The Genetics of Primary Immunodeficiency in Children

Jesmeen S Maimaris
Student No: 15115943

A thesis submitted for the degree PhD 2015-2020

Primary Supervisor: Prof. Adrian Thrasher
Secondary Supervisor: Dr. Chiara Bacchelli

UCL Great Ormond Street Institute of Child Health

DECLARATION

I, Jesmeen Maimaris confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

For my beautiful and talented daughters,

The grass is greener where you water it

ABSTRACT

Studies of children with recurrent infection demonstrate that primary immunodeficiency (PID) has a significant genetic component. In PID, over 300 genes of high penetrance inherited mostly in autosomal recessive manner have already been identified. However, many children, including those with early onset immunodeficiency have not received a genetic diagnosis, despite use of targeted sequencing methods.

I performed bioinformatic analysis in whole genome sequencing data in patients with immunodeficiency. These analyses were initially in a small cohort of affected children at Great Ormond Street in whom a genetic diagnosis was not known. I devised and utilised bioinformatic programs to identify novel genetic variants in this cohort. I evaluated the performance of whole genome sequence analyses with targeted gene panel analyses, which is the most utilised method of genetic diagnosis.

To expand my analysis, I looked at a larger cohort of young people and adults with immunodeficiency as part of the large national collaborative project NIHR Bioresource Rare Diseases BRIDGE-PID project. I quantified the burden of rare coding variation in a case cohort compared to controls and used rare variant association analysis to identify potential novel candidate genes in primary immunodeficiency. The final chapter focuses on 2 novel genetic variants found in the cohort and our initial functional testing to verify genetic diagnosis.

The work presented in this thesis demonstrates novel genetic causes of immunodeficiency and their functional implications. The results of my work have improved understanding of the genetic architecture of primary immunodeficiencies and has clinical utility in the diagnosis and subsequent treatment of immunodeficiency.

IMPACT STATEMENT

My research into whole genome sequencing in primary immunodeficiency has utilised novel sequencing methods in a population of patients severely affected by immunodeficiency.

Throughout my research I have managed several bioinformatic and molecular projects around the use of whole genome sequencing such as its efficacy against current practice and functional use in diagnosis and management of patients with primary immunodeficiency. These have included translational genetic studies to determine causality of implicated genes in patient blood and cell lines.

My results have found genetic diagnoses for many patients. This has reduced diagnostic uncertainty and, in many cases, changed treatment trajectory. From commencing this project to its conclusion, our use of genomic data to inform diagnosis and management of patients with immunodeficiency has dramatically increased with 86 genes routinely testing. This has thrown up challenges in interpretation of genetic variants of which little functional evidence exists. I envisage my work to add to the evidence to the genetic basis of immunodeficiency.

Research publications from this work:

Thaventhiran JED, Lango Allen H, Burren OS, Rae W, Greene D, Staples E, Zhang Z, Farmery JHR, Simeoni I, Rivers E, Maimaris J et al. Whole-genome sequencing of a sporadic primary immunodeficiency cohort Nature 06 May 2020

Farmery JHR, Smith ML, Lynch AG, Huissoon A, Furnell A, Mead A, Levine AP, Manzur A, Thrasher A, Greenhalgh A, et al. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data (vol 8, 1300, 2018)

Tuijnenburg P, Lango Allen H, Burns SO, Greene D, Jansen MH, Staples E, Stephens J, Carss KJ, Biasci D, Baxendale H, et al. Loss of function *NFKB1* variants are the most common monogenic cause of CVID in Europeans. Journal of Allergy and Clinical Immunology 02 Mar 2018

Davies G, Cheung M, Gilmour K, Maimaris J, Curry J, Furmanski A, Sebire N, Halliday N, Mengrelis K, Adams S, et al. Thymus transplantation for complete DiGeorge syndrome: European experience Journal of Allergy and Clinical Immunology 08 Apr 2017

Prepared for submission:

Maimaris J et al. Novel EP300 mutation associated with immunodeficiency and polycythaemia rubra vera in a patient with Rubenstein-Taybi Type 2.

Sprenkeler E et al. Granulocyte colony-stimulating factor receptor expression is not essential for neutrophil differentiation but impacts emergency neutrophil responses.

ACKNOWLEDGEMENTS

I would like first and foremost, to thank my wonderful supervisor Adrian Thrasher. I first came to him scientifically homeless and I'm incredibly grateful to be given the opportunity to take on such an ambitious, forward thinking project. He is a truly inspiring leader, and I am incredibly proud to have been his student.

I would also like to gratefully acknowledge the many scientists and friends who have helped, taught and encouraged me along the way, in particular: Giorgia Santilli, Claire Booth, Fang Zhang, and Karen Buckland at Institute of Child Health; Elizabeth Ralph and Nadia Shahid at Great Ormond Street Hospital; Carl Anderson and Eva Serra at the Wellcome Sanger Institute, Cambridge; Hana Lango Allen and Chris Penkett at NIHR Bioresource, University of Cambridge and at the Royal Free Institute of Immunity and Transplantation: Vanessa Daza Cajigal, Steve Hanson, and Adriana DaSilva. I would also like to acknowledge Sophie Hambleton who granted me use of data collected from the whole exome sequencing program at University of Newcastle.

Thanks also to the many clinicians who provided samples to the NIHR as well as to the patients and families involved in the research. I thank the Rosetrees Trust and University College London who funded my research.

I would particularly like to acknowledge three brilliant and inspiring women who have guided and mentored me throughout my PhD and beyond: Kimberly Gilmour, Chiara Bacchelli and Siobhan Burns. In particular Siobhan has supervised and encouraged me to develop new ideas and given me a long lead to experiment and explore.

I’m fortunate to have parents who have never held me back. I see now that their own dreams and ambitions were curtailed to allow me my own, and I love them so much for this. Finally, I thank my husband and my two wonderful daughters Maya and Zoe, for their love, smiles and support.

ABBREVIATIONS

1000G	1000 Genomes
ACMG	American College of Medical Genetics and Genomics
AIS	Autoimmunity syndrome
BAM	Binary-compressed SAM
bp	Base pair
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CCDS	Consensus Coding Sequence
Chr	Chromosome
CID	Combined immunodeficiency
CNV	Copy number variation
CVID	Common variable immunodeficiency
dbSNP	Database of Single Nucleotide Polymorphisms
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
EBV	Epstein Barr virus
ECL	Enhanced chemiluminescence
ExAC	Exome Aggregation Consortium
g	Grams
GATK	Genome Analysis Toolkit
GERP	Genomic Evolutionary Rate Profiling
GRCh	Genome Reference Consortium Assembly
GTP	Guanosine triphosphate
GWAS	Genome-wide Association Study
gVCF	Genome Variant Call Format file
HAT	Histone acetyltransferase
HPC	High Performance Computing
HLH	Haemophagocytic Lymphohistiocytosis
HPO	Human Phenotype Ontology
HWE	Hardy-Weinberg Equilibrium
Ig	Immunoglobulin
kb	Kilo bases
MAF	Minor Allele Frequency

MDT	Multi-disciplinary team
mg	Milligrams
ml	Millilitres
mM	Millimolar
mRNA	Messenger ribonucleic acid
NBT	Nitroblue tetrazolium test
NEMO	Nuclear Factor-kappa-B essential modulator
NGS	Next Generation Sequencing
NIHR	National Institute of Health Research
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
ng	Nanograms
NGS	Next generation sequencing
OMIM	Online Mendelian Inheritance in Man
PBMC	Peripheral Blood Mononuclear Cells
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PERL	Practical Extraction and Report Language
PID	Primary Immunodeficiency Diseases
PTV	Protein truncating variant
QC	Quality control
RNA	Ribonucleic acid
TNF-α	Tumour Necrosis Factor -alpha
SKAT	Sequence Kernel Association Test
SKAT-O	Optimised Sequence Kernel Association Test
SIFT	Sorting Intolerant from Tolerant
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
UTR	Untranslated regions
UV	Ultraviolet light
VCF	Variant Call Format
V(D)J	Variable, Diversity, Joining
VEP	Variant Effect Predictor
VUS	Variants of Unknown Significance
WES	Whole Exome Sequencing

WGS	Whole Genome Sequencing
µl	Microlitres
µg	Micrograms

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Primary Immunodeficiency Diseases

1.1.1 Clinical Overview

Primary Immunodeficiency Diseases (PID) are a large heterogeneous group of inherited disorders that result in defects in immune system development and/or function. They describe a deficit of the immune system, whereby the patient is unable to mount a response to a pathogen, resulting in an increased susceptibility to infection. Examples include failure of B cell development, the absence of complement components, or impaired granulocyte function. PID are themselves encompassed by the wider group: inborn errors of immunity. This group include diseases of immune dysregulation such as autoinflammation. PID variably affects both adults and children, and is often associated with substantial morbidity including complications of severe infections, malignancy and premature mortality.

PID can vary in severity from benign antibody deficiencies, to the more severe and rare disorders such as severe combined immunodeficiency (SCID) which can cause an absence of both B and T cells. PID affect approximately 15,000 people throughout Europe [1]. In the UK, the prevalence of PID is 3.5 in 100,000. Most are individually rare with prevalence of 1 in 100,000 or less [2].

PID is a life-limiting disease and individual prognosis depends on the clinical features, in particular the presence of disease-related complications which can significantly reduce life expectancy [3]. The mainstay of treatment is replacement immunoglobulin (IVIG). The use of IVIG treatment has improved outcomes in children and adults with PID with an expected survival of 58% 45 years after diagnosis [3]. However, in patients with severe immunodeficiencies the most common treatment is haematopoietic stem cell transplantation (HSCT), which can be curative

but may entail iatrogenic complications and cause significant morbidity, and mortality [4-6].

1.1.2 Classification of primary immunodeficiency

Our immune cells originate primarily from haematopoietic development in the bone marrow, and specifically from the haematopoetic stem cell (HSC). Stem cells differentiates to eventually form all of our immune cells. Primary immunodeficiency is caused by disruption in key genes of haematopoetic development (particularly severe forms), where disturbance of either number and/or function of differentiated cells derived from the HSC, or effector cells controlled by the haematopoetic process result in PID.

