WHOLE EXOME SEQUENCING ANALYSIS OF MALAYSIAN MONOZYGOTIC TWIN SUSPECTED WITH PRIMARY IMMUNODEFICIENCY DISEASES

HAMIDAH BINTI ABDUL GHANI

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

WHOLE EXOME SEQUENCING ANALYSIS OF MALAYSIAN MONOZYGOTIC TWIN SUSPECTED WITH PRIMARY IMMUNODEFICIENCY DISEASES

HAMIDAH BINTI ABDUL GHANI

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

INSTITUTE OF BIOLOGICAL SCIENCES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate : **HAMIDAH BINTI ABDUL GHANI**

I.C/Passport No

Matric No : **SGR160016**

Name of Degree : **MASTER OF SCIENCE**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

WHOLE EXOME SEQUENCING ANALYSIS OF MALAYSIAN MONOZYGOTIC TWIN SUSPECTED WITH PRIMARY IMMUNODEFICIENCY DISEASES

Field of Study :**BIOINFORMATICS AND WHOLE EXOME SEQUENCING**

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:

WHOLE EXOME SEQUENCING ANALYSIS OF MALAYSIAN MONOZYGOTIC TWIN SUSPECTED WITH PRIMARY IMMUNODEFICIENCY DISEASES

ABSTRACT

Primary immunodeficiency diseases (PID) are rare genetic diseases with more than 200 different forms of PID sharing almost similar symptoms. A pair of female Malay descendant monozygotic twin (MZ twin) was presented with recurrent upper respiratory tract infections, bicytopenia associated hepatosplenomegaly, which the clinicians suspected with PID. However, the definite diagnosis was not clear due to the complexity of the disease phenotypes. Whole exome sequencing analysis was applied to identify the disease-causative gene mutations for this MZ twin. A total numbers of 84 and 81 millions of paired-end reads were received with 99.26% and 99.39% were mapped back to the human reference genome hg38. We identified a compound heterozygous case of *CD21* gene at position c.1916G>A and c.2012G>A, suggesting that both patients had *CD21* deficiency. The pathogenic missense mutation were confirmed using Sanger sequencing. The study revealed that both patients suffering common variable immunodeficiency (CVID) due to the deficiency in *CD21* gene. We had also conducted gene network analysis for *CD21* gene to understand the gene-gene relationship among the CVID-related genes.

Keywords: primary immunodeficiency diseases, whole exome sequencing, common variable immunodeficiency disease

ANALISIS PENJUJUKAN PENUH EKSON KEPADA KEMBAR MONOZIGOTIK WARGA MALAYSIA DISYAKI DENGAN PENYAKIT IMUNODEFISIENSI PRIMER

ABSTRAK

Penyakit imunodefisiensi primer (PID) merupakan penyakit yang jarang ditemui dan mempunyai lebih daripada 200 jenis yang kesemuanya berkongsi simptom yang hampir sama. Sepasang kembar monozigotik (kembar MZ) perempuan berketurunan Melayu yang mempunyai simptom jangkitan pada saluran paru-paru dan "bicytopenia associated hepatosplenomegaly" disyaki menghidapi penyakit imunodefisiensi primer telah dikaji. Walaubagaimanapun, diagnosis tepat untuk kembar MZ tidak dapat ditentukan berpunca daripada kerumitan fenotip berdasarkan simptom klinikal. Analisis penjujukan penuh ekson telah dilakukan untuk mengenalpasti punca mutasi genetik bagi kembar MZ. Sejumlah angka 84 dan 81 juta jujukan telah diterima dengan sebanyak 99.26% dan 99.39% telah berjaya dipetakan kembali kepada rujukan genom manusia hg38. Kes heterozigot majmuk gen *CD21* telah berjaya diperoleh akibat mutasi bes DNA dan mengesahkan bahawa kembar MZ menghidapi kedefisienan *CD21* pada posisi c.1916G>A dan c.2012G>A. Analisis penjujukan Sanger telah mengesahkan mutasi heterozigot majmuk pada gen *CD21*. Analisis WES menyatakan bahawa kembar MZ menghidapi "common variable immunodeficiency disease" (CVID) disebabkan oleh kekurangan *CD21*. Analisis jaringan bagi gen *CD21* turut berjaya diperoleh bagi memahami hubungkait gen antara satu sama lain.

Kata kunci: penyakit imunodefisiensi primer, analisis penjujukan penuh ekson, “common variable immunodeficiency disease”

ACKNOWLEDGEMENTS

A special thanks to my supervisor, Dr. Saharuddin Bin Mohamad for the guidance and enlightenment throughout this postgraduate journey. I would also like to express my gratitude to Dr Adiratna Mat Ripen for her support during this study and thank you to the members of Primary Immunodeficiency Diseases Unit from the Institute for Medical Research, Malaysia: Dr Farid, Dr Munirah, Miss Chear, Fariza, Milah, Husna, and Mastura. I would especially like to acknowledge the Bioinformatic group at the University of Malaya, including Miss Dewi, Abdul Kadir Mukarram, Yousri, Elham, Nava, Adib, Abeer, Kadoj, Raga, Matine, Tam, Mei Kee, Fateen, Badeer, Amyrul, Syahrul, Syakirah, Wai Mun, Farah Wahidah, Jes, and my whole supportive family, including Khuzairi and my cat (Manja). Last but not least, thank you to the OMICS International for acknowledgement and appointing me as their Editorial Board for the Journal of Genome (https://www.omicsonline.org/editorialboard-journal-genome.php).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

>	:	More than
<	:	Less than
%	:	Percentage
Ng	:	Nanogram
µL	:	Microlitre
mL	:	Millilitre
AD	:	Autosomal dominant
AR	:	Autosomal recessive
BP	:	Base pairs
CBC	:	Complete blood count
CNVs	:	Copy-number variants
CVID	:	Common variable immunodeficiency disease
DC	:	Dendritic cells
ESID	:	European Society for Immunodeficiencies
INDELS	:	Insertion-deletions
IUIS	:	International union of immunological societies
IVIG	:	Intravenous immunoglobulin
MAF	:	Minor allele frequency
MZ	:	Monozygotic twin
NGS	:	Next-generation sequencing
OMIM	:	Online Mendelian inheritance database
PCR	:	Polymerase chain reaction
PE	:	Paired-end

LIST OF SYMBOLS AND ABBREVIATIONS

PID	:	Primary immunodeficiency diseases
PPI	:	Protein-protein interaction
RAM	:	Random-access memory
RAPID	:	Resource of primary immunodeficiency diseases
SCID	:	Severe combined immunodeficiency
SNP	:	Single nucleotide polymorphisms
SNVs	:	Single nucleotide variations
V	:	Voltage
WES	:	Whole exome sequencing
WGS	:	Whole genome sequencing

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

Primary immunodeficiency diseases (PID) are rare genetic diseases that impair the immune system permitting infection and various health issues to strike easily. They are heterogeneous group of immune disorders with more than 200 genetically determined disorders and broad range of clinical manifestations. Most PID are inherited cases such as X-linked, autosomal dominant and autosomal recessive inheritance, although some maybe caused by *de novo* mutation of the disease causative gene (Picard et al., 2015). Many of the PID are fatal if not treated and managed appropriately. However, the diagnosis and the management of PID is not an easy task due to the polygenic nature of PID and the clinical symptoms are very similar among the diseases. Misdiagnosis of PID would delay proper treatment which can be life-threatening to these patients.

The importance of definite diagnosis of rare diseases, for instance, primary immunodeficiency diseases (PID) has constantly risen over the years. Lately, advancement of next-generation sequencing (NGS) has assisted researchers in their genetic investigations. Whole exome sequencing (WES) is a sequencing method that targets the protein-coding regions in a particular genome had been widely used in the identification of genetic mutation for a wide range of rare disorders (Mirabello et al., 2014; Kelsen et al., 2015). WES strength was demonstrated by the identification of a single hemizygous, missense variant in the *XIAP* gene, causing X-linked lymphoproliferative disease 2, a condition that may not be easily identified using traditional genetic techniques in PID (Kelsen et al., 2015). In short, WES technology is a powerful method that can be applied to identify causative gene mutation leading to a definite clinical diagnosis of PID.

1.2 Study Design and Objectives

A pair of female monozygotic twin siblings (P1 and P2) born to non-consanguineous parents of Malay ancestry were the subjects of this study. In the first few years of their life, they presented with a few episodes of upper respiratory tract infections which were resolved without requiring hospital admission. The patients were first hospitalized for upper respiratory tract infection (URTI) at the age of seven, which they presented with tachypnea, and were noted to have digital clubbing, bronchiectasis, bicytopenia associated hepatosplenomegaly and failure to thrive. However, bone marrow aspiration did not suggest any malignancy infiltration for both siblings. Both patients had low count of lymphocytes and low serum antibody level. Based on these findings, the patients were suspected with PID, without any definite diagnosis of the specific type of PID. This study was designed to utilize WES analysis to both patients in order to determine the causative gene and gene variation for the disease phenotype. Figure 1.1 outline the study design for this overall thesis. The objectives of this study are:

1. To perform sequence assembly and mapping for WES dataset derived from the twin patients suspected with PID.
2. To identify gene variations and mutations that could lead to a definite clinical diagnosis for both twin patients suspected with PID.

The hypothesis for this thesis is WES analysis of P1 and P2 will identify genes responsible for PID.

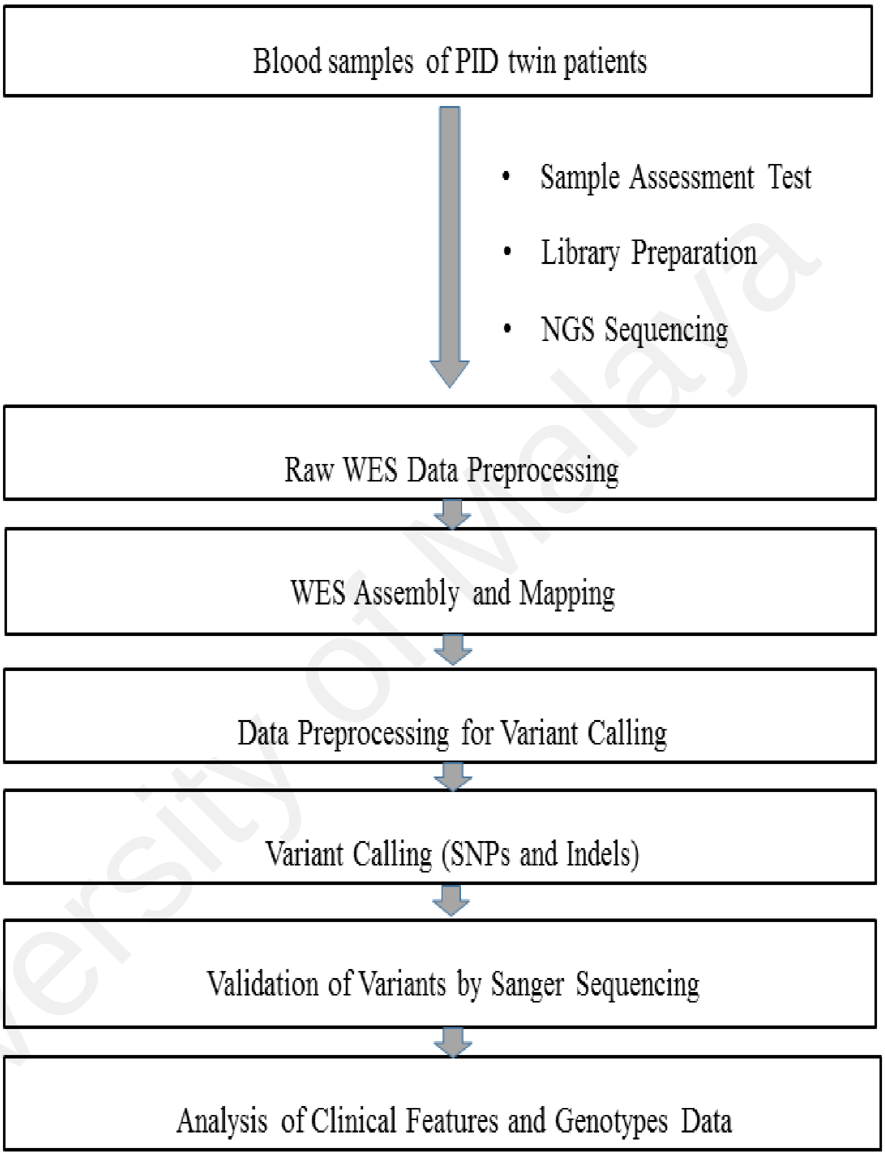

Blood samples of PID twin patients
Sample Assessment Test
Library Preparation
NGS Sequencing
Raw WES Data Preprocessing
WES Assembly and Mapping
Data Preprocessing for Variant Calling
Variant Calling (SNPs and Indels)
Validation of Variants by Sanger Sequencing
Analysis of Clinical Features and Genotypes Data

Figure 1.1: Hierarchical flow diagram for WES analysis

CHAPTER 2: LITERATURE REVIEW

2.1 Primary Immunodeficiency Diseases (PID)

The human immune system comprised of innate immunity and adaptive immunity as shown in Figure 2.1. Any defects in human immune cells will cause an inborn errors of immunity with more than 200 genetically determined disorders have led to broad range of clinical manifestations, including antibodies deficiency, susceptibility to infection and lymphoproliferation (Al-Herz et al., 2014). Patients with PID may develop allergy, autoimmune, inflammatory diseases and cancer (O'shea, Holland, & Staudt, 2013). In 1922, PID was first described by W. Schultz with various notable parameters, for instance, severe neutropenia (low levels of neutrophils) in several adult patients (Ferenzi, 1962). Globally, PID is considered as rare disease, however the incidence of PID vary among populations (Kirkpatrick & Riminton, 2007). In 2014, data from the European Society for Immunodeficiences (ESID) online database (www.esid.org) showed the higher absolute number of PID patients in France (n=5,426), followed by Spain (n=1,573), Italy (n=1,120), Netherlands (n=743) and Poland (n=560). However, national registry for PID is not established yet in Malaysia, so the PID incidence estimation in Malaysia is unknown. Consanguinity marriage has been reported with increase PID incidence such as in Kuwait, 44% of the 128 PID patients was reported with parental consanguinity (Al-Herz, Naguib, Notarangelo, Geha, & Alwadaani, 2011). Additionally, significant odds ratio of parental consanguinity compared to the healthy control is revealed by the meta-analysis and systemic review in PID study conducted by (Hadizadeh, Salehi, Khoramnejad, Vosoughi, & Rezaei, 2017).

Some laboratory methods are used to diagnose PID such as complete blood count (CBC) test, flow cytometry, vaccine response test, turbidimetric assay of antibodies and genetic diagnosis. CBC technique is one of the key contributions to the evolution of medical laboratory analysis during the 19th century, because quantitative analysis enables clinicians for their decision-making to rule out the disease (Verso, 1964). The main task is to monitor abnormalities in white blood cells (leukocytes), red blood cells (erythrocytes), and platelets (thrombocytes) for the further treatment. An increased number of neutrophils (white blood cells) could mean the person has an infection and a higher number of erythrocytes may suggest a person with chronic obstructive pulmonary disease (COPD) (Rodriguez et al., 1979). The benefits of CBC technique have driven the medical understanding of establishing a baseline for the health condition based on the comparison of reference range used.

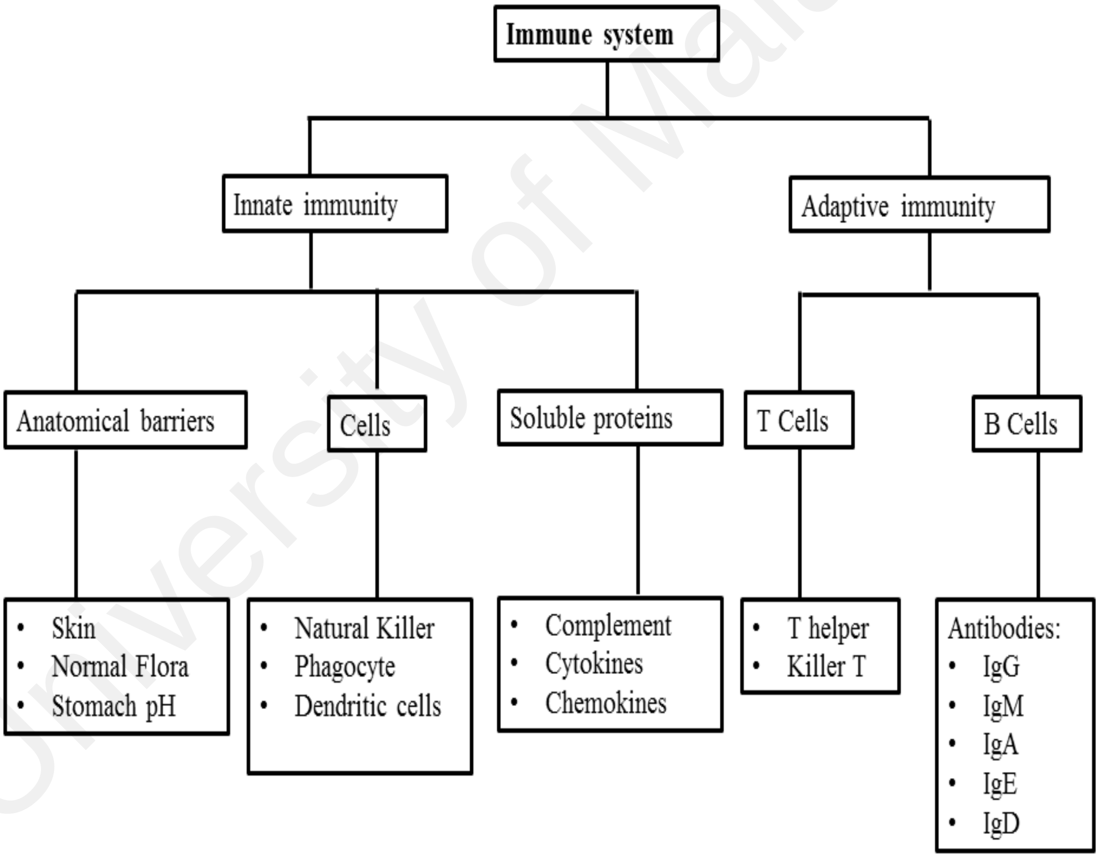


Figure 2.1: The human immune system (Simon, Hollander, & McMichael, 2015)

Flow cytometry extends the PID diagnostic procedures from the basic blood panels quantification to the fully automated immunophenotyping analysis. It depends heavily on the optical fibre principles, fluorescence staining and reaction of antigen and antibody for the enumeration of T-, B- and NK cells. The percentages and absolute cell numbers of T-, B- and NK cells can provide valuable information regarding the type of immunodeficiency e.g. DOCK8 deficiency is characterized by having an elevated serum IgE levels, reduced T and B cells, and decreased serum IgM levels (Aydin et al., 2015).

Turbidimetry method using a centrifugal analyzer is a rapid, quantitative, and accurate technique to detect immunoglobulin levels of the three major classes (IgG, IgA and IgM) in a human serum (Hills & Tiffany, 1980). A low level of immunoglobulin usually are suspected with hypogammaglobulinaemia (Kutukculer & Gulez, 2009). Many of PID cases are reported with hypogammaglobulinaemia. Example of PID cases with hypogammaglobulinaemia are the selective IgA deficiency (serum IgA is less than 5 mg/dL with normal IgG and IgM levels), *IL21* deficiency, *ADA* deficiency and *CD40* ligand deficiency (Conley, Notarangelo, & Etzioni, 1999). Determination of IgG functional antibody responses is sometimes useful (Maynard, Scott, Nahm, & Ladenson, 1986). Depending on the condition of PID, serum immunoglobulin tests can be ordered periodically to monitor and evaluate disease progression.

Vaccine response test aims to assess the humoral immune response using a single booster dose of diphtheria–tetanus toxoid, conjugated Hemophilus influenzae type B (Sekinaka et al.) and Pneumovax® (Ameratunga et al., 2016). Example applications of vaccine response test in PID are the common variable immune deficiency (CVID), severe combined immunodeficiency (SCID), and congenital thrombocytopenia (Wiskott-Aldrich syndrome) (Cunningham-Rundles, 1989). The failure of specific antibody production suggests an immune deficiency. The vaccine response are different, for instance, Pneumovax® is response to carbohydrate antigens, while the vaccinations against hepatitis B, tetanus, diphtheria, and Hib responses to protein antigens (Astronomo & Burton, 2010; Pichichero & Passador, 1997). The drawbacks of vaccine response test are vaccine hesitancy among the patient's family and vaccine are costly, thus, incurring financial burdens especially in the developing countries.

Last but not least, the genetic diagnosis is used to identify the causative gene in the patients suspected with PID. Targeted genes panel are usually assessed during the test, along with conventional Sanger sequencing method for well-established disease. In X-linked and autosomal recessive cases of chronic granulomatous disease (CGD), the genetic diagnosis has been applied by targeted five different genes (*CYBB*, *CYBA*, *NCF1*, *NCF2*, and *NCF4*) (Roos & Boer, 2014). However, the presence of variants with unknown clinical significance often complicates the problem. Major disadvantages of genetic diagnosis using conventional methods are time-consuming, expensive, general practitioner need to undertake serial testing of the gene panels and lower diagnostic rates due to locus heterogeneity (Tischer & Kaufer, 2012; Weir, 1990). An improvement of the robustness genetic diagnosis has been found using NGS technology (See Section 2.4).

2.2 Clinical classification of PID

Most of the PID cases are due to inherited genetic defects or sporadic diseases in children and adults (Costa-Carvalho et al., 2014) . Early detection of PID is key for the management. The main challenge for the detection is due to the failure to recognize general PID condition. However, many of the PID symptoms are similar to each other and in some cases the symptoms similarity extended to the non-PID cases. For this reason, National Primary Immunodeficiency Resource Center applied a list of 10 warning signs for PID to facilitate clinician or general practitioner in their immunological investigations as listed below in Table 2.1 (Reda, El-Ghoneimy, & Afifi, 2013):

Table 2.1: List of PID warning signs.

No.	Clinical Features
1.	Four or more new ear infections within one year
2.	Two or more serious sinus infections within one year
3.	Two or more months on antibiotics with little effect
4.	Two or more pneumonias within one year
5.	Failure of an infant to gain weight
6.	Recurrent, deep skin or organ abscesses
7.	Persistent thrush in mouth or fungal infection on skin
8.	Need for intravenous antibiotics to clear infections
9.	Two or more deep-seated infections including septicemia
10.	A family history of PID

The International Union of Immunological Societies (IUIS) PID expert committee has classified PID into nine main categories in 2015 as listed in the Table 2.2 and an approximately 200 gene mutations have been identified to be associated with PID (Some genes are listed in the Table 2.3) (Picard et al., 2015):

Table 2.2: Main categories of PID.

No.	Clinical Features
1.	Immunodeficiency affecting cellular and humoral immunity
2.	Combined immunodeficiency with associated or syndromic features
3.	Predominantly antibody deficiencies
4.	Diseases of immune dysregulation
5.	Congenital defects of phagocyte
6.	Defects in intrinsic and innate immunity
7.	Autoinflammatory disorders
8.	Complement deficiencies
9.	Phenocopies of PID

Table 2.3: List of PID-related genes.

ADA	*CFP*	*CD27*	*KRAS*	*ROC4*
AK2	*COH1*	*CD3D*	*I5*	*RHOH*
ARTEMIS	*COLEC11*	*CD3E*	*ICOS*	*SPINK5*
ATM	*CTSC*	*CD3E*	*IKAROS*	*SH2D1A*
AP3B1	*C7*	*CD3G*	*IL10*	*RTEL1*
AIRE	*C8A*	*CD40LG*	*IL10RA*	*SH3BP2*
AICDA	*C8B*	*CD8A*	*IL10RB*	*SLC29A3*
AIRE	*C8G*	*CIITA*	*MAGT1*	*PSTP1P1*
AP3B1	*C9*	*CORO1A*	*MTHFD1*	*POLE1*
ACTB	*CD27*	*C/EBPE*	*NOLA3*	*SLC46A1*
ACP5	*CD40*	*C1*	*HOIL1*	*RNASEH2A*
ACT1	*CFB*	*C16ORF5*	*IL17RA*	*SP110*
ADAR1	*CHD7*	*CD46*	*IRAK4*	*PIK3R1*
BLNK	*CASP10*	*CD59*	*MEFV*	*RMRP*
C3	*CIAS1*	*CFD*	*NLRP12*	*SAMHD1*
C1R	*CD19*	*CFH*	*PIK3CD*	*SBDS*
C1QC	*CARD9*	*CFHR1-5*	*PMS2*	*STAT1*
C4B	*DNMT3B*	*CFI*	*PNP*	*STAT5B*
CARD11	*CSF2RA*	*G6PC3*	*PRKDC*	*STIM1*
APOL1	*CR2 (CD21)*	*G6PT1*	*PTPRC*	*STK4*
BK1	*CARD11*	*GM-CSF*	*PIK3CD*	*TAZ*
C1S	*CD20*	*HAX1*	*PRF1*	*THBD*
C1QA	*CARD14*	*IFN*	*PRKCD*	*TLR3*
C2	*CD81*	*IFNGR2*	*PLCG2*	*TRAF3*
BLM	*C1QB*	*IL12B*	*PSMB8*	*TRIF*
BTK	*CD79*	*IL12RB1*	*RAC2*	*TCN2*
C4A	*CSF2RA*	*ITGB2*	*ROBLD3*	*UNC93B1*

2.3 PID Risk Factors

For PID control to advance systematically, it requires health care priority setting, which need an understanding of the PID problems that exist in a country, thus, provides an availablility to address them. Given the uncertainty about the risk factors of PID, familial aggregation should be considered for which evidence suggests an PID origin. Additionally, families with multiple affected relatives are prone to share disease-risk alleles (Fang, Abolhassani, Lim, Zhang, & Hammarström, 2016; Grant et al., 2001). Family aggregation have documented the clustering of certain PID, such as CVID (Vorechovský, Litzman, Lokaj, & Sobotkova, 1991), *LRBA* deficiency (Bratanič et al., 2017), and Chronic Granulomatous Disease (CGD) (Windhorst & Soothill, 1969). Claire (2016) presents an evidence of common environment factors for PID such as malnutrition, which is responsible for an impaired immune priming by dendritic cells (DC) and monocytes (Bourke, Berkley, & Prendergast, 2016). In addition, genetic susceptibility to opportunistic infections such as cytomegalovirus, Epstein–Barr virus or John Cunningham increases the risk for the person with PID (Schmidt, Grimbacher, & Witte, 2018).

