GENETIC VARIATION ANALYSIS OF A PRIMARY IMMUNODEFICIENCY PATIENT VIA WHOLE-EXOME SEQUENCING APPROACH

BADER ABDUL KADER EL FARRAN

FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2021

GENETIC VARIATION ANALYSIS OF A PRIMARY IMMUNODEFICIENCY PATIENT VIA WHOLE-EXOME SEQUENCING APPROACH

BADER ABDUL KADER EL FARRAN

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2021

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **BADER ABDUL KADER EL FARRAN**

Matric No: **SMA180039/17199021/1**

Name of Degree: **MASTER OF SCIENCE**

Title of Dissertation:

GENETIC VARIATION ANALYSIS OF A PRIMARY IMMUNODEFICIENCY PATIENT VIA WHOLE-EXOME SEQUENCING APPROACH

Field of Study: **BIOINFORMATICS**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

GENETIC VARIATION ANALYSIS OF A PRIMARY IMMUNODEFICIENCY PATIENT VIA WHOLE-EXOME SEQUENCING APPROACH

ABSTRACT

Primary Immunodeficiency (PID) refers to a collection of diseases whereby the production of antibodies is negatively affected, or the cellular defenses of the immune system do not operate appropriately. A Malaysian patient (P1) was initially suspected with a PID type known as Hyper IgM syndrome. However, the immunological workup was not compatible with Hyper IgM syndrome. Hence, a Whole-exome sequencing (WES) analysis was conducted to look for mutations in PID-related genes. P1's raw data were mapped to four different versions of the human reference genome to compare the results and determine which one is the best for this analysis. Once the variants were called from P1's data they were annotated to search for mutations. As a result, a novel mutation was detected in the Nuclear factor-kappa-B-inhibitor alpha (NFKBIA) gene, which is responsible for regulating the Nuclear factor-kappa-B (NFKB) gene. It is a single nucleotide polymorphism (SNP) (NFKBIA:NM_020529:exon1:c.A94T:p.S32C) at codon 94 (c.A94T) of P1's NFKBIA gene which resulted in the mutation of the serine residue (Ser32) to a cysteine residue (Cys32). This SNP lies in the destruction motif of the NFKBIA protein, which may have led to the impairment of NFKB activation in P1, which could explain the symptoms of the patient. Since this is a novel mutation, it warrants future investigation to find out what such mutation exactly does to the NFKBIA protein structure and how it affects its interaction with NFKB.

Keywords: P1, Primary Immunodeficiency, Hyper IgM Syndrome, Whole-Exome Sequencing, Nuclear factor-kappa-B-inhibitor alpha.

GENETIC VARIATION ANALYSIS OF A PRIMARY IMMUNODEFICIENCY PATIENT VIA WHOLE-EXOME SEQUENCING APPROACH

ABSTRAK

Keimunodefisienan primer (PID) merujuk kepada penyakit yang mampu menjejaskan pengeluaran antibodi atau merencatkan operasi tindak balas imun dalam tubuh badan. Seorang pesakit Malaysia (P1) pada mulanya disyaki mempunyai PID yang dikenali sebagai sindrom Hyper IgM. Walau bagaimanapun, pemeriksaan imunologi menunjukkan bahawa ia tidak sesuai dengan sindrom Hyper IgM. Oleh itu, analisis penjujukan keseluruhan exome (WES) dijalankan bagi mencari mutasi pada gen yang berkaitan dengan PID. Data mentah P1 dipetakan ke empat versi yang berbeza genom rujukan manusia untuk membandingkan keputusannya dan bagi menentukan apa yang terbaik untuk analisis ini. Setelah varian dipanggil dari data P1, mereka diberi anotasi untuk mencari mutasi.Penyelidikan ini telah mengenal pasti satu mutasi novel pada gen Nuclear factor-kappa-B-inhibitor alpha (NFKBIA) yang berfungsi untuk mengawal gen Nuclear factor-kappa-B (NFKB). Ia merupakan polimorfisme nukleotida tunggal (SNP) (NFKBIA:NM_020529: exon1: c.A94T: p.S32C) di kodon 94 (c.A94T) gen NFKBIA yang menyebabkan residu serin (Ser32) termutasi kepada residu sisteina (Cys32). SNP ini terletak di motif pemusnah pada protein NFKBIA yang berpontensi untuk mengakibatkan kemerosotan aktivasi NFKB dalam P1, dan ini menjelaskan simptom yang terdapat pada pesakit. Oleh sebab ia merupakan mutasi yang novel, penyelidikan masa depan diperlukan untuk mengenalpasti kesan mutasi ke atas struktur protein NFKBIA dengan lebih terperinci dan bagaimana mutasi tersebut mempengaruhi interaksinya dengan NFKB.

Kata kunci: P1, Keimunodefisienan primer (PID), sindrom Hyper IgM, penjujukan keseluruhan exome (WES), Nuclear factor-kappa-B-inhibitor alpha.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards my supervisor Associate Prof. Dr. Saharuddin Bin Mohamad and Dr. Adiratna Mat Ripen who is the Principal Investigator of the WES project, Allergy & Immunology Research Centre, Institute for Medical Research, Malaysia (IMR). I am grateful for their mentorship throughout this project. It would not have been possible if it was not for their guidance.

I would like to acknowledge the IMR for proceeding with the whole-exome sequencing and providing me with the sample dataset.

I would like to thank my friend Chiow Mei Yee for helping me in generating the Sanger sequencing results.

I am grateful to the Medical Research and Ethics Committee (MREC) from the National Institutes of Health, Malaysia (NIH) for giving ethical approval for this project due to the involvement of human samples.

I would like to dedicate this work to my parents, brothers, sister, and uncle. They have supported me all the time throughout this journey.

Finally, I would like to thank all my friends who have been by my side and gave me moral support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

A	:	Adenine
BQSR	:	Base quality score recalibration
BWA	:	Burrows-Wheeler Aligner
Cys	:	Cysteine
D	:	Damaging/Disease-causing
DNA	:	Deoxyribonucleic acid
EDA	:	Anhidrotic ectodermal dysplasia
GOF	:	Gain-of-function
het	:	Heterozygous
hg19D	:	Codename for the reference genome hg19 plus decoy sequences (hs37d5)
hg38NAD	:	Codename for the reference genome hg38 without alternate sequences, plus decoy sequences (GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set)
HIV	:	Human immunodeficiency virus
hom	:	Homozygous
INDEL	:	Insertion-deletion
NFKB	:	Nuclear factor-kappa-B
NFKBIA	:	Nuclear factor-kappa-B-inhibitor alpha
NGS	:	Next-generation sequencing
P1	:	The codename for the sample patient
PBMC	:	Peripheral blood mononuclear cell
PCR	:	Polymerase chain reaction
Ser	:	Serine

SNP	:	Single nucleotide polymorphism
T	:	Thymine
VCF	:	Variant call file
VQSR	:	Variant quality score recalibration
WES	:	Whole-exome sequencing
WGS	:	Whole-genome sequencing

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview of the Research

Primary immunodeficiency (PID) is a class of diseases whereby the immune system is damaged either by not being able to produce enough antibodies to fight off infections or the cellular defenses do not work appropriately (Sheikhbahaei et al., 2016). The word “Primary” in PID indicates the fact that it is caused intrinsically by DNA damage since birth in contrast to Secondary immunodeficiency whereby the word “Secondary” denotes the fact that it is caused by extrinsic factors such as the Human Immunodeficiency Virus (HIV).

PID consists of nine groups and each group has many types which add up to approximately 400 types of PIDs in total (Tangye et al., 2020).

The groups of PIDs include:

A) Immunodeficiencies affecting cellular and humoral immunity.

B) Combined Immunodeficiencies with associated or syndromic features.

C) Predominantly Antibody deficiencies.

D) Diseases of immune dysregulation.

E) Congenital defects of phagocyte number, function, or both.

F) Defects in Intrinsic and Innate immunity.

G) Auto-inflammatory disorders.

H) Complement deficiencies.

I) Phenocopies of PIDs.

PIDs are rare and they remain elusive. Diagnosis of such rare diseases is advancing throughout time but there are still more to be uncovered. Next-generation sequencing (NGS) is a newer sequencing technology that is non-Sanger-based. It has a significantly higher speed than its counterpart, Sanger-sequencing. NGS had to tackle the obstacle of the difficulty to switch from a Sanger-based approach to a faster and cheaper non-Sanger-based approach, as biology depended on Sanger-sequencing for 30 years (Schuster, 2008). NGS is cost and time-efficient as it requires much less time to sequence an entire genome at a much lower price. NGS has proven itself to be useful in clinical uses. It has enabled us to identify genetic variants easily. An NGS test in the clinical sense can be devised to capture a handful of selected genes by NGS gene panel, the exome by whole-exome sequencing (WES), or the genome by whole-genome sequencing (WGS). WES is the technique whereby the coding region of the DNA (1-2% of the entire genome) is targeted for sequencing while the rest is disregarded. WES has been a major player in clinical diagnosis in terms of identifying genetic variants that could be linked to causes of diseases (Retterer et al., 2016). This can be advantageous over WGS in terms of saving time and money. WGS is the technique whereby the entire genome is sequenced. This could be useful if we are looking for mutations that could be outside of the exome in non-coding regions. WES, when compared to WGS, is cost and time-efficient as the DNA-coding sequences are targeted for capturing and sequencing (1-2% of the human genome). Around 85% of mutations related to genetic diseases are found in the exome (Rabbani et al., 2014). WES can be applied to identify harmful genetic mutations that are PID-related. Figure 1.1 presents the types of NGS done in the clinical field.

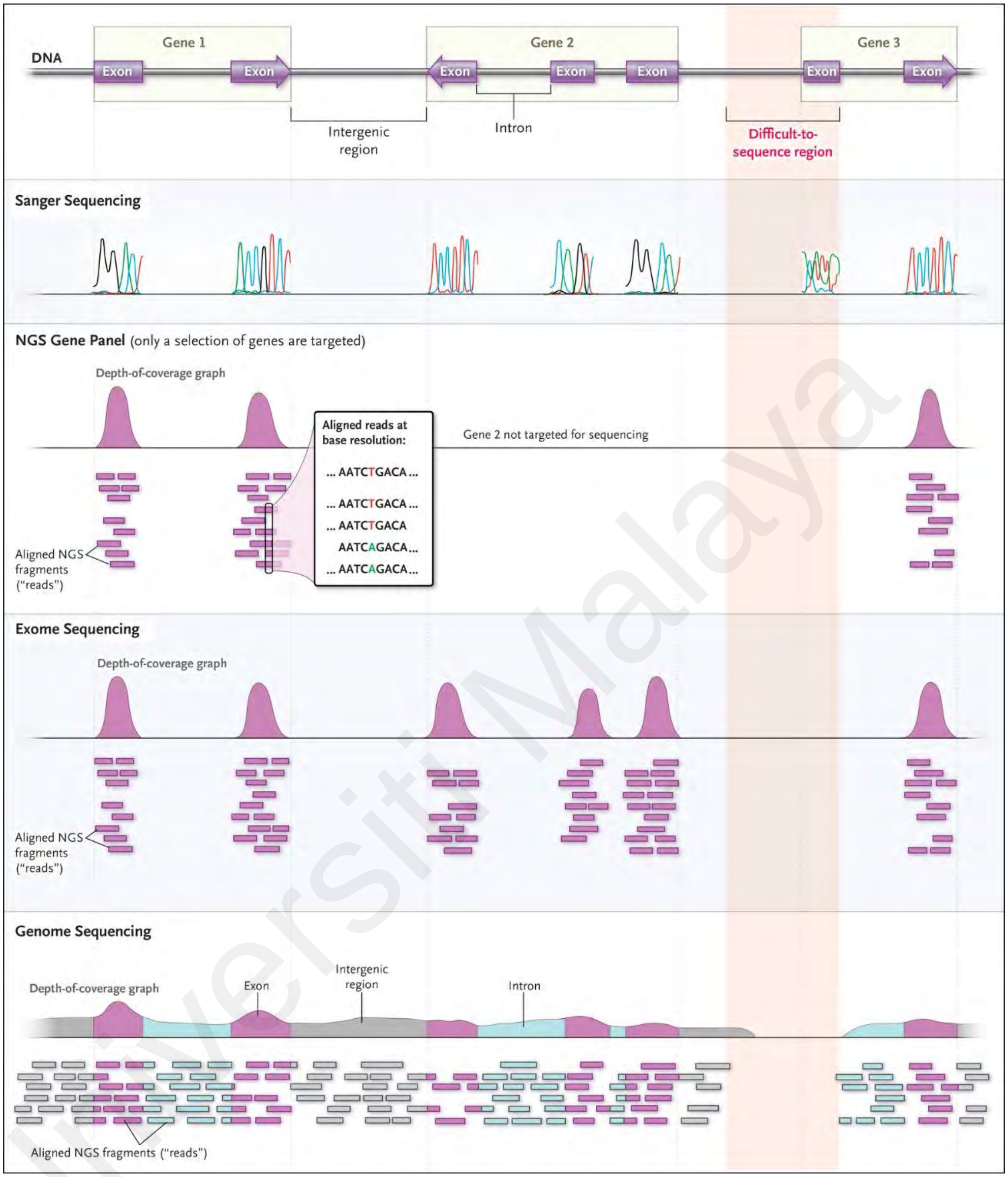


Figure 1.1: A diagram that illustrates the NGS types that are applied in the clinical field. The top shows an example DNA sequence. Sanger sequencing is typically used nowadays to confirm mutations that have been detected by NGS techniques. NGS gene panel refers to the technique where only a few genes are selected for sequencing. Exome sequencing refers to the technique that targets the exonic regions for sequencing. Genome sequencing refers to the technique whereby the entire DNA is sequenced. (Reproduced with permission from (Adams & Eng, 2018), Copyright Massachusetts Medical Society.).

1.2 The Case and the Objectives

The patient (P1) of this case study is a male who was initially suspected with a PID type known as Hyper IgM syndrome. However, the immunological workup was not compatible with Hyper IgM syndrome according to the observation. The patient was presenting symptoms that include fever, infections, septicemic shock, hepatosplenomegaly, and dysmorphic features. Hence, a WES analysis has been performed to locate and uncover harmful mutations that are PID-related and that may lead to a diagnosis. The patient's raw data were processed and mapped to four different versions of the human reference genome to select the best version for this study.