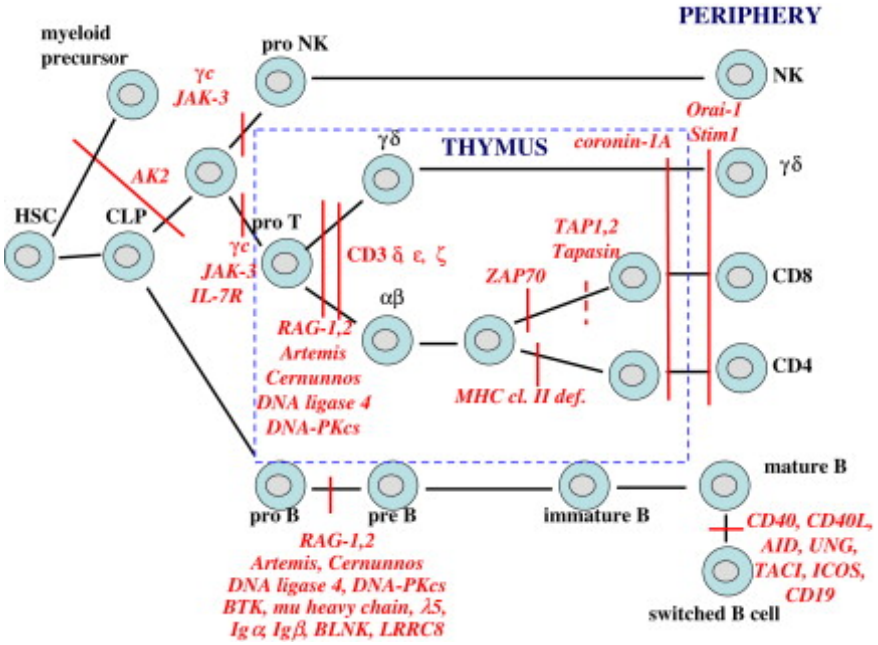
The broad spectrum of PID is reflected in the wide range of cells and tissues, which comprise the immune system. To date, 320 genes have been associated with PID as categorised by the International Union of Immunological Societies (IUIS) PID expert committee (See Appendix A) [7]. They can be broadly divided into PID diseases into the following categories:

1. Severe combined B and T Cell immunodeficiencies
2. Combined B and T Cell immunodeficiencies
3. Antibody deficiencies
4. Diseases of immune dysregulation
5. Congenital defects of phagocyte development and function
6. Defects in innate immunity
7. Autoinflammatory disorders
8. Defects of the complement cascade

1.1.1.1 Severe combined immunodeficiency

Severe combined immunodeficiency (SCID) occurs when there is disrupted development of both B and T cells (Figure 1.1). In the absence of newborn screening, patients typically present between 2 to 4 months with failure to thrive and sequelae of severe bacterial infections or opportunistic infections such as *Pneumocystis jiroveci*. Clinical signs include small or absent thymus and evidence of lymphocytic infiltration of skin akin to graft-versus-host disease.

Figure 1.1: Summary of genetic causes of SCID by aberration of T and B- cell development from the haematopoietic stem cell (HSC) and common lymphoid progenitor (CLP), genetic causes in red (Adapted from Notarangelo (2010)[8]



Immunophenotypic analysis of circulating lymphoid cells allows a useful method to sub-classify SCID. SCID can be caused by CD3+ T lymphopenia alone, where B cells are present but may be minimally functional. So called T-B+NK+ SCID defects are encountered in *IL7Rα* (IL-7 receptor subunit α) deficiency, *CORO1A* deficiency, and *FOXN1* disease. The most common form of SCID is caused by mutations in the

common gamma-chain receptor gene (*IL2RG*) found on the X chromosome. This receptor present on T, NK, mast cells and dendritic cells, signals multiple cytokines to mediate 3 major signalling immune pathways which have far-reaching consequences in immune function: the PI3K-Akt pathway, the RAS-MAPK pathway and the JAK-STAT pathway. T-B-NK+ SCID defects are typically associated with defects in DNA recombination and repair enzymes critical for V(D)J recombination of the T and B cell receptor e.g. *RAG1*, *RAG2*, *DCLRE1C*, *IL7RA*, *ZAP70*, *FOXN1*, *LIG4*. Hypomorphic missense mutations in some of these genes can result a partial loss of gene function with V(D)J recombination activity called Omenn's syndrome with immunodeficiency and characteristic features of erythroderma, and chronic diarrhoea caused by lymphocytic infiltration into skin, gut, liver and spleen. Autoreactive T cell clones via transplacental maternal-foetal engraftment may cause chronic skin inflammation, lymphadenopathy and chronic diarrhoea. Where natural killer cells are also absent (T-B-NK-), the likely diagnoses are adenosine deaminase (ADA) deficiency where there is toxicity to T, B and NK cells from high levels of abnormal purine degradation by-products, or *AK2* defects causing reticular dysgenesis.

In general, SCID is genetically well-defined with mostly biallelic causes arising in highly conserved genes. These monogenic causes make gene therapy an attractive option for treatment of SCID. However, genetically undefined SCID does present in non-consanguineous families with no apparent family history of immunodeficiency, suggesting *de-novo*, somatic mutations or epigenetic causes are possible.

1.1.1.2 Combined immunodeficiency

Combined immunodeficiency (CID) is a category of PID whereby both T cell and B cell functioning may be impaired but not completely absent. This gives rise to a spectrum of immunodeficiency disorders where the clinical features are generally milder than SCID with later onset of symptoms. Susceptibility to bacterial infections are indicative of B cell defects primarily, while infections with diverse pathogens including viral and fungal infections, are usually indicative of combined cellular and humoral defects.

Most genetically defined combined immunodeficiencies may be immunophenotypically categorised as shown in Table 1.1.

Table 1.1: Combined immunodeficiency disorders (CID) with characteristic immuno-phenotype

Type of CID	Immunotype	Genetic Mutations (Inheritance)	Clinical and Immunophenotypic Features
Low CD4 T cell subset	MHC Class II deficiency	RFX5 (AR) RFXANK (AR) RFXAP (AR) CIITA (AR)	Respiratory infections, Diarrhoea, Liver disease
	Intact MHC Class II expression	MAGT1 (XR) UNC119 (AD) LCK (AR)	Autoimmune features, Immune dysregulation
Low CD8 T cell subset	Normal Ig levels MHC Class I deficiency	TAP1 (AR) TAPBP (AR) TAP2 (AR) B2M (AR)	Pyoderma gangrenosum, Vasculitis
	Normal Ig levels Intact MHC Class I expression	ZAP70 (AR)	Immune dysregulation
		CD8A (AR)	Mild immunodeficiency
Low B cell count	High IgE Low IgM Low CD27+ memory B cells	DOCK8 (AR)	Eczema, Warts
		STK4 (AR)	EBV lymphoproliferation
		IL21 (AR)	Early onset colitis
	Low switched memory B cells	MAP3K14 (AR)	*Crytosporidium* infections
	Defective proliferation of T cells	MSN (XR)	Fluctuating neutropenia
Low Ig levels	High IgM	CD40 (AR) CD40LG (XR)	Opportunistic infections, Liver and biliary tract disease
	Impaired NK cell degranulation	DOCK2 (AR)	Early onset invasive bacterial and fungal infections
	Poor T cell proliferation	CARD11 (AR/AD)	Opportunistic infections
	Low memory T and T reg cells	BCL10 (AR)	Gastroenteritis
	Absent T reg and γδ T cells	IKBKB (AR/AD)	Opportunistic infections
	Variable T-cell dysfunction	ICOS (AR)	Autoimmunity, Granulomatous disease
	Poor T cell proliferation	TFRC (AR)	Thrombocytopenia
Poor specific antibody response	Low cytokine production	IL21R (AR)	Chronic cholangitis, Biliary and hepatic cirrhosis
	Impaired T cell proliferation	MALT1 (AR)	Inflammatory gastrointestinal disease, Variable dysmorphic features
High IgE	Absent IL-17 production by T cells	STAT3 (AD)	Retention of primary dentition, *S. aureus* infections
	Loss of PGM3 enzyme function	PGM3 (AR)	Severe atopy, some have skeletal abnormalities
	Reduced NK cells	SPINK5 (AR)	Congenital ichthyosis, Hair shaft defect
Neutropenia	Neutrophil migration defect	MKL1 (AR)	Cutaneous abscesses, Failure to thrive, Mild thromboctyopaenia

Genetic diagnosis is often aided by detecting associated syndromic features. Laboratory features such as thrombocytopenia may indicate a diagnosis of Wiskott-Aldrich syndrome, and may prompt investigation of *WAS* and other genes including *ARCP1B* and *WIPF1*, also known to cause immunodeficiency-associated thrombocytopenia [9]. Similarly, associated DNA-repair defects causing PID may present with growth retardation, syndromic facies, radiation sensitivity and propensity for malignancy. In these instances, mutations in genes critical to double-stranded DNA repair such as *ATM*, *NBS1*, *BLM*, *RNF168*, and *LIG4* may be found [10-14].

Combined immunodeficiency is also a main feature of the genetic syndromes velocardiofacial syndrome (DiGeorge syndrome) and CHARGE syndrome [15], where there is structural malformation of the thymus. Other genetic syndromes also variably present with immunodeficiency including Down syndrome, Turner syndrome, Seckel syndrome, Schimke immune-osteodysplasia and Wolf-Hirschhorn syndrome among others [15-18].

Ectodermal dysplasia syndromes are heterogeneous disorders in which there are structural abnormalities of organs arising from the ectoderm such as sweat glands, hair, skin and teeth. Immunodeficiency to various degrees have been described with some of these disorders; the precise mechanism of this is unclear. The most frequently encountered syndrome is caused by x-linked hypomorphic mutations in NEMO (*IKBKG*) which is characterised by susceptibility to invasive bacterial, fungal and mycobacterial infections of the respiratory tract, monocyte dysfunction, low IgG and IgA levels, and congenital features such as absent sweat glands, sparse hair and nail defects [19-21].

1.1.1.3 Antibody disorders

Genetic causes of agammaglobulinaemia are generally confined to genes in B-cell progenitor pathways in the bone marrow e.g. Igα deficiency caused by defects in *CD79A*. When B cell numbers are reduced but not entirely absent (>1%), a milder phenotype may present with recurrent infections with hypogammaglobulinaemia.