Evidence of the most comprehensive model of the genetic inheritance was proposed in the 1960s (Westerlund & Fairbanks, 2010). Much attention originally focused on autosomal dominant, autosomal recessive and X-linked cases, which can be used to filter and prioritize disease-causing variant in the particular disease (McKusick, 1976). In most cases, PID is inherited as monogenic disorder, however, delayed onset, incomplete penetrance, and/or different gene expressivity of the clinical features should be considered in the genetic analysis. Recently, genetic variant that is present for the first time in one family member (also known as *de novo* mutation) is also reported in the PID cases (Moya-Quiles et al., 2014; Picard et al., 2010; Russell et al., 2017).

2.3.1 Autosomal dominant

Autosomal dominant (AD) is referred to an affected person who inherits the disease or trait from any of an affected parent (heterozygous state). Dominant-negative mutations in the *STAT3* mutation, which is part of PID (known as AD form of Hyper-IgE syndrome), is introduced in 2007 (Minegishi et al., 2007). The syndrome was first described as "Job's syndrome" in two girls suffering from recurrent staphylococcal abscesses, pneumonia, and neonatal rash (Davis, Schaller, Wedgwood, & Harvard, 1966). The exact pathogenesis remains unknown. Several genes associated with PID have also been reported to have AD form of disease such as *TCF3*, *TWEAK*, *NFKB2*, *PIK3CD* and *CARD11* (Picard et al., 2015). Non-consanguineous marriage is often used as a filtering approach for AD pattern of a particular disease.

2.3.2 Autosomal recessive

Autosomal recessive (AR) are commonly employed to infer genetic pattern of homozygous or compound heterozygous mutations inherited from healthy heterozygous parents (carrier) (Abolhassani et al., 2014; Alkuraya, 2012). The basic idea behind the AR model is that 25% chance of child having normal genes (unaffected), a 50% chance of child carry normal and abnormal genes (carrier, unaffected) and another 25% chance of child having both abnormal genes (affected). Consanguineous marriage have led to the emergence of the AR pattern, which one of their potential drawback is that it tends to increase the risk of hereditary diseases in PID. Example pattern of AR inheritance in PID are Bloom syndrome (mutation in *BLM* gene), Comel-Netherton syndrome (mutation in *SPINK5* gene) and *CD19* deficiency (mutation in *CD19* gene) (German, Sanz, Ciocci, Ye, & Ellis, 2007; Komatsu et al., 2008; van Zelm et al., 2006).

2.3.3 X-linked cases

In the case of X-linked inheritance, males develop a disease if their only copy of the X chromosome express the disease-causing variant (Lederman & Winkelstein, 1985). While it predominately affects male, in PID it affects both gender (male and female), which female cases with X-linked recessive pattern are reported very rare and some study showed only 30% out of overall female proportions (de Saint Basile et al., 1999). Evaluation of the X-linked inheritance in female may require an additional testing in PID (Hollenbaugh et al., 1994). Examples of X-linked cases in PID are *IL2RG* deficiency (T−B+ Severe Combined Immunodeficiency (SCID)), *CD40* ligand deficiency, Wiskott-Aldrich syndrome (Serwas et al.), and *BTK* deficiency (Picard et al., 2015).

2.3.4 *De novo* mutation

In addition to the inherited genetic pattern, *de novo* mutations, which is the mutations occurring either in parental germline cells or at some point after fertilization process, contribute to human diseases (Veltman & Brunner, 2012). Many rare diseases reports patients did not inherit the disease from previous generations, thus, spontaneously *de novo* mutations may mostly explain their existance in the human population (Veltman & Brunner, 2012). The importance of *de novo* mutation in severe diseases, which is in more complex condition such as autism, has been well-understood (Girirajan et al., 2013). This evidence suggest that *de novo* mutation caused genetic defects in single gene for the complex disease rather than by common mutation in several genes. This has now been further supported by identifying *de novo* haplo-insufficiency in individuals affected with CHARGE syndrome due to *CHD7* defects in PID (Waleed Al-Herz et al., 2014). On average, the empirical recurrence risk of genetic complex diseases caused by *de novo* mutations has been reported as 1.3% in a family (Rahbari et al., 2016).

2.4 Whole Exome Sequencing

WES is an application of the NGS technology to sequence the entire known exome of human genome and determine the variations of all coding regions of known human genes (Yang et al., 2013). WES was introduced into clinical practice in 2009 by a collaboration of researchers from Department of Genome Sciences, Department of Pediatrics (University of Washington), Howard Hughes Medical Institute, and Agilent Technologies to keep the cost down while searching for candidates variants in protein coding regions of human genome (Sarah et al., 2009). Current WES technology is known to provide coverage of more than 95% of the exons, which covers 85% of disease-causing variants in Mendelian disorders (Rabbani, Tekin, & Mahdieh, 2014). Despite the fact that only 1.5% of the genome is sequenced, this approach still allows an identification of genetic variation for both Mendelian and multiple genes (polygenic) disease with cheaper costs as compared to whole genome sequencing (WGS) (Meynert, Ansari, FitzPatrick, & Taylor, 2014).

2.4.1 WGS versus WES

WGS have opened a route to a new dimension of molecular study, enabling the determination of the nucleotides (A, C, G, T) orders in the entire genome of the species. The aim of WGS is to find genetic alterations such as single nucleotide variants (SNVs), deletions, insertions, and copy number variants (CNVs) in the genome (Drmanac, 2012). In addition, changes in the noncoding regions of DNA known as introns and splice sites may also be detected. An alternative approach is to sequence protein-coding regions (known as WES technology), which is the area of DNA sequence when transcribed and translated into the proteins. WES comprise only about 1.5-2% of the genome compared to the WGS (Ku, Cooper, & Patrinos, 2016). WES are able to be sequenced with higher number depth of coverage at lower costs, which provides more confidence base calls for variant detection and downstream analysis that contribute to pathogenesis of the disease (Meynert et al., 2014). WES are often used in clinical setting to give greater confidence as well as minimize the cost and short turnaround time (Rehm et al., 2013). However, WES only focus on protein-coding areas of the genome, which is for some research study or genetics testing, WGS may be more advantageous. The major drawbacks of WES when compared to WGS is incapability to detect copy-number variations (CNVs) as it is beyond the targeted area. In addition, the detection of false-positive variants in single nucleotide polymorphisms (SNPs) study is higher for WES than for WGS as reported by (Belkadi et al., 2015).

2.4.2 WES technology

In this thesis, WES was conducted on Illumina platform for a pair of MZ twin, which delivers more accurate results and a high percentage of base calling above Q30 Phred score (Manley, Ma, & Levine, 2016). The general workflows for Illumina include four main steps (as desribed in Table 2.4 and Figure 2.2):

Table 2.4: Workflows for WES using Illumina platform.

Steps	Description
Library preparation	NGS library is prepared by DNA fragmentation process and ligation technique to the adapters
Cluster generation	NGS library is loaded into Illumina flow cell and hybridization process is occurred. Bridge amplification technique is used to amplify each bound fragment into a cluster group
Sequencing	For sequencing, four reversible terminator is used to detect each bases through fluorescently labeled nucleotides. The emission from each cluster is imaged and recorded
Data analysis	Bioinformatics software is used to analyse the data, along with the alignment and mapping process

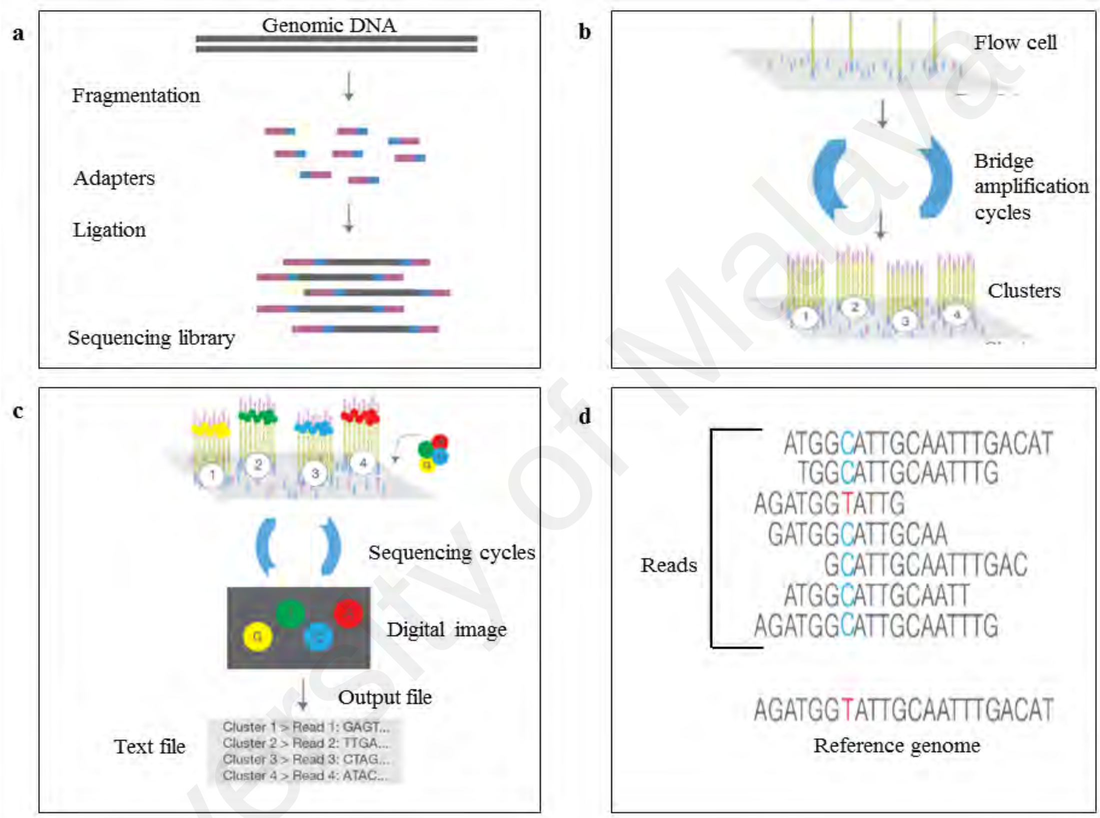


Figure 2.2: Illustrations on Illumina sequencing (a) library preparation (b) cluster generation (c) WES sequencing (d) alignment and data analysis (https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html**)**

A recent advancement in NGS technology for Illumina platform occurred with the evolution of paired-end (PE) sequencing as shown in Figure 2.3. PE reads are sequenced from the both ends of the DNA (forward and reverse) sequence during a library preparation and aligning the both reads as read pairs. Additionally, PE sequencing have higher accuracy and enables the detection of indels as compared to the single reads, along with the multiple numbers of reads being sequenced with the same run time and effort during library preparation (Tattini, D'Aurizio, & Magi, 2015). An explanation for the advantages of PE sequencing is it can increase the ability to remove PCR duplicates, that arises from PCR amplification during NGS library preparation and a higher number of SNPs calls from PE alignment (Tin, Rheindt, Cros, & Mikheyev, 2015).

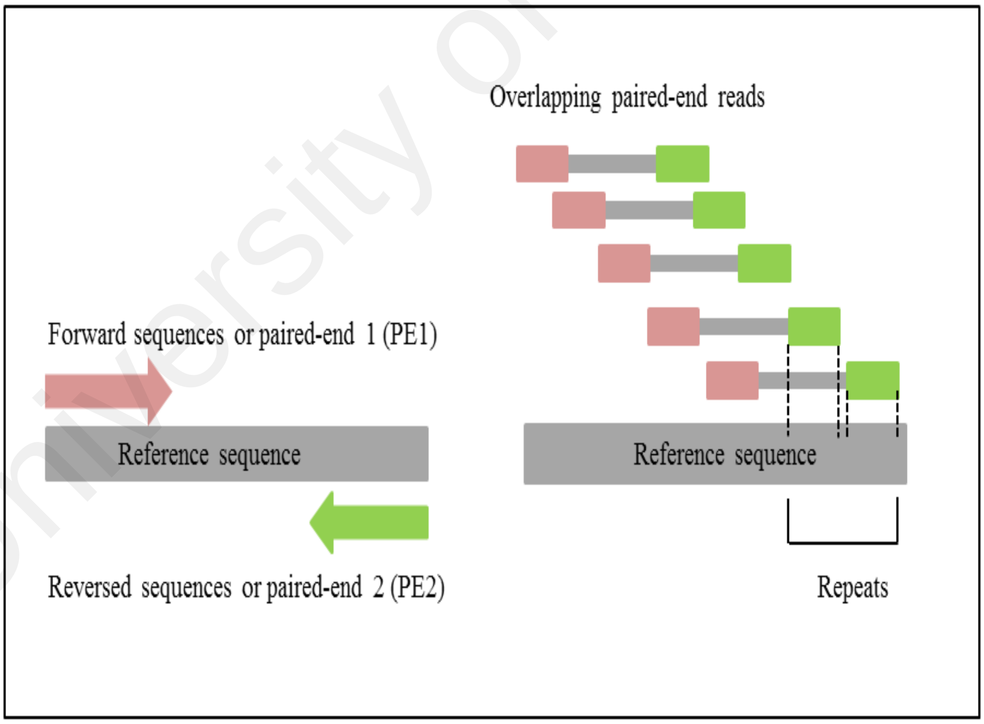

Overlapping paired-end reads
Forward sequences or paired-end 1 (PE1)
Reference sequence
Reversed sequences or paired-end 2 (PE2)
Reference sequence
Repeats

Figure 2.3: Diagrams on paired-end (PE) reads sequencing technology by Illumina (https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html**)**

2.4.3 WES application in rare diseases

The benefits of WES have driven the clinical understanding of human physiology, anatomy and subsequently human diseases to a new dimension of study and thus have made significant improvements in monitoring and health treatment. Understanding the pathogenesis of a disease depends heavily based on the discovery of disease-causing variant and studying the impact of pathogenic variants to the protein structure. WES is also used to identify rare, novel, as well as common variants in protein-coding area associated within complex and universal traits.

Previously, WES had been used to identify the causative variant of a rare form of inflammatory bowel disease (also known as IBD) in an infant (Worthey et al., 2011). For this case, conventional clinical diagnostics had failed to find an explanation for the patient's severe symptoms and clinicians needed to understand the possible cause of the symptoms before they could decide on how to treat the child accordingly. WES was first introduced and reported with four patients affected by an inherited disease called Freeman-Sheldon (Grant et al.) (S. B. Ng et al., 2009). FSS is a rare autosomal dominant disorder caused by the mutation in *MYH3* gene, which is already identified earlier using another methods.

Soon after, WES was done for four patients with Miller Syndrome (De Ligt et al.) with identification of a single candidate gene, known as *DHODH*. Sanger sequencing confirmed the *DHODH* mutations in another families with MS and the mutation was absent in the healthy control samples (Sarah et al., 2010). Similarly, WES was performed on the patients who had diamond blackfan anaemia (DBA). A mutation identified showed *GATA1* as a causative gene, encoding the hematopoietic transcription factor. WES was undertaken in additional study with family consanguinity of DBA and the researchers discovered the same location of splice site mutation as previous study.

This research provided insight into the fore WES capability and pathogenesis of DBA (Sankaran et al., 2012). This finding showed WES strength to identify disease-causing variant of rare disorders among affected individuals. The robust, reproducible and cost-effective approach of WES technology has led to the utility of WES as a new, reliable, and powerful diagnostic tool.

2.4.4 WES limitations

The drawbacks of WES technology lie again in the difficulty with an introduction of sequencing errors during library preparation by PCR, which is GC bias that act as the major source of unwanted variation in the sequencing results. However, PE sequencing method introduced by Illumina platform improves library preparation complexity and reduces the number of duplicate reads (Tin et al., 2015). The utility of WES in diagnosing rare diseases is promising (De Ligt et al., 2012), however, routine clinical diagnostics is often accompanied with severe complications, resulting to interpretive challenges as not all disease variant has been fully characterised in the literature study. Reliable analysis of the polygenic, novel or *de novo* mutations found through WES study will need additional experience and validation in the research setting before it reaches the clinic and patients, especially for diagnosis of complex and rare diseases (Katsanis & Katsanis, 2013).

Another limitation arises from the computer requirement and laboratory facility to run WES dataset. The computer requirement such as RAM and processors is often expensive to set up (B. Schmidt & Hildebrandt, 2017). In addition, issue on data storage for the cohort or pilot project on WES also need to be considered, which is recently cloud environment has been introduced (Kahn, 2011). Also, incomplete and insufficient computational framework for WES data analysis which may cause the delay in disease diagnosis is another potential drawback (Du et al., 2016). Although WES is widely utilised in clinical diagnostics, there is no gold standard provided for the analysis and identification of disease-causing variants (Rehm et al., 2013). Sometimes, causative genes may not be found in WES data due to disease variants is predisposed into non-coding regions and outside the targeted regions of WES (Cirulli & Goldstein, 2010). In this circumstance, WGS study should be applied. Furthermore, WES failed to detect large indels (>50 bp), thus, third generation sequencing with long-reads sequence such as PacBio technology can be considered (Pajusalu et al., 2018; Rhoads & Au, 2015).

2.4.5 Network Analysis

Network study is used to analyse biological dynamics such as protein interactions, biochemical reactions, and also gene regulatory mechanisms (Kitano, 2002a). The concept of network theory is introduced in the 18th century by Leonard Euler, but is not used for biological networks study until the evolution of the computer systems (Debnath, 2009). Previous biological experiments showed that proteins, as the main function of biological mechanism, plays a role in the determination of the phenotype of all organisms (Gonzalez & Kann, 2012). Next, with the most recent understanding of molecular biology, proteins are assumed not to work alone as individual protein but have different interactions with another protein and other biological molecules (*e.g.* DNA, RNA) that control metabolic level and signaling pathways, cellular processes, and another system biology in a cell (Kitano, 2002b). Therefore, studies of protein-protein interactions and biological network are fundamental to identify and understand their role within the cell.

Previous study have revealed that the structure of native proteins and their transition states can be interpreted using the network method (Vendruscolo, Dokholyan, Paci, & Karplus, 2002). Similarly, other researchers have contributed to an identification of functional residues, which is helped in the understanding of protein dynamics (Emerson

& Gothandam, 2012). In addition, proteins have shown as the part of biological network to comprehend protein interactions (del Sol, Fujihashi, Amoros, & Nussinov, 2006). There have been more detailed reviews on protein as networks (Rual et al., 2005), protein dynamics (Kidera & Go, 1990) and the linking of topological characteristics to the protein folding (Del Sol, Fujihashi, & O'meara, 2005).

2.4.6 Protein-protein interaction in functional network

The structure and nature of protein interaction networks is one of the best interest and is a considerable subject in understanding of system biology, which is due to the availability of the dataset of protein interactions for majority of protein study (Lehne & Schlitt, 2009). Protein-protein interaction (PPI) networks act as a powerful tool for exploring protein functions and disease-gene relationship (Lv et al., 2015). This method may allow researchers to link the genes with their associated biological pathways and the corresponding diseases, which is used to improve the biomedical applications (Barabási, Gulbahce, & Loscalzo, 2011). Major advantages of studying PPI is it may assign unknown roles to uncharacterised proteins, availability of detailed information within a signalling pathway, and define the relationships between genes that arises from multi-molecular complexes (Rao, Srinivas, Sujini, & Kumar, 2014). Network analysis may underline the existence of genes interaction harboring molecular alterations in PID (Xue et al., 2014). However, one potential drawback of PPI is the underlying mechanisms of complex and rare diseases, which comes from the interaction of multiple genetic with the environmental factors, cannot be revealed by such approaches (Furlong, 2013).

2.4.7 Disease module using network analysis

Recently, there has been increasing area to exploit the concepts of network biology to understand disease module (Vidal, Cusick, & Barabási, 2011). Disease module can be explained by existence of specific regions for the disease proteins based on the network study (Goh et al., 2007). The identification of these modules is the starting point for the discovery of drug targets. For example, network approaches to identify disease module have successfully been shown in Parkinson's disease and progressive supranuclear palsy (Eckert, Tang, & Eidelberg, 2007; Santiago & Potashkin, 2014). Additionally, disease module approaches have provided deep insights into the molecular level for underlying diseases mechanism associated with PID including CVID (Choi, Fernandez, Maecker, & Butte, 2017; Farmer et al., 2017). However, the drawback for the identification of disease modules is laborious and time-consuming for the subnetwork due to complex biological networks interpretation. In addition, technological biases in NGS approaches, along with the different software and database used for interaction study and molecular profiling can decrease the accuracy of analysis. Another potential drawback is our limitation in understanding and interpreting the biological knowledge raised from the complex diseases such as PID.

In general, most cases of genetic disease-causing in PID are still rarely studied in Malaysia. In this thesis, WES are performed with the more than 20,000 coding-genes for functional genomics analysis and to enhance our understanding of PID in the aspect of genetics, immunology and clinical using bioinformatics approach. Major contributions of this thesis are the bioinformatics approach and an identification of definite diagnosis in MZ twin for the better healthcare in clinical settings of rare diseases such as PID.

CHAPTER 3: METHODOLOGY

Chapter 3 is organized in the following way. First, all the materials used are listed in Section 3.1. Next, ethics statement for working with human subject and methods for DNA extraction from blood and WES, including buffy coat separation, lymphocyte subset enumeration by flow cytometry, WES assessments, library preparations, computer requirement, customized WES protocol, variants validation and network analysis are described in Section 3.2.

3.1 Materials

QIAamp® DNA Mini Kit (Qiagen, CA), all PCR components (Applied Biosystems, US), DNA markers, deionized water and 1X TBE buffer (Promega, US).

3.2 Method

In this thesis, two different WES datasets were used to evaluate the genetic diagnosis presented in this study for the driving clinical diagnostics applications of this work.

3.2.1 Subjects and ethics statement

The study protocol was approved by the Medical Research and Ethics Committee (MREC), Malaysia (Ethic approval no: NMRR-16-892-31023). Two participants selected for this study were children with suspected disorders of the PID. These children were MZ twin without definite diagnosis of the type of PID. Informed written consent was obtained from their families for research purposes. A volume of 10 mL of whole blood from each participant was obtained in EDTA blood tube.

3.2.2 Buffy Coat Separation

Total lymphoprep (3.5 mL) was added on a plain tube. Blood sample was diluted with PBS (Ratio 1:1; blood (2.5 mL): PBS (2.5 mL)). The diluted blood was layered on the lymphoprep. The tube was spun at 3,000 x g for 15 minutes. The buffy coat layer was separated (rich with the leukocytes). Next, the tube was washed with PBS at 2,500 x g for 5 minutes. The supernatant was discarded. In addition, vortex was used to break the pellet. Tube was washed again with PBS at 2,500 x g for 5 minutes. The supernatant was discarded. The pellet (200 µL) was transferred into an eppendorf tube. For DNA extraction, eppendorf tube was spun at 3,000 x g for two minutes. Next, protease (20 µL) was added to the tube and vortex. Buffer AL (200 µL) was added and vortex briefly. Tube was incubated at 56 ^{0}C for 10 minutes. After incubation, ethanol with 96-100% (200 µL) was added and vortex briefly. The sample was transferred into mini spin column with collection tube and was spun at 8,000 x g for one minute. The collection tube was discarded, Buffer AW1 (500 µL) was added to the tube and was spun at 8,000 x g for one minute. The collection tube was discarded, Buffer AW2 (500 µL) was added to the tube and was spun at 14,000 x g for three minutes. The spin column was placed into an eppendorf tube and it was spun at 14,000 x g for one minute to remove any carry-over. Next, the spin column was transferred into a clean eppendorf tube and Buffer AE (200 µL) was added to the tube. The sample was incubated at room temperature for five minutes and was spun at 8,000 x g for one minute. Lastly, the eppendorf tube was placed in hot plate for five minutes to denature the DNA and [DNA] was determined using NanoDrop.