There lies a knowledge gap in the field of PIDs. Diagnosis is still lagging in developed and undeveloped countries alike (El-Sayed & Radwan, 2020). This has to do with the fact that PIDs are extremely rare and our understanding of PIDs remains elusive. WES has been effective in diagnosing many cases of PID and has always led us to discover novel variants that are PID-related. Hence, the objectives and activities of this project are:

1. To map and improve the quality of the raw data by carrying out trimming, alignment of reads to a reference genome, file processing, filtering, and quality control for the WES raw data of the patient.
2. To determine gene variations or mutations that could give rise to a decisive diagnosis of PID in the patient by carrying out variant calling, annotation of the called variants, and interpretation of the annotation.
3. To determine the best possible reference genome for this analysis by mapping to four different versions of the human reference genome and comparing the results.

1.3 Thesis Organization

Chapter 2 is the Literature Review whereby the history of PID is discussed followed by the prevalence of PID which shows how rare PIDs are. The patient was initially suspected with Hyper IgM syndrome. Hence, Hyper IgM was discussed. Furthermore, treatment options of PID were described. As technology advanced, a new form of sequencing technique known as NGS has emerged and its history was included in the chapter. The application of an NGS technique known as WES was discussed on how it tackles PID with some of its limitations.

Chapter 3 is the Methodology whereby the machines which were used to conduct the WES for P1's DNA were mentioned followed by the mentioning of the institution that provided the raw data for this research project. Next, the Bioinformatics analysis pipeline was described and illustrated. Then, four different versions of the human reference genome were mentioned. This was followed by the commands and scripts that were used to run the Bioinformatics analysis pipeline. The last step of the bioinformatics analysis pipeline was the annotation of the called variants. The interpretation of the annotation is the next step and it was described.

Chapter 4 is the Results and it illustrates the findings of the Bioinformatics analysis pipeline, the comparison of the results of mapping the raw data to 4 different versions of the human reference genome, the interpretation of the annotated variant calls, and the Sanger sequencing validation of the detected mutation. The detected mutation is a single nucleotide polymorphism (SNP) in P1's Nuclear factor-kappa-B-inhibitor alpha (NFKBIA) gene.

Chapter 5 is the Discussion and it supports the results with previously done literature. First, it describes the importance of the study. This is followed by discussing the significance of the results. It discussed the best version out of the 4 versions of the human reference genome that is suitable for this Bioinformatics analysis pipeline. Next, it discussed the role of the gene (NFKBIA) in which the mutation was detected and genes that are relevant to it. Then, the mutation of the NFKBIA gene was discussed along with a possible scenario that may most probably occur based on literature.

Chapter 6 is the Conclusion and it is the final chapter that summarizes the thesis by briefly discussing the reason why this research is conducted. This was followed by a summary of the methods and the results and their discussion. Finally, this concluded that the findings warrant further investigation of the said mutation.

CHAPTER 2: LITERATURE REVIEW

2.1 History of PID

The history of PIDs began in 1952 when Agammaglobulinemia was discovered by Bruton (Bruton, 1952). However, some patients were observed in 1922 with symptoms in their pharynxes (Schultz, 1922) that were later categorized as PID. Disorders affecting the immune system like ataxia-telangiectasia (Syllaba, 1926) and Wiskott Aldrich syndrome (Familiärer, 1937) were discovered in 1926 and 1937 respectively. It was difficult to detect, categorize, and tackle such diseases. With the advancement of technology and cracking the code of DNA, the image of PID was beginning to become clearer as scientists can look for variants in the DNA that may be linked to PIDs. To this day, more and more PIDs are being discovered and categorized (Ochs and Hitzig, 2012). Over the years, technology advanced and scientists discovered approximately 400 types of PIDs which are spread out into nine groups (Tangye et al., 2020).

2.2 Prevalence of PID

PIDs are very rare as there are around 6 million people who may have a PID globally, though only 30 – 60 thousand people are registered with a PID (Bousfiha et al., 2013) and there are yet many secrets to unravel about them.

In Malaysia, according to a study, the prevalence of PID is 0.37 per 100,000 population. This prevalence rate is lower compared to other countries. This may have to do with the fact that Malaysia lacks a national registry for PID and detection strategies. 119 PID patients were included and studied in Malaysia. Figure 2.1 illustrates the prevalence of different PID types in Malaysian patients.

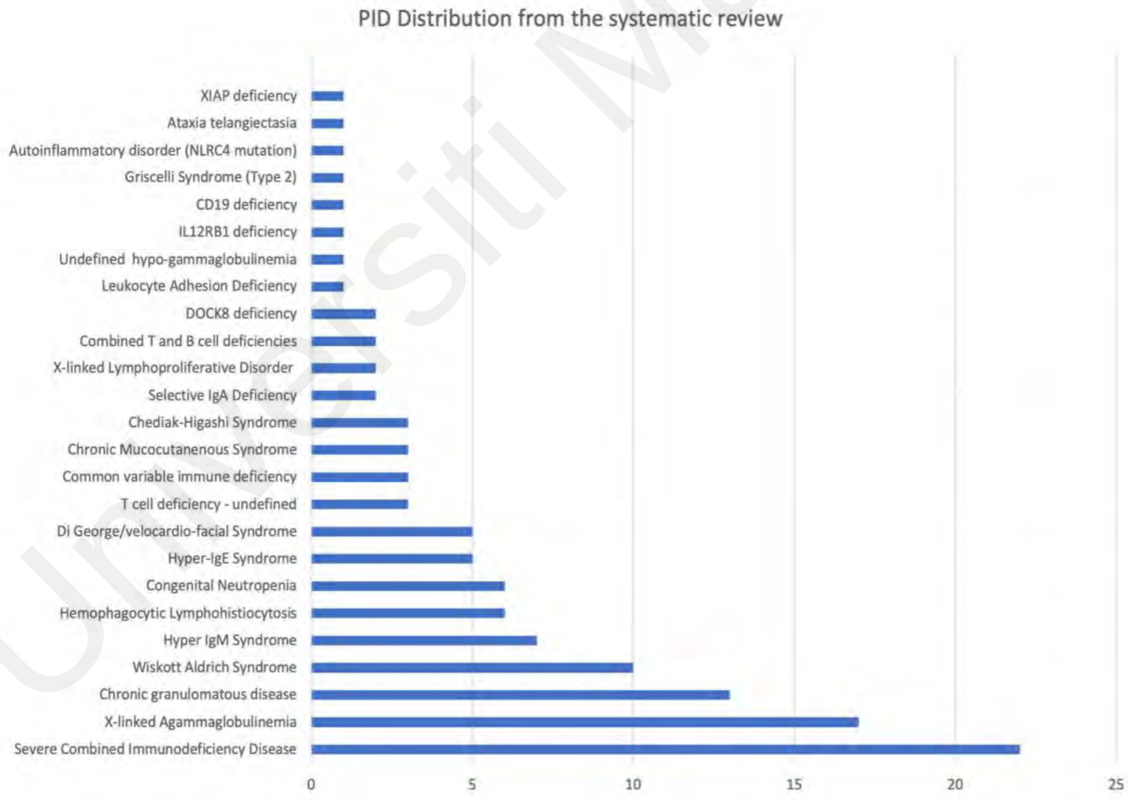


Figure 2.1: PID diagnosis distribution from the systematic review of Malaysian PID patients (Abd Hamid et al., 2020).

2.3 Hyper IgM Syndrome

Hyper IgM Syndrome is a rare PID that is X-linked and occurs due to a mutation in the CD40 Ligand gene (Levy et al., 1997). Levels of IgM will be normal or increased while levels of IgG and IgA will be decreased. Other forms of IgM Syndromes are not X-linked, but rather Autosomal Recessive and they are linked to mutations in the following genes: AICDA, UNG, INO80, and MSH6 (Tangye et al., 2020). Figure 2.2 illustrates an analysis of Hyper IgM patients' HEP-2 cells.

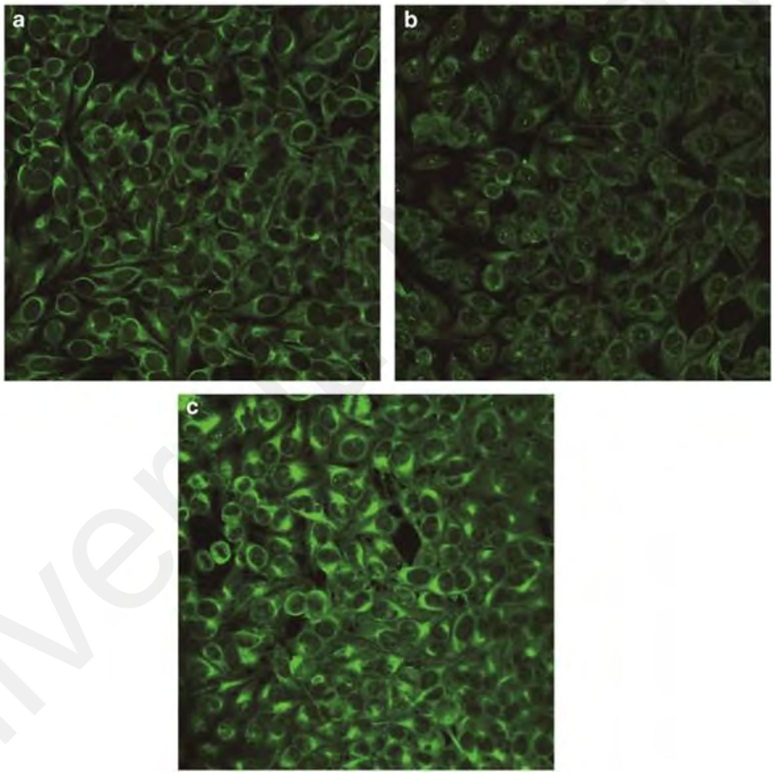


Figure 2.2: Immunofluorescence analysis of Hyper IgM patients' sera in HEP-2 cells. (a) Typical rim staining pattern. (b) Nuclear dot staining pattern. (c) Rim- and dot-staining pattern. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: [Springer Nature] (Barbouche et al., 2018), copyright(2018)).

2.4 Treatment of PID

PID is very rare and is comprised of many different types. Due to that, treating PID patients comes with many difficulties (Yong et al., 2010). One of the main treatments of PID include immunoglobulin replacement therapy, the procedure depends on the nature of the mutation in the gene (McCusker & Warrington, 2011).

2.4.1 Immunoglobulin Replacement Therapy

Above 50% of the types of PIDs include antibody deficiency (Kobrynski, 2012) which is why Immunoglobulin replacement therapy has been one of the standard treatments for PID patients.

Immunoglobulin replacement therapy has been administered in three different injection methods that include intramuscular, intravenous, and subcutaneous.

Intramuscular injection involves the procedure whereby the injection delivery goes widely rooted within a muscle that allows the substance to be absorbed faster by blood vessels.

Intravenous injection is the injection whereby the needle is inserted straight into the vein and the delivery goes directly into the bloodstream.

Subcutaneous injection involves the strategy whereby the substance is transferred into the tissues that lie between the skin and muscle.

In the initial stage of discovering the immunoglobulin replacement therapy (around the 1950s), the injection method was intramuscular, whereby a weekly dose of immunoglobulin is administered to treat the PID patients. This remained the standard method until the intravenous method of immunoglobulin injection was adopted around the 1980s due to the observed improvement of results when the intravenous method was adopted. Although this method had negative side effects, certain agents were added to reduce the effects (Skoda-Smith et al., 2010).

The intravenous method was used since the 1980s, the subcutaneous method was considered a second choice when the effects of the intravenous method were intolerable. However, scientists have been experimenting with the subcutaneous method and found it to be a more feasible method than the intravenous one. However certain PID patients may suffer from negative side effects from the subcutaneous method that they do not in the intravenous method (Skoda-Smith et al., 2010).

2.5 History of NGS

DNA sequencing technologies have been developed to tackle many diseases including PID. It started in 1977 when Fredric Sanger and Walter Gilbert developed the Sanger sequencing technology. Sanger sequencing, ever since, has been the major standard for sequencing DNA (Sanger et al., 1977). In recent years, NGS has started and it was proven a useful technique to sequence the DNA and look for mutations that may be the reasons behind certain diseases. NGS is cheaper and faster than Sanger sequencing. One of the main drawbacks of NGS is that it is less accurate than Sanger sequencing. However, it still produces significant results, and the NGS findings such as genetic variants are often confirmed by the Sanger sequencing of the genetic variation site. The prominent applications of NGS include WGS, WES, and RNA-Seq. Figure 2.3 illustrates a general pipeline used for NGS.

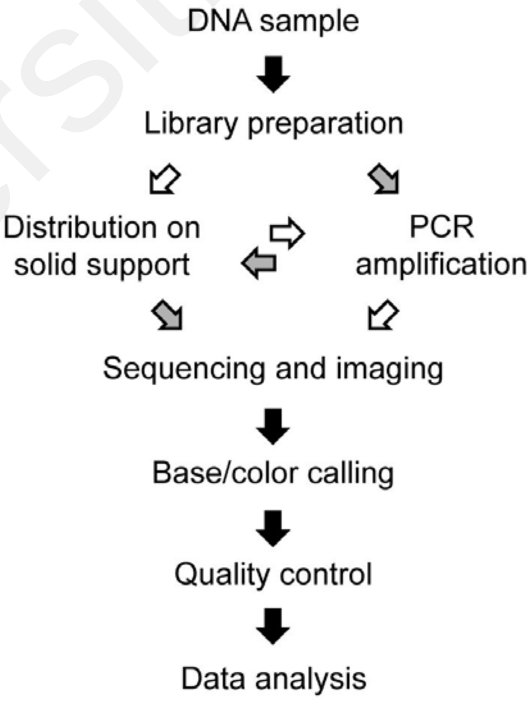


Figure 2.3: A typical NGS pipeline (Berglund et al., 2011).

2.6 Application of WES in Tackling PID

WES has been a cutting edge technology that enabled us to detect genetic variations that could be related to PIDs. Thanks to WES, there has been an outburst in the discovery of novel gene defects that are related to PID (Conley & Casanova, 2014). Figure 2.4 demonstrates the skyrocketing of the discovery of PID-related genetic mutations.