Common variable immunodeficiency (CVID) is the most frequently encountered symptomatic primary immunodeficiency in adults [22]. ESID criteria define this as hypogammaglobulinemia, specifically low IgG level with low IgA or low IgM, with poor polysaccharide vaccination response with absence of known causes of hypogammaglobulinaemia. Those with CVID are prone to developing autoimmune complications, granulomatous disease and have an increased chance of developing malignancy, particularly lymphoma [22, 23]. More detailed immunophenotyping has revealed a deficiency of switched memory B cells ($CD27^{+}IgD^{-}$,IgM^{-}), and further such tests allow for the opportunity for further classification and refinement of clinical management [24, 25]. The current consensus is that CVID is a heterogenous collection of different genetically determined diseases affecting B cell differentiation and development. Selective IgA deficiency is the most common immunodeficiency worldwide with an estimated 1 in 600 affected, but is excluded from this study, as those affected are asymptomatic or have sub-clinical pathology [26].

Pathogenic variants have been found in other genes critical to B cell receptor signalling (examples include *CD19*, *CD21*, *CD81*, *CD20* and PLC-γ2) or B-cell activation pathways (both in receptors: *ICOS*, *TNFRSF13B* or TACI, *TNFRSF13C* or BAFF-R, *MSH5* and ligands: *TNFSF13B* or BAFF, *TNFSF13* or APRIL, *TNFSF12* or TWEAK) [7, 23, 27, 28]. Genetic variants in genes causing CVID with autoimmune complications have also been found in *LRBA*, *CTLA4* and *PIK3CD*; these are the most common monogenic forms of CVID are notable for presenting in childhood [7, 29, 30].

Approximately 10% of CVID patients have first-degree relatives that also have either CVID or selective IgA deficiency, suggesting an inherited susceptibility to B cell dysfunction that may also be determined by cumulative environmental factors or shared non-Mendelian genetic factors [31]. In contrast to SCID or CID, single gene disorders are diagnosed in 2-10% of CVID cases [23, 28]. *NFKB1* gene mutations are the most commonly encountered monogenic cause of CVID [32]. Although biallelic *TNFRSF13B* mutations are also a known cause, the more common monoallelic *TNFRSF13B* mutations are currently thought to be a risk factor for developing CVID [33].

1.1.1.4 Diseases of immune dysregulation

Immune dysregulation is seen in many inherited immunodeficiency disorders and these features often overshadow immunodeficiency to impact significantly upon morbidity and mortality. Haemophagocytic lymphohistiocytosis (HLH) is a multisystem disorder of hyperinflammation resulting in immune dysregulation and tissue damage [34, 35]. HLH can be classified as primary (due to genetic disorder) or secondary. The age of onset can vary widely and infection may also trigger clinical presentation. Despite the genetic basis, it appears the condition may 'lie dormant' and then subject to an unknown trigger, causes an unregulated immune response leading to hypercytokinaemia and haemophagocytosis. To date, 5 genes have been associated with primary HLH (*PRF1*, *UNC13D*, *STX11*, *STXBP2*, *RAB27A*) [34-37].

Epstein-Barr virus (EBV) susceptibility to immune dysregulation has a genetic basis as demonstrated in genes causing X-linked lymphoproliferative disease (*SH2D1A*, *XIAP*), *MAGT1* deficiency among others [35, 38]. These genetic disorders may also present with primary HLH with EBV viraemia. Primary genetic diseases causing specific immunodysregulatory colitis by deficiency of the IL-10 pathway (*IL10*, *IL10RA*, *IL10RB*, *NFAT5*) are also well described [39-41].

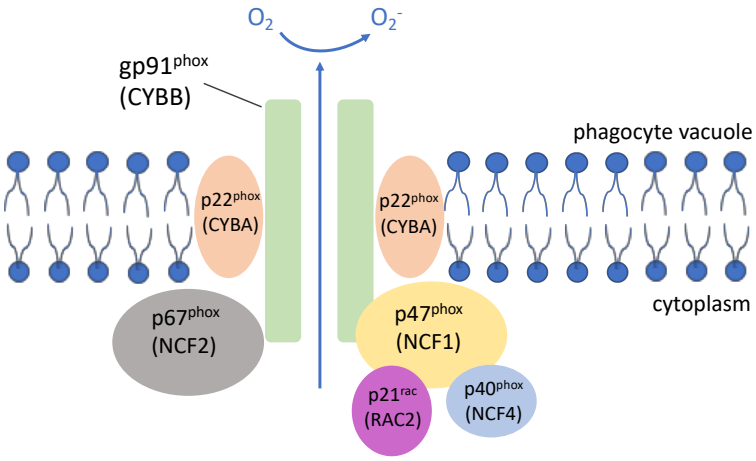
Apoptosis is a crucial mechanism in the immune system to halt an appropriate proliferative response to a pathogen. Deficiency in key proteins of the apoptosis mechanism cause profound autoimmune disease as in autoimmune lymphoproliferative syndrome (FAS (*TNFRSF6*), FAS-ligand (*TNFSF6*), *CASP10*, *CASP8*, *FADD*) [42-44]. Somatic mutations of *KRAS* and *NRAS* are also found in patients with ALPS [45, 46]. These rare variants are limited to cells in haematopoetic lineage, and absent in germline cells.

Autoimmunity occurs when there is failure to establish central or peripheral control of thymic or peripheral T regulatory cells. Breakdown of central tolerance in the thymus can be caused by mutations in *AIRE* causing autoimmunity-polyendocrinopathy-candidiasis-ectodermal dysplasia (APECED) syndrome [47]. Disruption of peripheral tolerance can cause autoimmune syndromes such as the immune dysregulation syndrome (IPEX) caused by mutations in the *FOXP3* gene [48]. Mutations of the CD28 costimulatory pathway genes such as *CTLA4*, *ICOS* and *CD28* are also implicated in multisystemic autoimmunity [49]. *CTLA4* deficiency, characterised by autoimmune cytopenia, enteropathy and granulomatous interstitial lung disease and *LRBA* deficiency, characterised by early-onset autoimmune enteropathy, cytopenia and interstitial lung disease and hypogammaglobulinaemia, are noteworthy examples [30, 49, 50].

1.1.1.5 Congenital defects of phagocyte development and function

Chronic granulomatous disease (CGD) describe specific defects in the intracellular oxidative burst of phagocytes, with both autosomal recessive (*CYBA*, *NCF1*, *NCF2, NCF4*) as well as X-linked genetic causes described (*CYBB*) [51, 52] (Figure 1.2). These genes are well defined, but more recently *CYBC1* has been implicated, by interfering with production of the p91 phox protein (*CYBB*) [53]. Other genes play a role in neutrophil killing of pathogens including *RAC2, GATA2, CEBPE, CSF2RA, CSF2RB* and *G6PD* [54-59].

Figure 1.2: Components of NAPDH oxidase complex of phagocytes with respective coding gene in brackets. Activation of NADPH oxidase depends on successful assembly of components during activation, and therefore gene defects cause failure of oxidation, resulting in chronic granulomatous disease (CGD)



Congenital defects of neutrophil number or function result from isolated pathogenic defects in neutrophil differentiation genes (*ELANE*, *GFL1*, *HAX1*) [60, 61]. Leucocyte adhesion defects are rare autosomal recessive disorders caused by genetic mutations in *ITGB2*, *SLC35C1*, and *FERMT3* that are allow phagocytes to tether and migrate to sites of infection [62-65]. Multisystem genetic syndromes may in addition, also feature congenital phagocyte defects e.g. Schwachman-Diamond syndrome, cystic fibrosis and Cohen syndrome [66-68].

1.1.1.6 Defects in Innate Immunity

Cells of the innate immune system protect against invasive bacterial and fungal infections. Serious bacterial infections including meningitis and sepsis with common pathogens such as *Streptococcus pneumoniae* and *Staphylococcus aureus* may indicate underlying innate immune defects such as *IRAK4* and *MYD88* deficiencies [69].

Genetic predisposition to viral infections may result from defects in the Toll-like-receptor 3 (TLR3) pathway including *TLR3*, *TRAF3* and *UNC93B1* [70]. However, there is often cross-over with other immunodeficiency, with *STAT1* deficiency, *STAT2* deficiency and WHIM syndrome (due to gain of function mutations in *CXCR4*) all causing predisposition to viral infections as well as other clinical and immunophenotypic features [71-73].

Predisposition to mucocutaneous candidiasis and other persistent fungal infections can indicate defects in the IL-17 signalling pathway (*IL17F*, *IL17RA*) and *STAT1* [74, 75]. Deleterious mutations in *CARD9* have also been shown to cause invasive fungal infections [76].

Mendelian susceptibility to mycobacterial disease (MSMD) can be caused by genetic mutations within diverse immunological pathways. More severe phenotypes present in *IFNGR1* and *IFNGR2* deficiencies while more milder phenotypes may be caused by IL-12, Tyk2 and *IRF8* deficiencies [77-80].

1.1.1.7 Autoinflammatory disorders

Diagnostic clues in the clinical presentation may also indicate specific pathway defects, allowing investigators to narrow down a set of genes that may be responsible for disease. A predominance of inflammatory features such as recurrent fever, periodic arthralgia and multiorgan involvement may be indicative of defects of pyrin (coded for by *MEFV*) or related inflammasome related pathways affecting particular genes such as *TNFRSF1A* and *NLRP3* [81, 82]. Sterile inflammation of the skin and joints are features of inborn errors of immunity caused by *PSTPIP1* mutations, and rare *NOD2* mutations [83, 84]. Type 1 interferonopathies such as the genetically heterogeneous Aicardi-Goutieres syndromes are very rare disorders usually presenting in early childhood with severe neurological encephalopathy and

microcephaly. However, type 1 interferonopathies may also be more restricted in scope such as early onset vasculopathy caused by mutations in *TMEM173*, due to constitutive activation of interferon-β production [85, 86].