3.2.3 TBNK test by flow cytometry

The patient's whole blood (50 µL) was stained with two cocktails containing monoclonal antibodies in a BD Trucount tube respectively: BD Multitest™ CD3 FITC (clone SK7) / CD8 PE (clone SK1) / CD45 PerCP (clone 2D1(HLe-1))/ CD4 APC (clone SK3) and BD Multitest™ CD3 FITC (clone SK7) / CD16+CD56 PE (clone B73.1 and Clone NCA16.2) / CD45 PerCP(clone2D1(HLe-1)) / CD19 APC (clone SJ25C1). The tubes were incubated for 15 minutes in dark. Then, the mixture was lysed with BD FACSLysing solution and incubated for 10 minutes. The percentage and absolute count of the lymphocyte subsets were analyzed using BD FACSDIVA™ software on a BD FACS Canto II flow cytometer (Becton Dickinson, USA).

3.2.4 WES Sample Assessment Test

Peripheral blood mononuclear cells (PBMC) were isolated from 10 mL of blood samples using the standard Ficoll-Paque centrifugation procedures. Genomic DNA was extracted using QIAamp® DNA Mini Kit (Qiagen,USA) following manufacturer's instructions. The quality of DNA samples was tested using 1% agarose gel electrophoresis and Qubit 2.0 Fluorometer (Life Technologies, USA) was used for DNA quantification.

3.2.5 Library Preparations

Library preparations for WES were conducted and sequencing for each captured library was performed to ensure that each sample meets the desired average sequencing depth 100x coverage using Agilent SureSelect Human All Exon V5 (Agilent, Santa Clara, CA) by targeting 50Mb of exonic sequence. The qualified genomic DNA sample was randomly fragmented into fragments with a base pair peak of 150 to 200 bp, and then adapters were ligated to both ends of the resulting fragments. The adapter-ligated templates were purified by the Agencourt AMPure SPRI beads (Beckman Coulter, Fullerton, CA, USA) and fragments with insert size about 200 bp were excised. Extracted DNA was amplified by ligation-mediated polymerase chain reaction (LM-PCR), purified and hybridized to the SureSelect Biotinylated RNA Library (BAITS) for enrichment. Hybridized fragments were bound to the streptavidin beads whereas non-hybridized fragments were washed out after 24 hours. Captured LM-PCR products were subjected to Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA) to estimate the magnitude of enrichment. Each captured library was then loaded on Illumina HiSeq4000 platform (Illumina, San Diego, CA) and high-throughput sequencing with paired-end reads of 101 bp were performed. Raw image files were processed by Illumina base calling Software 1.7 for base calling with default parameters.

3.2.6 Computer Requirement for Bioinformatics Analysis

In practice, a perfect hardware set of super computer with their processors is not available, so optimization must be made for some level of computer hardware, along with reliable internet connection and data storage. In this study, super computer with 32 GB of RAM is used to run bioinformatics analysis as shown in Table 3.1.

Table 3.1: Hardware requirement for the WES data analysis.

Hardware	**Recommendation**
Operating system	Ubuntu 14.04 LTS
RAM	32.0 GB
Bits	64
Hard disk	1.0 TB
PC	HP EliteDesk 800 G1 Small Form Factor PC
Processor	Intel(R) Core i5-4590 CPU @ 3.30 GHz
External	500.0 GB
Internet	Reliable and fast

Additionally, WES comprises comprehensive analysis steps and the process is a combination of several programs and databases for big data analysis. A subsequent task is to find genetic variants associated with the PID phenotypes. This is done mainly by exploring open-source bioinformatics software, for example, FastQC, BWA, SAMtools, GATK and wANNOVAR as listed in Table 3.2. In addition, Figure 3.1 illustrates basic pipeline used for genetic variants identification in WES, which is often started with sequence quality analysis.

The FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) aims to evaluate sequence reads quality from the WES raw data. The main task is to detect and monitor abnormalities, such as distributions of base qualities, GC content, and over-representation of adapters. In addition, FastQC can also directly guide homopolymer errors, which rises due to high number of off-peak signal intensities. In general, sequence quality is performed to discover anomalies that originated from the Illumina sequencer or library material used during the WES preparation, therefore reducing false variant calls error during variant calling analysis.

Table 3.2: Software used during WES analysis.

Software	**Purposes**
FastQC (Andrews, 2010)	Sequence quality control
Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009)	WES assembly and reference mapping
Sequence Alignment/Map (SAMtools) (Li et al., 2009)	Files conversions, indexing, sorting, viewing data
Picards (https://broadinstitute.github.io/picard/)	Sequence quality analysis and mark duplicates
Genome Analysis Toolkit (GATK) (McKenna et al., 2010)	Realignment, variants calling
wANNOVAR (Chang & Wang, 2012)	Variants annotation

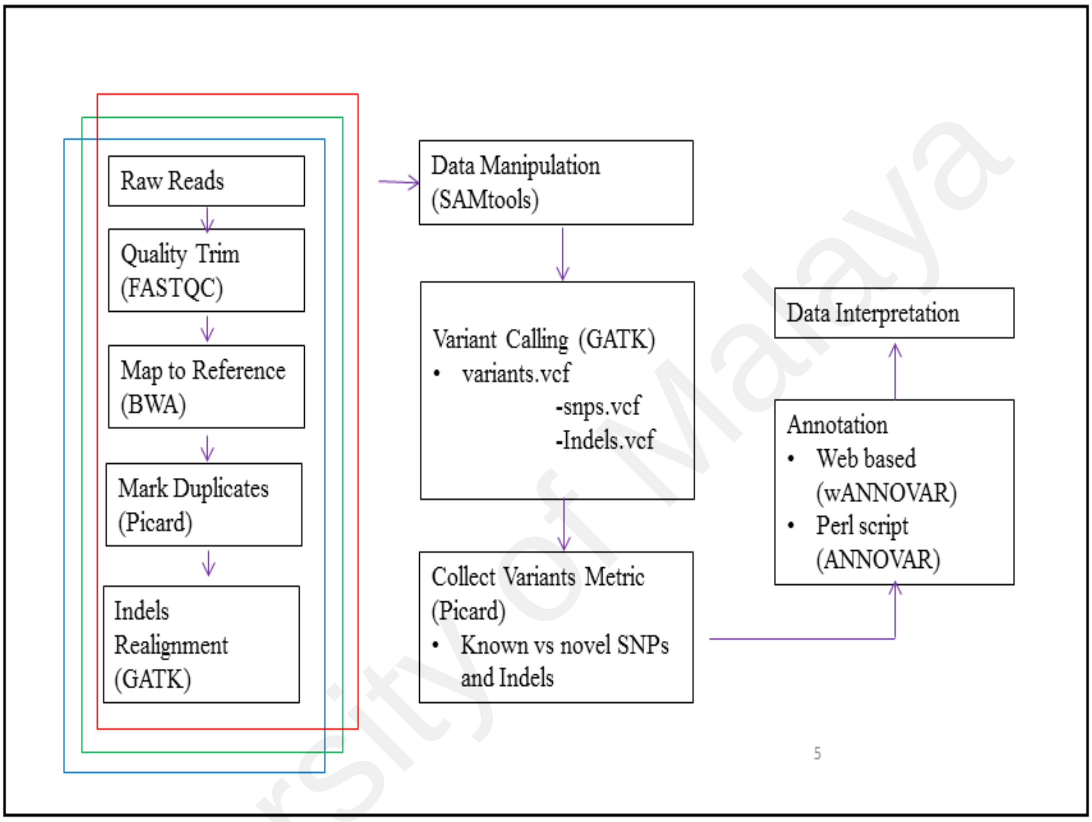


Figure 3.1: Established protocol for WES data analysis

Alignment is the process of mapping short DNA reads to a reference human genome in this study. Each of the millions of PE reads from the WES dataset should be matched with the genomic positions within the human reference sequences. This is an important computational step before any further analysis, especially for variant calling. One major drawback of alignment and mapping process is a unique reads versus non-unique reads mapping, heavy volume of PE sequences as up to millions of reads and variation in their base quality. Additionally, mapping process is considered as computationally challenging and time consuming step (Day-Williams & Zeggini, 2011). Any technical errors during alignment and mapping to the reference human genome will be affecting the downstream step. In this study, clean reads (after removal and filtered out all bad reads) were aligned and mapped to the 3.2 Gb FastQ files of GRCh38/hg38 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000001405.27) using the Burrows-Wheeler Aligner (BWA) software. SAMtools was used for sorting, indexing, and statistical metrics purposes, along with Picard-tools (http://picard.sourceforge.net/) to mark the read duplicates, collect sequencing artifact-metrics and jumping library metrics and from the WES dataset.

3.2.7 Customized protocol established for WES

Many approaches have now been developed for WES, which originally developed from open-source bioinformatics tools. The software version and hardware will affect the run time during WES analysis, thus when utilising a different version and algorithm, estimation of their run time is crucial to monitor the progress. In this study, customized WES protocol, along with the run time estimated for each dataset with 32 GB of RAM as described in Table 3.3 is used as a part of novelty made during this study. The advantage of this step is to monitor the process and the method can be slightly modified for the latest version of any software used in this study.

Table 3.3: Semi-automatic steps involved for WES analysis.

Steps	**Command-line**	**Run time**
FastQC	`fastqc  WES1_1.fastq.gz WES1_2.fastq.gz`	6 mins
Raw data	`(copy all raw data needed first in directory (e.g. jar files, human reference genome)`	10 mins
Data trimming	`java -jar trimmomatic-0.36.jar PE -phred33 WES1_1.fastq.gz WES1_2.fastq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:30`	15 mins
FastQC	`fastqc output_forward_paired.fastq.gz output_reverse_paired.fastq.gz`	6 mins
Index	`bwa index -p hg38.fa –a bwtsw hg38.fa`	1 hour
Mapping sequencing reads against a reference genome	`bwa mem –M –t 4 hg38.fa output_forward_paired.fq.gz output_reverse_paired.fq.gz > output.sam`	1 hour
Creating an indexed reference sequence	`samtools faidx hg38.fa`	15secs
Conversion of SAM to BAM file	`samtools view –h –b -S output.sam  > output.bam`	20 mins
Sort a BAM file	`samtools sort output.bam sort`	20 mins
Creating a BAM index file	`samtools index sort.bam`	1 min
Alignment Statistics	`samtools flagstat sort.bam`	2 mins
Running IGV	`Igv`	Depends
Visualize the coordinates	`samtools view –h output.bam | head -30`	1 sec
Convert SAM to BAM	`picard-tools SortSam I=output.sam O=sorted.bam SORT_ORDER=coordinate CREATE_INDEX= true`	30 mins
Add read groups	`picard-tools AddOrReplaceReadGroups I=sorted.bam o=sort_RG.bam RGID= 1 RGLB= library1 RGPL= illumina RGPU= FCH522FBBXXX RGSM= human`	10 mins
Quick check added RG BAM files	`samtools view -H sort_RG.bam`	1 sec
Mark Duplicates	`picard-tools MarkDuplicates I=sort_RG.bam O=mark.bam METRICS_FILE=metrics.txt`	12 mins
Building BAM Index	`picard-tools BuildBamIndex I=mark.bam`	2 mins

Table 3.3, continued.

Steps	Command-line	Run time
Create Sequence Dictionary	`picard-tools CreateSequenceDictionary R=hg38.fa O=hg38.dict`	1 min
Create realignment targets	`java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R hg38.fa -I mark.bam -o targetintervals.list`	23 mins
Perform local alignment of reads around indels	`java -jar GenomeAnalysisTK.jar -T IndelRealigner -R hg38.fa -I mark.bam –targetIntervals targetintervals.list –o mark2.bam`	20 mins
Variant Calling	`java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R hg38.fa -I mark2.bam -stand_call_conf 30 -stand_emit_conf 10 -o variants.vcf`	~1 hours
Extraction of SNPs	`java -jar GenomeAnalysisTK.jar -T SelectVariants -R hg38.fa -V variants.vcf -selectType SNP -o snps.vcf`	30 secs
Extraction of indels	`java -jar GenomeAnalysisTK.jar -T SelectVariants -R hg38.fa -V variants.vcf -selectType INDEL -o indels.vcf`	30 secs
Collect variant calling metrics	`java -jar picard.jar CollectVariantCallingMetrics INPUT= variants.vcf OUTPUT=variants_metrics1 DBSNP=Homo_sapiens_assembly38.dbsnp138.vcf.gz` *dbsnp138 data were downloaded from the GATK resources bundle	4 mins
wANNOVAR database	(http://wannovar.wglab.org/) Input: snps.vcf and indels.vcf files Reference genome: hg38 Results duration: 1 day	10 mins
Total run time for each dataset		388 mins

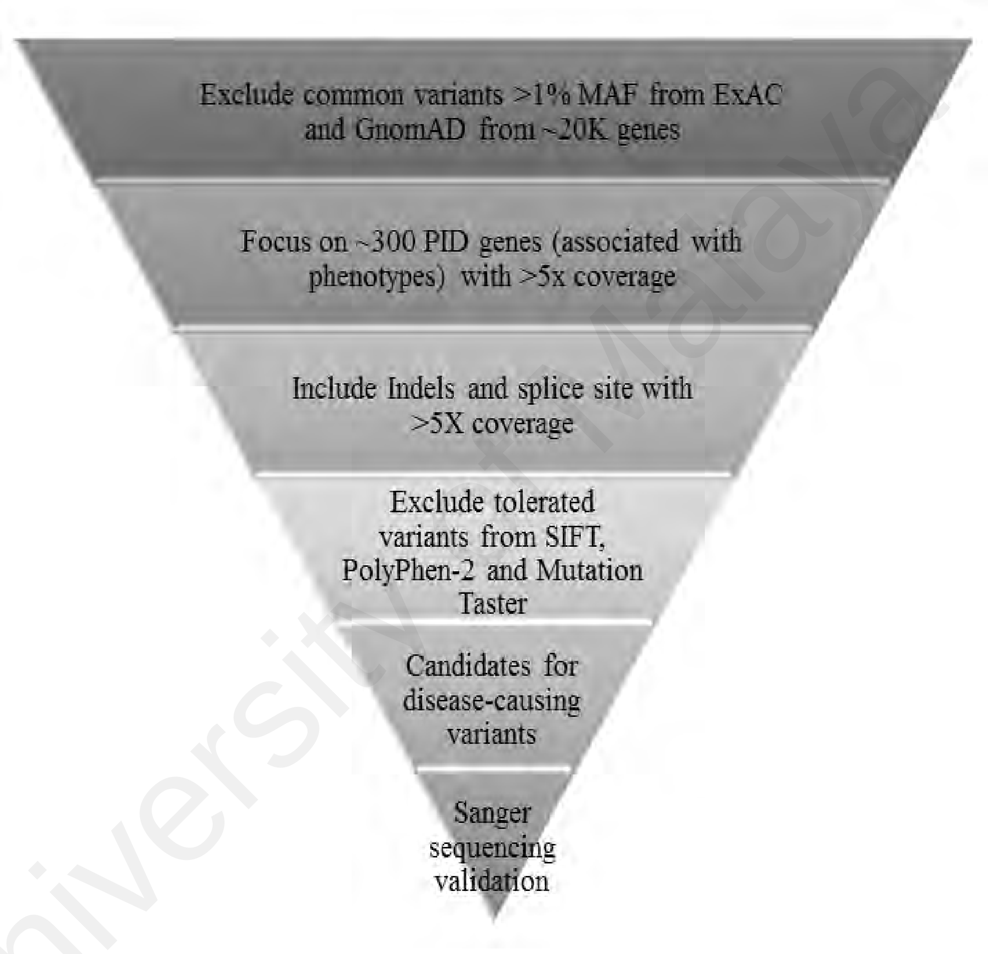
Mins=minutes

Variant calling is the next step taken after mapping process. Since the sequences are already aligned with the reference genome, the genetic variants can be analysed. These variants may act as the disease-causing variant, common variant and could be PCR noise without affecting clinical features characteristics. Variant call format (VCF) is the files format for storing information on sequence variation including SNPs and indels (Danecek et al., 2011). The major drawback in variant calling step is the difficulty to differentiate between true variants and false positive variant that rises from WES sequencing errors. In this study, novel approaches is introduced to filter WES data for prioritize true variant, which efficiently capture false positive variant for high sensitivity and specificity step in identifying disease-causing variant (as shown in Figure 3.2).

Variants quality are filtered out using the following criteria: (i) SNPs quality equal to or larger than Phred score 30; sequencing depth larger than 5;(iii) local realignment process using Genome Analysis Toolkit (GATK) to minimize mismatch bases; (iv) evaluate detailed variant calling metrics using Picard-tools. Variant calling (SNPs and indels) are performed using GATK. Lastly, the genotypes of SNPs and indels are annotated using wANNOVAR database. Variant calling steps and annotation procedure are repeated for the accuracy of the analysis.

3.2.8 Variants Validation

Classification of variants based on the following criteria: (i) whether they were previously reported in literature review; (ii) the types of the variants (for example: missense or silent variants; (iii) the variant alleles frequency with less than 5% and 1% reported in the different population databases such as 1000 Genomes database (Consortium, 2015), NHLBI Exome Sequencing Project ESP6500 (http://evs.gs.washington.edu/EVS), Exome Aggregation Consortium (ExAC) (Karczewski et al., 2016) and Genome Aggregation Database (GnomAD); (iv) computational prediction functional score using Sorting Intolerant from Tolerant (SIFT) (P. C. Ng & Henikoff, 2003), Polymorphism Phenotyping v2 (PolyPhen-2) (Adzhubei, Jordan, & Sunyaev, 2013), and Mutation Taster (Schwarz, Rödelsperger, Schuelke, & Seelow, 2010).



Exclude common variants >1% MAF from ExAC and GnomAD from ~20K genes
Focus on ~300 PID genes (associated with phenotypes) with >5x coverage
Include Indels and splice site with >5X coverage
Exclude tolerated variants from SIFT, PolyPhen-2 and Mutation Taster
Candidates for disease-causing variants
Sanger sequencing validation

Figure 3.2: Established filtering approach used for prioritize genetic variants

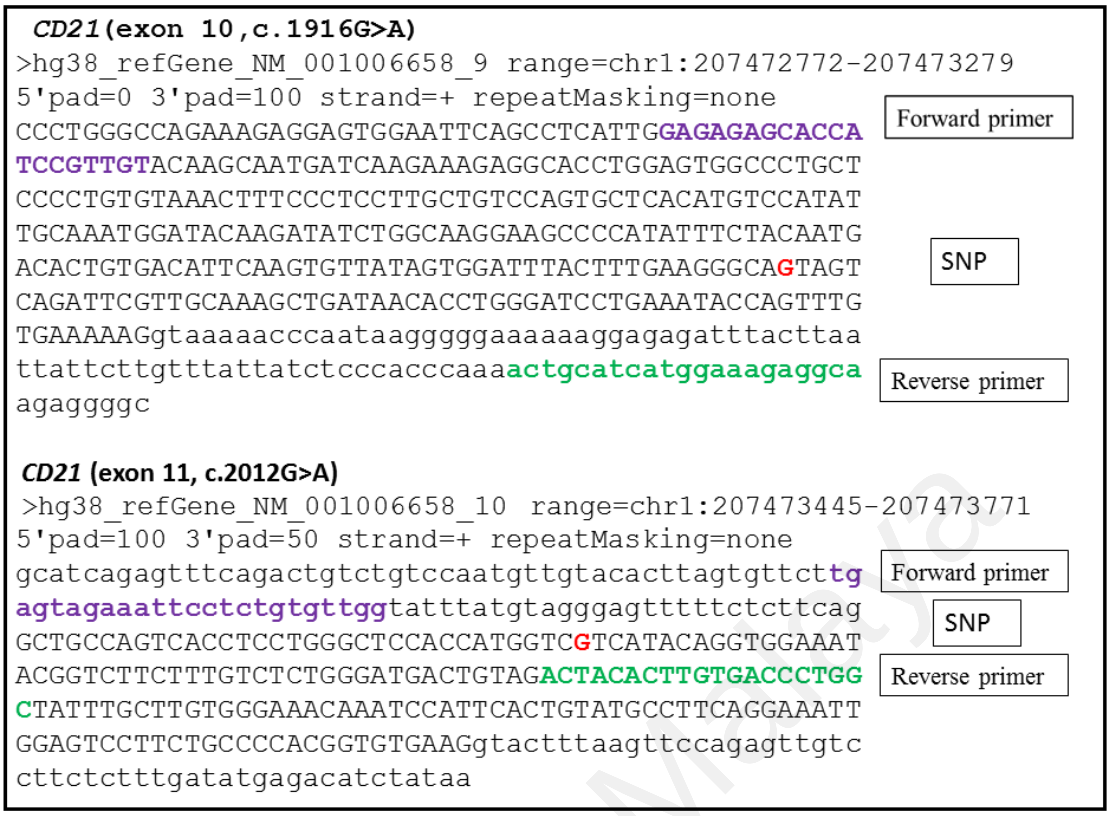
The candidate were validated using Sanger sequencing. Two pairs of primer for *CD21 (CR2)* were designed using a combination of the University of California Santa Cruz (UCSC) Table Browser and Primer3 v.0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0/) as shown in Figure 3.3 and the other primer set for selected variants were available in Appendix A. In addition, the PCR components were listed in the Table 3.5. The conditions of the PCR cycle were: 94 °C for 3 minutes, followed by 35 cycles of 94 °C for 30 seconds, 56 °C for 30 seconds, 72 °C for 20 seconds and the last extension cycle of 72 °C for 8 mins. No PCR replicates and positive samples are used due to limited DNA samples. Sanger sequencing was performed on the PCR products and the results were viewed using Unipro UGENE v1.29 software (http://ugene.net).

Table 3.4: Primer sequences used in this study.

Name	**Types**	**Sequence (5' to 3')**	**Annealing Temperature**	**Product Size (bp)**
CD21 (exon 10, c.1916G>A)	Forward	GAGAGAGCACCATCCGTTGT	52 ºC	362
	Reverse	TGCCTCTTTCCATGATGCAGT		
CD21 (exon 11,c.2012G>A)	Forward	TGAGTAGAAATTCCTCTGTGTTGGT	56 ºC	153
	Reverse	GCCAGGGTCACAAGTGTAGT		

Table 3.5: PCR components.

Component	**Volume (µL)**
10x PCR Buffer II	5.00
Forward Primer (10 µM)	1.25
Reverse Primer (10 µM)	1.25
dNTP (10 mM)	2.00
$MgCl_2$ (25 mM)	4.50
dh_2O	34.75
AmpliTaq (5U/µL)	0.25
Genomic DNA	1.00
Final volume	50.00



```
 CD21(exon 10,c.1916G>A)
>hg38_refGene_NM_001006658_9 range=chr1:207472772-207473279
5'pad=0 3'pad=100 strand=+ repeatMasking=none
CCCTGGGCCAGAAAGAGGAGTGGAATTCAGCCTCATTGGAGAGAGCACCA
TCCGTTGTACAAGCAATGATCAAGAAAGAGGCACCTGGAGTGGCCCTGCT
CCCCTGTGTAAACTTTCCCTCCTTGCTGTCCAGTGCTCACATGTCCATAT
TGCAAATGGATACAAGATATCTGGCAAGGAAGCCCCATATTTCTACAATG
ACACTGTGACATTCAAGTGTTATAGTGGATTTACTTTGAAGGGCAGTAGT
CAGATTCGTTGCAAAGCTGATAACACCTGGGATCCTGAAATACCAGTTTG
TGAAAAAGgtaaaaacccaataaggggaaaaaaggagagatttacttaa
ttattcttgtttattatctcccacccaaaactgcatcatggaaagaggca
agaggggc

CD21 (exon 11, c.2012G>A)
 >hg38_refGene_NM_001006658_10  range=chr1:207473445-207473771
5'pad=100 3'pad=50 strand=+ repeatMasking=none
gcatcagagtttcagactgtctgtccaatgttgtacacttagtgttcttg
agtagaaattcctctgtgttggtatttatgtagggagtttttctcttcag
GCTGCCAGTCACCTCCTGGGCTCCACCATGGTCGTCATACAGGTGGAAAT
ACGGTCTTCTTTGTCTCTGGGATGACTGTAGACTACACTTGTGACCCTGG
CTATTTGCTTGTGGGAAACAAATCCATTCACTGTATGCCTTCAGGAAATT
GGAGTCCTTCTGCCCCACGGTGTGAAGgtactttaagttccagagttgtc
cttctctttgatatgagacatctataa
```

Figure 3.3: Primer sequences for *CD21* gene. Purple and green colors represent the forward and reverse primer sequences and red color highlighted the deletion site of *CD21* gene.

3.2.9 Network analysis

CD21 was subjected to network analysis using GeneMania server (Warde-Farley et al., 2010) by using default parameter (network weighting: query-dependent weighting, network weighting code: automatic_select, number of gene results: 20, and number of attribute results: 10). Two parts of GeneMANIA algorithm were used which is a linear regression-based and a label propagation algorithm. A linear regression-based calculates a single composite functional association network from multiple data sources whereas a label propagation algorithm used for predicting gene function given the composite functional association network. Each set of genes function were further analysed to identify their gene-gene interactions.

CHAPTER 4: RESULTS

Chapter 4 is organized in the following way. First, Section 4.1 describes the clinical features and experimental analysis of the MZ twin, followed by the family pedigree analysis in Section 4.2. Section 4.3, 4.4 and 4.5 describes the experimental results for the preparation of WES dataset. Next, Section 4.6 shows the sequence quality of WES dataset. Section 4.7 deals with the overall WES statistics, followed by the variant calling information in Section 4.8 and 4.9. Section 4.10 presents the validation results of WES using Sanger sequencing technique (damaging variants, benign variants, false positive variants and indels). Section 4.11 shows the network analysis among the CVID-related genes.

