dominant disease truly have X-linked Retinitis Pigmentosa" [14]. Within the realm of personalized medicine, numerous fronts exist involving genetic sequencing that allow for the improvement of a population's overall health, but unintended consequences may appear through ethical dilemmas about how to address a potential disease.

Knowledge of these genetic mutations can be incredibly useful in preventive medicine though, as common inherited genetic conditions can be detected at early stages rather than once symptoms have become clinically apparent [15]. This allows for more proactive medical practice that can lead to increased life expectancy and greater likelihood of recovery. There is also a vast volume of data and information that exists within the human genome, and policies to maintain the security of this information are still being formed, meaning that there is not complete proficiency as to how this information can or should be managed.

Conclusion

NGS technologies have been pivotal in clinical interactions. Through the process of alignment, genotyping, annotation, and interpretation, this human genome could be analyzed to determine loci signalling disease from SNVs. When the SNPs from this individual's genome located on Chromosome 12 were further analyzed through the NCBI, OMIM, and GWAS databases, I found that the variant of HNF1A would cause the disease of maturity onset diabetes of the young. Personal genomics involves a computational method of parsing through a genome to search for genetic variants. This allows for potential diagnosis of inherited diseases during the early stages of their existence.

This evolving technology constantly provides researchers with innovative insights into the human genome and allows them to understand how specific genotypes have contributed to physical traits in such organisms. It is critical to continue exploring the vast field of personal genomics to gain a more thorough understanding of its intricacies.

References

1. Warr, Amanda et al. 2015. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda, Md).* 5,8:1543-50.
2. Li, Bingshan et al. 2013. Identifying rare variants associated with complex traits via sequencing. *Current protocols in human genetics.* 1:26.
3. Locke, Jonathan M et al. 2018. The Common *HNF1A* Variant I27L Is a Modifier of Age at Diabetes Diagnosis in Individuals With HNF1A-MODY. *Diabetes.* 67:9.
4. Pevsner, Jonathan. 2015. Bioinformatics and Functional Genomics. 3:621-695.
5. Brown, Tom et al. 2005. Next Generation Sequencing. ATDBio.
6. Mardis, Elaine. 2013. Next Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*. 6:287-303.
7. Kaminska, Joanna et al. 2017. Multiple sclerosis - etiology and diagnostic potential. *Postepy Hig Med Dosw.* 0:551-563.
8. National Heart, Lung, and Blood Institute. 2018. Coronary Heart Disease.
9. Jose Ma. J. Alvir, Jeffrey A. Lieberman, Allan Z. Safferman, et al. 1993. Clozapine-Induced Agranulocytosis -- Incidence and Risk Factors in the United States. *The New England Journal of Medicine.* 329(1):162-167.
10. Pavić, Tamara et al. 2018. Maturity onset diabetes of the young due to *HNF1A* variants in Croatia. *Biochemia medica.* 28(1):2.
11. Ada Hamosh wt al. 2011. HNF1 Homeobox A; HNF1A. *Online Mendelian Inheritance in Man.*
12. Magaña-Cerino, Jesús Miguel et al. 2016. Identification and functional analysis of c.422_423InsT, a novel mutation of the *HNF1A* gene in a patient with diabetes. *Molecular genetics & genomic medicine.* 5(1):50-65.
13. Pabinger, Stephan et al. 2012. A survey of tools for variant analysis of next-generation sequencing data. *Briefings in Bioinformatics.* 15(2):256-278
14. Koboldt, Daniel et al. 2013. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell.* 155.27-38.
15. Khromykh, Alina and Benjamin D Solomon. 2015. The Benefits of Whole-Genome Sequencing Now and in the Future. *Molecular syndromology.* 6(3): 108-9.