1.1.1.8 Defects of the complement cascade

The complement system is a group of 14 evolutionary-conserved proteins, comprising key proteins of the complement cascade and associated regulatory proteins and their receptors. The system functions in 3 interacting pathways, the classical, alternative and mannose-binding lectin pathway. Defects of the late components (C5-C9, properdin, Factor D) often causing increased susceptibility to encapsulated bacteria especially *Neisseria* [87-92]. While typically classified as PIDs, some genetic variants in complement genes present without immunodeficiency e.g. systemic lupus erythematous (C1q/r/s, C2 and C4) and atypical haemolytic uraemic syndrome (C3, Factor B/H/I) [88, 93-96]. *C3* and *CFB* genetic mutations causing atypical haemolytic uraemic syndrome have a gain of function effect, and biallelic loss in CFH can cause C3 glomerulopathy. In practice, a genetic diagnosis is not mandated as protein and biochemical tests can accurately pinpoint both individual complement level and functional defects. However, genetic diagnosis can reveal underlying inheritance patterns and allow family members to access early detection and treatment.

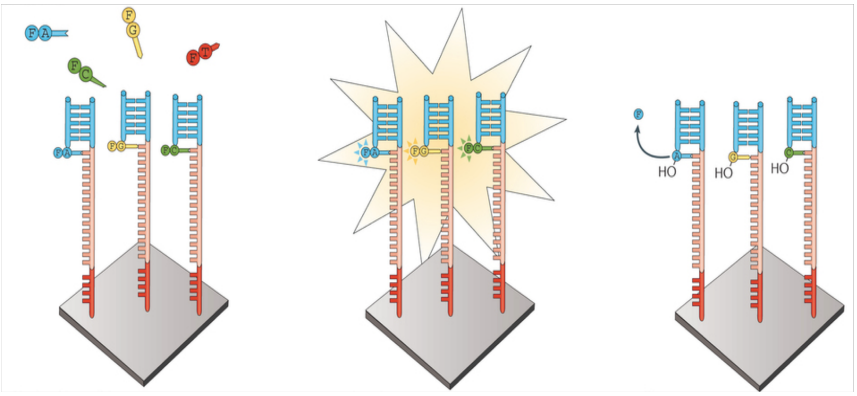
1.2 Whole Genome Sequencing

1.2.1 Overview

Whole genome sequencing (WGS) is the methodology used to determine the full DNA sequence of the entire genome. The analysis of the whole genome began with Sanger sequencing almost 2 decades ago (International Human Genome Consortium 2004). Since then, next generation sequencing (NGS) technology has developed for large scale sequencing projects, such that it would accommodate a higher throughput and be more cost effective. It illustrates the shift from sequencing one fragment at a time to generating thousands of sequences concurrently in a faster time frame. Since its inception, it has dramatically accelerated biological research by allowing comprehensive analysis of genomes to become more routine and undertaken in less time.

For this project, NIHR Bioresource Rare Disease (NIHRBRD) collaborated with the biotechnology company Illumina to provide the sequencing technology and platform. The NGS method 'sequencing by synthesis' was used (Figure 1.3). This chemistry relies on 4 fluorescently labelled nucleotides with reversible termination property which naturally compete for incorporation into DNA polymerase. These deoxynucleotides are present in the flow cell in which DNA samples are loaded.

Figure 1.3: 'Sequencing by synthesis' method which relies upon the reversible termination properties of fluorescent labelled nucleotides as illustrated in Goodwin et al (2016)([97]



Nucleotide addition
-fluorophore labelled hybridisation

Imaging
-slides imaged in laser channels
-fluorophore quenched as base incorporated

Cleavage
-fluorophores cleaved
-3-OH regenerated for new cycle

Traditionally, DNA samples would undergo a library preparation step whereby DNA is sheared into smaller pieces (100-400bp in length) and PCR amplified using solid phase bridge amplification. However, newer methods used for this project are PCR-free, such that the library preparation step only incorporates shearing of DNA, purification of size selected fragments and dual indexing of fragments. The absence of a PCR step affords fewer gaps in coverage, greater uniformity of coverage and better coverage in areas which are GC-rich which can be difficult to sequence [98].

The library fragments graft onto the flow cell by their adapter and are clonally generated in a method called bridge amplification. This generates multiple copies of a single DNA molecule arranged spatially on the surface of the flow cell. Subsequently primers hybridise onto the fragments and sequencing reactions occur in parallel while the fluorescence generated from incorporated nucleotides are detected by the machine. Where paired end reads are used, the original template is used to synthesise a complementary strand. As the length of the fragment is known,

variations in length can be characterised and as such, structural variation can be identified.

1.2.2 Use of Whole Genome Sequencing for novel gene discovery

The advent of NGS technologies has made it possible to perform WGS to identify novel disease predisposition genes in rare Mendelian disorders [99, 100]. Exome sequencing is the most commonly used NGS technique to identify novel disease-causing genes in PID, although these successes have largely been restricted to highly penetrant single gene disorders [75, 101, 102]. There remains a large number of patients in which exome sequencing has not found a causative gene due to inherent limitations. Exome sequencing covers ~ 1% of the genome [103]. It does not capture all exons, nor non-coding regulatory regions. Structural variants such as copy number variants (CNVs) are also not readily detectable in WES; CNVs in exonic areas are very rare and generally extend outwith the targeted region [104, 105]. Additionally, variants within captured regions may be missed either due to insufficient coverage or due to errors in the mapping/variant calling algorithms. WGS is better than WES in finding rare variants within exons [106]. This is, in large part, due to reliance of WES to PCR-amplify targeted regions [107]. While this increases the number of reads which overlap mid-exon, there are inevitably fewer at the beginning and end of the targeted region. Current whole genome sequencing (WGS) technology allows ~97% of the genome to be covered [108]. The remainder likely comprise of highly repetitive regions, which will likely remain refractory to high throughput sequencing using current technologies. Despite the focus of protein-coding regions in novel gene discovery, intronic mutations have been described in PID [101, 109]. The potential to uncover pathogenic variants in unexplored areas within known disease genes as well as novel disease variants, is therefore significantly enhanced by using WGS. It is estimated that the cost of sequencing a whole genome is ~$1000 [110]. This makes the diagnosis and understanding of the genetic basis of rare diseases through WGS increasingly attainable and cost-effective.

1.3 Mechanisms of Genetic Disease

1.3.1 Mendelian disorders

Genetic variation can contribute to the development of human disease. Conditions arising from single gene defects are known as Mendelian disorders, whereas those requiring multiple interacting genetic variants are categorised as complex genetic disorders. Highly penetrant rare genetic variants in Mendelian disorders are sufficient to cause disease regardless of the genetic or environmental background.

The inheritance of single gene disorders varies with autosomal recessive forms frequently encountered in early onset PID. Autosomal recessive conditions require both alleles of a gene to be mutated. The same variant may be found in both alleles, or each allele may contain a different mutation, known as compound heterozygosity. As each allele is inherited from a different parent, usually both parents are carriers of a mutated allele with a 25% risk of the child developing the disease. In PID, autosomal recessive disorders result from a 'loss-of-function' effect, where there is either reduced activity or absence of functional protein.

Autosomal dominant inheritance occurs when a single copy of the genetic variant on one allele of germline DNA is sufficient to cause disease. The disease is manifest in heterozygotes, with a 50% change of passing the causative allele to offspring.

Some autosomal dominant conditions are known to cause 'gain-of-function' effects, where the mutant allele results in a protein which interferes with normal protein e.g. in Hyper-IgE syndrome, heterozygous missense mutations in *STAT3* can cause disease by interference of dimerization of STAT3 protein, resulting in impaired JAK/STAT signalling [111]. Gain of function effects may also arise from a gene product

with a new function that overrides normal protein function, termed neomorphic mechanism of disease – so called dominant negative.

Autosomal dominant conditions may also cause disease by haploinsufficiency, imitating a loss of function mechanism. This occurs when one wild-type allele does not provide enough of the functional protein such that disease manifests. This is seen in the example of monoallelic genetic variants in the nuclear-factor kappa B 1 gene (*NFKB1*) commonly causing combined variable immunodeficiency [32, 112]. These mutations may also be termed hypomorphic mutations.

X-linked recessive disorders in PID are caused by mutations in genes on the X chromosome, and hemizygous males with one copy of the genetic variant are affected. It is possible for females to develop X-linked conditions; in recessive disorders, this is due to X-inactivation, where the permanent inactivation of one X chromosome in female cells can reveal disease in the other allele. Males affected with X-linked recessive disorders pass the mutant allele to their daughters (who are carriers), but will not pass the allele to their sons.

Most genetically defined PID are single gene disorders with an autosomal recessive or X-linked recessive mode of inheritance. A so-called 'somatic second hit' has also been described in many conditions, particularly in relation to tumour suppressor genes and cancer predisposition. Mitochondrial or Y linked genetic inheritance has yet to be identified in PID.

Genetic mutations associated with PID are almost all germline mutations. This means that they occur in gametes, every cell in the organism is affected, and are heritable. Exceptions are now being discovered such as somatic mutations of *KRAS* and *NRAS*

which are known to cause some forms of autoimmune lymphoproliferative syndrome (ALPS) [46, 113]. Somatic mutations originate in body cells and are not heritable.

Somatic mosaicism is the presence of a genetically distinct cell population within a tissue, and may be caused by spontaneous DNA mutations within a cell. The extent of genetic PID caused by somatic mosaicism is unclear. However, high-depth sequencing of particular gene regions have demonstrated somatic *NLRP3* mosaicism in 17.7% of sequencing reads in an affected patient where germline *NLRP3* mutations are absent [114].

1.3.2 Disease gene identification

Next generation sequencing has had a significant impact in monogenic disorders, and in PID, has identified new pathways important for immune function. These advances improve management of rare disorders but also enhance understanding of critical pathogenic mechanisms underlying complex disease. Making a genetic diagnosis can determine subsequent management decisions and affect long term prognosis. Genetic testing has implications for family members who may be at risk of PID, particularly for autosomal dominant conditions, where late-onset of symptoms may occur. It may also allow carriers of disease to be identified for counselling and possible intervention such as preimplantation genetic diagnosis.