4.1 Clinical Features and Laboratory Analysis of the MZ Twin

A pair of MZ twin siblings (P1 and P2) born to non-consanguineous parents of Malay ancestry (Figure 4.1) were the subjects of this report. The first few years of their life, they presented with few upper respiratory tract infection symptoms which resolved without requiring hospital admission. P1 were first hospitalized for upper respiratory tract infection at the age of seven, which she presented with tachypnea, and were noted to have digital clubbing, bronchiectasis, bicytopenia associated hepatospelomegaly and failure to thrive. Bone marrow aspiration did not suggest any malignancy infiltration for P1. Her sputum was positive for Tuberculosis Polymerase chain reaction (PCR) but her Tuberculosis culture were negative. Sputum for fungal and other bacteria examinations was negative. Her human immunodeficiency virus status was negative. Her twin sister (P2) was assessed and evaluated despite she did not have recurrent symptoms to suggest recurrent infection except for the skin rashes and upper respiratory tract infection (URTI). P2 was also found to have bronchiectasis, hepatosplenomegaly, anemia and failure to thrive. Further investigations revealed P2 had similar diagnosis as P1 and was treated accordingly. Lymphocyte enumeration test was conducted using flow cytometry, revealed reduced cell count for T cells, B cells and NK cells. P1 and P2 also showed low serum IgG (4.22 and 4.16 (Normal value=5.2-15.6)) and IgA (<0.02 (Normal value=0.54-3.6)), while the serum IgM level was normal (Table 4.1). The parents and 3 other siblings were not affected (Figure 4.1). Taken together, these findings suggest that the patients were suspected with Common Variable Immunodeficiency (CVID) without any definite diagnosis of the type of PID. Vaccine test was not available during the time of study. P1 and P2 were treated as Tuberculosis with failure to thrive and were given long term oxygen therapy, nutritional rehabilitation and regular immunoglobulin infusion (IVIG). WES analysis was utilized to both patients to determine the causative gene and gene variation for the disease phenotype. At the age of 10, P1 passed away due to severe lung infections and followed by P2 at the age of 11 years old.

Table 4.1: Clinical and laboratory findings.

Features	P1	P2
Age at presentation	7 yr	7 yr
Sex	Female	Female
Geographic location	Malaysia	Malaysia
Race	Malay	Malay
Infection	URTI	URTI
Lymphoproliferation	Hepatosplenomegaly	Hepatosplenomegaly
Lung disease	Bronchiectasis	Bronchiectasis
Bone marrow aspiration	Not suggestive for any malignancy infiltration	
Others	Anaemia, digital clubbing, persistent tacyhpnea, failure to thrive	
Treatment	IVIG	IVIG
Outcome	Passed away at age 10	Passed away at age 11
Immunoglobulins	Low IgG and IgA	Low IgG and IgA
IgG (g/L)	4.22 (N=5.2-15.6)	4.16 (N=5.2-15.6)
IgA (g/L)	<0.02 (N=0.54-3.6)	<0.02 (N=0.54-3.6)
IgM (g/L)	1.27 (N=0.13-2.4)	9.97 (N=0.13-2.4)
Lymphocyte populations	Progressive lymphopenia (low T cell subsets, low B cells and low NK cells)	Progressive lymphopenia (low T cell subsets, low B cells and low NK cells
T cells (x10^{6}/L)	671 (N=1,400-2,000)	1,066 (N=1,400-2,000)
B cells (x106/L)	26 (N=300-500)	82 (N=300-500)
CD4^{+} (x106/L)	463 (N=700-1,000)	726 (N=700-1,000)
CD8^{+} (x106/L)	235 (N=600-900)	295 (N=600-900)
NK cells (x106/L)	79 (N=200-600)	170 (N=200-600)

yr=year, IVIG=Intravenous immunoglobulin therapy, URTI=Upper respiratory tract infections, N=Normal value, NK=Natural killers, values in parentheses represent age-matched reference values.

The progresses on laboratory findings of serum immunoglobulin were monitored as shown in Table 4.2. P1 and P2 had decreased production and redistribution of lymphocytes, which indicates an immunodeficiency disease. In addition, C3 and C4 component were also being monitored as both played a role in the human immune defenses for the activation complement pathway against bacterial infections (Noris & Remuzzi, 2013).

Table 4.2: Progressive low of serum immunoglobulins.

Name	**IgG (g/L)** **N=5.2-15.6**	**IgA (g/L)** **N=0.54-3.6**	**IgM (g/L)** **N=0.13-2.4**	**C3**	**C4**
P1	6.84	<0.02	0.99	1.50	0.37
P1	4.89	<0.02	0.84	1.11	0.24
P1	7.68	<0.02	3.10	Insufficient	Insufficient
P1	6.08	<0.02	1.76	0.85	0.17
P1	4.95	<0.02	0.85	0.76	0.19
P1	3.96	<0.02	0.54	-	-
P1	5.31	<0.02	0.45	0.87	0.15
P1	4.70	<0.02	0.68	-	-
P1	5.19	<0.02	5.80	-	-
P1*	4.16	<0.02	9.97	-	-
P1	5.81	<0.02	>15.00	0.51	0.21
P1	10.24	<0.02	>7.63	-	-
P1	8.57	<0.02	>15.26	-	-
P1	8.87	<0.02	>7.63	-	-
P1	10.92	<0.02	>15.26	-	-
P2	4.68	<0.02	4.13	0.81	0.21
P2*	4.22	<0.02	1.27	-	-
P2	5.11	<0.02	1.36	0.88	0.10
P2	5.38	<0.02	2.61	-	-

*Samples taken for WES; N=reference value according to age; ‘-’ =data is not available

4.2 Family Pedigree Analysis and Lymphocyte Subset Enumeration Test

Mode of inheritance can be roughly divided into Mendelian and non-Mendelian patterns (Botstein & Risch, 2003). Based on the observation of family pedigree for this MZ twin as shown in Figure 4.1, two possible outcomes can be withdraw;1) the variant follows the Mendelian inheritance (e.g. autosomal recessive), or 2) the variant does not follow Mendelian pattern of inheritance (*de novo* mutation, the mutation is not inherited).

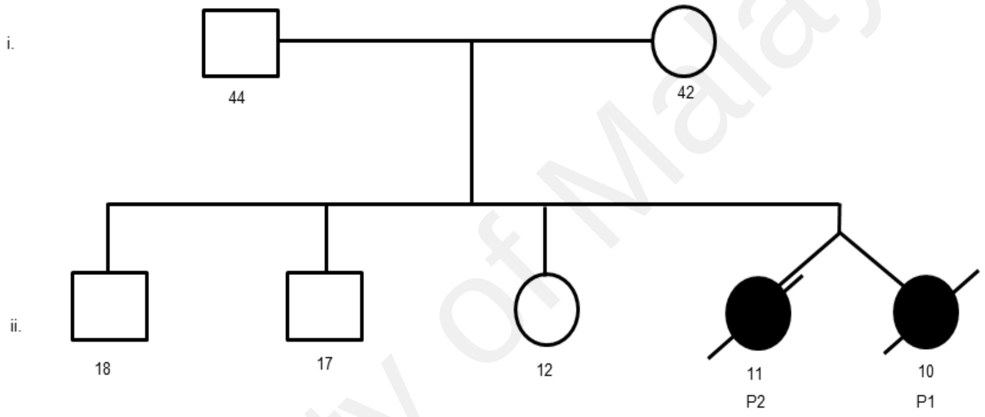


Figure 4.1: Family pedigree and PID test of the affected MZ twin. Pedigree showing PID patients; Roman numerals i and ii indicates generations, rectangles represent males, circles represent females, filled circles indicates affected MZ twin and crossed-out circles represent the deceased. Samples from unaffected family members were not available for genotyping. The values under each symbol represent the present age of each family member in years.

4.3 DNA Extraction

Genomic DNA extraction was performed using standard laboratory protocol. Samples quality for exome library were low, however, the samples can still be used for WES as shown in Table 4.3.

Table 4.3: Information of extraction genomic samples.

Sample	Mass DNA	Volume (µL)	[DNA] ng/µL	OD$_{260/280}$	0D$_{260/230}$
P1	0.91	150	60.5	1.97	1.52
P2	0.67	150	44.8	2.02	1.38

4.4 Sample Assessment of DNA Extraction

Genomics sample assessment were done before the samples were sent for WES using method of concentration determination and sample integrity test as shown in Table 4.4.

4.4.1 Method of concentration determination

Qubit Fluorometer was used for the DNA concentration determination.

Table 4.4: Determination of exome library type sample using Qubit Fluorometer.

Sample	Total Mass DNA	Volume (µL)	dsDNA concentration (ng/µL)	Test Result
P1	0.68	131	5.2	Sample is slight degraded; The concentration & total mass is too low
P2	0.84	131	6.4	Sample is slight degraded; The concentration & total mass is too low

4.4.2 Method of sample integrity test

This method is essential in order to ensure sample integrity and maintain a good quality before the samples being subjected for WES. Both samples were slightly degraded as shown in Figure 4.2.

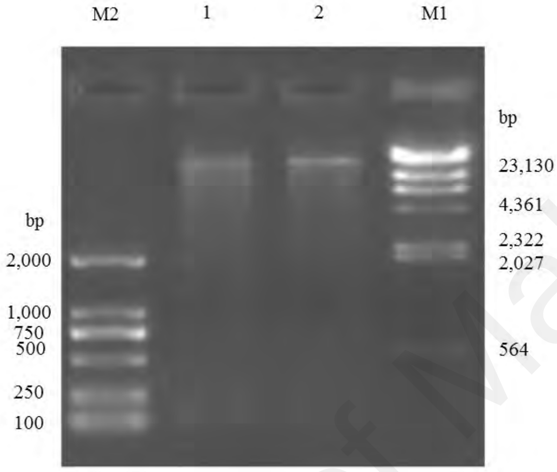


Figure 4.2: PCR results for the both samples using 1% agarose gel; Voltage used: 150 V; Electrophoresis Time: 40 min; M1 markers: λ-Hind III digest (Takara,US) and for the M2 markers: 1 kb plus DNA ladder (Tiangen, China)

4.5 Library Preparation

Library preparation exome kits contain optimally formulated materials selected through Illumina technology and implement a highly optimized “bead” strategy that facilitate sample preparation of DNA for enable WES library construction efficiency and facilitate automation. In this study, the Agilent Sure Select Human All Exons V5 kits had been known to offer higher success rates compared to conventional library preparation and were carefully designed to capture exonic regions of the genomic samples for P1 and P2. Preparation of high quality libraries with a good quality of yield is an important step in NGS workflows. Detailed information on exome kits used for this study was described in Table 4.5.

Table 4.5: Information on detailed exome kits used for this study.

Features	**P1**	**P2**
Exome-kits	Agilent Sure Select Human All Exons V5	Agilent Sure Select Human All Exons V5
Probes types	Biotinylated RNA baits	Biotinylated RNA baits
Probe design	Non-overlapping, paired ends reads used to fill gaps	Non-overlapping, paired ends reads used to fill gaps
Targeted regions	50 Mb	50 Mb
Targeted genes : 20,313	99.26% (~20,162 genes)	99.39% (~20,189 genes)
Targeted coverage	100X	100X

4.6 Sequence Quality of Exome Data

To ensure the sequence quality performance of our raw data, a comparison before and after trimming the dataset were performed. Figure 4.3 shows the quality scores at each base position in the FastQ Illumina output files using the BoxWhisker type plot. The components of the plot were explained as follows: the central red line represent the median value, the yellow box shows the inter-quartile range (25-75%), and mean quality represented in blue line. Bases in between Phred score 30-40 were considered as good quality data, which may be used for further analysis step.

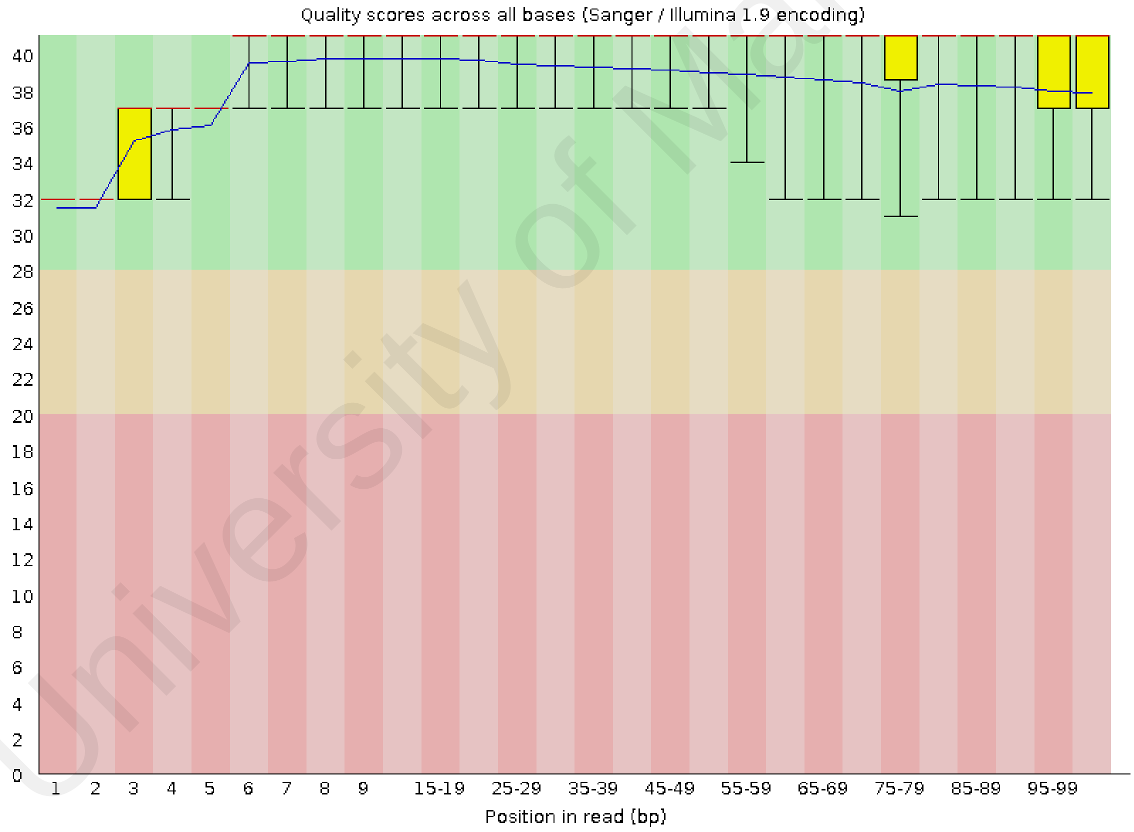


Figure 4.3: Quality scores across all bases. (a) Quality of PE1 reads before trimming

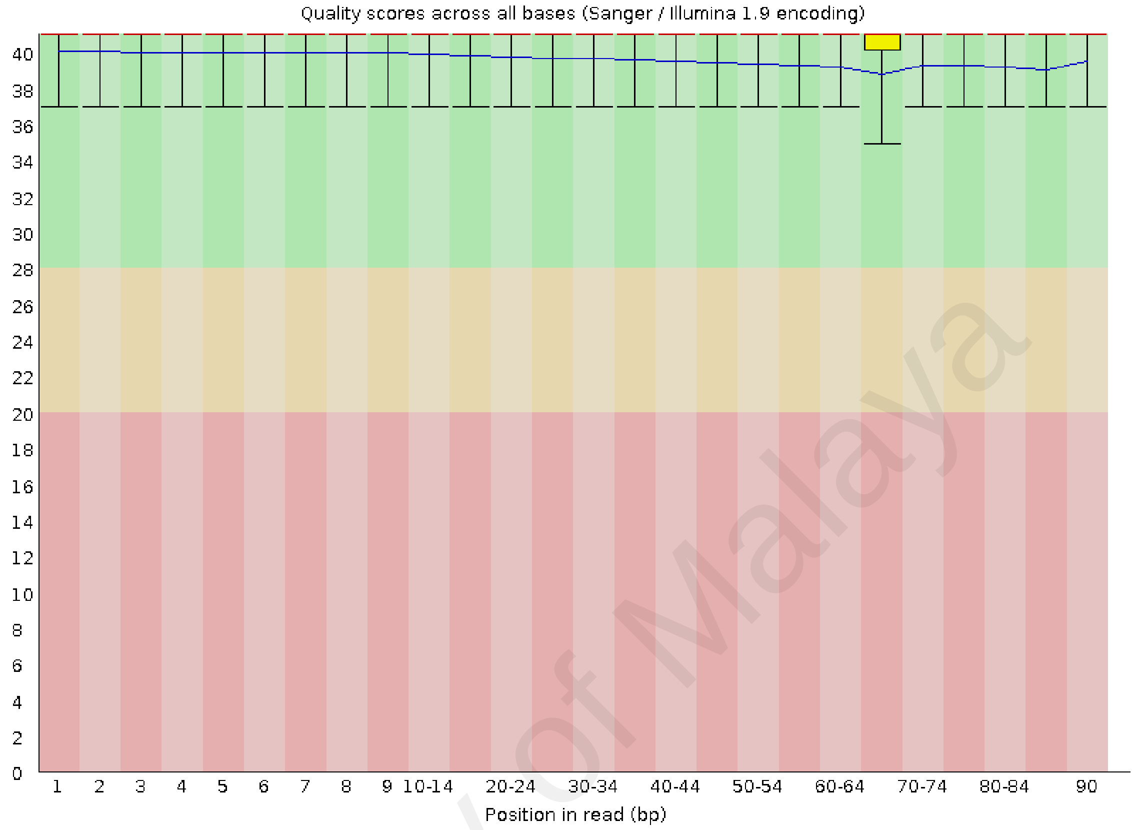


Figure 4.3: Quality scores across all bases**.** (b) Quality of PE2 reads before trimming

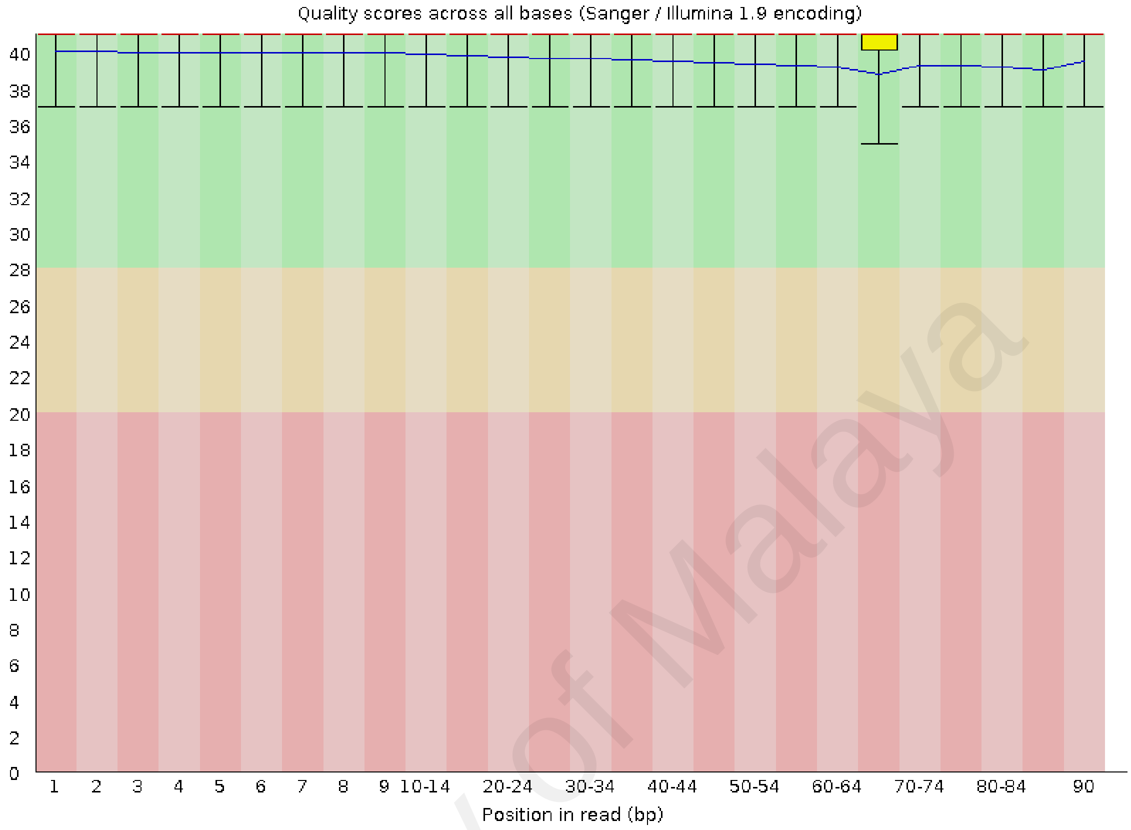


Figure 4.3: Quality scores across all bases. (c) Quality of PE1 reads after trimming

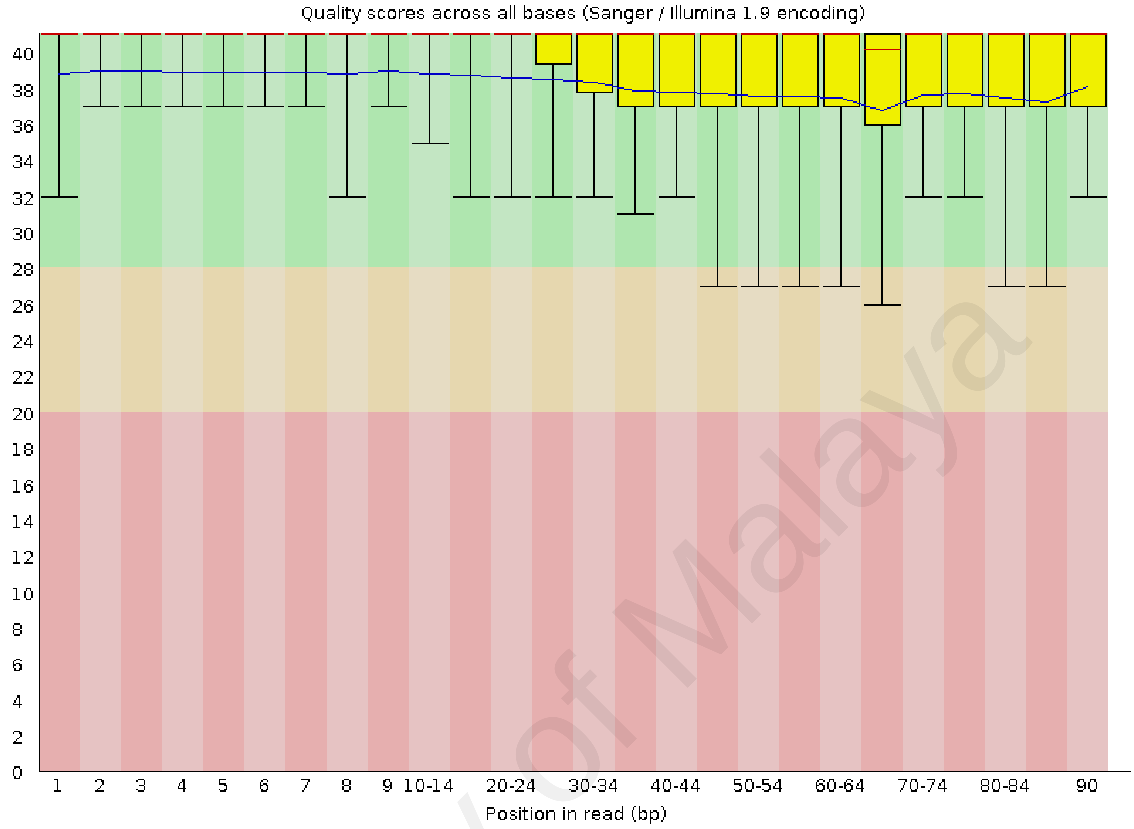


Figure 4.3: Quality scores across all bases. (d) Quality of PE2 reads after trimming

Figure 4.4 compares the average sequence quality and the distributions of the average quality were plotted in the graph, which enables to represent low quality bases across all reads. In addition, the y-axis represents the number of reads and the x-axis represents the mean quality score. The dataset for P1 and P2 showed good score after trimming compared to before trimming dataset.