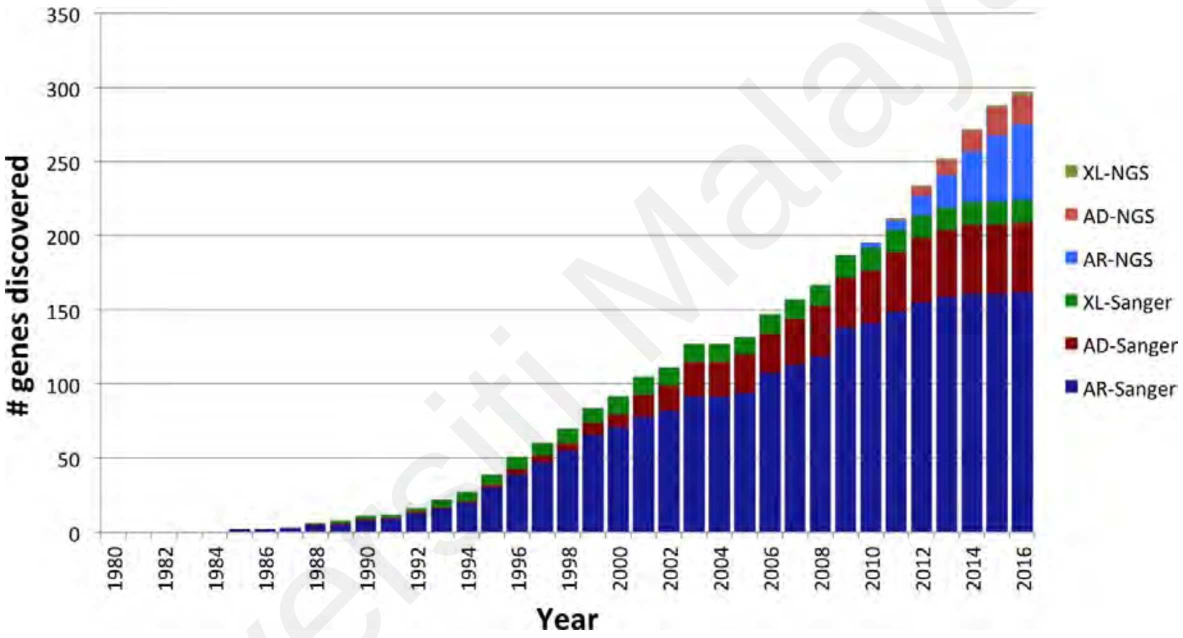


Figure 2.4: The skyrocketing of the discovery of genetic mutations related to PIDs (especially when WES has been applied) (Meyts et al., 2016).

The raw data which is generated by WES undergo mapping, post-processing, variant calling, and annotation of variants will be one of the final steps to detect the cause of PID (Rudilla et al., 2019).

2.7 WES Limitations

Despite the cost and time efficiency of WES, it still comes with limitations. Theoretically speaking, the WES procedure should capture 100% of the exome. However, this is not the case, as the WES technology is still not advanced enough to capture 100% of the exonic regions. Research has demonstrated that WGS is more reliable than WES in terms of looking for variants, as WGS identified more variants and came with fewer errors than WES (Belkadi et al., 2015).

Another limitation emerges from the fact that not every time causative variants can be detected from the WES analysis. This may be caused by the fact that WGS offers more accurate detections of variants as it does not have a bias caused by probe sequences from WES. As a result, WES will miss some variants (Warr et al., 2015). Sometimes, the causative variants may be located in non-coding regions of the DNA.

After a WES procedure is complete, a Bioinformatics analysis pipeline would take place and that would require strong computational power. Assembling a computer with a powerful processor and memory capacity is quite expensive (Schmidt & Hildebrandt, 2017). This may hinder labs that do not have sufficient funds from performing analyses.

CHAPTER 3: METHODOLOGY

3.1 Whole-Exome Sequencing

The DNA was seized from the peripheral blood mononuclear cells (PBMC) of the patient via the QIamp DNA Blood Mini Kit (Qiagen, Germany). Agilent SureSelect Human All Exon V5 was the exome capture kit that was used to attain the coding sequences from the DNA. After that, the DNA was sequenced using the Illumina Hiseq 4000 machine.

3.2 Acquisition of Data

The raw data of the P1 patient was acquired, in FASTQ (Appendix B) format, from IMR for research. The patient was suspected of PID. PID symptoms include (but are not limited to): Inflammation, recurrent infections, blood disorders, and delayed growth and development. P1 presented symptoms that include dysmorphic features and more as mentioned above in the introduction.

The raw data were subjected to the bioinformatics analysis pipeline, which will be explained further in detail, to look for harmful mutations that are the likely cause of the patient's symptoms.

3.3 Computer Specifications

Bioinformatics analyses usually require strong specifications in the computer that will be used. These strong computers are referred to as supercomputers. Table 3.1 presents the specifications of the computer that was used for this project.

Table 3.1: The specifications of the computer that was used for this research.

Hardware and Software	Specifications
Memory	94.4 GiB
Processor	Intel® Xeon(R) CPU X5550 @ 2.67GHz x 4
Graphics	Quadro FX 3800/PCIe/SSE2
OS type	64-bit
Disk	882.9 GB
Operating System	Ubuntu 16.04 LTS

3.4 Bioinformatics Analysis Pipeline

Trimmomatic (Bolger et al., 2014) was used to trim off 10 bases from the reads to improve their quality as that removes unnecessary overlaps that include low qualities. Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) was utilized to align the reads to the human reference genome. File processing via Picard (http://broadinstitute.github.io/picard) and Genome Analysis Toolkit (GATK) 4.0 (McKenna et al., 2010) was done to reduce errors and increase the confidence score in the data. Next, the Mapping percentage, and reads genomic origin have been viewed using Qualimap (García-Alcalde et al., 2012). Then, variant calling was done also via GATK 4.0 to call the SNPs and insertion-deletions (INDEL) and list them in a variant call file (VCF) file (Appendix B). After that, wANNOVAR (the web version of ANNOVAR) (Chang & Wang, 2012) was used for the annotation of the VCF file to retrieve the names of genes with either SNPs or INDELs and get the specifications of each. Figure 3.1 illustrates the pipeline of the Bioinformatics analysis.

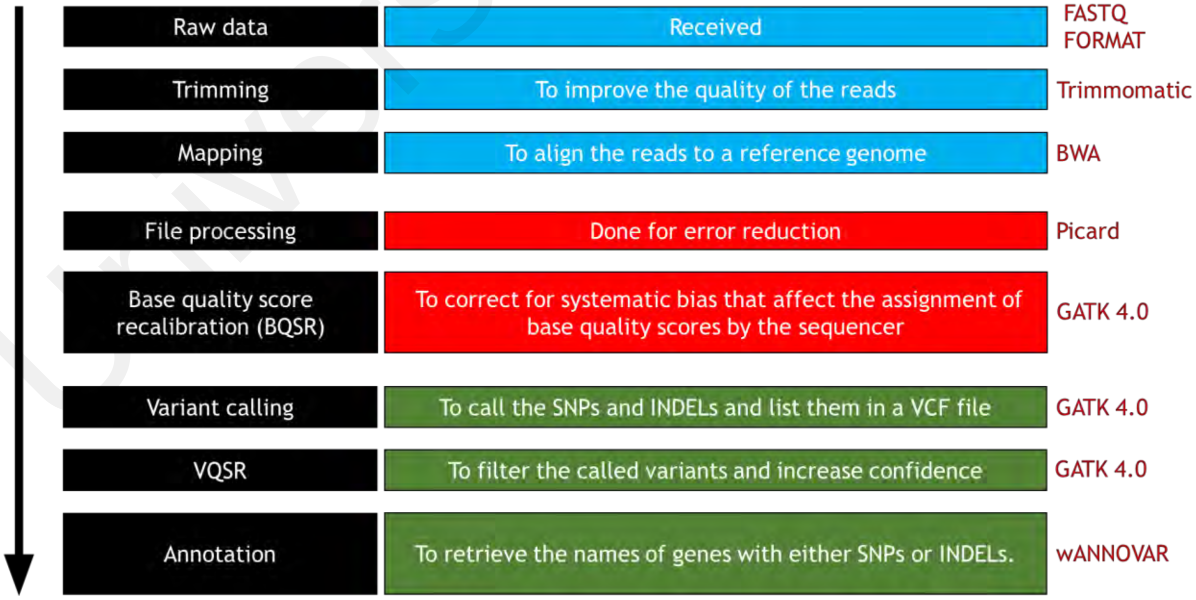

Raw data
Received
FASTQ FORMAT
Trimming
To improve the quality of the reads
Trimmomatic
Mapping
To align the reads to a reference genome
BWA
File processing
Done for error reduction
Picard
Base quality score recalibration (BQSR)
To correct for systematic bias that affect the assignment of base quality scores by the sequencer
GATK 4.0
Variant calling
To call the SNPs and INDELs and list them in a VCF file
GATK 4.0
VQSR
To filter the called variants and increase confidence
GATK 4.0
Annotation
To retrieve the names of genes with either SNPs or INDELs.
wANNOVAR

Figure 3.1: The flow of the Bioinformatics analysis pipeline.

3.5 The Reference Genomes Used for Mapping

A reference genome is an assembled database of nucleic acid sequences. It is a digital representation of the genome of a certain organism. In this project, we are dealing with the human genome. When NGS analysis is applied, mostly, one of the initial steps is mapping the raw reads generated by the sequencer to a reference genome. Four different versions of the human reference genome have been used in the bioinformatics analysis pipeline. The patient's data have been mapped to hg19, hs37d5 (hg19 plus decoy sequences) codenamed hg19D, hg38, and GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set (hg38 without alternate sequences, plus decoy sequences) codenamed hg38NAD.

3.6 Commands and Scripts that are Used to Run the Software Packages

The script for trimming the raw reads via Trimmomatic was written as:

```
{java -jar /Path/to/trimmomatic-0.39.jar PE -phred33
/Path/to/WES_1.fastq.gz
/Path/to/WES_2.fastq.gz
/Path/to/WES_1_paired.fq.gz
/Path/to/WES_1_unpaired.fq.gz
/Path/to/WES_2_paired.fq.gz
/Path/to/WES_2_unpaired.fq.gz
HEADCROP:10}
```

whereby,

- java -jar /Path/to/trimmomatic-0.39.jar denotes the directory in which Trimmomatic software was located;
- PE indicates that Trimmomatic will work with paired strands;
- phred33 (Appendix B) specifies the base quality encoding;
- /Path/to/WES_1.fastq.gz denotes the path to the forward pair input;
- /Path/to/WES_2.fastq.gz denotes the path to the reverse pair input;

- /Path/to/WES_1_paired.fq.gz denotes the directory of the output of the trimmed forward pair;

- /Path/to/WES_1_unpaired.fq.gz denotes the directory of the output of the unpaired reads of the forward pair;

- /Path/to/WES_2_paired.fq.gz denotes the directory of the output of the trimmed reverse pair;

- /Path/to/WES_2_unpaired.fq.gz denotes the directory of the output of the unpaired reads of the reverse pair;

- HEADCROP:10 indicated that 10 bases were removed from the beginning regardless of the quality;

The script for indexing the reference genome via SAMtools was written as:

```
{/Path/To/samtools faidx /Path/To/hg.fa}
```

whereby,

- /Path/To/samtools denotes the directory in which SAMtools is located;

- faidx is the command which will perform the indexing of a reference sequence (the human genome in this case);

- /Path/To/hg.fa denotes the path to the human reference genome file.

The script for creating sequence dictionary for the reference genome via Picard was written as:

```
{java -jar /Path/To/picard.jar CreateSequenceDictionary R=/Path/To/hg.fa
O=/Path/To/hg.dict}
```

whereby,

- java -jar /Path/To/picard.jar denotes the directory in which Picard is located;

- CreateSequenceDictionary is the command that creates a dictionary for a reference sequence (the human reference genome in this case), which may be required for other tools;

- R=/Path/To/hg.fa denotes the directory of the human reference genome;

- O=/Path/To/hg.dict denotes the directory of the output file.

The script for mapping via BWA was written as:

```
{/Path/to/bwa mem -M -t 4 /Path/to/hg.fa /Path/to/WES_1_paired.fq.gz
/Path/to/WES_2_paired.fq.gz > /Path/to/WES_Mapped.sam}
```

whereby,

- /Path/To/bwa denotes the directory in which BWA is located;
- mem refers to an algorithm for local alignment;
- -M mark shorter split hits as secondary (for Picard compatibility);
- -t 4 indicated that the number of threads used is 4;
- /Path/To/hg.fa denotes the directory in which the human reference genome is located;
- /Path/To/WES_1_paired.fq.gz **and** /Path/To/WES_2_paired.fq.gz denotes the path to the forward and reverse pair respectively;
- /Path/to/WES_Mapped.sam denotes the directory for the output file.

The script for converting the SAM (Appendix B) mapping output to a BAM (Appendix B) output via SAMtools was written as:

```
{/Path/To/samtools     view -h -b -S
/Path/To/WES_Mapped.sam >
/Path/To/WES_Mapped.bam}
```

whereby,

- /Path/To/samtools denotes the directory in which SAMtools is located;
- view is a command to convert the mapping file to a different format according to the chosen options;
- -h includes the header in the output;
- -b indicates that the output will be in BAM format;
- -S indicates that the input is a SAM file;
- /Path/To/WES_Mapped.sam denotes the directory of the SAM output file that will be converted to BAM;
- /Path/To/WES_Mapped.bam denotes the directory of the BAM output file.

The script for replacing read groups via Picard was written as:

```
{java -jar /Path/To/picard.jar AddOrReplaceReadGroups
I=/Path/To/WES_Mapped.bam o=/Path/To/RG_WES_Mapped.bam RGID=1
RGLB=library1   RGPL=illumina   RGPU=K00171   RGSM=human }
```

whereby,

- java -jar /Path/To/picard.jar denotes the directory in which Picard is located;
- AddOrReplaceReadGroups is the command for replacing read groups in the input file with a one whole new read group in the output file which contains the read groups from the input;
- I=/Path/To/WES_Mapped.bam denotes the directory of the input file;
- o=/Path/To/RG_WES_Mapped.bam denotes the directory of the output file;
- RGID=1 indicates the read group ID;
- RGLB=library1 indicates the choice of the required library;
- RGPL=illumina indicates the required platform;
- RGPU=K00171 indicates the platform unit;
- RGSM=human indicates the group sample name.