The human DNA sequence varies by approximately 0.1-0.4% of nucleotides between unrelated genomes, with most sequence variation seen in the non-coding space [115]. Rare diseases which arise from hereditary or sporadic mutations typically affect a small percentage of the population. Deleterious mutations are subject to evolutionary selection pressure, and are therefore, more likely to be coding

mutations that have a negative impact on protein function. This explains the lower frequency of these variants in the genome. Historically, identification of causal genetic variants querying them in literature or public databases. From sequencing studies, prioritisation of variants is typically performed through sequential filtering until the number of variants to be tested is reduced to a manageable size. Typically, these are on population allele frequency and variants with clear functional consequences (Table 1.2). With increasing availability of large public genome datasets, selection of a mean allele frequency of 1% is a permissive approach [116]. In genome association studies, more common variants may be prioritised and thus, annotation of non-coding regions is useful to identify proximity to transcription factor binding sites and phylogenetic mammalian-conserved regions.

Table 1.2: Variant prioritisation for candidate disease-causing variants in coding regions (Adapted from Meyts et al) [117]

Levels of variant analysis	Criteria
Genetic Hypothesis	Mode of inheritance
	Clinical penetrance
	Genetic heterogeneity
Variant-specific	Low allele frequency (<1% for rare diseases)
	Predicted damaging protein impact
Gene-level	Encodes protein relevant to phenotype
	Expressed in cell-type relevant to phenotype
	Gene under strong purifying selection pressure
Experimental Validation of causal relationship between genotype and phenotype	

In-vitro functional studies are conducted to assess the effect of the variant. For variants that affect coding sequences of genes, protein levels and or an output of function may be measured. For variants that lie in regulatory regions, transcriptional effects can be measured using mRNA. However, if variants occur far from known genes or further functional confirmation is required, transgenic or knockout animal

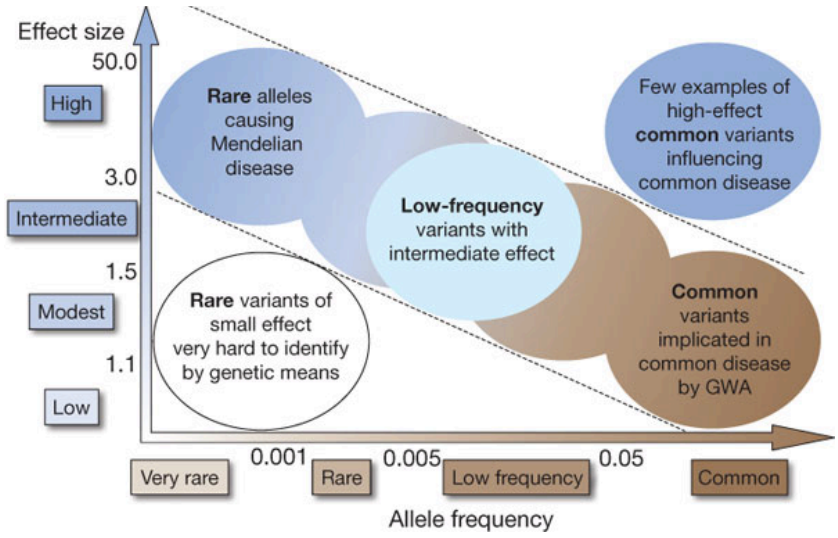
studies may be used to demonstrate *in-vivo* effects of a genetic variant. More recent use of induced pluripotent stem cells to model genetic variants has allowed detailed study of *in-vitro* functional effects [118, 119].

1.3.3 Genome wide association studies

Genome wide association studies (GWAS) are based on the statistical hypothesis concerning the co-occurrence of common variants (minor allele frequency >5%) with disease, termed the 'common disease common variants' hypothesis [120, 121]. A genetic variant or risk allele is associated with a disease if it appears at a significantly higher frequency in those affected compared to control subjects [121, 122]. GWAS is a statistical method which requires nor assumes *a priori* knowledge of location or function.

Success in the identification of novel genetic predisposition factors rely upon understanding of the underlying genetic architecture. Relatively common diseases, such as Crohn's disease or Type 1 diabetes mellitus, have been shown to have a substantial proportion of heritability attributable to common variants [123, 124]. These variants may represent modest increases in risk, typically 1.5 fold or less (Figure 1.4).

Figure 1.4: Components of genetic architecture (Taken from Manolio et al (2009)) [125]



While association studies are an attractive option for mapping of susceptibility loci, in most cases large numbers of samples of both cases and controls are required to reach statistical significance which can be a considerable challenge in the study of rare diseases. However, there are some notable exceptions to this. The power of a study to detect a statistically significant association with disease is reliant on the frequency and penetrance of the variant, and its individual disease risk effect with correction for multiple testing.

Low-frequency alleles which may have an intermediate effect tend to have a lower effect size and have reduced mean allele frequency (MAF <0.05). The 'common-disease, rare-variant' hypothesis suggests that rare alleles (MAF <0.05) may contribute to the missing heritability of complex disease [125-127]. However, the significance of variants in non-coding regions may be difficult to assess functionally. Similarly, two or more loci that interact to cause a positive disease association, known as epistasis is also challenging to study using association methods, and are reliant on large sample sizes [128].

1.4 Genetic studies in primary immunodeficiency

Historically, immunopathology was identified when an affected protein or cell type was recognised to be deficient. Thereafter, combined pedigree and linkage analysis was used to map disease genes particularly in consanguineous families, or in families where multiple generations of boys were found such as *CYBB* on the X chromosome, causing chronic granulomatous disease. Genetic linkage analysis was however laborious and time-consuming, particularly as sequencing was often undertaken manually. Both the development of high-throughput next-generation sequencing (NGS) and the map of the human genome available from 2003, have facilitated the rapid increase in PID gene discovery. With technological advancements and automation, the price of NGS sequencing has fallen, allowing initially specific regions of a gene to be sequenced, to more agnostic tests now within reach, such as sequencing of the whole genome.

Whole exome sequencing in a large cohort of affected patients has led to the discovery of PID genes, initially *FADD* in a series of patients with autoimmune lymphoproliferative disease (ALPS) in 2010 [129]. Since then, many large-scale projects have replicated this design to find genes in a phenotypically-restricted cohort. To date, there are few large-scale studies demonstrating the efficacy of whole genome sequencing in PID [130]. The studies that have been reported generally focus on screening patient or familial cohorts for variants in PID genes using gene panel sequencing or whole exome sequencing [131-134].

Utility of sequencing varies according to availability and cost. In the UK, clinical immunologists have several opportunities to submit patients' DNA for research-led national NGS sequencing projects (e.g. NIHR Bioresource BRIDGE) which usually have a diagnostic arm.

Worldwide, focused targeted gene panel testing for specific PID-related genes appears to be the most utilised, particularly as it is the most cost-efficient with results from a local referral centre and typical costs in range of $250-500 per sample [135].

1.4.1 PID as a complex disease

Complex diseases arise due to a combination of genetic and environmental factors. In immunodeficiency, it is widely thought that acquiring infection may uncover or precipitate symptomatic immunodeficiency and HLH, particularly as seen in Epstein-Barr virus (EBV) [136, 137]. For a given disease, the relative contributions of genetic and environmental factors may vary between populations. Advances in molecular biology have enabled us to identify specific genetic factors underlying complex disease, using techniques from karyotyping, linkage analysis to more specific interrogation of specific candidate genes and whole genome sequencing. High penetrance predisposition genes, mutations in which are rare in the population have been identified using sequencing studies while association studies have allowed identification of common low-penetrance predisposition loci. By consideration of both, an architecture of genetic susceptibility to PID, may be built. However, given that most PID does not have a genetic diagnosis, and is heterogenous in onset and features, it is clear that we have only identified the genetic factors accounting for a small but reasonable, proportion of the genetic risk. The remainder of genetic risk may lie in aggregate consequences of multiple genes, faulty interactions between genes, or epigenetic factors.

1.5 Overview of this Thesis

In this introduction, I have reviewed the basis of genetic disease and genetic studies in PID. My hypothesis is that the application of WGS to a sample of the PID population is capable of discovering rare genetic variants that cause disease. I utilised bioinformatic techniques, clinical interpretation and laboratory bench analyses in this thesis. The outline of the thesis is as follows:

Chapter 2: Materials and Methods used in this thesis

Chapter 3: Analysis of frequently utilised NGS methods: targeted gene panel sequencing, whole exome sequencing and WGS to determine their efficacy in obtaining a genetic diagnosis in PID.

Chapter 4: Comprehensive WGS-based screening of gene known to cause PID in a paediatric cohort with previously undiagnosed PID.

Chapter 5: Novel gene discovery in PID by bioinformatic pathway analysis and rare variant gene association study to find novel variants that may contribute to PID.

Chapter 6: Two case studies of patients that have a novel genetic variant found by analyses in Chapter 5. The first is the case of a novel mechanism of Rac2 haploinsufficiency caused by a heterozygous mutation in *RAC2* in a man with combined immunodeficiency. The second case study is of a novel *EP300* gene mutation in a patient with a novel PID syndrome.

Chapter 7: Final discussion

2 MATERIALS AND METHODS

The whole genome sequencing data analysed in this thesis is part of the NIHR Bioresource Rare Disease BRIDGE-PID project, a multi-centre collaborative projects established by the BRIDGE consortium in 2012 with the aim to find genetic causes of rare diseases across caused by inborn errors of immunity. My contribution included bioinformatic analysis of WGS data, including annotating BAM and VCF sequencing files, designing variant analysis pipeline, statistical analysis and quality control of data submitted from Great Ormond Street Hospitals NHS Foundation Trust, and latterly from Royal Free NHS Foundation Trust. In this, I was supervised by Dr Chiara Bacchelli, UCL and Dr Hana Lango Allen, University of Cambridge. The WGS sequencing data is housed in a secure repository hosted by the University of Cambridge, which I was able to remotely access, with special permission from Professor Ken Smith, University of Cambridge.

I conducted laboratory experiments at UCL, both at Institute of Child Health under the supervision of Professor Adrian Thrasher, and at Institute of Immunity and Transplantation, Royal Free Hospital NHS Foundation Trust, under the supervision of Dr Siobhan Burns.