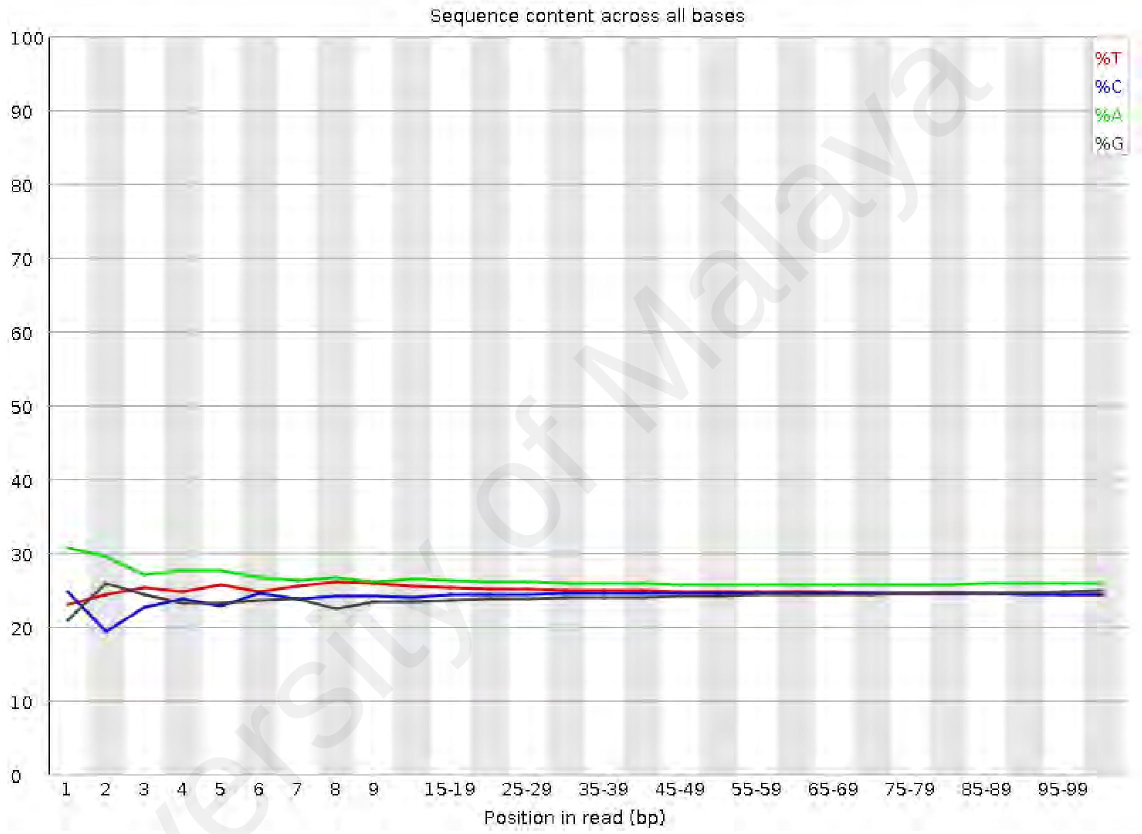


Figure 4.4: Sequence content across all bases. (a) Sequence content of PE1 reads before trimming

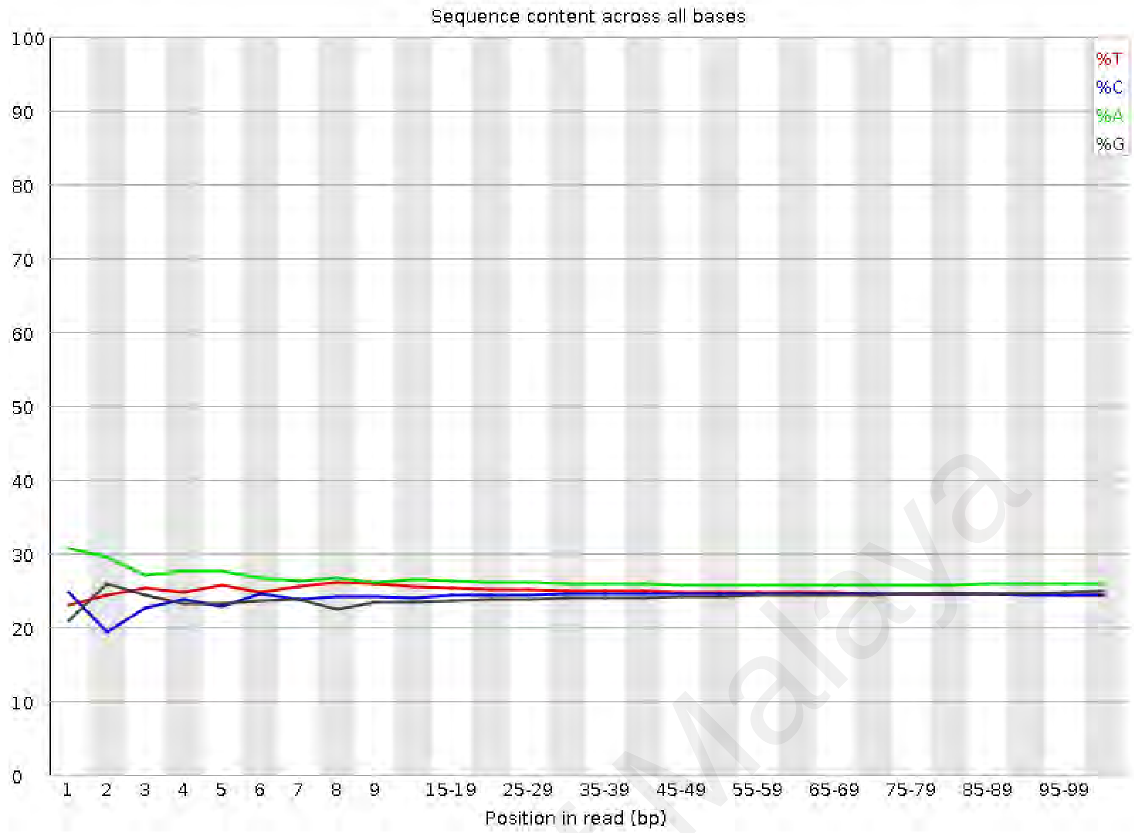


Figure 4.4: Sequence content across all bases. (b) Sequence content of PE2 reads before trimming

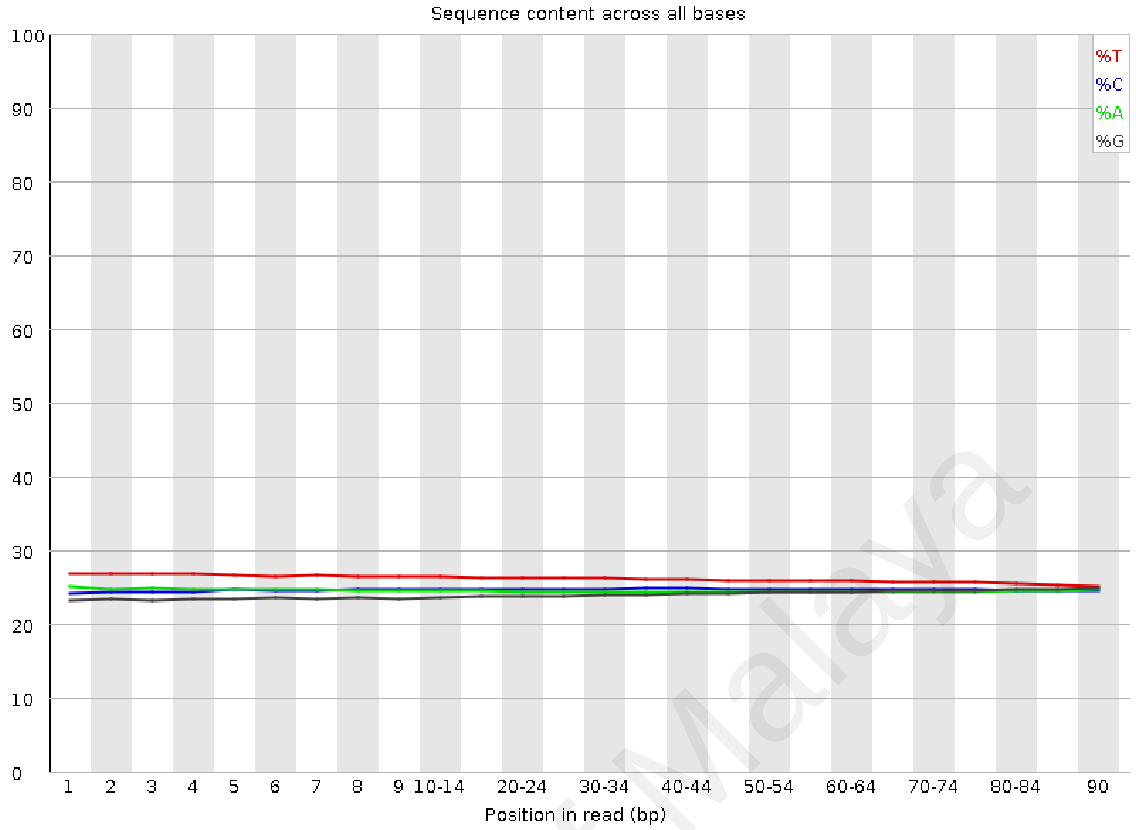


Figure 4.4: Sequence content across all bases. (c) Sequence content of PE1 reads after trimming

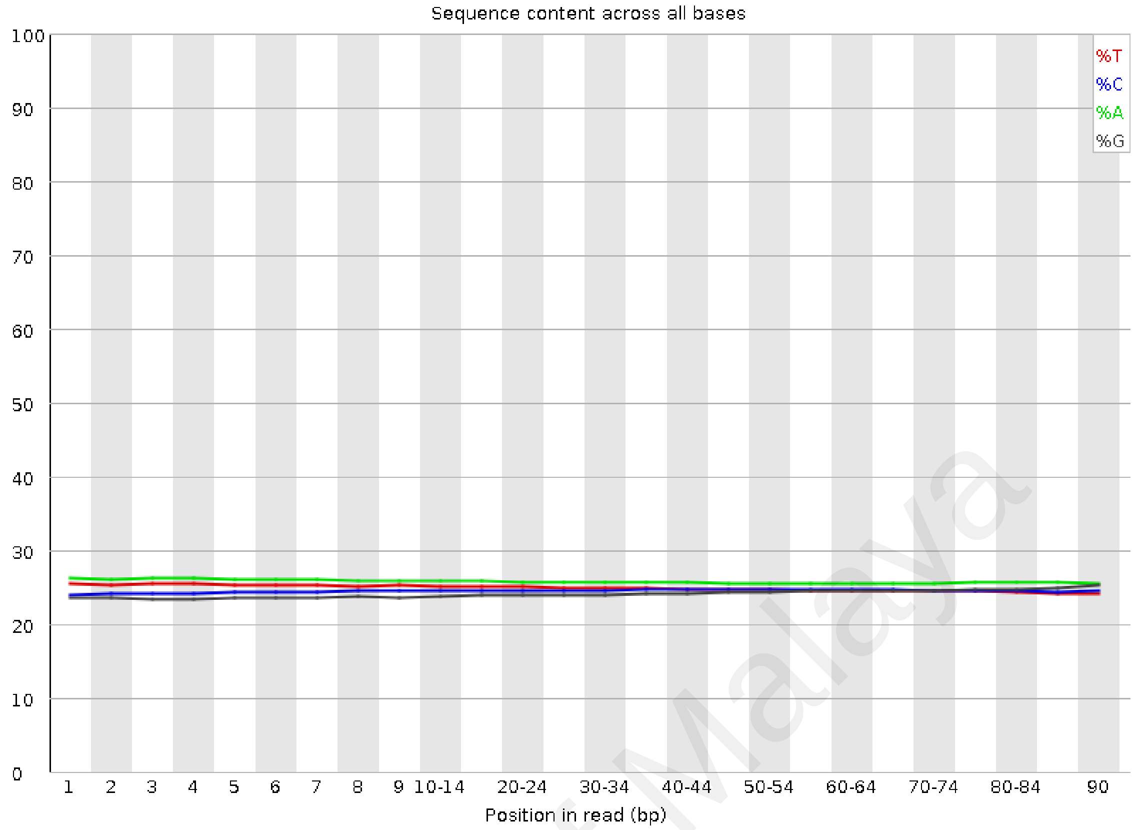


Figure 4.4: Sequence content across all bases. (d) Sequence content of PE2 reads after trimming

Additional quality checks for QC content across all bases were performed as shown in Figure 4.5. Generally, GC content were assumed to equally distributed across the genome (50% of GC and 50% of AT), which the central peak represents the overall GC content for the P1 and P2 dataset. In conclusion, WES dataset for P1 and P2 were considered as a reliable and good quality prior to the library preparation.

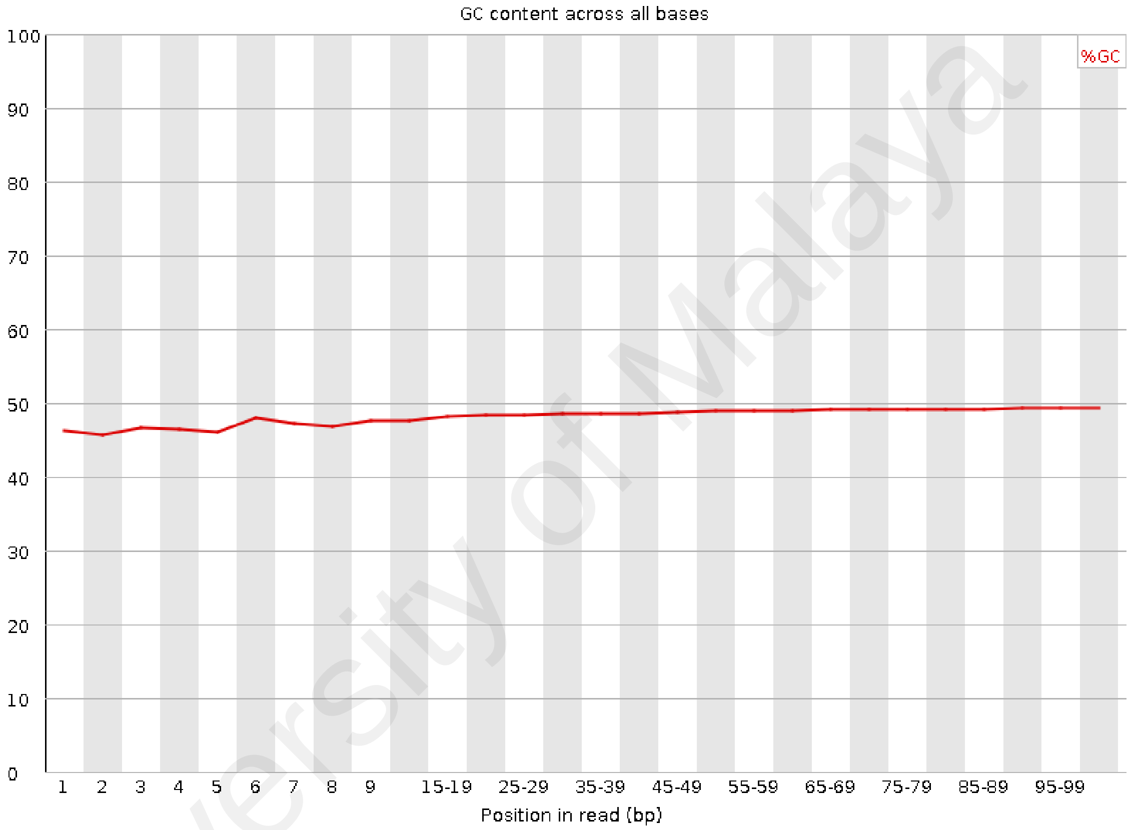


Figure 4.5: GC content across all bases. (a) GC content of PE1 reads before trimming

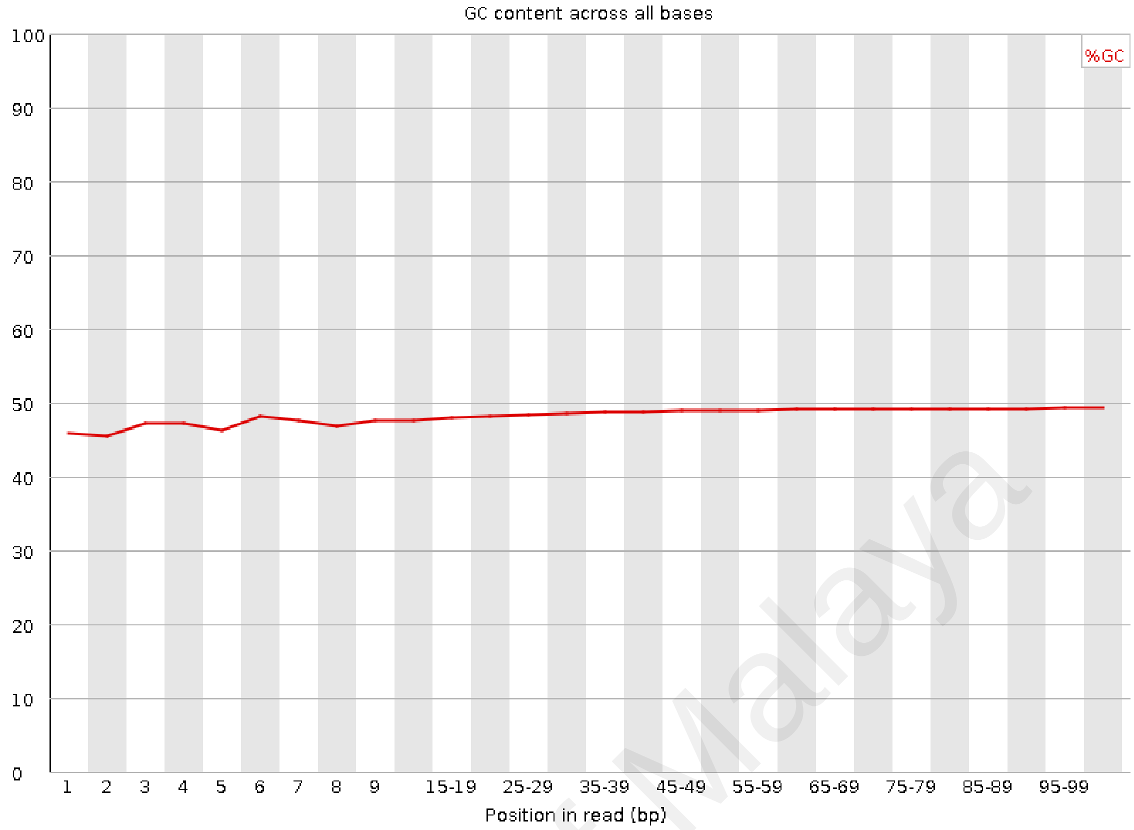


Figure 4.5: GC content across all bases. (b) GC content of PE2 reads before trimming

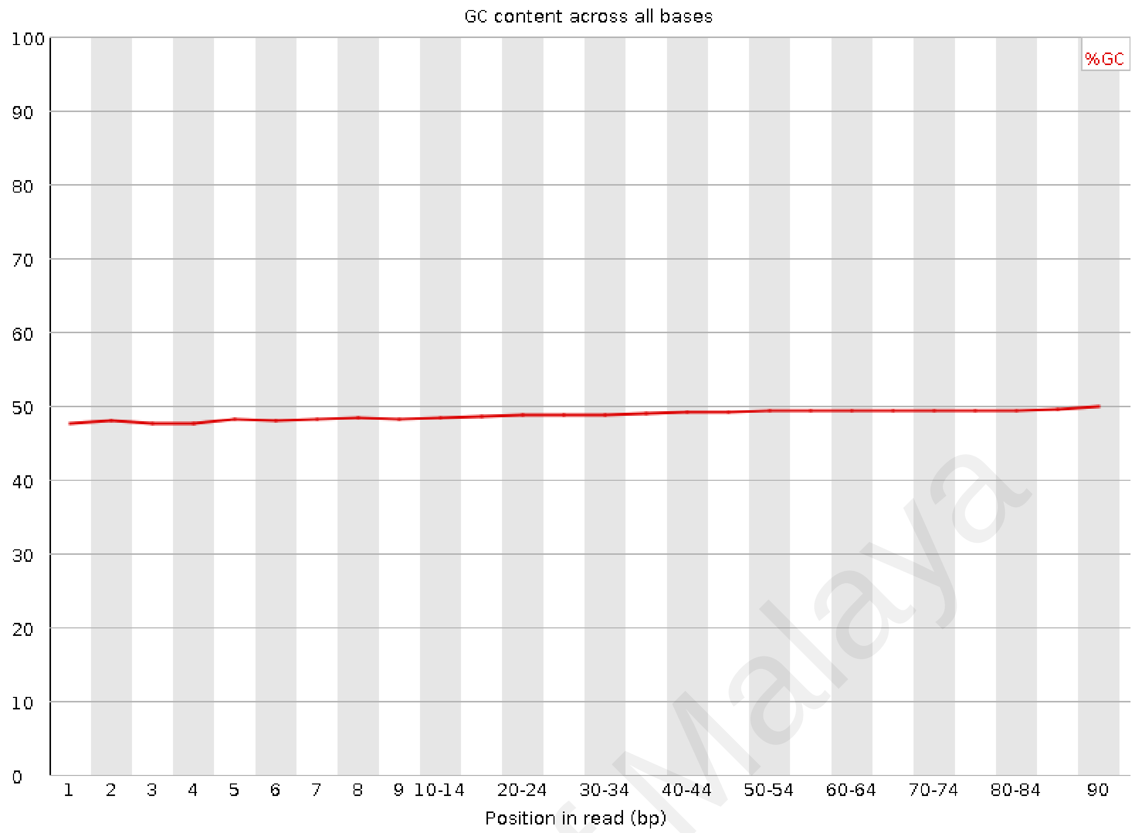


Figure 4.5: GC content across all bases**.** (c) GC content of PE1 reads after trimming

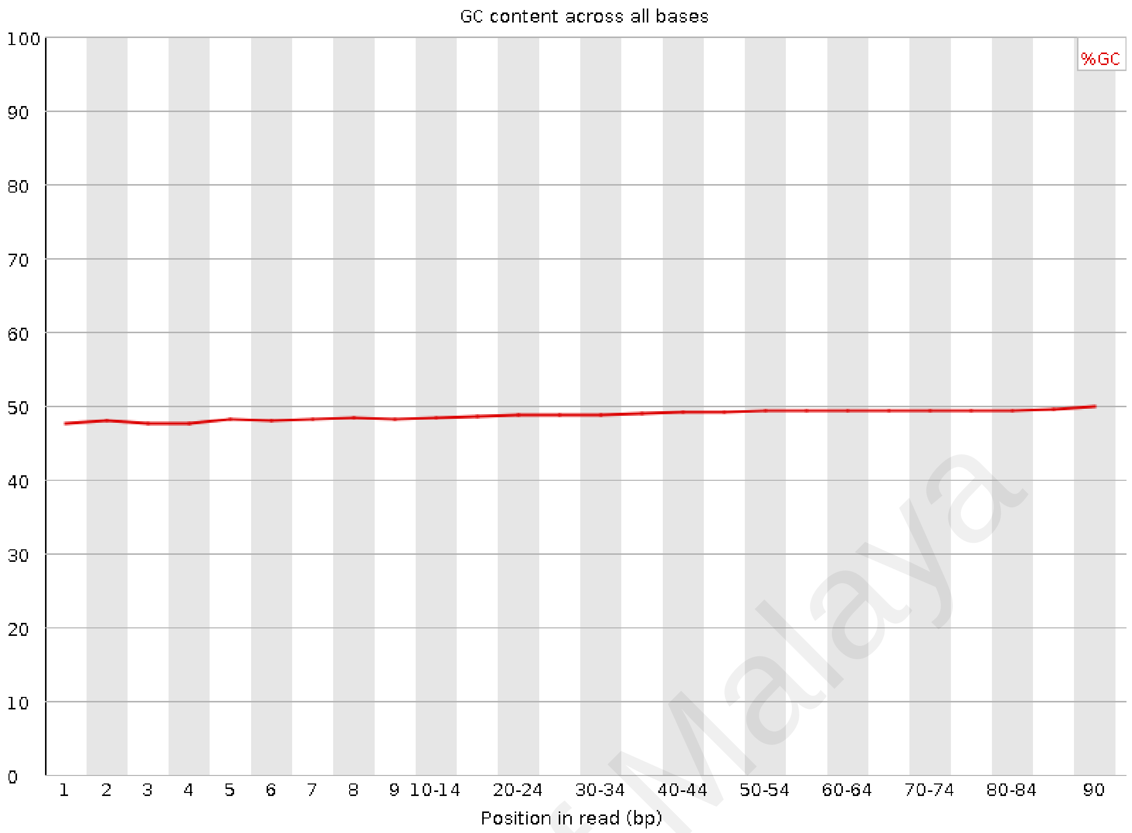


Figure 4.5: GC content across all bases. (d) GC content of PE2 reads after trimming

4.7 Exome Sequencing Statistics

A total numbers of 84,933,915 and 81,842,083 paired-end reads were received for both patients. The WES dataset were processed, aligned and mapped to human reference genome hg38 (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/) with 99.26% and 99.39% were mapped to the reference genome for P1 and P2, respectively. We focused on SNPs and indels as this is the two largest types of human genetic variation (Consortium, 2012). In general, WES variant calling analysis found >250,000 SNPs and <50,000 indels in both P1 and P2 as shown in Figure 4.6. WES able to capture 20,162 genes in P1 and 20,189 genes in P2. Detailed WES metrics were summarized in Table 4.6.

Table 4.6: Detailed on WES results generated from bioinformatics analysis.