The script for sorting the mapping file by coordinates via Picard was written as:

```
{ java -jar /Path/To/picard.jar SortSam
I=/Path/To/RG_WES_Mapped.bam
O=/Path/To/SS_RG_WES_Mapped.bam
SORT_ORDER=coordinate}
```

whereby,

- java -jar /Path/To/picard.jar denotes the directory in which Picard is located;

- SortSam is a command used for sorting SAM or BAM files;

- I=/Path/To/RG_WES_Mapped.bam denotes the directory of the input file;

- O=/Path/To/SS_RG_WES_Mapped.bam denotes the directory of where the output file will be saved;

- SORT_ORDER=coordinate indicates that the input will be sorted by coordinates;

The script for marking duplicates via Picard was written as:

```
{ java -jar /Path/To/picard.jar   MarkDuplicates
I=/Path/To/SS_RG_WES_Mapped.bam
O=/Path/To/MD_SS_RG_WES_Mapped.bam
M=/Path/To/WES_Mapped_Marked.txt}
```

whereby,

- java -jar /Path/To/picard.jar denotes the directory in which Picard is located;

- MarkDuplicates is the command which is used for the identification of duplicate reads. The duplicate reads are errors that may arise from polymerase chain reaction (PCR) library construction;

- I=/Path/To/SS_RG_WES_Mapped.bam denotes the directory of the input file;

- O=/Path/To/MD_SS_RG_WES_Mapped.bam denotes the directory of the output file;

- M=/Path/To/WES_Mapped_Marked.txt denotes the directory of the output where the duplication metrics will be written.

The script for base recalibration via GATK 4.0 was written as:

```
{/Path/To/gatk BaseRecalibrator
-I /Path/To/MD_SS_RG_WES_Mapped.bam
-R /Path/To/hg.fa
--known-sites /Path/To/00-All.vcf
-O /Path/To/WES_Mapped_Marked.bam.table}
```

whereby,

- /Path/To/gatk denotes the directory in which GATK 4.0 is located;

- BaseRecalibrator is the command which is used for generating recalibration tables for Base Quality Score Recalibration (BQSR), which is the next script;

- -I /Path/To/MD_SS_RG_WES_Mapped.bam denotes the directory of the input file;

- -R /Path/To/hg.fa denotes the directory in which the human reference genome is located;

- --known-sites /Path/To/00-All.vcf denotes the directory of the file that contains all known polymorphic sites according to the reference genome;

- -O /Path/To/WES_Mapped_Marked.bam.table denotes the directory of the output file.

The script for applying BQSR via GATK 4.0 was written as:

```
{/Path/To/gatk ApplyBQSR
-R /Path/To/hg.fa
-I /Path/To/MD_SS_RG_WES_Mapped.bam
--bqsr-recal-file /Path/To/WES_Mapped_Marked.bam.table
-O /Path/To/WES_BQSR.bam}
```

whereby,

- /Path/To/gatk denotes the directory in which GATK 4.0 is located;
- ApplyBQSR is the command which used for applying base quality score recalibration whose aim is to amend systemic bias of the sequencer which affects the appointment of quality scores for the bases;
- -R /Path/To/hg.fa denotes the directory in which the human reference genome is located;
- -I /Path/To/MD_SS_RG_WES_Mapped.bam denotes the directory of the input file;
- --bqsr-recal-file /Path/To/WES_Mapped_Marked.bam.table denotes the directory of the input table;
- -O /Path/To/WES_BQSR.bam denotes the directory of the output file.

The script for gene counting via the Python htseq-count (Anders et al., 2015) command was written as:

```
{htseq-count    -f bam    -r pos    -s no    -m union
/Path/To/WES_BQSR.bam
/Path/To/hg.knownGene.gtf    >>    /Path/To/GeneCounting_WES.txt}
```

whereby,

- htseq-count denotes the command which is used for counting reads in features;
- -f bam denotes that the format of the input file was BAM.
- -r pos   denotes that the data was sorted by position of alignment.
- -s no denotes that it was not strand-specific.
- -m union denotes that the mode used was union.
- /Path/To/WES_BQSR.bam denotes the directory of the input file.
- /Path/To/hg.knownGene.gtf   denotes the directory of the GTF file (Appendix B).
- >>   /Path/To/GeneCounting_WES.txt   denotes the directory of the output file.

The script for calling germline variants via GATK 4.0 was written as:

```
{/Path/To/gatk   --java-options "-Xmx4g"   HaplotypeCaller
-R /Path/To/hg.fa
-I /Path/To/WES_BQSR.bam
-O /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf}
```

whereby,

- /Path/To/gatk denotes the directory in which GATK 4.0 is located;

- --java-options "-Xmx4g" denotes that the amount of utilized RAM was 4GB;

- HaplotypeCaller is the command that is used to call germline variants (SNPs and INDELs) from the mapped and processed BAM file;

- -R /Path/To/hg.fa denotes the directory in which the human reference genome is located;

- -I /Path/To/WES_BQSR.bam denotes the input file directory;

- -O /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf denotes the output file directory.

The script for generating a variant recalibration model via GATK 4.0 was written as:

```
{/Path/To/gatk   --java-options "-Xmx40g"   VariantRecalibrator
--use-jdk-deflater true   --use-jdk-inflater true
-R /Path/To/hg.fa   -V /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf
--resource:hapmap,known=false,training=true,truth=true,prior=15.0
ResourseFiles/hapmap_3.3.hg.vcf
--resource:omni,known=false,training=true,truth=false,prior=12.0
ResourseFiles/1000G_omni2.5.hg.vcf
--resource:1000G,known=false,training=true,truth=false,prior=10.0
ResourseFiles/1000G_phase1.snps.high_confidence.hg.vcf
--resource:dbsnp,known=true,training=false,truth=false,prior=2.0
ResourseFiles/dbsnp.hg.vcf
-an QD   -an MQ   -an MQRankSum   -an ReadPosRankSum   -an FS   -an SOR
-mode SNP
-O /Path/To/HC_WES_BQSR.bam.vcf.recal
--tranches-file /Path/To/HC_WES_BQSR.bam.vcf.tranches}
```

whereby,

- /Path/To/gatk denotes the directory in which GATK 4.0 is located;

- --java-options "-Xmx40g" denotes that the amount of utilized RAM was 40GB;

- VariantRecalibrator is the command that is used for the goal of appointing a calibrated probability to each variant call in a call set for the purpose of filtering the variants with higher accuracy;

- --use-jdk-deflater true denotes the use of a Java class;

- --use-jdk-inflater true denotes the use of a Java class;

- -R /Path/To/hg.fa denotes the directory in which the human reference genome is located;

- -V /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf denotes the input directory of the VCF file;

- --resource:hapmap,known=false,training=true,truth=true,prior=15.0 ResourseFiles/hapmap_3.3.hg.vcf denotes a VCF sites file for probability calculation purposes and the input directory of the file;

- --resource:omni,known=false,training=true,truth=false,prior=12.0 ResourseFiles/1000G_omni2.5.hg.vcf denotes a VCF sites file for probability calculation purposes and the input directory of the file;

- --resource:1000G,known=false,training=true,truth=false,prior=10.0 ResourseFiles/1000G_phase1.snps.high_confidence.hg.vcf denotes a VCF sites file for probability calculation purposes and the input directory of the file;

- --resource:dbsnp,known=true,training=false,truth=false,prior=2.0 ResourseFiles/dbsnp.hg.vcf denotes a VCF sites file for probability calculation purposes and the input directory of the file;

- -an QD denotes a respective name of annotation in the recalibration model;

- -an MQ denotes a respective name of annotation in the recalibration model;

- -an MQRankSum denotes a respective name of annotation in the recalibration model;

- -an ReadPosRankSum denotes a respective name of annotation in the recalibration model;

- -an FS denotes a respective name of annotation in the recalibration model;

- -an SOR denotes a respective name of annotation in the recalibration model;

- -mode SNP denotes the mode of recalibration;

- -O /Path/To/HC_WES_BQSR.bam.vcf.recal denotes the output of the recalibration model file;

- --tranches-file /Path/To/HC_WES_BQSR.bam.vcf.tranches denotes the directory of the tranches output file.

The script for applying the VQSR based on the recalibration and tranches files via GATK 4.0 was written as:

```
{/Path/To/gatk --java-options "-Xmx40g" ApplyVQSR -R /Path/To/hg.fa
-V /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf
-O /Path/To/WES_VQSR.vcf     -ts-filter-level 99.0
--tranches-file /Path/To/HC_WES_BQSR.bam.vcf.tranches
--recal-file /Path/To/HC_WES_BQSR.bam.vcf.recal
--mode SNP}
```

whereby,

- /Path/To/gatk denotes the directory in which GATK 4.0 is located;

- --java-options "-Xmx40g" denotes that the amount of utilized RAM was 40GB;

- ApplyVQSR is the command that is used for applying filtering of called variants in accordance with the recalibration model;

- -R /Path/To/hg.fa denotes the directory in which the human reference genome is located;

- -V /Path/To/HaplotypeCaller_WES_BQSR.bam.vcf denotes the input VCF file directory;

- -O /Path/To/WES_VQSR.vcf denotes the output file directory;

- -ts-filter-level 99.0 denotes the truth sensitivity level;

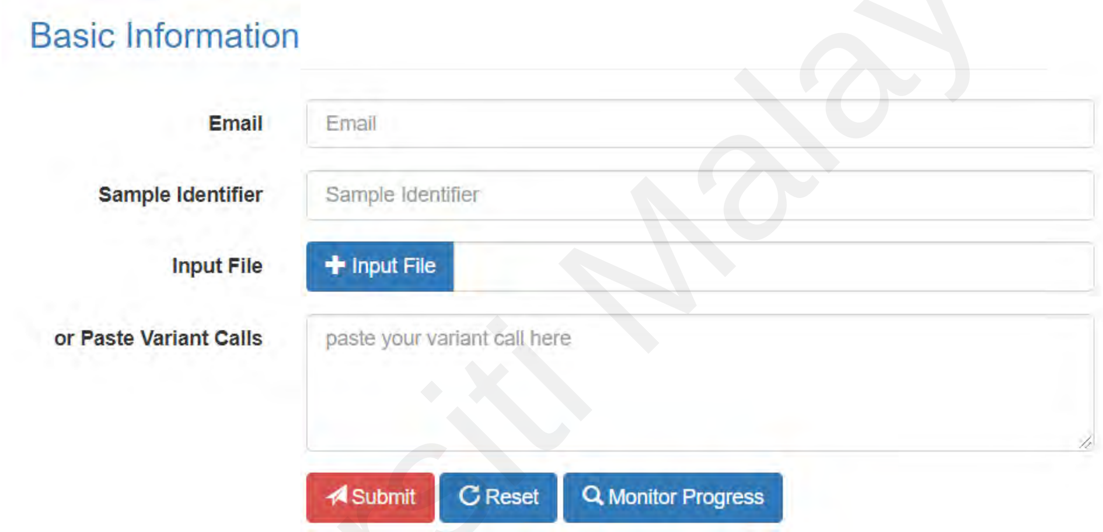
- --tranches-file /Path/To/HC_WES_BQSR.bam.vcf.tranches denotes the directory of the input tranches file;

- --recal-file /Path/To/HC_WES_BQSR.bam.vcf.recal denotes the directory of the input recalibration model file;

- --mode SNP denotes the mode of recalibration.

3.7 Annotation of the Called and Recalibrated Variants

The recalibrated VCF file (WES_VQSR.vcf) has been annotated via the web tool wANNOVAR. Figure 3.2 illustrates a part of the webpage whereby the email is inputted to receive the results, the sample identifier would be the name of the output file, and the input file would be inserted in the third slot.



Basic Information

Email: Email

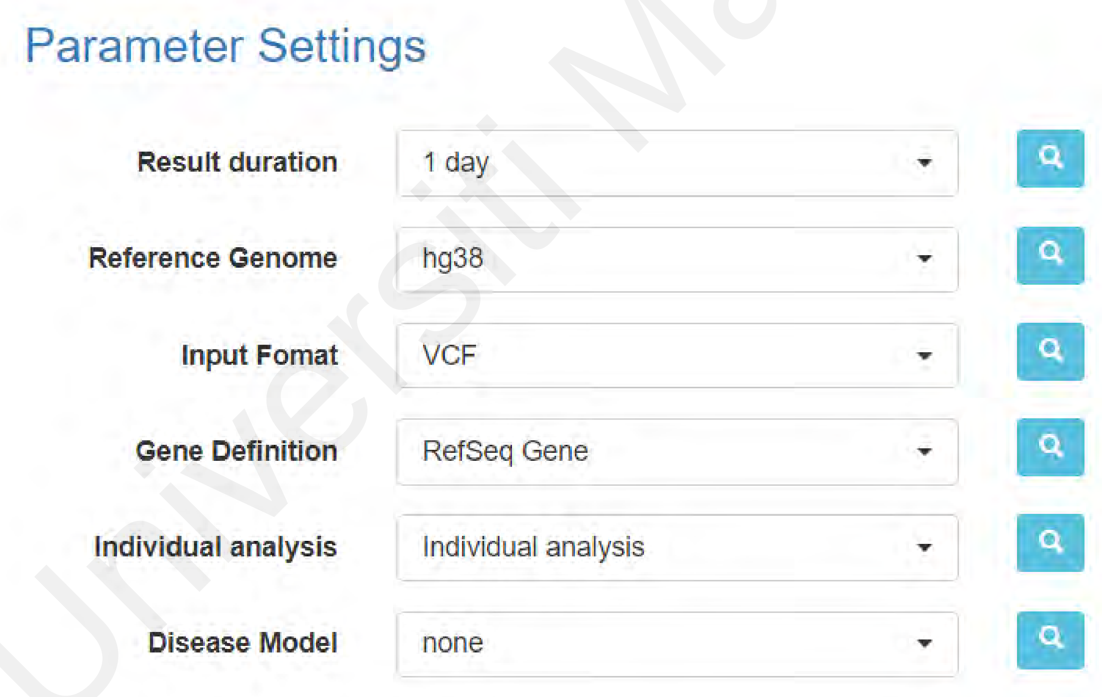
Sample Identifier: Sample Identifier

Input File: + Input File

or Paste Variant Calls: paste your variant call here

Figure 3.2: Illustration of the input setup of wANNOVAR.