2.1 WGS analysis in BRIDGE-PID

2.1.1 BRIDGE-PID patient recruitment

The NIHR BioResource BRIDGE-PID project is one of 14 NIHR Rare Disease projects established by the BRIDGE consortium in 2012. Children with a genetically undefined primary immunodeficiency disease at Great Ormond Street Hospital were recruited to the NIHR BioResource BRIDGE-PID study if they fulfil one of the criteria:

i. Primary immunodeficiency disorder (PID): Patients with recurrent and/or unusual microbial infections suggestive of severely defective innate or cell-mediated immunity.

ii. Common variable immunodeficiency (CVID): Patients show a failure of production of specific antibodies or documented failure of specific antibody production.

iii. "Extreme" autoimmunity: Patients with aggressive autoimmune disease with an early age onset. This includes patients with clinical features of inborn errors of immunity or chronic lymphoproliferation with autoimmunity, but without mutations in classically associated genes.

Where possible, parents were recruited concurrently to the study. Recruitment to the study required a clinical proforma to be completed, a consent form signed by the individual if over 18 years or by the parent, if the patients is under 18 years old, and a whole blood sample or DNA (previously extracted from whole blood sample). In all cases, clinical proformas were completed by the lead clinical scientist or named consultant in the child's care. Each individual family was assigned a number and the

pseudonymised demographic, clinical and any other molecular information was recorded in a database, physically located at the University of Cambridge.

2.1.2 WGS workflow

WGS was carried out at the University of Cambridge with DNA samples processed in batches of 94 in order to include positive and negative control samples. Library preparation was done using Illumina TruSeq DNA PCR-Free HT sample preparation kit (Illumina Inc., San Diego, CA, USA) on the Beckman Biomek FX automated workstation (Beckman Coulter Inc., Brea, CA, USA). Libraries were sequenced as 150bp paired-end reads on the Illumina Hiseq X instrument.

BAMs (~430 GB/sample) and gVCFs (~3 GB/sample) were generated by the NIHR BioResource core pipeline team. Reads were aligned to the GRCh37 build of the human reference genome using BWA v0.6.223. Realignment around indels and base call quality recalibration was performed using GATK v2.3_924. Variants were called by the Illumina ISAAC variant caller v2.0.17. Structural variants including copy number variation were called with both Canvas and Manta callers.

Remote data access was provided by NIHR BioResource and the University of Cambridge. Data is hosted at University of Cambridge High Performance Computing facility (HPC). Shared genotype data was accompanied by high level phenotype data (gender, type of rare disease, and ethnicity).

2.1.3 Sequencing metrics

Prior to analysis of genetic variants, I conducted a series of quality control assessments on the VCF files to make sure the sequencing data were of high quality. Quality control assessments are required to detect problems such as sample contamination, sample mis-labelling or failed sequencing runs. Sequencing errors may occur due to factors such as depth of sequencing, read length and method preparation, all of which can act as barriers to the success of genetic studies. For some of the analyses, I focussed on SNVs as they are the most plentiful variant and represent a large proportion of known pathogenic variants identified in PID.

I examined the following sequencing metrics:

- Population stratification using principal component analysis
- Variant quality with transition/transversion (Ts/Tv)
- Het/alt ratio
- Strand bias
- Average depth of coverage of variants

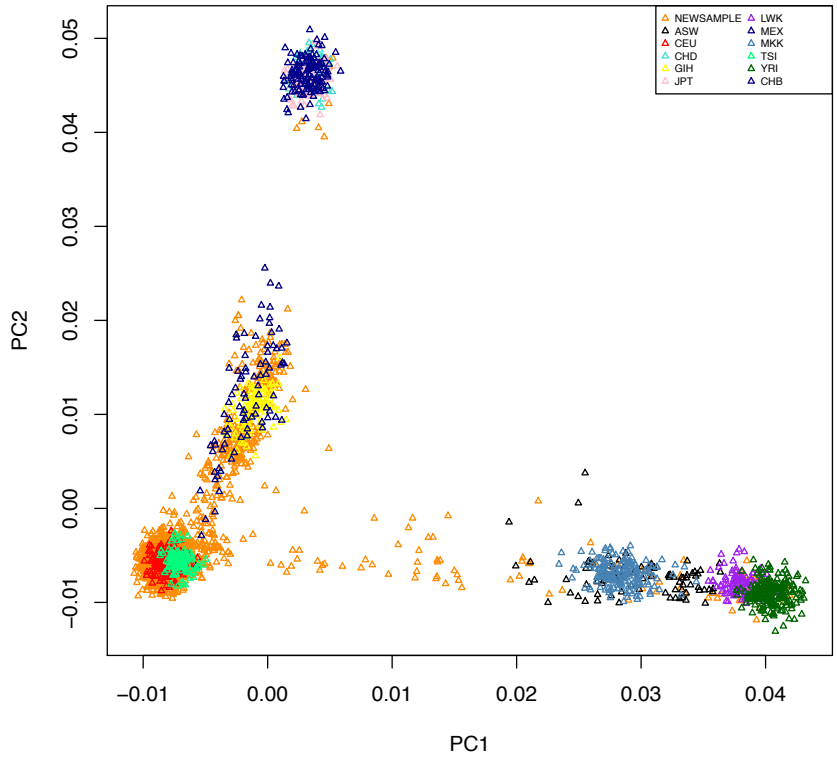
<u>Principal component analysis (PCA)</u>

Principal component analysis is a method to identify axes or components of variability, to simplify the data into groups, holding most of the variation. When using genomic data, the majority of variation correlates with ancestry differences in the samples. Ancestry analysis is useful to undertake to identify the underlying population structure, highlight groups which differ at level of minor allele frequency, and identify outliers that may require either further analysis or elimination from the sample population.

To undertake the PCA, I used standard analysis commands in bcftools and PLINK. Initially, SNVs from the sample population were separated from multi-allelic calls and indels. Independent SNVs were then restricted to biallelic SNVs in autosomes with a minor allele frequency >10%. These were compared to HapMap3 markers, and the markers in common were further 'pruned' for linkage disequilibrium (LD) (HapMap3 markers obtained from: https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html Accessed in March 2015). This left a total of 18,740 SNVs for analysis. The PCA analysis was carried out with Eigenstrat software and generated in R [138].

The PCA analysis showed that 38% of the WGS samples were not of European ancestry (Figure 2.1). The largest non-European overlap was with HapMap samples from Gujurati Indian origin, suggesting that a significant group of the patient population had South Asian ancestry.

Figure 2.1: Principal Component Analysis showing cohort population HapMap markers (NEWSAMPLE in orange) overlapping other key HapMap populations defined in separate colours (ASW: Americans of African Ancestry in USA, CEU: Northern and Western European ancestry in USA, CHD: Chinese ancestry in USA, GIH: Gujarati Indians in USA, JPT: Japanese in Tokyo, Japan, LWK: Luhya population in Kenya, MEX: Mexican ancestry in Los Angeles, MKK: Masai ancestry in Kenya, TSI: Toscani in Italy, YRI: Yoruba in Nigeria, CHB: Han Chinese in China)



Transition/transversion (Ts/Tv) ratio

The transition/transversion (Ts/Tv) ratio is used in large-scale next generation sequencing projects as a SNV quality control test. Transitions are mutations of a purine base to another purine, or pyramidine base to another pyramidine, while transversion SNVs are mutations between purine and pyramidine bases.

A mean Ts/Tv ratio of 2.85 (+/- 2.77-2.93) was found across coding regions, and 2.06 (+/- 2.02-2.09) across the whole genome (Figure 2.2A). These ratios were within the expected values for exome and for whole genome datasets [139, 140].

Het/nrHom Ratio

The Het/nrHom ratio is a ratio of heterozygous to non-reference homozygous variants calculated by deriving the inbreeding co-efficient (F) using BCFtools to estimate the proportion of the genome that is autozygous. Sample relatedness causes segments of chromosomes to be identical-by-descent (IBD), with children of consanguineous parents having a higher number of homozygous IBD variants. When conducted to validate rare disease variant studies, offspring of consanguineous parents have higher number of homozygous alternative calls because a higher proportion of their genome contains a larger proportion of alleles that represent physical copies of each other (autozygous) or contain a larger proportion of ancestral alleles that are IBD. Genomes from patients that were known to be consanguineous, had a F value > 0.0156 and the Het/Hom ratio was <2 (Figure 2.2B).