	P1	**P2**
Total reads	84,933,915	81,842,083
Total bases (bp)	8,493,391,500	8,184,208,300
Mapped reads to hg38	99.26% (84,308,897)	99.39% (81,342,537)
Targeted genes	20,162	20,189
GC content (%)	48.80	50.53

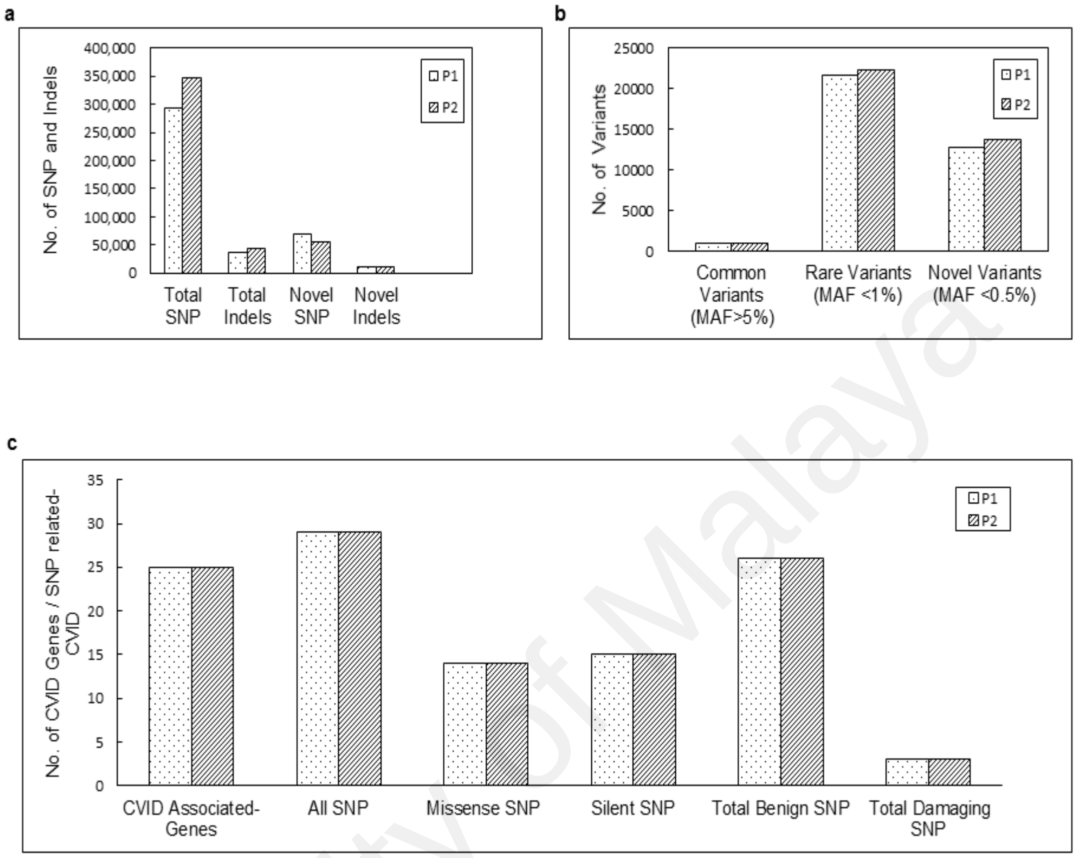


Figure 4.6: Exome sequencing metrics. (a) Number of SNPs and indels found in MZ twin exome data (b) Number of variants indicated as common, rare and novel between MZ twin exome data based on Minor Allele Frequency in ExAC database (c) Number of CVID-related genes with various types of variants found in this study

4.8 Variant Calling Information

Further analyses for variant calling were revealed 69,611 novel SNPs in P1 and 55,800 in P2, whereas 11,126 novel indels in P1 and 10,141 novel indels in P2 WES dataset as listed in Table 4.7. In addition, common variants, rare and novel variants were also calculated using GATK and wANNOVAR tools.

Table 4.7: Variant calling information for both dataset.

Properties	P1	P2
Total SNPs[a]	293,959	347,980
Total indels[a]	37,210	42,690
Novel SNPs[a]	69,611	55,800
Novel indels[a]	11,126	10,141
Common Variants (MAF >5%)[b]	951	934
Rare Variants (MAF <1%)[b]	21,584	22,307
Novel Variants (MAF <0.5%)[b]	12,770	13,620

[a] GATK analysis
[b] wANNOVAR annotation analysis based on ExAC database

4.9 Variant Analysis

Following sequence alignment and variant calling, we filtered patient exomes to identify specific PID diagnosis for these twin patients. We screened 240 PID genes as listed in Table 2.3 based on standard classification of PID (Picard et al., 2015). The genetic variation analysis revealed 112 known SNPs with missense variants. Since the clinical symptoms resemble CVID, we narrowed the analysis to the 31 genes known to be associated to CVID (Table 4.8). We identified 17 missense SNPs and 15 silent SNPs for the CVID-associated genes (Table 4.9 and 4.10). The information on depth of coverage for 31 CVID-related genes was available through the Table 4.11. The missense SNP were subjected to damaging variant prediction by SIFT, PolyPhen2 and Mutation Taster prediction software (Table 4.9). Only 3 of the missense SNPs were predicted to be damaging variant by at least one of the prediction software (Table 4.9). However, although predicted as damaging by PolyPhen2 software, the variations on *FANCA* gene were not likely to be the causative variation for the disease, because of high frequency of occurrence for the variation as shown by MAF analysis and it is does not fit the genetic inheritance pattern for *FANCA* gene (Table 4.9). Another two SNPs for *CD21* (also known as *CR2*) is likely to be disease-causing gene as it is fit the genetic pattern for *CD21* deficiency (compound heterozygous case) and predicted to be damaging by PolyPhen2 software.

Table 4.8: CVID-related genes.

No.	CVID-related genes	WES Analysis on P1	WES Analysis on P2
1.	*CD19* (Kanegane et al., 2007)	rs2904880	rs2904880
2.	*CD27* (Ahn & Cunningham-Rundles, 2009)	rs2532502	rs2532502
3.	*CR2 (CD21)* (Patuzzo et al., 2013)	rs17615, rs17616	rs17615, rs17616
4.	*CD86* (Denz et al., 2000)	rs2681417	rs2681417
5.	*ADAM28* (J. H. Park, E. S. Resnick, & C. Cunningham-Rundles, 2011)	rs7814768	rs7814768
6.	*SDK1* (Keller & Jyonouchi, 2013)	rs138116831	rs138116831
7.	*CTLA4* (Schubert et al., 2014)	rs231775	rs231775
8.	*STXBP2* (Maffucci et al., 2016)	rs6791	rs6791
9.	*FANCA* (Sekinaka et al., 2017)	rs11646374, rs2239359, rs17232910, rs7195066, rs9282681	rs11646374, rs2239359, rs17232910, rs7195066, rs9282681
10.	*FANCE* (Sekinaka et al., 2017)	rs4713867	rs4713867
11.	*PRKCD* (J A Bogaert et al., 2016)	rs2306574	rs2306574
12.	*PIK3CD* (Sheikhbahaei et al., 2016)	rs11121484	rs11121484
13.	*NFKB1* (Bryant & Tangye, 2016)	rs1609993	rs1609993
14.	*NFKB2* (Liu et al., 2014)	rs4919633	rs4919633
15.	*TNFSF12 (TWEAK)* (H.-Y. Wang et al., 2013)	rs3803798	rs3803798
16.	*ITGAM* (Maggadottir et al., 2015)	rs1143682	rs1143682
17.	*MS4A1 (CD20)* (Joon H. Park, Elena S. Resnick, & Charlotte Cunningham-Rundles, 2011)	rs2070770	rs2070770
18.	*PLCG2* (Aderibigbe, Priel, Lee, & et al., 2015)	rs1143685	rs1143685
19.	*PIK3R1* (Elgizouli et al., 2016)	rs3730090	rs3730090
20.	*RAC2* (Alkhairy et al., 2015)	rs2239774	rs2239774
21.	*BLK* (Compeer et al., 2015)	rs2306234	rs2306234
22.	*IL21* (Salzer et al., 2014)	rs4833837	rs4833837
23.	*CARD11 (CARMA1)* (Tampella et al., 2011)	rs1124581	rs1124581
24.	*LRBA* (Lopez-Herrera et al., 2012)	rs1782360	rs1782360
25.	*LRBA* (Lopez-Herrera et al., 2012)	Novel deletion	Novel deletion
25.	*IL21R* (Bogaert et al., 2016)	-	-
26.	*TNFSF13B (BAFF)* (Kopecký & Lukešová, 2007)	-	-

Table 4.8, continued.

No.	CVID-related genes	WES Analysis on P1	WES Analysis on P2
27.	*TNFRSF13C (BAFFR)* (Warnatz et al., 2009)	-	-
28.	*VAV1* (Bogaert et al., 2016)	-	-
29.	*XIAP* (Gulez et al., 2011)	-	-
30.	*ICOS* (Bogert et al., 2016)	-	-
31.	*IKZF1 (IKAROS)* (Yong, Salzer, & Grimbacher, 2009)	-	-

Symbols '-' = No mutation

Table 4.9: Missense variants.

Gene	**Chr**	**Exon**	**DNA changes**	**AA changes**	**Zyg**	**1**	**2**	**3**	**MAF (%)**
									GnomAD_ exome_ALL
CD19 **(rs2904880)**	16	3	c.520C>G	p.Leu174Val	Hom	T	B	P	0.7183
CD27 **(rs2532502)**	12	6	c.698A>G	p.His233Arg	Hom	T	B	P	0.9928
CD86 **(rs2681417)**	3	3	c.217G>A	p.Val73Ile	Hom	T	B	P	0.9211
CD21(CR2) **(rs17615)**	1	10	c.1916G>A	p.Ser639Asn	Het	T	D	P	0.2609
CD21 (CR2) (rs17616)	1	11	c.2012G>A	p.Arg671His	Het	T	D	P	0.2597
LRBA **(rs1782360)**	4	23	c.3269C>G	p.Ala1090Gly	Het	T	B	P	0.1245
ADAM28 **(rs7814768)**	8	22	c.2293G>A	p.Val765Met	Hom	T	B	P	0.9875
SDK1 **(rs671694)**	7	7	c.383A>G	p.His128Arg	Het	T	B	P	0.7519
SDK1 **(rs138116831)**	7	15	c.2161G>A	p.Val721Ile	Het	T	B	P	0.0008
CTLA4 (rs231775)	2	1	c.49A>G	p.Thr17Ala	Het	T	B	P	0.415
STXBP2 **(rs6791)**	19	18	c.1567A>G	p.Ile523Val	Hom	T	B	P	0.6342
FANCA **(rs1800282)**	16	1	c.17T>A	p.Val6Asp	Het	D	B	N	0.0769
FANCA **(rs11646374)**	16	14	c.1235C>T	p.Ala412Val	Het	T	B	P	0.0657
FANCA **(rs2239359)**	16	16	c.1501G>A	p.Gly501Ser	Het	T	B	P	0.5015
FANCA **(rs17232910)**	16	22	c.1927C>G	p.Pro643Ala	Het	T	B	P	0.0669
FANCA **(rs7195066)**	16	26	c.2426G>A	p.Gly809Asp	Het	T	B	P	0.471
FANCA **(rs9282681)**	16	40	c.3982A>G	p.Thr1328Ala	Het	T	B	P	0.0653

Chr= Chromosome, AA= Amino Acid, Zyg=Zygosity, 1=SIFT, 2=Poly-phen2, 3=Mutation Taster, Hom= Homozygous, Het= Heterozygous, T= Tolerated, D= Damaging, P= Polymorphism, B=Benign N= Neutral, MAF= Minor Allele Frequency, Blue color=disease-causative gene

Table 4.10: Silent variants.

Gene	Chr	Exon	DNA changes	AA changes	Zygosity	MAF (%)
						GnomAD_exome_ALL
PRKCD **(rs2306574)**	3	15	c.1441C>T	p.Leu481Leu	Hom	0.7511
PIK3CD **(rs11121484)**	1	22	c.2808C>T	p.Tyr936Tyr	Het	0.1658
NFKB2 **(rs4919633)**	10	12	c.1269A>G	p.Pro423Pro	Hom	0.9976
NFKB1 **(rs1609993)**	4	12	c.1140T>C	p.Ala380Ala	Hom	0.9398
TNFSF12 **(*TWEAK*)** **(rs3803798)**	17	7	c.600G>C	p.Ala200Ala	Hom	0.5484
ITGAM **(rs1143682)**	16	20	c.2499G>A	p.Thr833Thr	Hom	0.3499
MS4A1 **(*CD20*)** **(rs2070770)**	11	3	c.216C>T	p.Ile72Ile	Het	0.0717
PLCG2 **(rs1143685)**	16	2	c.174T>C	p.Ala58Ala	Het	0.7272
PIK3R1 **(rs3730090)**	5	3	c.87C>T	p.Phe29Phe	Hom	0.0569
RAC2 **(rs2239774)**	22	2	c.81C>G	p.Ala27Ala	Hom	0.1526
BLK **(rs2306234)**	8	9	c.843T>C	p.Phe281Phe	Het	0.8357
TNFRSF13B ***(TACI)*** **(rs11078355)**	17	5	c.831T>C	p.Ser277Ser	Het	0.4932
IL21 **(rs4833837)**	4	3	c.234C>T	p.Cys78Cys	Hom	0.7428
CARD11 ***(CARMA1)*** **(rs1124581)**	7	25	c.3276A>G	p. Arg1092Arg	Hom	0.509
FANCE ***(rs4713867)***	6	2	c.387A>C	p.Pro129Pro	Het	0.721
FANCA **(rs1800331)**	16	13	c.1143G>T	p.Thr381Thr	Het	0.0661
FANCA **(rs17226980)**	16	30	c.2901C>T	p.Ser967Ser	Het	0.0665
FANCA **(rs1800358)**	16	37	c.3654A>G	p.Pro1218Pro	Het	0.1005
FANCA **(rs11649210)**	16	38	c.3807G>C	p.Leu1269Leu	Het	0.0924

Chr= Chromosome, AA= Amino Acid, Hom= Homozygous, Het= Heterozygous, N/A=Not available, MAF= Minor Allele Frequency, No prediction available from SIFT, Poly-Phen2 and Mutation Taster tools

Table 4.11: The depth of coverage on WES data.

No.	**Gene** **(Missense mutations)**	**Depth of Coverage (DP)**		**Variant ID**
		P1	**P2**	
1.	*CD19*	18	74	rs2904880
2.	*CR2* (exon 10)	19	55	rs17615
3.	*CR2* (exon 11)	23	16	rs17616
4.	*CD27*	12	13	rs2532502
4.	*CD86*	22	15	rs2681417
5.	*LRBA* (SNPs)	51	29	rs1782360
6.	*ADAM28*	19	5	rs7814768
7.	*SDK1*	15	11	rs671694
8.	*CTLA4*	35	16	rs231775
9..	*STXBP2*	N/A	45	rs6791
10.	*FANCA*	23	31	rs1800282
11.	*FANCA*	64	92	rs7190823
12.	*FANCA*	13	23	rs11646374
13.	*FANCA*	4	5	rs2239359
14.	*FANCA*	15	16	rs17232910
15.	*FANCA*	30	22	rs7195066
16.	*FANCA*	17	19	rs9282681
	Gene **(Silent mutations)**	**Depth of Coverage (DP)**		**Variant ID**
		P1	**P2**	
17..	*PRKCD*	33	33	rs2306574
18..	*PIK3CD*	26	49	rs11121484
19.	*NFKB2*	39	39	rs4919633
20.	*NFKB1*	14	8	rs1609993
21.	*TNFSF12 (TWEAK)*	22	62	rs3803798
22..	*ITGAM*	59	78	rs1143682
23.	*MS4A1 (CD20)*	9	35	rs2070770
24.	*PLCG2*	30	66	rs1143685
25.	*PIK3R1*	14	89	rs3730090
26.	*RAC2*	21	25	rs2239774
27.	*BLK*	20	46	rs2306234
28.	*TNFRSF13B (TACI)*	19	31	rs11078355
29.	*IL21*	8	6	rs4833837
30.	*CARD11 (CARMA1)*	7	13	rs1124581
31.	*FANCE*	19	89	rs4713867

N/A= Not Available

4.10 Validation of Variants using Sanger Sequencing

Damaging variants were validated by Sanger sequencing. Figure 4.7 showed Sanger results on damaging variant candidates, which are listed as part of the missense mutations in the previous section (Table 4.9).

4.10.1 Damaging gene candidates

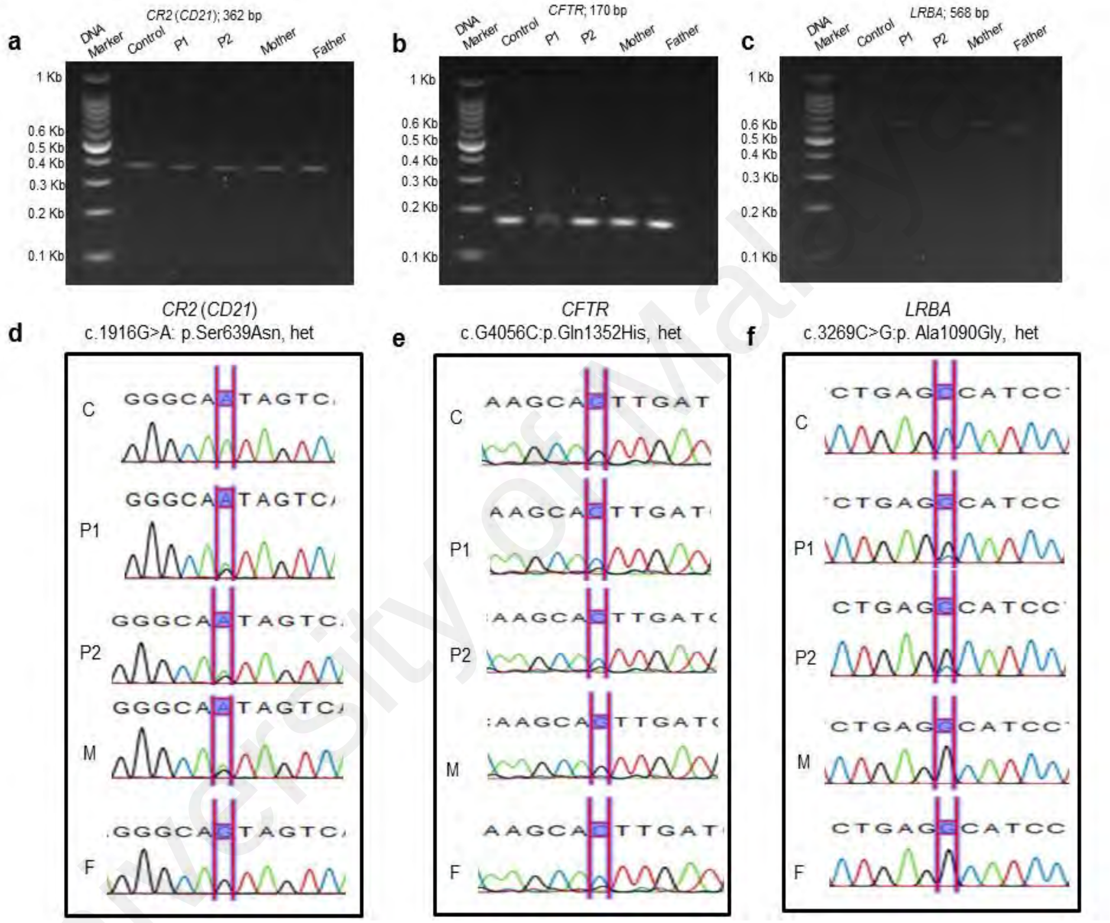


Figure 4.7: Sanger sequencing of CVID-associated genes. Three heterozygous mutations were identified in *CD21*, *CFTR*, and *LRBA* genes. (a), (b), and (c) showed PCR results of *CD21*, *CFTR* and *LRBA* genes. (d),(e), and (f) showed Sanger sequencing results of each genes stated above on normal healthy control (C), P1, P2, mother (M) and father (F).

4.10.2 Benign variants associated with CVID-related genes

Variants in *CTLA4*, *STXBP2* and *CD19* genes were considered as benign variants and randomly selected for validation using Sanger sequencing as shown in Figure 4.8.

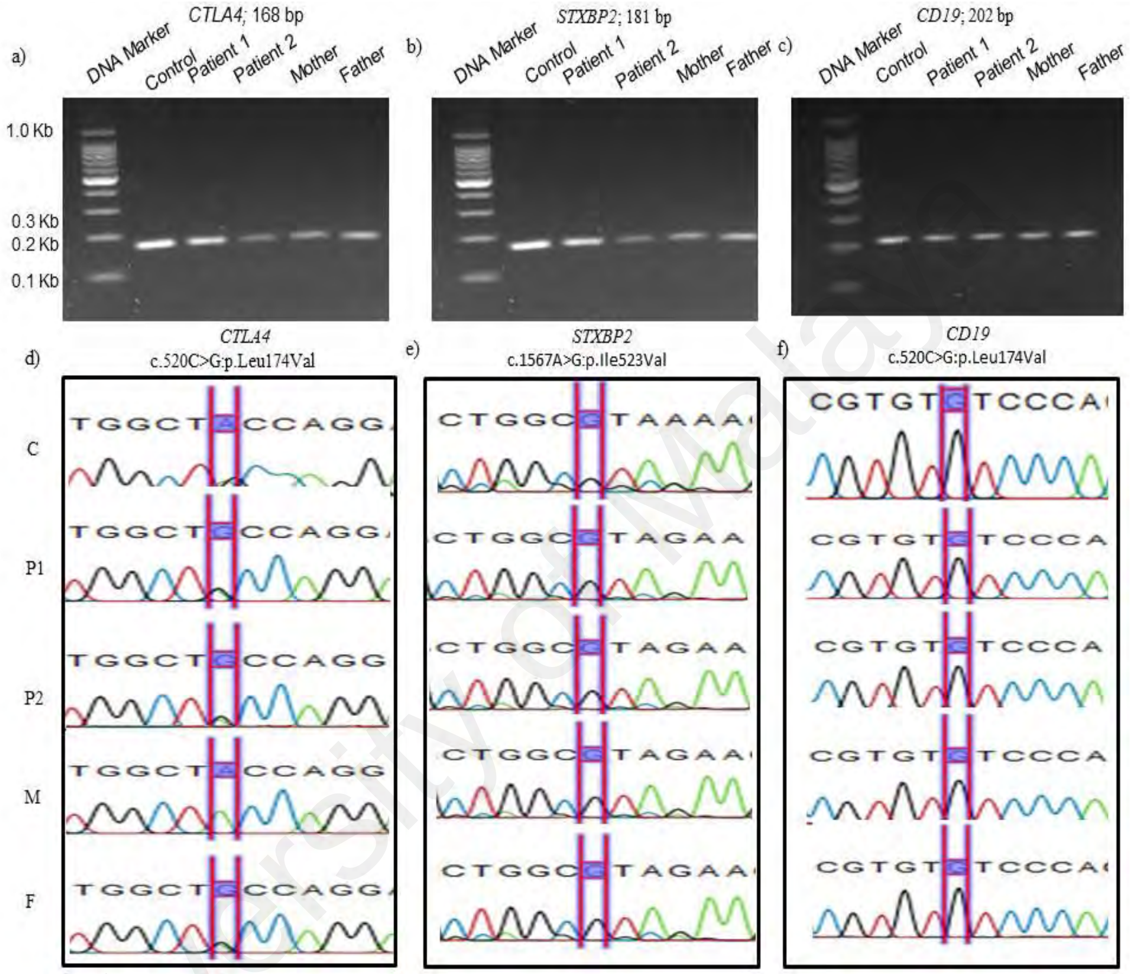
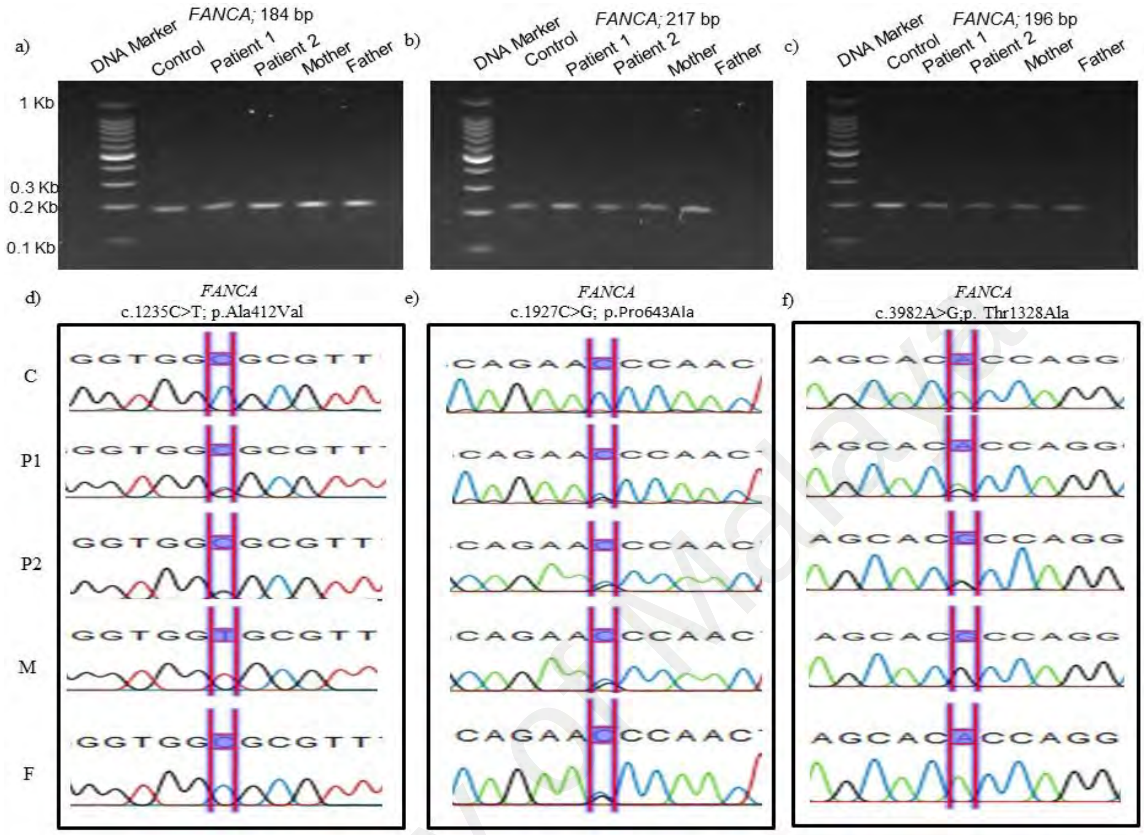


Figure 4.8: Sanger sequencing results. Three heterozygous variants were identified in *CTLA4*, *STXBP2,* and *CD19.* (a), (b), and (c) showed PCR results of *CTLA4*, *STXBP2,* and *CD19* genes. (d), (e), and (f) showed Sanger sequencing results of *CTLA4*, *STXBP2,* and *CD19* on normal healthy control (C), P1, P2, mother (M) and father (F).