The parameter settings illustrated in Figure 3.3 determine the filtering of annotation. The result duration denotes how many days the result would last on the website before it is deleted. The reference genome was selected according to the reference genome the data was mapped to. The input format was set to VCF. The gene definition would determine the database of which the annotation will be based on. The individual analysis determines that only the variants that are present in the VCF file would be considered for annotation. The disease model would do filtering to find variants that are only related to a certain disease, however, none was selected so that it allows the selection of all variants for annotation.



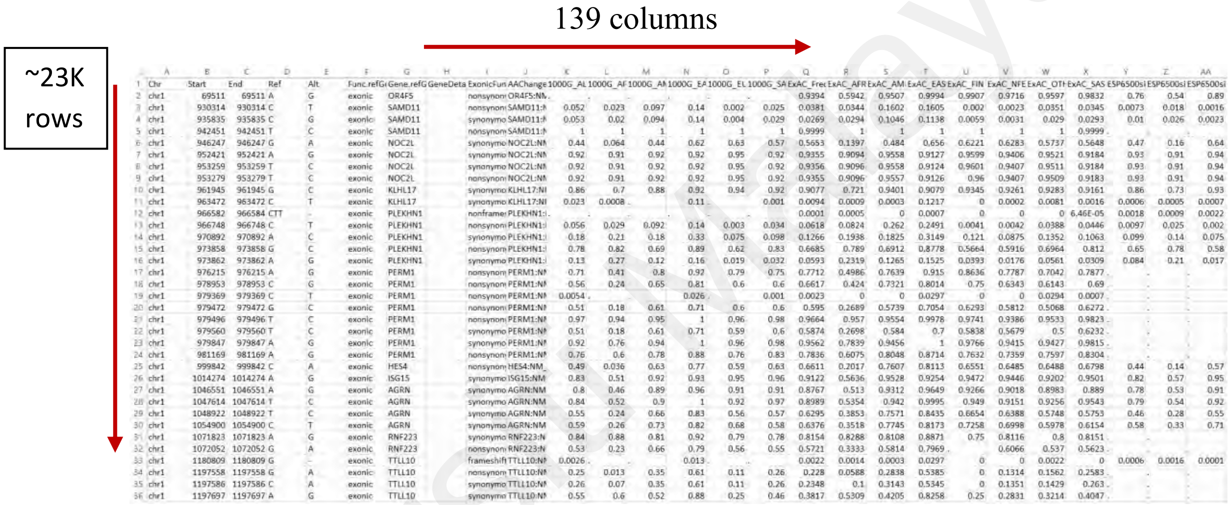
Parameter Settings

Result duration	1 day
Reference Genome	hg38
Input Fomat	VCF
Gene Definition	RefSeq Gene
Individual analysis	Individual analysis
Disease Model	none

Figure 3.3: The parameter settings of wANNOVAR.

3.8 Interpretation of the Variants' Annotation

After the result generation from wANNOVAR, a file in CSV format can be downloaded or viewed on the website itself. The downloaded file shows 139 columns, each portraying a different kind of annotation, and approximately 23,000 rows each representing an annotated variant in a particular gene. Figure 3.4 illustrates how the CSV file looks like once opened.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Chr	Start	End	Ref	Alt	Func.refG	Gene.refG	GeneDeta	ExonicFun	AAChange	1000G_AL	1000G_AF	1000G_AM	1000G_EA	1000G_EU	1000G_SA	ExAC_Frec	ExAC_AFR	ExAC_AM	ExAC_EAS	ExAC_FIN	ExAC_NFE	ExAC_OTH	ExAC_SAS	ESP6500si	ESP6500si	ESP6500si
2	chr1	69511	69511	A	G	exonic	OR4F5		nonsynon	OR4F5:NM.	.	.	.	.	.	.	0.9394	0.5942	0.9507	0.9994	0.9907	0.9716	0.9597	0.9832	0.76	0.54	0.89
3	chr1	930314	930314	C	T	exonic	SAMD11		nonsynon	SAMD11:N	0.052	0.023	0.097	0.14	0.002	0.025	0.0381	0.0344	0.1602	0.1605	0.002	0.0023	0.0351	0.0345	0.0073	0.018	0.0016
4	chr1	935835	935835	C	G	exonic	SAMD11		synonymo	SAMD11:N	0.053	0.02	0.094	0.14	0.004	0.029	0.0269	0.0294	0.1046	0.1138	0.0059	0.0031	0.029	0.0293	0.01	0.026	0.0023
5	chr1	942451	942451	T	C	exonic	SAMD11		nonsynon	SAMD11:N	1	1	1	1	1	1	0.9999	1	1	1	1	1	1	0.9999	.	.	.
6	chr1	946247	946247	G	A	exonic	NOC2L		synonymo	NOC2L:NM	0.44	0.064	0.44	0.62	0.63	0.57	0.5653	0.1397	0.484	0.656	0.6221	0.6283	0.5737	0.5648	0.47	0.16	0.64
7	chr1	952421	952421	A	G	exonic	NOC2L		synonymo	NOC2L:NM	0.92	0.91	0.92	0.92	0.95	0.92	0.9355	0.9094	0.9558	0.9127	0.9599	0.9406	0.9521	0.9184	0.93	0.91	0.94
8	chr1	953259	953259	T	C	exonic	NOC2L		synonymo	NOC2L:NM	0.92	0.91	0.92	0.92	0.95	0.92	0.9356	0.9096	0.9558	0.9124	0.9601	0.9407	0.9511	0.9184	0.93	0.91	0.94
9	chr1	953279	953279	T	C	exonic	NOC2L		nonsynon	NOC2L:NM	0.92	0.91	0.92	0.92	0.95	0.92	0.9355	0.9096	0.9557	0.9126	0.96	0.9407	0.9509	0.9183	0.93	0.91	0.94
10	chr1	961945	961945	G	C	exonic	KLHL17		synonymo	KLHL17:NI	0.86	0.7	0.88	0.92	0.94	0.92	0.9077	0.721	0.9401	0.9079	0.9345	0.9261	0.9283	0.9161	0.86	0.73	0.93
11	chr1	963472	963472	C	T	exonic	KLHL17		synonymo	KLHL17:NI	0.023	0.0008	.	0.11	.	0.001	0.0094	0.0009	0.0003	0.1217	0	0.0002	0.0081	0.0016	0.0006	0.0005	0.0007
12	chr1	966582	966584	CTT	-	exonic	PLEKHN1		nonframe	PLEKHN1:I.	.	.	.	.	.	.	0.0001	0.0005	0	0.0007	0	0	0	6.46E-05	0.0018	0.0009	0.0022
13	chr1	966748	966748	C	T	exonic	PLEKHN1		nonsynon	PLEKHN1:I	0.056	0.029	0.092	0.14	0.003	0.034	0.0618	0.0824	0.262	0.2491	0.0041	0.0042	0.0388	0.0446	0.0097	0.025	0.002
14	chr1	970892	970892	A	C	exonic	PLEKHN1		synonymo	PLEKHN1;I	0.18	0.21	0.18	0.33	0.075	0.098	0.1266	0.1938	0.1825	0.3149	0.121	0.0875	0.1352	0.1063	0.099	0.14	0.075
15	chr1	973858	973858	G	C	exonic	PLEKHN1		nonsynon	PLEKHN1:I	0.78	0.82	0.69	0.89	0.62	0.83	0.6685	0.789	0.6912	0.8778	0.5664	0.5916	0.6964	0.812	0.65	0.78	0.58
16	chr1	973862	973862	A	G	exonic	PLEKHN1		synonymo	PLEKHN1:I	0.13	0.27	0.12	0.16	0.019	0.032	0.0593	0.2319	0.1265	0.1525	0.0393	0.0176	0.0561	0.0309	0.084	0.21	0.017
17	chr1	976215	976215	A	G	exonic	PERM1		nonsynon	PERM1:NM	0.71	0.41	0.8	0.92	0.79	0.75	0.7712	0.4986	0.7639	0.915	0.8636	0.7787	0.7042	0.7877	.	.	.
18	chr1	978953	978953	C	G	exonic	PERM1		nonsynon	PERM1:NM	0.56	0.24	0.65	0.81	0.6	0.6	0.6617	0.424	0.7321	0.8014	0.75	0.6343	0.6143	0.69	.	.	.
19	chr1	979369	979369	C	T	exonic	PERM1		nonsynon	PERM1:NM	0.0054	.	.	0.026	.	0.001	0.0023	0	0	0.0297	0	0	0.0294	0.0007	.	.	.
20	chr1	979472	979472	G	C	exonic	PERM1		nonsynon	PERM1:NM	0.51	0.18	0.61	0.71	0.6	0.6	0.595	0.2689	0.5739	0.7054	0.6293	0.5812	0.5068	0.6272	.	.	.
21	chr1	979496	979496	T	C	exonic	PERM1		nonsynon	PERM1:NM	0.97	0.94	0.95	1	0.96	0.98	0.9664	0.957	0.9554	0.9978	0.9741	0.9386	0.9533	0.9823	.	.	.
22	chr1	979560	979560	T	C	exonic	PERM1		synonymo	PERM1:NM	0.51	0.18	0.61	0.71	0.59	0.6	0.5874	0.2698	0.584	0.7	0.5838	0.5679	0.5	0.6232	.	.	.
23	chr1	979847	979847	A	G	exonic	PERM1		synonymo	PERM1:NM	0.92	0.76	0.94	1	0.96	0.98	0.9562	0.7839	0.9456	1	0.9766	0.9415	0.9427	0.9815	.	.	.
24	chr1	981169	981169	A	G	exonic	PERM1		nonsynon	PERM1:NM	0.76	0.6	0.78	0.88	0.76	0.83	0.7836	0.6075	0.8048	0.8714	0.7632	0.7359	0.7597	0.8304	.	.	.
25	chr1	999842	999842	C	A	exonic	HES4		nonsynon	HES4:NM_	0.49	0.036	0.63	0.77	0.59	0.63	0.6611	0.2017	0.7607	0.8113	0.6551	0.6485	0.6488	0.6798	0.44	0.14	0.57
26	chr1	1014274	1014274	A	G	exonic	ISG15		synonymo	ISG15:NM	0.83	0.51	0.92	0.93	0.95	0.96	0.9122	0.5636	0.9528	0.9254	0.9472	0.9446	0.9202	0.9501	0.82	0.57	0.95
27	chr1	1046551	1046551	A	G	exonic	AGRN		synonymo	AGRN:NM	0.8	0.46	0.89	0.96	0.91	0.91	0.8767	0.513	0.9312	0.9649	0.9266	0.9018	0.8983	0.889	0.78	0.53	0.91
28	chr1	1047614	1047614	T	C	exonic	AGRN		synonymo	AGRN:NM	0.84	0.52	0.9	1	0.92	0.97	0.8989	0.5354	0.942	0.9995	0.949	0.9151	0.9256	0.9543	0.79	0.54	0.92
29	chr1	1048922	1048922	T	C	exonic	AGRN		synonymo	AGRN:NM	0.55	0.24	0.66	0.83	0.56	0.57	0.6295	0.3853	0.7571	0.8435	0.6654	0.6388	0.5748	0.5753	0.46	0.28	0.55
30	chr1	1054900	1054900	C	T	exonic	AGRN		synonymo	AGRN:NM	0.59	0.26	0.73	0.82	0.68	0.58	0.6376	0.3518	0.7745	0.8173	0.7258	0.6998	0.5978	0.6154	0.58	0.33	0.71
31	chr1	1071823	1071823	A	G	exonic	RNF223		synonymo	RNF223:N	0.84	0.88	0.81	0.92	0.79	0.78	0.8154	0.8288	0.8108	0.8871	0.75	0.8116	0.8	0.8151	.	.	.
32	chr1	1072052	1072052	G	A	exonic	RNF223		nonsynon	RNF223:N	0.53	0.23	0.66	0.79	0.56	0.55	0.5721	0.3333	0.5814	0.7969	.	0.6066	0.537	0.5623	.	.	.
33	chr1	1180809	1180809	G	-	exonic	TTLL10		frameshif	TTLL10:NM	0.0026	.	.	0.013	.	.	0.0022	0.0014	0.0003	0.0297	0	0	0.0022	0	0.0006	0.0016	0.0001
34	chr1	1197558	1197558	G	A	exonic	TTLL10		nonsynon	TTLL10:NM	0.25	0.013	0.35	0.61	0.11	0.26	0.228	0.0588	0.2838	0.5385	0	0.1314	0.1562	0.2583	.	.	.
35	chr1	1197586	1197586	C	A	exonic	TTLL10		synonymo	TTLL10:NM	0.26	0.07	0.35	0.61	0.11	0.26	0.2348	0.1	0.3143	0.5345	0	0.1351	0.1429	0.263	.	.	.
36	chr1	1197697	1197697	A	G	exonic	TTLL10		synonymo	TTLL10:NM	0.55	0.6	0.52	0.88	0.25	0.46	0.3817	0.5309	0.4205	0.8258	0.25	0.2831	0.3214	0.4047	.	.	.

Figure 3.4: The raw annotation file in CSV format.

Many of the columns that included irrelevant information were disregarded. Only the columns with the imperative information were regarded. This included the annotation of variants in all the known genes. In other words, the rows included variations in all the known genes. The PID-related genes were listed on the side. Using the VLOOKUP function in excel, the rows were reduced to the point of showing only the variants in PID-related genes. Further filtering was done to determine which of the variants could be the harmful and causative ones for the symptoms in the patient.

CHAPTER 4: RESULTS

4.1 Trimming Results

FastQC (Andrews, 2010) was utilized to check the quality scores of the reads. In the figures below (4.1a, 4.1b, 4.1c, and 4.1d), the y-axis represents the quality scores. The higher the better quality. There are three colors in the background of the graph. Green for very good quality, yellow for reasonable quality, and red for poor quality.