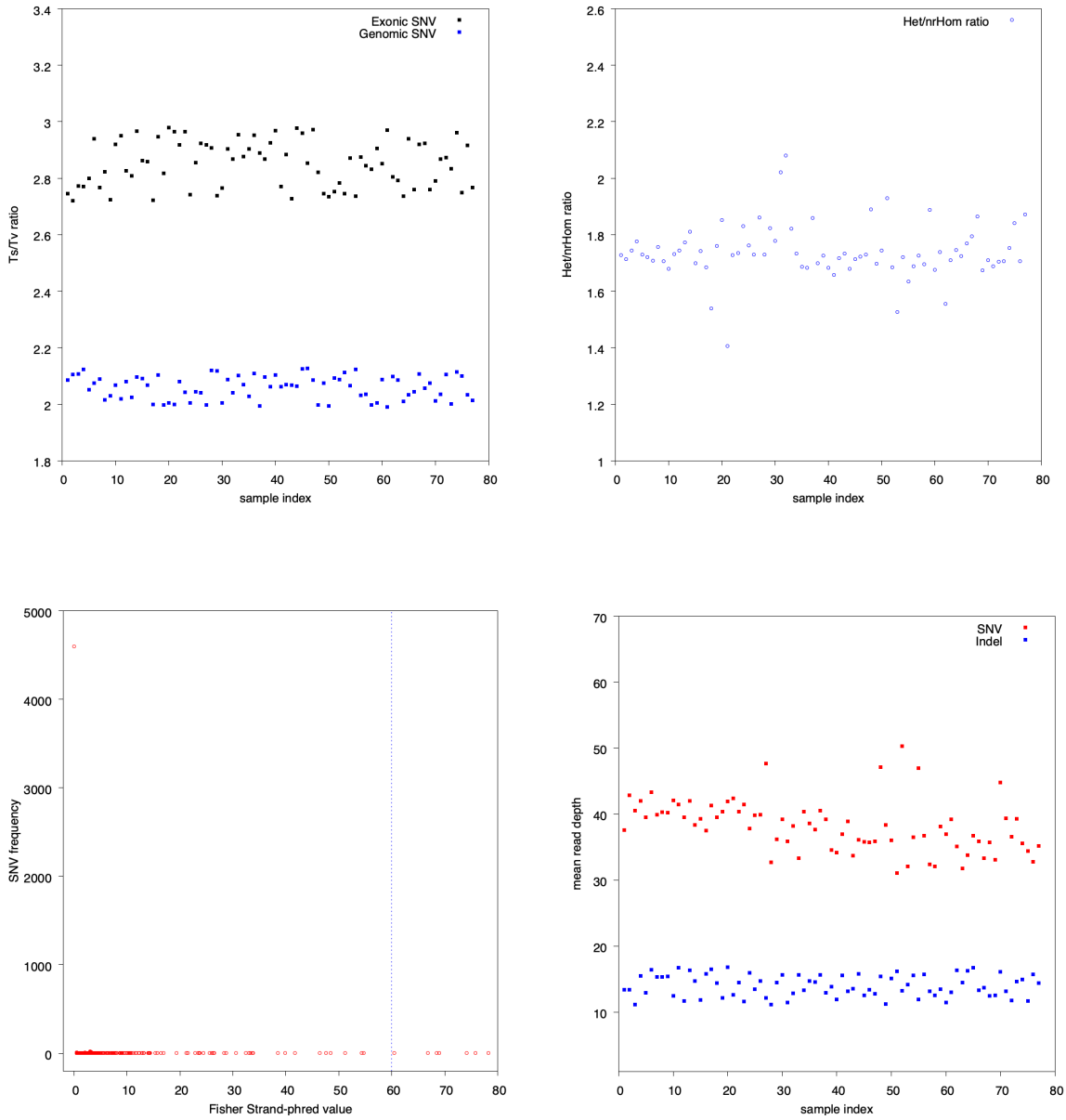
Strand bias

Strand bias quantifies the genotype of a variant being seen on only the forward or only the reverse strand. Positive values denote significant strand bias and are associated with lower values of Ts/Tv ratio, therefore they are more likely to indicate false positive variant calls. Strand bias was calculated using GATK Haplotype caller Fisher Strand analysis, based on Fisher's exact test. The output is Phred scaled p-value with higher FS values indicating the greater likelihood of strand bias. GATK best practices guidelines suggest discarding variants with an FS Phred score greater than 60; with 7 SNVs were filtered out prior to analysis in this experiment [141]. The vast majority of SNVs had a FS value of 0, suggesting no strand bias occurred (Figure 2.2C).

Average depth of sequencing

A useful metric is to assess the sequencing efficacy is the average depth of coverage of variants. Per-sample average depth of variants were sought using Bcftools version statistics output. Low quality DNA or adaptor contamination in the library preparation step can cause a lower than expected average depth count [142]. All samples in the cohort met Illumina's sequencing standard of an average SNV depth >30 (Figure 2.2D).

Figure 2.2: Sequencing quality control metrics for the PID patient cohort. A) Transversion/Transition ratio for variants within coding regions and across non-coding regions. B) Het/nrHom ratio of each sample in the cohort. C) Strand bias analysis using Fisher-Strand values, with variants with FS value> 60 (dotted line) excluded. D) Average depth of coverage of SNVs and Indels per sample



Illumina's ISAAC variant caller was used to generate SNVs and indels from Illumina's gVCFs. gVCFs limited to the GOSH probands sample cohort were generated from the current version of the aggregated shared genotype data. This allowed for a much

smaller file and thus, streamlined data processing. The VCF was annotated with Ensembl Variant Effect Predictor version 78. Allele frequencies were sought at ExAC (http://exac.broadinstitute.org/),1000G (https://www.ncbi.nlm.nih.gov/SNP/) and data from generated from NIHR BioResource (BRIDGE) itself. Variants of unknown significance were studied throughout the BRIDGE cohorts, both in their presence and allele frequency in disease-relevant populations. A PERL script was written to efficiently process the annotated gVCF file and perform variant filtering, in order to prioritise pertinent variants in particular gene lists.

Finally, to assign pathogenicity, potential variants are scrutinised at a monthly multi-disciplinary meeting to assess the pathogenicity of the variant and its likelihood to cause the disease in the given patient. If agreed to be likely causative, the clinical genetics laboratory confirmed the variant via Sanger sequencing before issuing a confirmatory report. In addition, putative genetic variants are often confirmed experimentally to cause disease.

2.2 Analysis of gene-panel and whole exome sequencing in PID

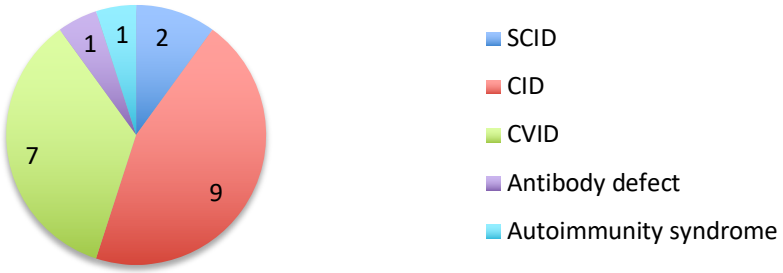
Whole exome sequencing genetic data files (BAM files) were provided by Professor Sophie Hambleton at Newcastle University. Targeted gene panel genetic data files (BAM files) were provided by the GOSH Genetics Laboratory, with approval from Dr Lucy Jenkins, Dr Shahnaz Bibi, and Dr Kimberly Gilmour. The gene panel was designed by the team at GOSH Molecular Genetics Laboratory, in particular Dr Suzanne Drury, Dr Shahnaz Bibi, Dr Christopher Boustred and Dr Kimberly Gilmour.

The TIGER gene panel covers the coding regions of 82 genes particularly relevant to severe early-onset PID e.g. severe combined immunodeficiency (SCID) and haemophagocytic lymphohistiocytosis (HLH). The panel has been used at Great Ormond Street Hospitals NHS Trust since 2013 where it began as a 34 gene panel.

2.2.1 Patient cohort

A total of 20 children with undiagnosed immunodeficiency underwent both targeted gene panel sequencing (TIGER panel) and whole genome sequencing via the BRIDGE-PID project. Figure 2.3 shows the phenotype categories represented in the 20-patient cohort. This is representative of a childhood immunodeficiency clinic population [2].

Figure 2.3: Patient phenotypes who had both TIGER gene panel and WGS

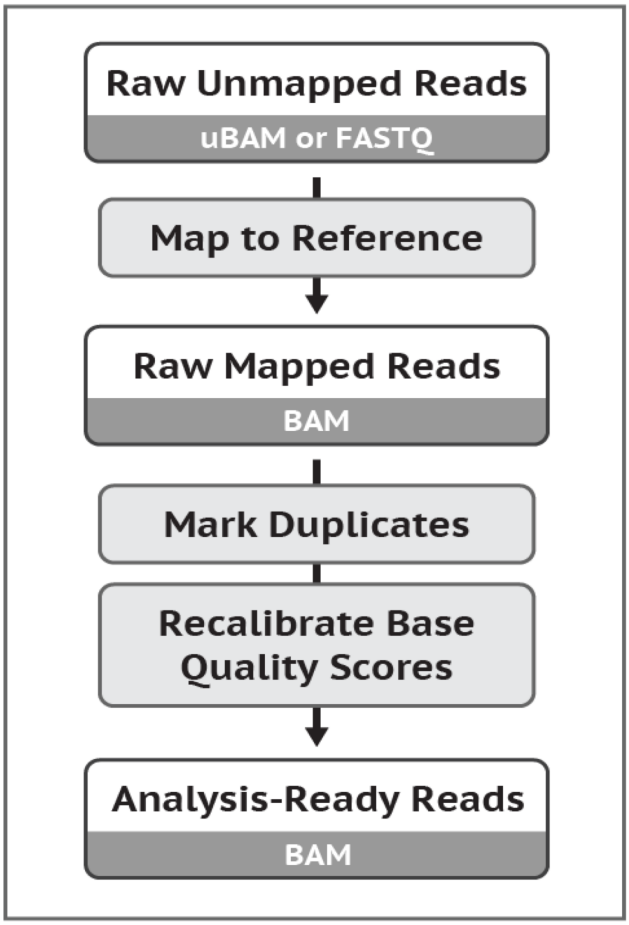


A separate cohort of 20 children with undiagnosed immunodeficiency underwent whole exome sequencing at Newcastle University. Phenotypic data was not collected.

2.2.2 Variant analysis using Genome Analysis Toolkit (GATK)

As variant calling differed between the different sequencing methodology, the Genome Analysis Toolkit (GATK) was used to perform variant calling analysis and assess coverage. The GATK Version 3.8 best practice guidelines were followed where possible [141, 143]. As part of the Best Practice workflow, data pre-processing steps were followed on all raw sequencing data (FASTQ files) (Figure 2.4).

Figure 2.4: Best Practice workflow of sequencing data files (figure from GATK Broad Institute: www.software.broadinstitute.org)



Reads for the 60 files from whole genome and targeted gene panel sequencing were aligned to the hg19/GRCh37 assembly of the human genome reference sequence with Burrows-Wheeler Aligner (BWA) 0.7.10, to produce aligned BAM files. For paired end reads as used in all 3 methodologies, all FASTQ file pairs are entered with the program aligning for each file pair separately and resulting SAM files are created from the merged alignments. BWA aln aligns reads to the reference while allowing for gaps in the alignments. BWA sampe takes both alignment runs and paired data and outputs a SAM file with alignment of paired end data and pair insert sizes. The SAM files are text-based data files which are stored in ASCII format. These were then transferred to readable BAM files. Command line text used is shown below:

bwa aln $reference_seq_file $pair_1.fastq > $name.run_1.sai

bwa aln $reference_seq_file $pair_2.fastq > $name.run_2.sa

bwa sampe $reference_seq_file $name.run_1.sai $name.run_2.sai $pair_1.fastq $pair_2.fastq > aligned.sam

samtools view –hbS $aligned.sam –o $aligned.bwa.bam

Duplicate reads were removed using the MarkDuplicates tool of Picard (Broad Institute v2.0.1). Command text below applied to WGS BAM files and TIGER BAM files:

java -jar $MarkDuplicates.jar

INPUT=$aligned.bwa.bam

OUTPUT=$name.dedup.bam

REMOVE_DUPLICATES=true

Duplicate reads may cause bias from PCR amplification artefacts in the library construction step in sequencing. Whole exome BAM files were provided with duplicates already removed. Remaining duplicates were sought using the same methodology in whole exome BAM files, to ensure that PCR duplicates had been detected, and removed. As shown below, a very small percentage remained (0.000008%) suggesting that PCR duplicates had been removed effectively (Table 2.1).