Three variants in *FANCA* gene were randomly selected (benign variants) for Sanger sequencing as described in Figure 4.9.



. **Figure 4.9:** Sanger sequencing on *FANCA* gene. (a), (b), and (c) showed PCR results of *FANCA* gene. Three heterozygous *FANCA* variants were confirmed by Sanger sequencing on normal healthy control (C), P1, P2, mother (M) and father (F) as shown in (d), (e), and (f).

4.10.3 Sanger sequencing of *CD21* deficiency

We analysed two damaging SNPs from *CD21* gene and identified that it was a compound heterozygous case of *CD21* gene at exon 10 and 11 causing nucleotide changes from G to A, at position c.1916G>A and c.2012G>A of chromosome 1 for MZ twin. The protein product of the altered *CD21* gene was predicted using wANNOVAR tool and revealed that both SNPs occurred at the central conserved core of Sushi/SCR/CCP domain of CD21. Both SNPs were confirmed by Sanger sequencing (Figure 4.10).

The depth of coverage for *CD21* gene (both SNPs) are given in the Table 4.11. The better WES result comes from a combination of coverage, genotypes-phenotypes relationship and validation of variants through Sanger sequencing. The largest improvement in understanding WES is to not treat the variants with less than 5x of coverage, but rather assume accuracy and sensitivity of WES with more than 8x coverage for the clinical exome study. This improves over potential errors such as false-positive and false-negative variants as stated in previous analyses.

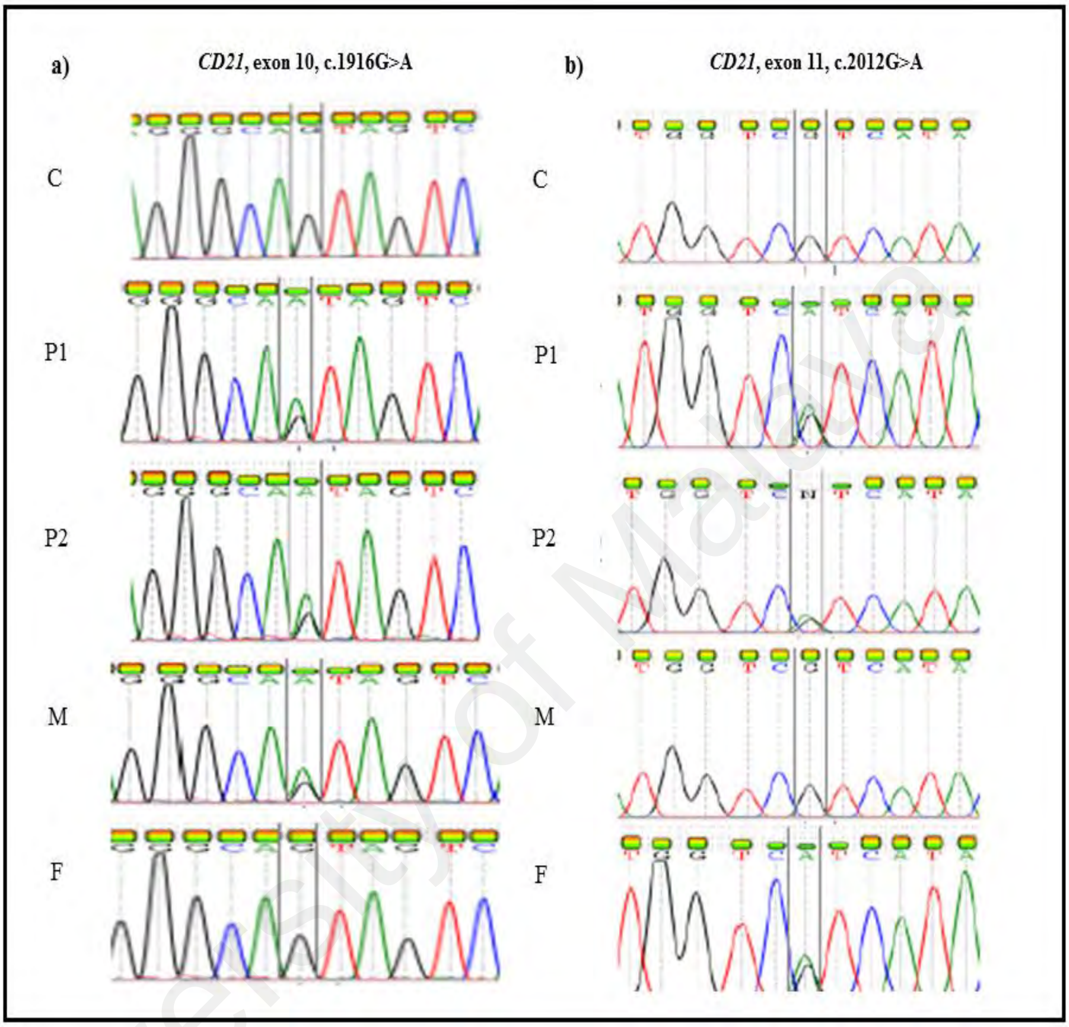


Figure 4.10: Sanger sequencing results of *CD21* (a) and (b) showed sequencing results for both SNPs in *CD21* gene on normal healthy control (C), P1, P2, mother (M) and father (F) of the subjects, vertical lines (highlighted in black color) represent the SNP site.

4.11 Network Analysis

In order to reveal the network landscape of *CD21* gene, the network interactions were analysed using *CD2*1 gene as the main input using GeneMania server as shown in the Figure 4.11. We identified 20 genes interacting with *CD21* gene (*CR1, FCER2, IFNA1, CD19, C3, MS4A1, CD72, KRTAP5-8, TCL1A, IL27, GPR18, SAMD10, BLK, CD27, CHI3L2, IFITM1, CD53, CCL21, CD22,* and *RGS13*). This interactions involving 67.61% physical interactions, 13.50% co-expression, 6.35% predicted network, 6.17% co-localization, 4.35% common pathway, 1.40% genetic interaction, and 0.59% shared protein domain. Detailed information of genes network was described in Table 4.12. Most of the genes were involved in lymphocyte proliferation (58.1%, $P_{adjusted}$= 3.85x10^{-2}) and leukocyte proliferation (45.2%, $P_{adjusted}$= 3.85 x10^{-2}). Among the 20 genes, three genes (*CD19, C3,* and *CD27*) were listed by Picard et. al (2015). *CD19* were involved in predominantly antibody deficiencies whereas *CD27* were reported to be associated with cellular and humoral immunity. *C3* were known to cause complement deficiencies such as defective humoral immune response.

Table 4.12: Genes network and their functions predicted by GeneMANIA.

Function	**FDR**	**Coverage**	**Genes**
Lymphocyte proliferation	3.85e-2	4/123	*CR2, IFNA1, MS4A1, IL27*
Leukocyte proliferation	3.85e-2	4/133	*CR2, IFNA1, MS4A1, IL27*
Adaptive immune system	3.85e-2	4/135	*CR1, IFNA1, IL27*
Mononuclear cell proliferation	3.85e-2	4/125	*CR2, IFNA1, MS4A1, IL27*
B cell activation	3.85e-2	4/119	*CR2, IFNA1, MS4A1, CD27*
Humoral immune response	3.85e-2	4/106	*CR1, IFNA1, C3, MS4A1*
Lymphocyte differentiation	4.26e-2	4/144	*CR2, IFNA1,CD27, IL27*
B cell proliferation	6.23e-2	3/54	CR2, *IFNA1, MS4A1*
B cell differentiation	8.81e-2	3/63	*CR2, IFNA1, CD27*
Leukocyte differentiation	1.74e-1	4/226	*CR2, IFNA1,CD27, IL27*

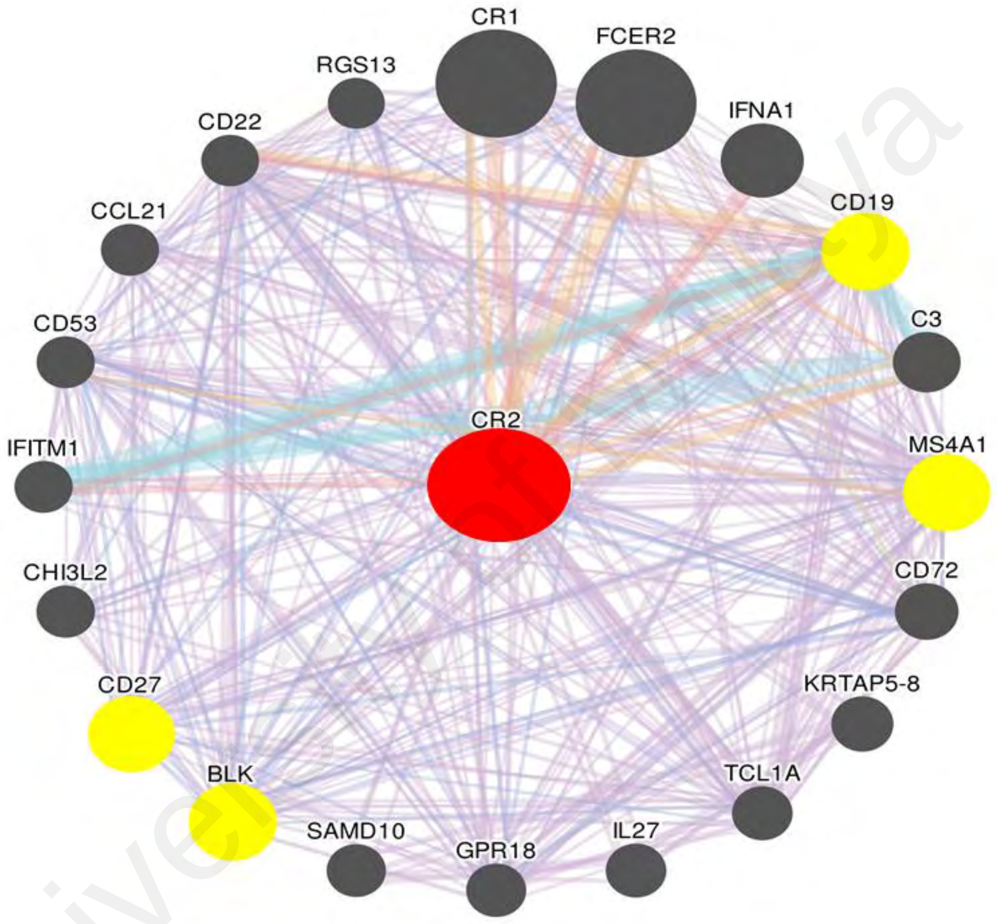


Figure 4.11: Gene-gene network interaction. Network interaction on *CD21* gene (red color). Yellow color indicates several CVID-associated genes found throughout the network interaction.

CHAPTER 5: DISCUSSION

Immune system is one of the key events that play a role in human defense mechanism under most circumstances. Any defects on human immune function, particularly a gene defect, are known to cause PID with the almost similar symptoms or clinical features as described in MZ twin (Section 4.1). Recently, more than 200 PID-genes has been reported (Picard et al., 2015). One of the emerging approaches to identify causative genetic mutations of PID is using a WES technology (Sarah et al., 2009). However, due to the limitation in the bioinformatics approaches to filter and analyses WES dataset, the successful rate of mutation detection in WES was reported within the range of 25 to 50% (Michael Niaki et al., 2014). Several WES analysis software and tools were tested and in the thesis, analysis tools that provide the best outcome were used to provide a WES data analysis pipeline to study both subjects.

Based on the Section 4.3, genomic DNA samples extraction for both MZ twin (P1 and P2) yields a good quality using commercial QIAamp® DNA Mini Kit. Determination of quality and quantity of DNA samples were done using Nanodrop (USA) as shown in the Table 4.3 and agarose gel electrophoresis (1% w/v) (Figure 4.3). In general, Nanodrop quantification will give more precise results when compared with the agarose gel electrophoresis technique. Principle of Nanodrop usage is based on the spectrometry method using $A_{260/280}$ and $A_{260/230}$ ratio as shown in Section 4.3.

$A_{260/280}$ ratio within the range from 1.80 to 2.00 represented a good quality of DNA, less contamination from the molecules such as carbohydrates and lipids. Next, $A_{260/230}$ ratio within the range from 1.80 to 2.00 showed a good quality of DNA, less contamination from the phenolic compounds and other organics. Therefore, overall genomic DNA samples were in good condition and quality. $A_{260/230}$ ratio (P2=1.38) showed slightly lower quality, however, the DNA samples for P2 still can be used for the next step processes.

WES dataset of MZ twin for sequence quality evaluation using FastQC and Trimmomatic software has been made as shown in Section 4.6. Directly removing the bad reads from WES dataset is challenging, in particular for the reads across all bases and the sequence contents, due to the PE sequencing and library preparations. Here, FastQC software was able to generate clean reads for both dataset for further steps, which is reads mapping to the hg38 and variant calling analysis. Mapping analysis revealed a total of 99.26% and 99.39% were achieved when mapped to the hg38 genome for MZ twin, which is considered as high quality WES dataset. Variant calling metrics was also successfully generated with more than 250,000 SNPs and 50,000 indels for MZ twin as described in the Section 4.7 and 4.8.

Here, we report a comprehensive WES analysis on a pair of MZ twin siblings suspected with PID to identify the causative genetic variation for their condition. The filtering results demonstrated likely disease-causing mutation in *CD21* gene, which is part of CVID. CVID is a primary immunodeficiency characterized by a decrease in serum IgG levels, a decrease in either IgA or IgM, and a poor response to vaccines in a child at least 2 years of age, after excluding other causes of hypogammaglobulinemia (Kelly, Tam, Verbsky, & Routes, 2013). Notably in our study, we have insufficient information to quantify genetic risk factor in this family as shown in Section 4.2, however, we considered for MAF available in exome and genome sequencing databases to support the evidence of variants pathogenicity in the population (patient and healthy control). The present case provides an opportunity to assess the genotype–phenotype relationship in a patient with CVID.

In this comprehensive WES study, mutation analysis of the 19 exons including the intron–exon boundaries of the entire *CD21* gene found two heterozygous missense SNPs (exon 10:c.1916G>A and exon 11:c.2012G>A) in a compound heterozygous fashion of the MZ twin patients (GenBank accession:NM_001006658).

The mutation c.1916G>A is a missense mutation that changed a serine residue of codon 639 to asparagine residue (p.Ser639Asn). This missense mutation p.Ser639Asn was not found in either in the father or normal control, although the patient's mother was heterozygous for the missense mutation c.1916G>A (maternal mutation). WES analysis and direct sequencing also revealed a heterozygous mutation at c.2012G>A in exon 11 of *CD21* gene. This SNP was found only in MZ twin and their father in heterozygous state (paternal mutation). The mutation c.2012G>A caused changes from arginine to histidine (p.Arg671His) of codon 671. These mutations were not found in 60 normal, unrelated Malaysian alleles (60 normal unrelated Malaysian individuals) by sequence analysis, and were predicted to be damaging SNP by PolyPhen-2. However, protein expression was unable to be conducted because both patients deceased at age 10 and 11 years old, in the middle of the study.

First identified as a complement receptor 2 (*CR2*/*CD21*) or EBV receptor in B cells and macrophages, *CD21* gene is a 145-kDa protein with 19 exons at chromosomal location of 1q32 (Wentink et al., 2015). The protein composed of 15 or 16 extracellular short consensus repeats (SCR), one transmembrane domain and an intracytoplasmic region (Jabs, Paulsen, Wagner, Kirchner, & Klüter, 1999). The physiological function of CD21 protein is as a coordinator for humoral response and complement system (Frank, 2012). A wide spectrum of clinical manifestation has been associated with *CD21* gene deficiency such as mild hypogammaglobulinemia, bronchitis, and recurrent infections (Rosain et al., 2017; Wentink et al., 2015). However, a clear genotype to phenotype correlation on *CD21* deficiency is yet to be determined.

On the clinical laboratory prospective, *CD21* deficiency can be characterized by reduced counts of memory B cells, low IgG and impaired anti-pneumococcal response (W. Al-Herz et al., 2014). Lymphocyte subset enumeration by flow cytometry indicated the low T, B and NK cell counts for both patients. Nonsense, frameshift, splice site and missense mutations have been discovered within multiple CD21 domains (Wentink et al., 2015). The SNPs site in both patients was identified to be in the short consensus repeats (SCRs) or Sushi domains of CD21. *CD21* deficiency has been reported to cause recurrent infections (particularly respiratory infections), hemolytic anemia, and low immunoglobulin levels by IgG in most as reported in the literature (Picard et al., 2015) and OMIM database (#120650 and #614699), which is consistent with our MZ twin clinical features.

Recent studies showed that most of the diseases are raised from the disease module, which is a particular set of disease-causing genes that grouped together, forming a network and contributed to the clinical phenotypes (Barabási, Gulbahce, & Loscalzo, 2010). In order to understand the gene-gene interaction among the genes associated with CVID, we conducted a gene-gene interaction network analysis on *CD21* gene. Our results showed that 4 out of 20 genes are involved in lymphocyte proliferation pathway and another 4 out of 20 genes are involved in leukocyte proliferation pathway. Any defects in this coordinated function of this network may disturb the cellular biological system related to immune system, thus, may cause a wide spectrum of clinical manifestation as seen in CVID cases. Although we identified some genes in the CVID-associated genes that interact with *CD21* gene, no specific signalling pathway was predicted in the gene-gene interaction network analysis.

In conclusion, this thesis described a comprehensive analysis based on WES dataset to identify a compound heterozygous case of *CD21* gene as a causative genetic variant in a pair of Malay descendent twin diagnosed with CVID. This damaging heterozygous missense mutations (exon 10:c.1916G>A and exon 11:c.2012G>A) of *CD21* gene would produce unrelated amino acids (p.Ser639Asn and p.Arg671His) that may cause CD21 protein deficiency in both patients. This study provides additional information on how to filter causative-genetic variants using WES approach and helps to understand the role of *CD21* gene in CVID in the broader spectrum of PID.

CHAPTER 6: CONCLUSION

This thesis has presented a comprehensive WES analysis of PID using MZ twin suspected with CVID. The framework combines a number of novel bioinformatics approaches and conventional experimental study for PID especially for the MZ twin. The main contributions of this thesis are summarized in the Section 3.3 (WES pipelines and approach), Section 4.10 (Sanger sequencing validation) and Section 4.11 (network analysis of 31 CVID-related genes). All the two objectives are identified, which are the backbone in this thesis. First, a total number of 99.26% and 99.39% of WES dataset for P1 and P2 are mapped back to the human reference genome (hg38). Second, an identification of damaging variant for the MZ twin, which are the missense mutation (c.1916G>A and c.2012G>A) at exon 10 and 11 of *CD21* gene and are confirmed by the Sanger sequencing as shown in the Section 4.10.4. This findings confirmed that both patients had compound heterozygous cases of *CD21* deficiency and the hypothesis were accepted. Network analysis of the *CD21* are also performed and analyzed to identify their gene-gene interactions.

WES study also described some of the major challenges involved with PID in the context of CVID. A novel pipelines and filtering analysis for WES, along with primer design for Sanger sequencing to identify disease-causing gene across over 200 different PID genes has been developed and thoroughly tested. Methodological with computational-aided have been made to the optimization of WES, which also allows an effective way of dealing with rare and complex diseases such as CVID, which includes SNPs and indels study in MZ twin suspected with PID. At the time of writing, several fully automatic WES pipelines for PID are under development, along with continued improvement to existing facilities.

For future direction, it would be crucial to evaluate and validate the clinical WES based on this comprehensive framework using functional studies (*in vitro* or *in vivo*) for the identification of disease-causing gene in rare diseases such as PID. This could become an important diagnostic tool for PID patients. The development of a large cohort WES for the assessment of genetic contributions in Malaysian PID patients has been started for this purpose. A different view of NGS technology using WES gives clinicians the ability to identify and perform best practices in genetic counselling and enhance the diagnostic yield. The findings of pathogenic variants, which can be performed using the presented bioinformatics pipelines with reduced computation time, could then be used to establish Malaysian PID database with automatic global data sharing for the mutation information. The promised WES will allow researchers to reach to the objective of precise medicine and better healthcare in our community.

REFERENCES

Aderibigbe, O. M., Priel, D., Lee, C., Ombrello M. J., Prajapati, V. H., Liang M. G., . . . Milner, J. D. (2015). Distinct cutaneous manifestations and cold-induced leukocyte activation associated with PLCG2 mutations. *JAMA Dermatology, 151*(6), 627-634.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(1), 7-20.

Ahn, S., & Charlotte, C. (2009). Role of B cells in common variable immune deficiency. *Expert Review of Clinical Immunology, 5*(5), 557-564.

Al-Herz, W., Bousfiha, A., Casanova, J. L., Chatila, T., Conley, M. E., Charlotte, C., . . . Holland, S. M. (2014). Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. *Frontiers in Immunology, 5(1)*, 162-189.

Al-Herz, W., Bousfiha, A., Chapel, H., Conley, M. E., Charlotte C., Etzioni, A., . . . Lennart. H. (2011). Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. *Frontiers in Immunology, 2(1)*, 54-68.

Al-Herz, W., Naguib, Kamal K., Notarangelo, L. D., Geha, R. S., & Alwadaani, A. (2011). Parental consanguinity and the risk of primary immunodeficiency disorders: report from the Kuwait National Primary Immunodeficiency Disorders Registry. *International Archives of Allergy and Immunology, 154*(1), 76-80.

Alkhairy, O. K., Rezaei, N., Graham, R. R., Abolhassani, H., Borte, S. K., . . . Qiang. H.(2015). RAC2 Loss-of-function Mutation in Two Siblings with Characteristics of Common Variable Immunodeficiency. *The Journal of Allergy and Clinical Immunology, 135*(5), 1380-1384.

Alkuraya, F. S. (2012). Discovery of rare homozygous mutations from studies of consanguineous pedigrees. *Current Protocols in Human Genetics*, *75*(1), 6-12.

Ameratunga, R., Storey, P., Barker, R., Jordan, A., Koopmans, W., & Woon, S. T. (2016). Application of diagnostic and treatment criteria for common variable immunodeficiency disorder. *Expert Review of Clinical Immunology, 12*(3), 257-266.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Antonio, S., Fujihashi, H., Amoros, D., & Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology, 2*(1), 17-25.

Antonio, S., Fujihashi, H., & O'meara, P. (2005). Topology of small-world networks of protein–protein complex structures. *Bioinformatics, 21*(8), 1311-1315.

Astronomo, R. D., & Burton, D. R. (2010). Carbohydrate vaccines: developing sweet solutions to sticky situations? *Nature Reviews Drug Discovery, 9*(4), 308-324.

Aydin, S. E., Kilic, S. S., Aytekin, C., Kumar, A., Porras, O., Kainulainen, L., . . . Karaca, N. (2015). DOCK8 deficiency: clinical and immunological phenotype and treatment options-a review of 136 patients. *Journal of Clinical Immunology, 35*(2), 189-198.

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2010). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics, 12*(1), 56-68.

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice, 98*(6), 236-238.

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.

Bogaert, J. A., Delfien, D., Melissa, L., Bart, V., Karim, D., Elfride, B., & Haerynck, F. (2016). Genes associated with common variable immunodeficiency: One diagnosis to rule them all? *Journal of Medical Genetics, 53*(9), 575-590.

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics, 33*(3s), 228-237.

Bourke, C. D., Berkley, J. A., & Prendergast, A. J. (2016). Immune dysfunction as a cause and consequence of malnutrition. *Trends in Immunology, 37*(6), 386-398.

Bratanič, N., Kovač, J., Pohar, K., Podkrajšek, K. T., Ihan, Alojz, B. T., & Stefanija, M. A. (2017). Multifocal gastric adenocarcinoma in a patient with LRBA deficiency. *Orphanet Journal of Rare Diseases, 12*(1), Article#131.

Bryant, V. L., & Tangye, S. G. (2016). The Expanding Spectrum of NFkB1 Deficiency. *Journal of Clinical Immunology, 36*(6), 531-532.

Chaitankar, V., Karakülah, G., Ratnapriya, R., Giuste, F. O., Brooks, M. J., & Swaroop, A. (2016). Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research, 55*(2), 1-31.

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

Choi, J., Fernandez, R., Maecker, H. T., & Butte, Manish J. (2017). Systems approach to uncover signaling networks in primary immunodeficiency diseases. *Journal of Allergy and Clinical Immunology, 140*(3), 881-884.