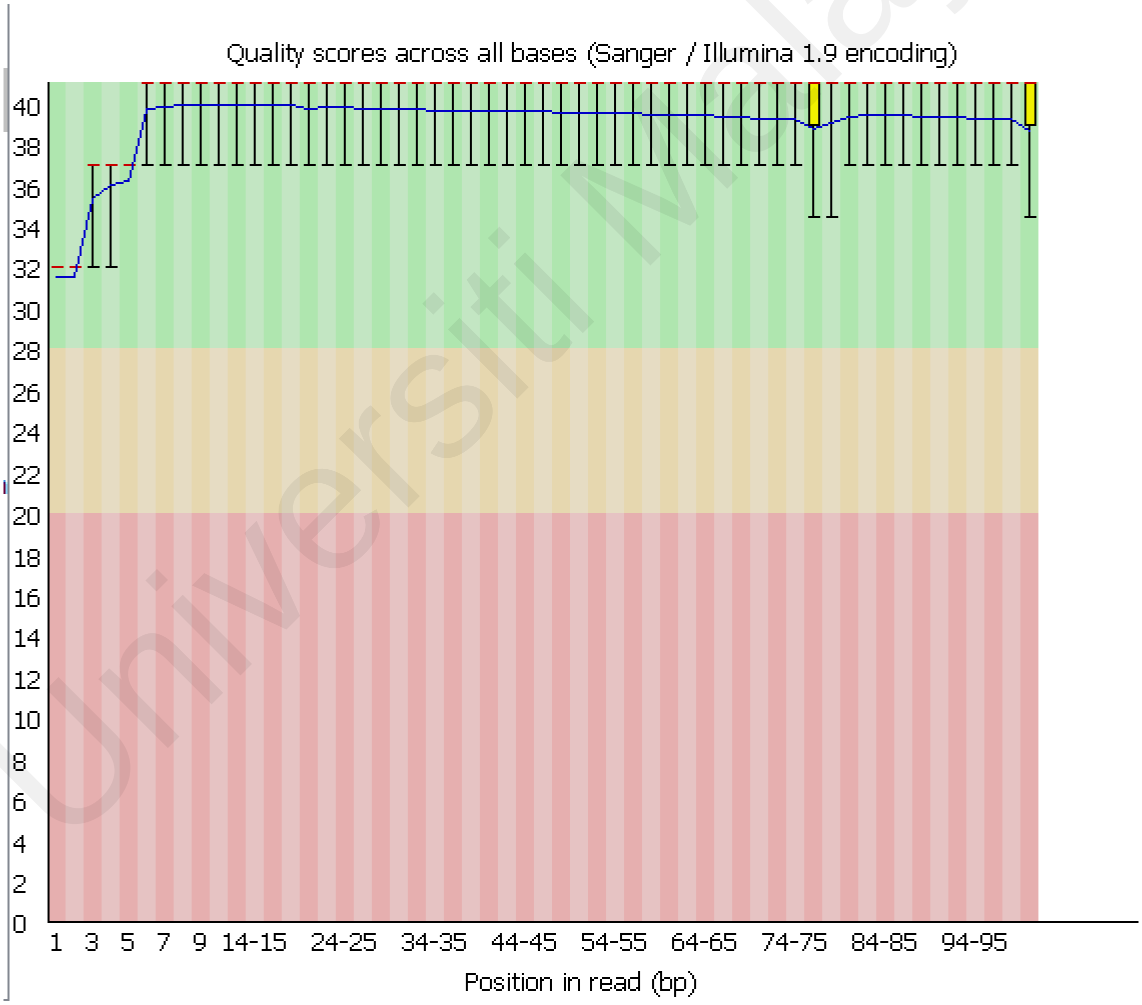


Figure 4.1a: Forward pair before trimming.

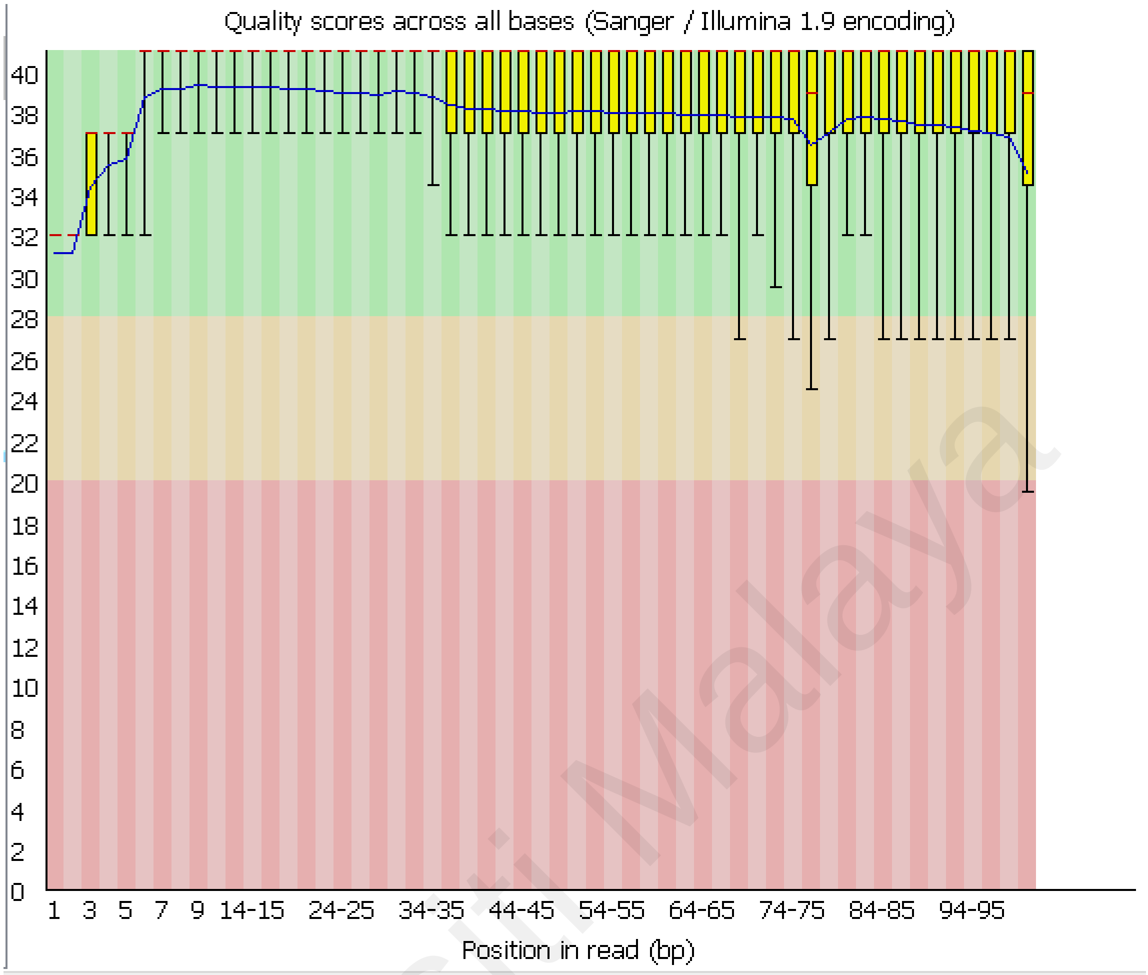


Figure 4.1b: Reverse pair before trimming.

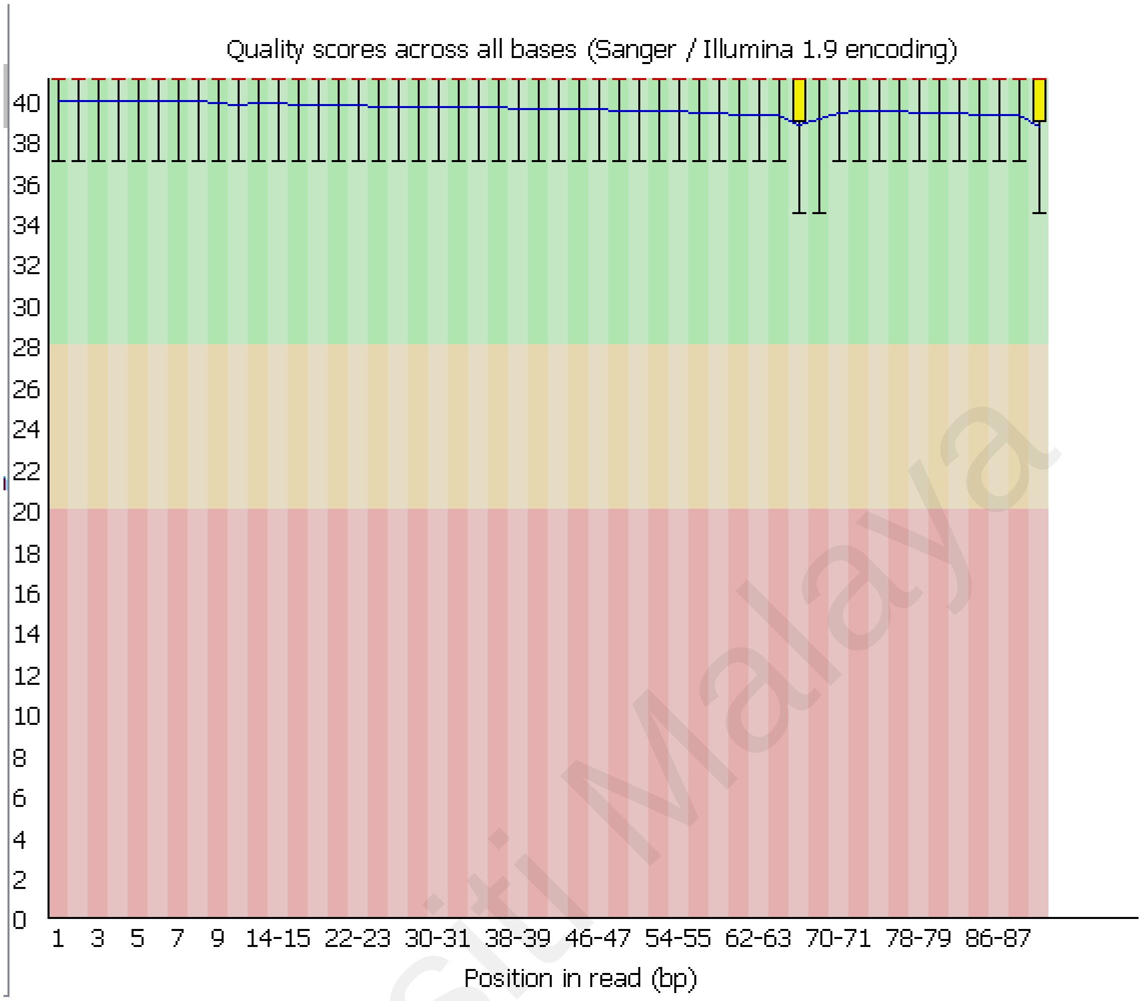


Figure 4.1c: Forward pair after trimming.

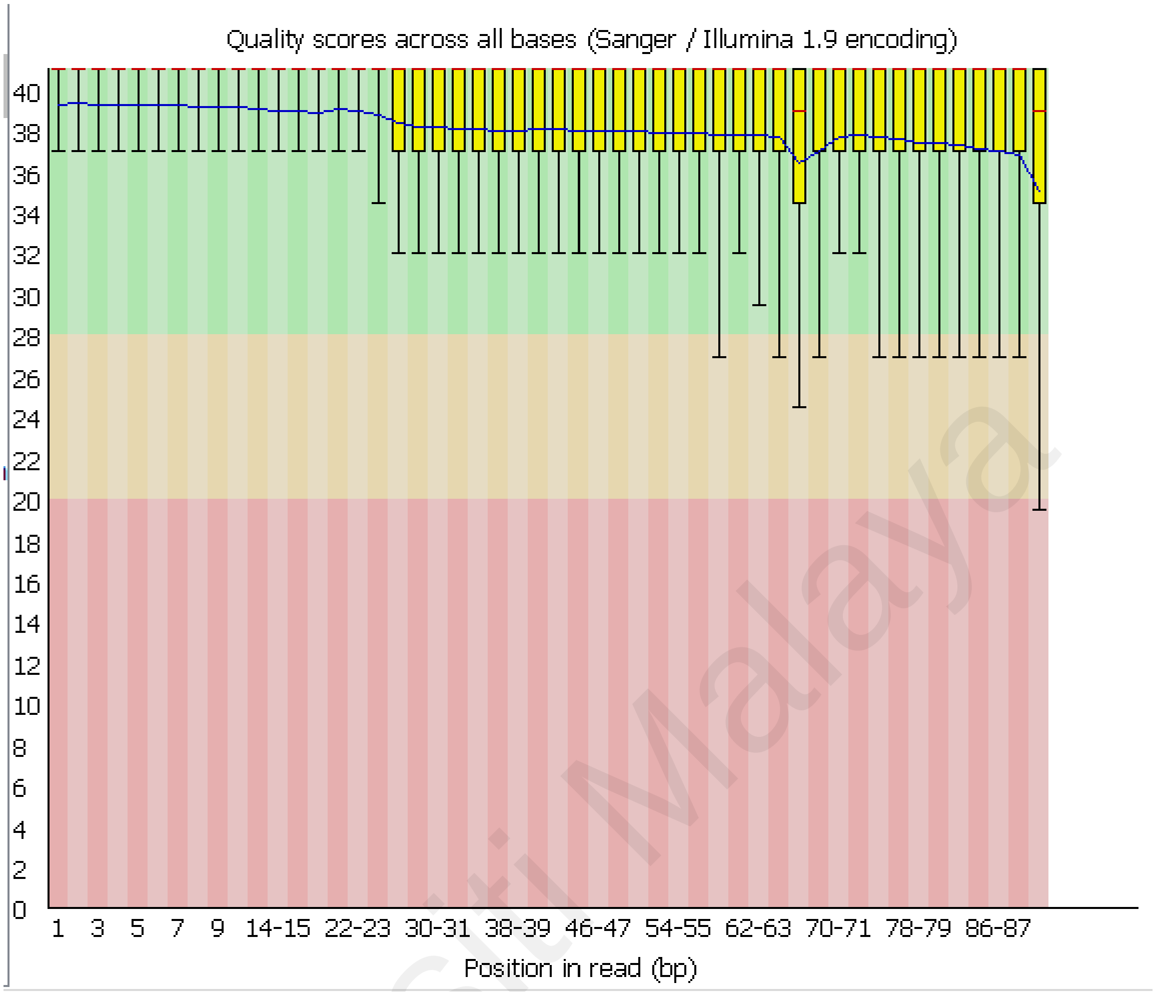


Figure 4.1d: Reverse pair after trimming.

Figures 4.1a and 4.1b illustrate the forward and reverse pairs before trimming 10 bases off the head, the first 10 bases of the reads with a slightly lower quality were still present (at the top left of the graphs). Figures 4.1c and 4.1d, on the other hand, show the forward and reverse pairs after trimming 10 bases off the head, the first 10 bases of the reads with a slightly lower quality were removed.

4.2 General Statistics

The results of the bioinformatics analysis pipeline have been compared when the patient's data were mapped to four different versions of reference genomes based on four categories: mapping percentage, reads genomic origin, gene counting, and the number of annotated variants. Green-labeled cells indicate the highest number, yellow-labeled cells indicate mid-range numbers, and red-labeled cells indicate the lowest number.

4.2.1 Mapping Percentage

Tables 4.1, 4.2, 4.3, and 4.4 describe the percentage of reads that mapped to the respective reference genome. As shown in the tables, when the raw data were mapped to the 4 different versions of the human reference genome, it is shown that all of them have a high mapping percentage, meaning that the utmost majority of the reads have mapped to the respective reference genome version. However, hg38NAD had the highest percentage.

Table 4.1: Mapping statistics of P1's data when mapped to hg19.

Number of reads	**66,476,469 100%**
Mapped reads	**66,391,615 99.87%**
Unmapped reads	**84,854 0.13%**

Table 4.2: Mapping statistics of P1's data when mapped to hg19D.

Number of reads	**66,473,788 100%**
Mapped reads	**66,399,287 99.89%**
Unmapped reads	**74,501 0.11%**

Table 4.3: Mapping statistics of P1’s data when mapped to hg38.

Number of reads	**66,475,603 100%**
Mapped reads	**66,398,985 99.88%**
Unmapped reads	**76,618 0.12%**

Table 4.4: Mapping statistics of P1’s data when mapped to hg38NAD.

Number of reads	**66,474,294 100%**
Mapped reads	**66,399,030 99.89%**
Unmapped reads	**75,264 0.11%**

4.2.2 Reads Genomic Origins

Tables 4.5, 4.6, 4.7, and 4.8 describe the percentage of reads that mapped to the exonic regions of the respective reference genome. The exonic row represents the percentage of reads that were mapped to exonic regions. When the raw data were mapped to hg38NAD, they had the biggest percentage of reads mapped to exonic regions.

Table 4.5: Reads genomic origins of P1’s data when mapped to hg19.

Exonic	22.5%
Intronic	66.3%
Intergenic	11.2%
Intronic/Intergenic overlapping exon	24.84%

Table 4.6: Reads genomic origins of P1's data when mapped to hg19D.

Exonic	24.34%
Intronic	64.77%
Intergenic	10.88%
Intronic/Intergenic overlapping exon	24.69%

Table 4.7: Reads genomic origins of P1's data when mapped to hg38.

Exonic	30.3%
Intronic	60.87%
Intergenic	8.82%
Intronic/Intergenic overlapping exon	24.16%

Table 4.8: Reads genomic origins of P1's data when mapped to hg38NAD.

Exonic	30.36%
Intronic	60.96%
Intergenic	8.68%
Intronic/Intergenic overlapping exon	24.2%

4.2.3 Gene Counting

Tables 4.9, 4.10, 4.11, and 4.12 describe the number of reads that are mapped to the respective reference genome. Reads aligned to features (genes) denotes the total reads that have been mapped to genes. No feature denotes reads that could not be mapped to genes. Ambiguous denotes reads that have been mapped to more than one feature and hence were not counted. Too low aQual denotes reads that had been skipped due to the default quality parameter. Not aligned denotes the reads that have not been aligned in the BAM file. The most reads aligned to features were found in hg38NAD.

Table 4.9: Gene counting results for P1 when mapped to hg19.

Reads aligned to features (genes)	**24,181,342**
No feature	**7,367,541**
Ambiguous	**637,823**
Too low aQual	**1,031,006**
Not aligned	**7,757**

Table 4.10: Gene counting results for P1 when mapped to hg19D.

Reads aligned to features (genes)	**24,574,409**
No feature	**7,108,943**
Ambiguous	**745,238**
Too low aQual	**793,316**
Not aligned	**3,563**

Table 4.11: Gene counting results for P1 when mapped to hg38.

Reads aligned to features (genes)	**24,293,915**
No feature	**5,411,241**
Ambiguous	**1,656,135**
Too low aQual	**1,859,829**
Not aligned	**4,349**

Table 4.12: Gene counting results for P1 when mapped to hg38NAD.

Reads aligned to features (genes)	**25,077,795**
No feature	**5,562,397**
Ambiguous	**1,725,631**
Too low aQual	**855,785**
Not aligned	**3,861**

4.2.4 Number of Annotated Exonic Variants

Table 4.13 describes the number of annotated exonic variants when the data of P1 was mapped to the respective reference genome. When the raw data were mapped to hg19, it was found that it had the biggest number of annotated exonic variants.

Table 4.13: Number of annotated exonic variants of P1's data when mapped to the four different versions of the human reference genome.

When mapped to hg19	24,141
When mapped to hg19D	23,524
When mapped to hg38	22,382
When mapped to hg38NAD	23,790

4.3 Variants Annotation

Table 4.14 shows the annotation of the variants which are the candidates for the causation of the symptom in P1. The Gene column represents the gene in which the mutation occurred. The Chr column represents the chromosome in which the mutation occurred. The Nucleotide change column shows the location of the SNP in the gene. The Inheritance column represents the known inheritance of the mutation in the gene. The Genotype column shows whether the mutation is heterozygous (het) or homozygous (hom). The Variant impact column shows whether it is a synonymous single nucleotide variant (SNV) or a nonsynonymous SNV. The AA change column represents the location of the mutation in the amino acid chain. SIFT, Polyphen, and Mutation Taster are the predictor tools that show whether the mutation is Damaging/Disease-causing (D) or not. The Frequency in gnomAD column represents the frequency of occurrence of the said mutation in the database. The coverage column portrays how many times the mutation has been covered during the sequencing.

Table 4.14: Annotation data presenting the SNPs in the three genes and the three prediction tools' scores for each SNP.

Gene	Chr	Nucleotide change	Inheritance	Genotype	Variant impact	AA change	SIFT	Polyphen	Mutation Taster	Frequency in gnomAD	Coverage
NFKBIA	14	c.A94T	AD	het	nonsynonymous SNV	p.Ser32Cys	D	D	D	.	66,56:122
TLR3	4	c.C2384T	AD or AR	het	nonsynonymous SNV	p.A795V	D	D	D	0.00005692	36,26:62
SAMD9L	7	c.T866C	AD	het	nonsynonymous SNV	p.F289S	D	D	D	0.018	102,93:195

4.4 Sanger Sequencing Results and Confirmation

Figure 4.2 illustrates the Sanger sequencing results and Figure 4.3 presents the PCR result which confirms the Sanger sequencing.

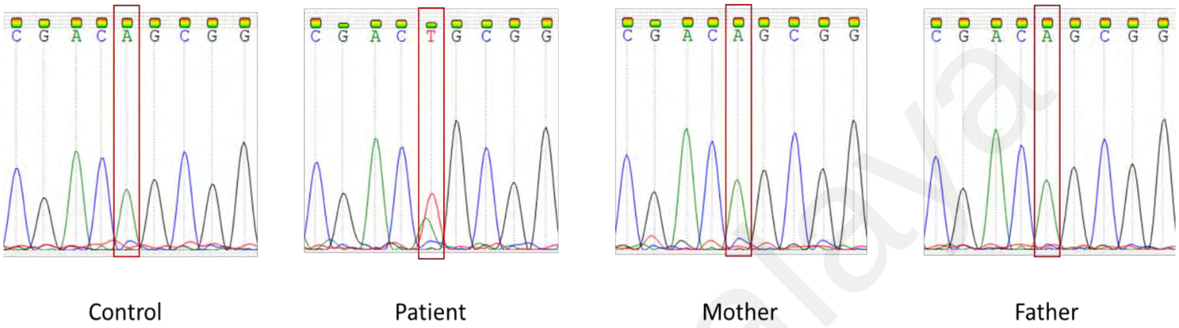


Figure 4.2: Sanger sequencing validation confirms the mutation of the Adenine base to a Thymine base (c.A94T) in the patient's NFKBIA gene that was detected by WES analysis. The mutation point is indicated by red rectangles. The control and both of the parents have the wildtype allele at the same point.

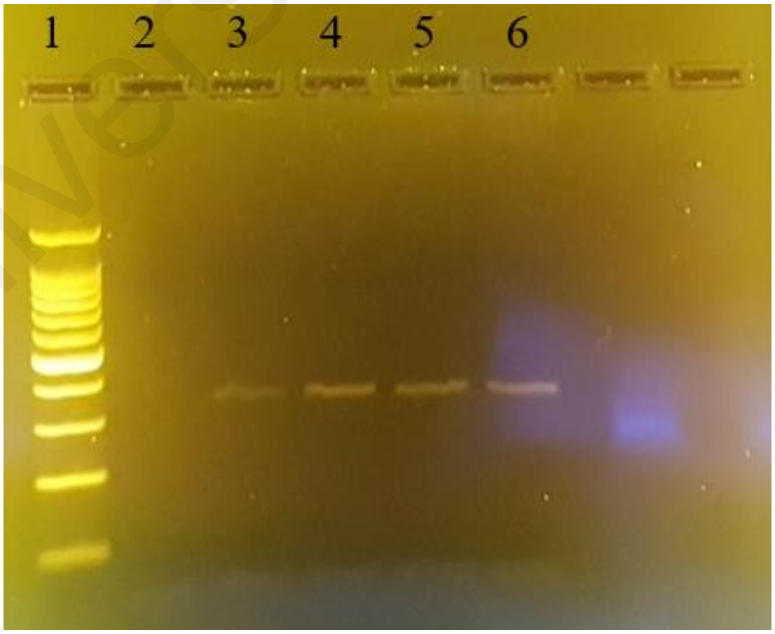


Figure 4.3: PCR result confirms the Sanger sequencing. Lane 1 is the BenchTop DNA ladder (100bp). Lane 2 is the negative control. Lane 3 is the control. Lane 4 is the patient's DNA. Lane 5 is the patient's mother's DNA. Lane 6 is the patient's father's DNA.

CHAPTER 5: DISCUSSION

5.1 Importance of the Study

The main objective of this study is to detect genetic variants that could be linked to the patient’s symptoms. The significance of WES analysis is that it often leads to the discovery of novel mutations that have not been reported before. This allows for detailed inspection to further our understanding of PIDs.

5.2 Trimming

To achieve the objective of detecting genetic variants, the raw reads must be mapped to align them to a reference genome. However, before mapping the reads, quality control was conducted to ensure a high quality of the reads. The results in figures (4.1a, 4.1b, 4.1c, and 4.1d) showed that the quality of the raw reads is high before and after the trimming since most of them lie in the green region in FastQC.

5.3 The Best Reference Genome for this Study

The mapping percentage appeared to be the highest when the samples were mapped to hg38NAD; reads genomic origins showed the highest percentage of exonic regions when mapped to hg38NAD; gene counting showed the highest number of reads mapped to features when the samples were mapped to hg38NAD; the number of annotated variants were the highest when the samples mapped to hg19. After the comparison of the four different versions of the human reference genome, it was found that hg38NAD had won in three categories out of the four. Thus, making it the most suitable human reference version for this Bioinformatics analysis pipeline.

5.4 Candidate Genes

After going through and filtering the data of the annotated variant calls, three variants in the NFKBIA, TLR3, and SAMD9L genes were detected as shown in Table 4.14. The NFKBIA variant has damaging scores from SIFT and Polyphen, and it was predicted to be disease-causing by MutationTaster. The NFKBIA variant was not reported in the gnomAD database, hence the nil frequency is shown in the annotation. Which tells us that this is a novel mutation. It is a heterozygous autosomal dominant mutation. The annotation shows that Adenine changed to Thymine at the location of the variant (A>T), which made the codon code for Cysteine instead of Serine (p.Ser32Cys). The WES coverage of this mutation was 122. This mutation was validated via Sanger sequencing and the results were visualized using Unipro UGENE (Okonechnikov et al., 2012). This confirms the mutation to be the most possible and the biggest candidate behind the patient's symptoms. The SNPs in the TLR3 and SAMD9L genes were considered improbable candidates as none of the associated features with the mutations were found in the patient.

5.5 Role of NFKB

Nuclear factor-kB (NFKB) is a transcription factor that is pervasive in terms of regulating genes that encode cytokines, chemokines, growth factors, cell adhesion molecules, and more (Chen et al., 1999). NFKB is comprised of a family of associated transcription factors that includes five genes NFKB1, NFKB2, RelA, c-Rel, and RelB.

One of NFKB's functions is regulating the immune response to infections. If the NFKB complex is damaged (due to a harmful mutation), a severe form of PID would be caused in the patient (Klemann et al., 2019).

The regulation of NFKB is highly important as faulty regulation can cause cancer or immune deficiency. One of the main regulators of NFKB is an inhibitor known as NFKBIA.

5.6 Role of NFKBIA

The NFKBIA gene codes for an inhibitor that regulates the expression of the NFKB gene. While inactive, NFKB is bound to NFKBIA and is in the cytoplasm. The NFKBIA Inhibits the NFKB complex by cornering the REL dimers in the cytoplasm. As a result, this will conceal the nuclear localization signals. When it is activated; due to certain triggers such as cytokines, chemokines, immune responses, and more; the NFKBIA is unbound by becoming phosphorylated which results in its ubiquitination and degradation. The RelA, in turn, is translocated into the nucleus to activate the transcription and perform the NFKB's function (Scherer et al., 1995; Barnes & Karin, 1997).

5.7 Hyper IgM

The patient was initially suspected of Hyper IgM syndrome. Hyper IgM syndrome is a PID in which the patient will have elevated IgM levels and reduced IgA and IgG levels. It is associated with a harmful mutation in the CD40 gene. NFKBIA mutations are also associated with elevated IgM levels and reduced IgA and IgG levels (Tangye et al., 2020). However, the immunological workup of the patient was observed and was found not to be compatible with Hyper IgM syndrome because the clinicians observed normal levels of CD40 expressions in the patient. Hence, the WES analysis was conducted to look for mutations in all of the known PID-related genes. As a result, a mutation (SNP) in the NFKBIA gene of the patient was detected and will be discussed further. No harmful mutation was found in any of Hyper IgM syndrome-related genes.