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics, 11*(6), 415-425.

Compeer, E. B., Janssen, W., Annet, R. K., Marielle, G., Joris M. M., & Boes, M. (2015). Dysfunctional BLK in common variable immunodeficiency perturbs B-cell proliferation and ability to elicit antigen-specific CD4(+) T-cell help. *Oncotarget, 6*(13), 10759-10771.

Conley, M. E., Notarangelo, L. D., & Etzioni, A. (1999). Diagnostic criteria for primary immunodeficiencies. *Clinical Immunology, 93*(3), 190-197.

Consortium, Genomes Project. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-64.

Consortium, Genomes Project. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-71.

Costa-Carvalho, B. T., Grumach, A. S., Franco, J. L., Espinosa-Rosales, F. J., Leiva, L. E., King, A., . . . Sorensen, R. U. (2014). Attending to warning signs of primary immunodeficiency diseases across the range of clinical practice. *Journal of Clinical Immunology, 34*(1), 10-22.

Charlotte, C. R. (1989). Clinical and immunologic analyses of 103 patients with common variable immunodeficiency. *Journal of Clinical Immunology, 9*(1), 22-33.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Davis, S. D., Schaller, J., Wedgwood, R. .J, & Harvard, M. D. (1966). Job's syndrome: recurrent," cold", staphylococcal abscesses. *The Lancet, 287*(7445), 1013-1015.

David, W. F., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Lopes, C. T. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research, 38*(suppl_2), W214-W220.

Day-Williams, A. G., & Zeggini, E. (2011). The effect of next-generation sequencing technology on complex trait research. *European Journal of Clinical Investigation, 41*(5), 561-567.

Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*(9), 418-426.

Debnath, L. (2009). The legacy of Leonhard Euler–a tricentennial tribute. *International Journal of Mathematical Education in Science and Technology, 40*(3), 353-388.

Denz, A., Eibel, H., Illges, H., Kienzle, G., Schlesier, M., & Peter, H. H. (2000). Impaired up-regulation of CD86 in B cells of "type A" common variable immunodeficiency patients. *Clinical and Experimental Immunology*, *30*(4), 1069-1077.

Drmanac, R. (2012). The ultimate genetic test. *Science, 336*(6085), 1110-1112.

Du, C., Pusey, B. N., Adams, C. J., Lau, Christopher, C. B., William P. G., William A., . . . Adams, D. R. (2016). Explorations to improve the completeness of exome sequencing. *BioMed Central Medical Genomics, 9*(1), 56-67.

Eckert, T., Tang, C., & Eidelberg, D. (2007). Assessment of the progression of Parkinson's disease: a metabolic network approach. *The Lancet Neurology, 6*(10), 926-932.

Elgizouli, M., Lowe, D. M., Speckmann, C., Schubert, D., Hülsdünker, J., Eskandarian, Z., . . . Grimbacher, B. (2016). Activating PI3Kδ mutations in a cohort of 669 patients with primary immunodeficiency. *Clinical and Experimental Immunology, 183*(2), 221-229.

Emerson, I. A., & Gothandam, K. M. (2012). Network analysis of transmembrane protein structures. *Physica A: Statistical Mechanics and its Applications, 391*(3), 905-916.

Erwin, D. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research, 322*(1), 12-20.

Fang, M., Abolhassani, H., Lim, C. K., Zhang, J., & Hammarström, L. (2016). Next Generation Sequencing Data Analysis in Primary Immunodeficiency Disorders– Future Directions. *Journal of Clinical Immunology, 36*(1), 68-75.

Farmer, J. R, Ong, M. S., Barmettler, S., Yonker, L. M., Fuleihan, R., Sullivan, K. E., . . . Walter, J. E. (2017). Common variable immunodeficiency (CVID) non-infectious disease endotypes redefined using unbiased network clustering in large electronic datasets. *Frontiers in Immunology, 8*, Article#1740.

Ferenzi, G. W. (1962). The significance of neutropenia. *Medical Clinics of North America, 46*(1), 245-252.

Frank, Michael M. (2012). CD21 deficiency, complement, and the development of common variable immunodeficiency. *Journal of Allergy and Clinical Immunology, 129*(3), 811-813.

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics, 29*(3), 150-159.

Geneviève, D. B., Tabone, M. D., Durandy, A., Phan, F., Fischer, A., & Françoise, L. D. (1999). CD40 ligand expression deficiency in a female carrier of the X-linked hyper-IgM syndrome as a result of X chromosome lyonization. *European Journal of Immunology, 29*(1), 367-373.

German, J., Sanz, M. M., Ciocci, S., Tian, Y. Z., & Ellis, N. A. (2007). Syndrome-causing mutations of the BLM gene in persons in the Bloom's Syndrome Registry. *Human Mutation, 28*(8), 743-753.

Girirajan, S., Dennis, M. Y., Baker, C., Malig, M., Coe, B. P., Campbell, C. D., . . . Cheng, Z. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *The American Journal of Human Genetics, 92*(2), 221-237.

Goh, K., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences, 104*(21), 8685-8690.

Gonzalez, M. W., & Kann, M. G. (2012). Protein interactions and disease. *PLoS Computational Biology, 8*(12), Article#e1002819.

Grant, S. F. A., Thorleifsson, G., Frigge, M. L., Thorsteinsson, J., Gunnlaugsdóttir, B., Geirsson, Á. J., . . . Valsson, Júlíus. (2001). The inheritance of rheumatoid arthritis in Iceland. *Arthritis & Rheumatology, 44*(10), 2247-2254.

Gulez, N., Aksu, G., Berdeli, A., Karaca, N., Tanrıverdi, S., Kutukculer, N., & Azarsiz, E. (2011). X-Linked Lymphoproliferative Syndrome and Common Variable Immunodeficiency May Not Be Differentiated by SH2D1A and XIAP/BIRC4 Genes Sequence Analysis. *Case Reports in Medicine, 2011*, 121258.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., . . . Sakaguchi, A. Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature, 306*(5940), 234-238.

Hassan, A., Wang, N., Asghar, A., Nima, R., Lee, Y. N., Francesco, F., . . . Lennart, H. (2014). A hypomorphic recombination-activating gene 1 (RAG1) mutation resulting in a phenotype resembling common variable immunodeficiency. *Journal of Allergy and Clinical Immunology, 134*(6), 1375-1380.

Hadizadeh, H., Salehi, M., Khoramnejad, S., Vosoughi, K., & Rezaei, N. (2017). The association between parental consanguinity and primary immunodeficiency diseases: A systematic review and meta-analysis. *Pediatric Allergy and Immunology, 28*(3), 280-287.

Hills, L. P., & Tiffany, T. O. (1980). Comparison of turbidimetric and light-scattering measurements of immunoglobulins by use of a centrifugal analyzer with absorbance and fluorescence/light-scattering optics. *Clinical Chemistry, 26*(10), 1459-1466.

Hollenbaugh, D., Wu, L. H., Ochs, H. D., Nonoyama, S., Grosmaire, L. S., Ledbetter, J. A, . . . Aruffo, A. (1994). The random inactivation of the X chromosome carrying the defective gene responsible for X-linked hyper IgM syndrome (X-HIM) in female carriers of HIGM1. *The Journal of Clinical Investigation, 94*(2), 616-622.

Jabs, WJ, Paulsen, M, Wagner, HJ, Kirchner, H, & Klüter, H. (1999). Analysis of Epstein–Barr virus (EBV) receptor CD21 on peripheral B lymphocytes of long-term EBV− adults. *Clinical and Experimental Immunology, 116*(3), 468-473.

Jamra, R. A., Wohlfart, S., Zweier, M., Uebe, S., Priebe, L., Ekici, A., . . . Fakher, M. (2011). Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. *European Journal of Human Genetics, 19*(11), 1161-1166.

Joep, L., Willemsen, Marjolein H., Bregje W. M. V. B., Kleefstra, T., Yntema, H. G., Kroes, T., . . . Gilissen, C. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine, 367*(20), 1921-1929.

Kahn, S. D. (2011). On the future of genomic data. *Science, 331*(6018), 728-729.

Kanegane, H., Agematsu, K., Futatani, T., Sira, M. M., Suga, K., Sekiguchi, T., . . . Miyawaki, T. (2007). Novel mutations in a Japanese patient with CD19 deficiency. *Genes And Immunity, 8*, 663-670.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, Douglas M., Kavanagh, D., . . . Cummings, B. B. (2016). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research, 45*(D1), D840-D845.

Katsanis, S. H., & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics, 14*(6), 415-426.

Keerthikumar, S., Raju, R., Kandasamy, K., Hijikata, A., Ramabadran, S., Balakrishnan, L., . . . Somanathan, D. S. (2008). RAPID: resource of Asian primary immunodeficiency diseases. *Nucleic Acids Research, 37*(suppl_1), D863-D867.

Keller, M. D., & Jyonouchi, S. (2013). Chipping away at a mountain: Genomic studies in Common Variable Immunodeficiency. *Autoimmunity Reviews, 12*(6), 687-689.

Kelly, B. T., Tam, Jonathan S., Verbsky, J. W., & Routes, J. M. (2013). Screening for severe combined immunodeficiency in neonates. *Clinical Epidemiology, 5*, 363-369.

Kelsen, J. R., Dawany, N., Moran, C. J., Petersen, B. S., Sarmady, M., Sasson, A., ... & Rappaport, E. (2015). Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease. *Gastroenterology*, *149*(6), 1415-1424.

Kidera, A., & Nobuhiro, G. (1990). Refinement of protein dynamic structure: normal mode refinement. *Proceedings of the National Academy of Sciences, 87*(10), 3718-3722.

Kirkpatrick, P. & Riminton, S. (2007). Primary immunodeficiency diseases in Australia and New Zealand. *Journal of Clinical Immunology, 27*(5), 517-524.

Kitano, H. (2002a). Computational systems biology. *Nature, 420*(6912), 206-210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science, 295*(5560), 1662-1664.

Komatsu, N., Saijoh, K., Jayakumar, A., Clayman, G. L., Tohyama, M., Suga, Y., . . . Takehara, K. (2008). Correlation between SPINK5 gene mutations and clinical manifestations in Netherton syndrome patients. *Journal of Investigative Dermatology, 128*(5), 1148-1159.

Kopecký, O., & Lukešová, Š. (2007). Genetic defects in common variable immunodeficiency. *International Journal of Immunogenetics, 34*(4), 225-229.

Ku, C. S., Cooper, D. N., & Patrinos, G. P. (2016). The rise and rise of exome sequencing. *Public Health Genomics, 19*(6), 315-324.

Kutukculer, N., & Gulez, N. (2009). The outcome of patients with unclassified hypogammaglobulinemia in early childhood. *Pediatric Allergy and Immunology, 20*(7), 693-698.

Lederman, H. M., & Winkelstein, J. A. (1985). X-linked agammaglobulinemia: an analysis of 96 patients. *Medicine, 64*(3), 145-156.

Lehne, B., & Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. *Human Genomics, 3*(3), 291-298.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International, 2*(4), 64-73.

Liu, Y., Hanson, S., Gurugama, P., Jones, A., Clark, B., & Ibrahim, M. A. (2014). Novel NFKB2 mutation in early-onset CVID. *Journal of Clinical Immunology, 34*(6), 686-690.

Lv, Q., Ma, W., Liu, H., Li, J., Wang, H., Lu, F., . . . Shi, T. (2015). Genome-wide protein-protein interactions and protein function exploration in cyanobacteria. *Scientific Reports, 5*, Article#15519.

Maffucci, P., Filion, C. A., Boisson, B., Itan, Y., Shang, L., Casanova, J. L., & Charlotte, C. R. (2016). Genetic Diagnosis Using Whole Exome Sequencing in Common Variable Immunodeficiency. *Frontiers in Immunology, 7*, Article#220.

Maggadottir, S. M., Li, J., Glessner, J. T., Li, Y. R., Wei, Z., Chang, X., . . . Hakonarson, H. (2015). Rare variants at 16p11.2 are associated with common variable immunodeficiency. *The Journal of Allergy and Clinical Immunology, 135*(6), 1569-1577.

Manley, L. J., Ma, D., & Levine, S. S. (2016). Monitoring Error Rates In Illumina Sequencing. *Journal of Biomolecular Techniques: JBT, 27*(4), 125-128.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry, 6*, 287-303.

Maynard, Y., Scott, M. G., Nahm, M. H., & Ladenson, J. H. (1986). Turbidimetric assay of IgG with use of single monoclonal antibodies. *Clinical Chemistry, 32*(5), 752-757.

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

McKusick, V. A. (1976). *Mendelian inheritance in man* (5th ed.). Baltimore: John Hopkins University Press.

Menno, Z. C., Reisli, I., Mirjam, B., Castaño, D. N., Carel, J. M., Maarten, D. J. D., . . Jacques, D. J. M. (2006). An antibody-deficiency syndrome due to mutations in the CD19 gene. *New England Journal of Medicine, 354*(18), 1901-1912.

Meynert, A. M., Ansari, M., FitzPatrick, D. R., & Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BioMed Central Bioinformatics, 15*(1), 247-258.

Michael, N., Julia, P., Margaret H., Andrea, K., Sylvie, P., Leah, S., . . . Louanne H. (2014). Clinical whole-exome sequencing: are we there yet? *Genetic in Medicine*. *16*(9), 717-719.

Minegishi, Y., Saito, M., Tsuchiya, S., Tsuge, I., Takada, H., Hara, T., . . . Stojkovic, O. (2007). Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome. *Nature, 448*(7157), 1058-1062.

Mirabello, L., Macari, E. R., Jessop, L., Ellis, S. R., Myers, T., Giri, N., ... & Yeager, M. (2014). Whole-exome sequencing and functional studies identify RPS29 as a novel gene mutated in multi-case Diamond-Blackfan anemia families. *Blood*, *124*(1), 24-32.

Moya-Quiles, M. R., Bernardo-Pisa, M. V., Menasalvas, A., Alfayate, S., Fuster, J. L., Boix, F., . . . Álvarez-López, M. R. (2014). Severe combined immunodeficiency: first report of a de novo mutation in the IL2RG gene in a boy conceived by in vitro fertilization. *Clinical Genetics, 85*(5), 500-501.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research, 31*(13), 3812-3814.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., . . . Nickerson, D. A. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics, 42*(1), 30-35.

Ng, S. B, Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., . . . Eichler, E. E. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461*(7261), 272-276.

Noris, M., & Remuzzi, G. (2013). Overview of Complement Activation and Regulation. *Seminars in Nephrology, 33*(6), 479-492.

O'shea, J. J., Holland, S. M., & Staudt, L. M. (2013). JAKs and STATs in immunity, immunodeficiency, and cancer. *New England Journal of Medicine, 368*(2), 161-170.

Pajusalu, S., Pfundt, R., Vissers, L. E. L. M., Kwint, M. P., Reimand, T., Õunap, K., . . . Jayne, H. K. (2018). Identifying long indels in exome sequencing data of patients with intellectual disability. *bioRxiv*, 244756.

Park, J. H., Resnick, E. S., & Cunningham-Rundles, C. (2011). Perspectives on common variable immune deficiency. *Annals of the New York Academy Science, 1246*, 41-49.

Patuzzo, G., Mazzi, F., Vella, A., Ortolani, R., Barbieri, A., Tinazzi, E., . . . Lunardi, C. (2013). Immunophenotypic Analysis of B Lymphocytes in Patients with Common Variable Immunodeficiency: Identification of CD23 as a Useful Marker in the Definition of the Disease. *ISRN Immunology*, 1-8.

Picard, C., Al-Herz, W., Bousfiha, A., Casanova, J. L., Chatila, T., Conley, M. E., . . . Klein, C. (2015). Primary immunodeficiency diseases: an update on the classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. *Journal of Clinical Immunology, 35*(8), 696-726.

Picard, C., Horst, V. B., Ghandil, P., Chrabieh, M., Levy, O., Arkwright, P. D., . . . Krause, J. C. (2010). Clinical features and outcome of patients with IRAK-4 and MyD88 deficiency. *Medicine, 89*(6), 403-425.

Pichichero, M. E., & Passador, S. (1997). Administration of combined diphtheria and tetanus toxoids and pertussis vaccine, hepatitis B vaccine, and Haemophilus influenzae type b (Sekinaka et al.) vaccine to infants and response to a booster dose of Hib conjugate vaccine. *Clinical Infectious Diseases, 25*(6), 1378-1384.

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

Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Saeed, A. T., . . . Smith, Blair. (2016). Timing, rates and spectra of human germline mutation. *Nature Genetics, 48*(2), 126-133.

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *International Journal of Proteomics, 2*(1), 1-12.

Reda, S. M., El-Ghoneimy, D. H., & Afifi, H. M. (2013). Clinical predictors of primary immunodeficiency diseases in children. *Allergy, Asthma & Immunology Research, 5*(2), 88-95.

Rehm, H. L., Bale, S. J., Pinar,B. T., Berg, J. S., Brown, K. K., Deignan, J, L., . . . Lyon, E.. (2013). ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine, 15*(9), 733-747.

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278-289.

Rodriguez, J. R., Seals, J. E., Radin, A., Lin, J. S., Mandl, I., & Turino, G. M. (1979). Neutrophil lysosomal elastase activity in normal subjects and in patients with chronic obstructive pulmonary disease. *American Review of Respiratory Disease, 119*(3), 409-417.

Rosain, J., Miot, C., Lambert, N., Rousselet, M. C., Pellier, I., & Picard, C. (2017). CD21 deficiency in 2 siblings with recurrent respiratory infections and hypogammaglobulinemia. *The Journal of Allergy and Clinical Immunology: In Practice*, *5*(6), 1765-1767.

Roos, D., & Boer, M. (2014). Molecular diagnosis of chronic granulomatous disease. *Clinical & Experimental Immunology, 175*(2), 139-149.

Rual, J. F., Venkatesan, K., Hao, T., Tomoko, H. K., Dricot, A., Li, N., . . . Nono, A. G. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature, 437*(7062), 1173-1178.

Russell, M. A., Pigors, M., Houssen, M. E., Manson, A., Kelsell, D., Longhurst, H., & Morgan, N. G. (2017). A novel de novo activating mutation in STAT3 identified in a patient with common variable immunodeficiency (CVID). *Clinical Immunology*, *187*, 132-136.

Salzer, E., Kansu, A., Sic, H., Majek, P., Ikinciogullari, A., Dogu, F. E., . . . Boztug, K. (2014). Early-onset inflammatory bowel disease and common variable immunodeficiency-like disease caused by IL-21 deficiency. *Journal of Allergy and Clinical Immunology, 133*(6), 1651-1659..

Sankaran, V. G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J. A., Beggs, A. H., . . . Lander, E. S. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *The Journal of Clinical Investigation, 122*(7), 2439-2443.

Santiago, J. A., & Potashkin, J. A. (2014). A network approach to diagnostic biomarkers in progressive supranuclear palsy. *Movement Disorders, 29*(4), 550-555.

Schäffer, A. A., Pfannstiel, J., Webster, A., David B., Plebani, A., Lennart, H., & Grimbacher, B. (2006). Analysis of families with common variable immunodeficiency (CVID) and IgA deficiency suggests linkage of CVID to chromosome 16q. *Human Genetics, 118*(6), 725-729.

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research, 20*(9), 1165-1173.

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

Schmidt, R. E., Grimbacher, B., & Witte, T. (2018). Autoimmunity and primary immunodeficiency: two sides of the same coin? *Nature Reviews Rheumatology, 14*(1), 7-18.

Schubert, D., Bode, C., Kenefeck, R., Hou, T. Z., Wing, J. B., Kennedy, A., . . . Grimbacher, B. (2014). Autosomal-dominant immune dysregulation syndrome in humans with CTLA4 mutations. *Nature Medicine, 20*(12), 1410-1416.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D.. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods, 7*(8), 575-576.

Sekinaka, Y., Mitsuiki, N., Imai, K., Yabe, M., Yabe, H., Mitsui-Sekinaka, K., . . . Nonoyama, S. (2017). Common Variable Immunodeficiency Caused by FANC Mutations. *Jornal of Clinical Immunology, 37*(5), 434-444.

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), 62-72.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*(10), 1135-1145.

Simon, A. K., Hollander, G. A., & Andrew, M. (2015). Evolution of the immune system in humans from infancy to old age. *Proceedings of Royal Society. B, 282*(1821), 20143085.

Sobreira, N. L. M., Cirulli, E. T., Avramopoulos, D., Wohler, E., Oswald, G. L., Stevens, E. L., . . . Maia, J. M. (2010). Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genetics, 6*(6), 991-996.

Tampella, G., Baronio, M., Vitali, M., Soresina, A., Badolato, R., Giliani, S., . . . Lougaris, V. (2011). Evaluation of CARMA1/CARD11 and Bob1 as candidate genes in common variable immunodeficiency. The *Journal of Investigational Allergology and Clinical Immunology, 21*(5), 348-353.

Tattini, L., Romina, D. R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology, 3*, 92-99.

Thiel, J., Kimmig, L., Salzer, U., Grudzien, M., Lebrecht, D., Hagena, T., ... & Gutenberger, S. (2012). Genetic CD21 deficiency is associated with hypogammaglobulinemia. *Journal of Allergy and Clinical Immunology*, *129*(3), 801-810.

Tin, M. M. Y., Rheindt, F. E., Cros, E., & Mikheyev, A. S. (2015). Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources, 15*(2), 329-336.

Tischer, B. K., & Kaufer, B. B. (2012). Viral bacterial artificial chromosomes: generation, mutagenesis, and removal of mini-F sequences. *BioMed Research International, 2012*.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36.

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics, 13*(8), 565-575.

Vendruscolo, M., Dokholyan, N. V., Paci, E., & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E, 65*(6), 1-4.

Verso, M. L. (1964). The evolution of blood-counting techniques. *Medical History, 8*(2), 149-158.

Vidal, M., Cusick, M. E., & Barabási, A. L. (2011). Interactome networks and human disease. *Cell, 144*(6), 986-998.

Vorechovský, I., Litzman, J., Lokaj, J., & Sobotkova, R. (1991). Family studies in common variable immunodeficiency. *Journal of Hygiene, Epidemiology, Microbiology, and Immunology, 35*(1), 17-26.

Vořechovský, I., Zetterquist, H., Paganelli, R., Koskinen, S., David, A., Webster, B., . . . Lennart, H. (1995). Family and linkage study of selective IgA deficiency and common variable immunodeficiency. *Clinical Immunology and Immunopathology, 77*(2), 185-192.

Wang, H. Y., Ma, C. A., Zhao, Y., Fan, X., Zhou, Q., Edmonds, P., . . . Jain, A. (2013). Antibody deficiency associated with an inherited autosomal dominant mutation in TWEAK. *Proceedings of the National Academy of Sciences, 110*(13), 5127-5132.

Wang, S., Haynes, C., Barany, F., & Ott, J. (2009). Genome-wide autozygosity mapping in human populations. *Genetic Epidemiology, 33*(2), 172-180.

Willem, H., Pijnenburg, Y. A. L., Strijers, R. L. M., Yolande, M., Wiesje M. F., Scheltens, P., & Stam, C. J. (2009). Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. *BioMed Central Neuroscience, 10*(1), 101-112.

Warnatz, K., Salzer, U., Rizzi, M., Fischer, B., Gutenberger, S., Böhm, J., . . . Eibel, H. (2009). B-cell activating factor receptor deficiency is associated with an adult-onset antibody deficiency syndrome in humans. *Proceedings of the National Academy of Sciences, 106*(33), 13945-13950.

Wentink, Marjolein WJ, Lambeck, Annechien JA, van Zelm, Menno C, Simons, Erik, van Dongen, Jacques JM, IJspeert, Hanna, . . . van der Burg, Mirjam. (2015). CD21 and CD19 deficiency: two defects in the same complex leading to different disease modalities. *Clinical Immunology, 161*(2), 120-127.

Westerlund, J. F., & Fairbanks, D. J. (2010). Gregor Mendel's classic paper and the nature of science in genetics courses. *Hereditas, 147*(6), 293-303.

Windhorst, D., & Soothill, J. F. (1969). Inheritance of chronic granulomatous disease. *The Lancet, 294*(7619), 543-544.

Worthey, E. A., Mayer, A. N, Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., . . . Veith, R. L. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine, 13*(3), 255-262.

Xue, J., Schmidt, S. V., Sander, J., Draffehn, A., Krebs, W., Quester, I., . . . Schmidleithner, L. (2014). Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity, 40*(2), 274-288.

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., . . . Niu, Z. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine, 369*(16), 1502-1511.

Yong, P. F., Salzer, U., & Grimbacher, B. (2009). The role of costimulation in antibody deficiencies: ICOS and common variable immunodeficiency. *Immunological Reviews, 229*(1), 101-113.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Published Article:

Hamidah Ghani. (2018). Clinical Exomes: Genetic Test for Pediatric Inborn Errors. [Editorial Article]. *Journal of Genome*, 1(2).

Published Conference Paper in Book Form:

Chear, C. T., Farid Baharin, Munirah Hishamshah, Asiah Kassim., **Hamidah Ghani**, Saharuddin Mohamad, & Adiratna Mat Ripen. (2017). *Whole exome sequencing reveals a high genetic heterogeneity on Common Variable Immunodeficiency*: *Proceedings of the 4th Asian Regional Conference on Systems Biology* (pp. 20-22). Putrajaya, Malaysia.