5.8 Novel Mutation in the NFKBIA Gene

In the case of P1, the NFKB complex is not damaged. However, the NFKBIA, one of the inhibitors of NFKB was damaged by an SNP (NFKBIA: NM_020529: exon1: c.A94T: p.S32C). The determined SNP in the NFKBIA gene is a non-synonymous, heterozygous mutation. No SNP has been detected before in that particular site (c.A94T) in the gnomAD database.

Since NFKBIA is an inhibitor for NFKB, if NFKBIA is mutated in a damaging manner, two possible scenarios will be present. Either the inhibitor is mutated in a way that prevents it from binding to NFKB to a point that NFKB loses control and overexpresses itself, or the inhibitor is mutated in a way the keeps it overly bound to NFKB to a point that NFKB activation is impaired and it will not be expressed when required. This raises the following question: which one of the previously mentioned scenarios occurred in the patient? The answer to that question is most probably the second scenario. According to research, the abnormal constitutive activation of NFKB is usually associated with specific cancers depending on the site of mutation (White et al., 2009; Bredel et al., 2011; Zhao et al., 2014). However, when it is a gain-of-function (GOF) mutation (as it is apparent in the P1), it is usually associated with anhidrotic ectodermal dysplasia (EDA) and PID (also as it appears in P1) (Lopez-Granados et al., 2008; Yoshioka et al., 2013; Sogkas et al., 2020).

5.9 Destruction Motif in NFKBIA

The novel mutation (NFKBIA:NM_020529:exon1:c.A94T:p.S32C) in the NFKBIA is located in the destruction motif of NFKBIA (Specifically the Serine 32 residue). The destruction motif is important for the phosphorylation in regulatory proteins (Wu et al., 2003). However, this Serine 32 residue mutated to a Cysteine in the patient.

5.10 Function of Serine32 in NFKBIA

As mentioned previously in the results, the Serine 32 mutated into a Cysteine in the patient. This prompts us to discuss some of the functions of Serine and cysteine in general, the function of Serine 32 in the inhibitor (NFKBIA), and what a Cysteine 32 mutation could do (since it was not reported before), to give more confirmation for the second scenario in the patient.

One of the Serine's important functions is that it gets phosphorylated to facilitate cell signaling. On the other hand, Cysteine is involved in the formation of disulfide bridges for crosslinking proteins.

In a normal NFKBIA protein, Serine 32 is imperative for the protein’s stability and the activation of NFKB (Traenckner et al., 1995). As the inhibitor is removed from the NFKB gene to activate it, the Serine 32 is phosphorylated for the ubiquitination of the NFKBIA protein which urges its degradation and results in the nuclear translocation of the NFKB protein to do its following function in the nucleus (Kawai et al., 2012).

In the case of the P1’s NFKBIA protein. The Serine 32 is mutated to a Cysteine. This means there will be no phosphorylation of that residue in the patient, which may lead to the difficulty of the inhibitor to get removed from the NFKB and get it activated, which confirms the second scenario in the patient instead of the first scenario.

Since this mutation (p.Ser32Cys) is not reported before, we cannot predict what Cysteine 32 is exactly going to do in the patient’s inhibitor protein and how it would cause the inhibitor (NFKBIA) to react with NFKB. This requires protein modeling and simulation to be done to investigate and predict how a Cysteine 32 in NFKBIA would function.

CHAPTER 6: CONCLUSION

Our knowledge in the field of PID diseases is ambiguous. Hence, this study was conducted to investigate the case of a PID patient in Malaysia. The patient was initially suspected of Hyper IgM Syndrome. The immunological workup was not compatible with Hyper IgM according to the observation. Hence, the WES analysis was conducted to look for mutations in the PID-related genes. The patient's data were mapped to four different versions of the human reference genome (hg19, hg19D, hg38, and hg38NAD). Each time when the data is mapped to the respective version of the human reference genome, the data underwent file processing, BQSR, gene counting, variant calling, VQSR, and annotation. hg38NAD won in three out of four categories in comparison. Hence, it was concluded to be the best reference genome to use in this Bioinformatics analysis pipeline. After interpreting the annotation results, a novel harmful mutation (SNP) has been detected in the NFKBIA gene of the patient (NFKBIA:NM_020529:exon1:c.A94T:p.S32C), an important regulator (inhibitor) of the NFKB gene. This mutation has not been detected before as it was never reported in databases that concern genetic mutations. The Serine32 residue, which lies in the destruction motif of NFKBIA, has mutated to a Cysteine residue. This could have led to the inability of NFKBIA to detach itself from the NFKB gene when required, which resulted in the impairment of the NFKB gene's activation, which could have also lead to the symptoms of the patient. This warrants further investigation to determine what this particular mutation exactly does.

REFERENCES

Abd Hamid, I. J., Azman, N. A., Hashim, I. F., Mangantig, E., Gennery, A. R., & Zainudeen, Z. T. (2020). Systematic review of Primary Immunodeficiency Diseases in Malaysia: 1979–2020. *Frontiers in Immunology*, *11*, Article#1923.

Adams, D. R., & Eng, C. M. (2018). Next-generation sequencing to diagnose suspected genetic disorders. *New England Journal of Medicine*, *379*(14), 1353-1362.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Barbouche, M. R., Chen, Q., Carbone, M., Ben-Mustapha, I., Shums, Z., Trifa, M., ... & Norman, G. L. (2018). Comprehensive review of autoantibodies in patients with hyper-IgM syndrome. *Cellular & Molecular Immunology*, *15*(6), 610-617.

Barnes, P. J., & Karin, M. (1997). Nuclear factor-κB—a pivotal transcription factor in chronic inflammatory diseases. *New England Journal of Medicine*, *336*(15), 1066-1071.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., ... & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*(17), 5473-5478.

Berglund, E. C., Kiialainen, A., & Syvänen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, *2*(1), 23.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bousfiha, A. A., Jeddane, L., Ailal, F., Benhsaien, I., Mahlaoui, N., Casanova, J. L., & Abel, L. (2013). Primary immunodeficiency diseases worldwide: more common than generally thought. *Journal of Clinical Immunology*, *33*(1), 1-7.

Bredel, M., Scholtens, D. M., Yadav, A. K., Alvarez, A. A., Renfrow, J. J., Chandler, J. P., ... & Ferrarese, R. (2011). NFKBIA deletion in glioblastomas. *New England Journal of Medicine*, *364*(7), 627-637.

Bruton, O. C. (1952). Agammaglobulinemia. *Pediatrics*, *9*(6), 722-728.

Chang, X., & Wang, K. (2012). wANNOVAR: annotating genetic variants for personal genomes via the web. *Journal of Medical Genetics*, *49*(7), 433-436.

Chen, F., Castranova, V., Shi, X., & Demers, L. M. (1999). New insights into the role of nuclear factor-κB, a ubiquitous transcription factor in the initiation of diseases. *Clinical Chemistry*, *45*(1), 7-17.

Conley, M. E., & Casanova, J. L. (2014). Discovery of single-gene inborn errors of immunity by next generation sequencing. *Current Opinion in Immunology*, *30*, 17-23.

El-Sayed, Z. A., & Radwan, N. (2020). Newborn screening for primary immunodeficiencies: the gaps, challenges, and outlook for developing countries. *Frontiers in Immunology*, *10*, Article#2987.

Familiärer, W. A. (1937). angeborener morbus Werlhofii. *Monatsschr Kinderheilkd*, *68*, 212-216.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., ... & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678-2679.

Kawai, T., Nishikomori, R., & Heike, T. (2012). Diagnosis and treatment in anhidrotic ectodermal dysplasia with immunodeficiency. *Allergology International*, *61*(2), 207-217.

Klemann, C., Camacho-Ordonez, N., Yang, L., Eskandarian, Z., Rojas-Restrepo, J. L., Frede, N., ... & Seidl, M. (2019). Clinical and immunological phenotype of patients with primary immunodeficiency due to damaging mutations in NFKB2. *Frontiers in Immunology*, *10*, Article#297.

Kobrynski, L. (2012). Subcutaneous immunoglobulin therapy: a new option for patients with primary immunodeficiency diseases. *Biologics: Targets & Therapy*, *6*, Article#277.

Levy, J., Espanol-Boren, T., Thomas, C., Fischer, A., Tovo, P., Bordigoni, P., ... & Sanders, E. A. M. (1997). Clinical spectrum of X-linked hyper-IgM syndrome. *The Journal of Pediatrics*, *131*(1), 47-54.

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

Lopez-Granados, E., Keenan, J. E., Kinney, M. C., Leo, H., Jain, N., Ma, C. A., ... & Jain, A. (2008). A novel mutation in NFKBIA/IKBA results in a degradation-resistant N-truncated protein and is associated with ectodermal dysplasia with immunodeficiency. *Human Mutation*, *29*(6), 861-868.

McCusker, C., & Warrington, R. (2011). Primary immunodeficiency. *Allergy, Asthma & Clinical Immunology*, *7*(S1), S11.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.

Meyts, I., Bosch, B., Bolze, A., Boisson, B., Itan, Y., Belkadi, A., ... & Bossuyt, X. (2016). Exome and genome sequencing for inborn errors of immunity. *Journal of Allergy and Clinical Immunology*, *138*(4), 957-969.

Ochs, H. D., & Hitzig, W. H. (2012). History of primary immunodeficiency diseases. *Current Opinion in Allergy and Clinical Immunology*, *12*(6), 577-587.

Okonechnikov, K., Golosova, O., Fursov, M., & Ugene Team. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166-1167.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*(1), 5-15.

Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., ... & McKnight, D. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*, *18*(7), 696-704.

Rudilla, F., Franco-Jarava, C., Martínez-Gallo, M., Garcia-Prat, M., Martín-Nalda, A., Rivière, J., ... & Irastorza, I. (2019). Expanding the clinical and genetic spectra of primary immunodeficiency-related disorders with clinical exome sequencing: expected and unexpected findings. *Frontiers in Immunology*, *10*, Article#2325.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463-5467.

Scherer, D. C., Brockman, J. A., Chen, Z., Maniatis, T., & Ballard, D. W. (1995). Signal-induced degradation of I kappa B alpha requires site-specific ubiquitination. *Proceedings of the National Academy of Sciences*, *92*(24), 11259-11263.

Schmidt, B., & Hildebrandt, A. (2017). Next-generation sequencing: big data meets high performance computing. *Drug Discovery Today*, *22*(4), 712-717.

Schultz, W. (1922). Über eigenartige Halserkrankungen. *Dtsch Med Wochenschr*, *48*, Article#1495.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16-18.

Sheikhbahaei, S., Sherkat, R., Roos, D., Yaran, M., Najafi, S., & Emami, A. (2016). Gene mutations responsible for primary immunodeficiency disorders: A report from the first primary immunodeficiency biobank in Iran. *Allergy, Asthma & Clinical Immunology*, *12*(1), Article#62.

Skoda-Smith, S., Torgerson, T. R., & Ochs, H. D. (2010). Subcutaneous immunoglobulin replacement therapy in the treatment of patients with primary immunodeficiency disease. *Therapeutics and Clinical Risk Management*, *6*, 1-10.

Sogkas, G., Adriawan, I. R., Ringshausen, F. C., Baumann, U., Schröder, C., Klemann, C., ... & Ernst, D. (2020). A novel NFKBIA variant substituting serine 36 of IκBα causes immunodeficiency with warts, bronchiectasis and juvenile rheumatoid arthritis in the absence of ectodermal dysplasia. *Clinical Immunology*, *210*, Article#108269.

Syllaba, L. (1926). Contribution a I'independance de I'athetose double idiopathique'et congenitale. Atteinte faniliate, syndrome dystrophique, signe du resean vasculaire conjonctival, integrite psychique. *Revue Neurologique*, *1*, 541-562.

Tangye, S. G., Al-Herz, W., Bousfiha, A., Chatila, T., Cunningham-Rundles, C., Etzioni, A., ... & Ochs, H. D. (2020). Human inborn errors of immunity: 2019 update on the classification from the international union of immunological societies expert committee. *Journal of Clinical Immunology*, *40*(1), 24-64.

Traenckner, E. M., Pahl, H. L., Henkel, T., Schmidt, K. N., Wilk, S., & Baeuerle, P. A. (1995). Phosphorylation of human I kappa B-alpha on serines 32 and 36 controls I kappa B-alpha proteolysis and NF-kappa B activation in response to diverse stimuli. *The EMBO Journal*, *14*(12), 2876-2883.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: current and future perspectives. *G3: Genes, Genomes, Genetics*, *5*(8), 1543-1550.

White, K. L., Vierkant, R. A., Phelan, C. M., Fridley, B. L., Anderson, S., Knutson, K. L., ... & Rider, D. N. (2009). Polymorphisms in NF-κB inhibitors and risk of epithelial ovarian cancer. *BMC Cancer*, *9*(1), 170.

Wu, G., Xu, G., Schulman, B. A., Jeffrey, P. D., Harper, J. W., & Pavletich, N. P. (2003). Structure of a β-TrCP1-Skp1-β-catenin complex: destruction motif binding and lysine specificity of the SCFβ-TrCP1 ubiquitin ligase. *Molecular Cell*, *11*(6), 1445-1456.

Yong, P. L., Boyle, J., Ballow, M., Boyle, M., Berger, M., Bleesing, J., ... & Nelson, L. (2010). Use of intravenous immunoglobulin and adjunctive therapies in the treatment of primary immunodeficiencies: a working group report of and study by the Primary Immunodeficiency Committee of the American Academy of Allergy Asthma and Immunology. *Clinical Immunology*, *135*(2), 255-263.

Yoshioka, T., Nishikomori, R., Hara, J., Okada, K., Hashii, Y., Okafuji, I., ... & Yasumi, T. (2013). Autosomal dominant anhidrotic ectodermal dysplasia with immunodeficiency caused by a novel NFKBIA mutation, p. Ser36Tyr, presents with mild ectodermal dysplasia and non-infectious systemic inflammation. *Journal of Clinical Immunology*, *33*(7), 1165-1174.

Zhao, Z., Zhong, X., Wu, T., Yang, T., Chen, G., Xie, X., ... & Du, Z. (2014). Identification of a NFKBIA polymorphism associated with lower NFKBIA protein levels and poor survival outcomes in patients with glioblastoma multiforme. *International Journal of Molecular Medicine*, *34*(5), 1233-1240.