Table 2.1: Duplicate reads metrics (CI-confidence interval)

	Targeted Panel Sequencing (TIGER)	WGS	WES
Unpaired reads	11136.85 (95% CI: 7451.71 - 14821.99)	6423.85 (95% CI: 5125.78 - 7721.92)	39920.4 (95% CI: 33233.92-46606.88)
Read Pairs	1284261.65 (95% CI: 1106209.06 - 1462314.24)	238283.65 (95% CI: 226203.94 - 250363.36)	23299949.75 (95% CI: C21704629.9-24895269.6)
Unmapped Reads	92543.95 (95% CI: 27608.17 - 157479.73)	4523.25 (95% CI: 3644.09 - 5402.41)	232041.3 (95% CI: 189402.01- 274680.59)
Unpaired Read Duplicates	5451.35 (95% CI: 3432.67 - 7470.03)	1347.3 (95% CI: 956.8 - 1737.8)	2.1 (95% CI: 1.45 - 2.75)
Read Pair Duplicates	33838.7 (95% CI: 26893.95 - 40783.45)	8577.9 (95% CI: 4449.89 - 12705.91)	189.8 (95% CI: 167.08 - 212.52)
% Duplication	0.028 (95% CI: 0.02-0.03)	0.04 (95% CI: 0.02 - 0.06)	0.000008 (95% CI: 0.0000074 – 0.0000862)
Estimated Library Size	48283097.65 (39507485.36 - 57058709.94)	87968753.4 (95% CI: 71892692.71 - 104044814.09)	1518445998503.7 (95% CI: 1.42×10^{12} - 1.61×10^{12})

The remaining reads were realigned around indels using the Genome Analysis Toolkit (GATK) tools RealignerTargetCreator, IndelRealigner and BaseRecalibrator, using GATK best practice guidelines [141, 143]. RealignerTargetCreator detects known indels within the BAM file when given a reference indel file as it is these variants that are most likely incorrectly mismatched and in boundaries, often incorrect SNVs can be called. The standard data set Human 1000G Standard Indels sets was used. IndelRealigner uses insertion-deletions (indels) targets to perform a local re-alignment. The command text used is shown below:

```
java -jar $GenomeAnalysisTK.jar -T RealignerTargetCreator
-R $reference_seq_file
-Input $name.dedup.bam
-known $human1000G_indels_file
-ouput $name.intervals

java -jar $GenomeAnalysisTK.jar -T IndelRealigner
-R $reference_seq_file
-Input $name.dedup.bam
-known $human1000G_indels_file
-targetIntervals $name.intervals
-output $realigned.bam
```

Base quality score recalibration (BQSR) is the process of detecting and subsequently minimising sequencing errors via the quality score assigned to each base. This step is essential to ensure accurate variant calling in later steps. A DNA sequencer will assign a quality score to each base when it is read (Phred score). The BQSR detects patterns of error throughout the sequence e.g. repetitive regions or at the ends of reads. When these areas are encountered, empiric steps are used to reassign quality scores, based on sites of known variation. The human dataset 1000 Genomes dbSNP SNV file was used as a template for pattern detection (Data uploaded from: http://www.internationalgenome.org Accessed March 2016). BQSR has been shown to be effective in improving variant calling by calling rare SNVs at higher variant detection sensitivity, and reducing the false discovery rates [144]. Command text which was used on each individual sequencing BAM files is shown below.

```
java -jar $GenomeAnalysisTK.jar -T BaseRecalibrator
-R $reference_seq_file
-Input $realigned.bam
-knownSites $dbsnp_file
-output $recalibrate.grp

java -jar $GenomeAnalysisTK.jar -T PrintReads
-R $reference_seq_file
-Inout $realigned.bam
-BQSR $recalibrate.grp
-output $recalibrated.bam
```

The recalibrated BAM files generated from whole genome and whole exome samples were then applied to a BED file of coding regions of genes known to cause PID [7]. The coding regions of the genes include 20 bases into the intronic region to detect splice variants. The TIGER gene panel files were applied to a BED file of coding regions of 82 target genes (TIGER genes).

2.2.3 Variant Detection Sensitivity

To determine the ability of the sequencing modalities to detect SNVs and indels, I performed independent assessments of variant detection sensitivity [134, 145]. BAM files were downsampled using Picard DownsampleSam (Broad Institute v2.0.1). This is the process of randomly selecting and retaining a set proportion of reads from the full BAM file. BAM files were downsampled to the following proportions: 0.05, 0.1,

0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 (representing 5-90% of the full BAM file reads). Command text shown below:

```
java -jar $DownsampleSam.jar
INPUT=$recalibrated.bam
PROBABILITY=$0.05-0.9 R=null
OUTPUT=$downsample_probability.bam
```

Duplicates were again detected and removed as previously described.

```
java -jar $MarkDuplicates.jar
INPUT=$downsample_probability.bam
OUTPUT=$downsample.probability.dedup.bam
REMOVE_DUPLICATES=true
```

Genotyping microarray experiments (also called SNP array) are carried out alongside next generation sequencing, providing an additional independent measure of call set quality. To test variant detection of WGS and WES in patients with PID, all SNVs and indels were called from within the coding regions of known genes implicated in PID +/- 20 bases. These were contrasted with the number of SNVs found by genotyping SNP array, that was run concurrently.

Diagnostic performance of targeted gene panel was tested by calling SNVs and indels within the coding regions of the 82 genes on the panel. GATK HaplotypeCaller was used to call variants from BAM files from gene panel, WGS and WES using default settings [141].

The sensitivity of WGS and TIGER gene panel variant detection was defined as the ability to correctly detect SNVs present in the SNV array genotyping file (Figure 2.5).

Figure 2.5: Variant sensitivity and specificity matrix

	SNV ARRAY GENOTYPING	
WGS/TIGER gene panel results	**Variant present**	**No variant**
Variant detected	A True positive	B False positive
Variant not detected	C False negative	D True negative
TOTAL	A+C	B + D

Variant detection sensitivity = A/(A+C)

Variant detection specificity = D/(B+D)

Variant detection sensitivity was calculated in each downsampled sequencing file. Read depths of variants in downsampled files were obtained concurrently. Heterozygous and homozygous SNVs and indels were considered separately.

2.2.4 Depth of coverage metrics

Read depth was calculated using GATK Depth of Coverage tools with default settings. Coverage metrics were sought across the targeted PID gene regions as well as known

PID disease variants. Disease variants within the respective gene regions were sought from the ClinVar database server; an archive of medically relevant disease variants and phenotypes (2017 ed: https://www.ncbi.nlm.nih.gov/clinvar).

2.2.5 Predicting the impact of splice donor variant

To evaluate the impact of splice-disrupting variants in WGS data, in-silico prediction tools were used to determine whether a genetic variant found in a 5' or 3' splicing consensus region, is likely to disrupt the exon-intron boundaries of the protein and affect RNA splicing.

I mainly assessed splicing using Alamut visual splicing tools (2017) which applies multiple predictors including ESE, NNSplice and MaxENTscan. MaxEntScan provides a score as a numerical measure of the strength of the splicing signal, and has been shown to have highest accuracy at predicting the effects of genetic variants at the 5' splice sites [146]. I generated scores according to MaxEnt for both the wild-type (WT) and mutant sequences, using the method described in Houdayer et al (2013). The greater the difference in scores, the greater the likelihood of a splicing defect.

2.3 Functional analysis of genetic variants

2.3.1 Cell separation techniques

Isolation of peripheral blood mononuclear cells (PBMC) was undertaken by density gradient centrifugation from heparinised peripheral blood. Whole blood was diluted 1:1 with Phosphate Buffered Saline (PBS) and then layered equal quantity of Ficoll-Paque (GE Healthcare). This solution was then centrifuged at 2300rpm for 30 minutes without brake in a tabletop centrifuge. PBMCs were aspirated from the serum interphase using a Pasteur pipette and washed with equal quantity of PBS.

2.3.2 Western blot

Patient and healthy control samples were collected on the same day and handled identically. Between $5x10^5$ and $1x10^6$ primary cells were pelleted at 1600prm for 5 minutes in a tabletop centrifuge, resuspended in 100ul Laemmlli buffer (contents) with 1ul protease inhibitor and incubated on ice for 20 minutes. Samples were then centrifuged at 16000rpm for 20 minutes and supernatant was transferred to fresh 1.5ml Eppendorf microfuge tubes.

When ready for use, samples were heated to 95C for 5 minutes. 10-15ul of the sample was loaded onto BioRad Mini-PROTEAN TGX precast 4-20% polyacrylamide gel in buffer. 5ul of PageRuler Plus prestained protein ladder (Thermo Scientific) was loaded into outer wells. The gel was electrophoresed in a BioRad Mini-Protean Tetra Cell at 200V for 15-20 minutes. The gel was trimmed, before semi-dry transfer using BioRad Trans-Blot Turbo System for 10 minutes. The nitrocellulose membrane was incubated for 2 hours at room temperature in blocking buffer (0.1% PBS/Tween, 5% BSA) and washed 3 times in 0.05% PBS/Tween. The primary western blot antibody (Cell Signalling Technology) was diluted 1:1000 with blocking buffer and incubated overnight at room temperature. After 3 washes with 0.05% PBS/Tween, the

membrane was incubated for 1 hour at room temperature with an appropriate secondary anti-HRP antibody diluted to 1:8000 with blocking buffer. The membrane was asked a final 3 times and 1 ml Clarity Western Luminol electrochemical luminesce (ECL) reagent was poured onto the membrane and left for 5 minutes at room temperature. Protein bands were visualised with a BioRad ChemiDoc MP imaging system on UV chemi-luminescent settings.

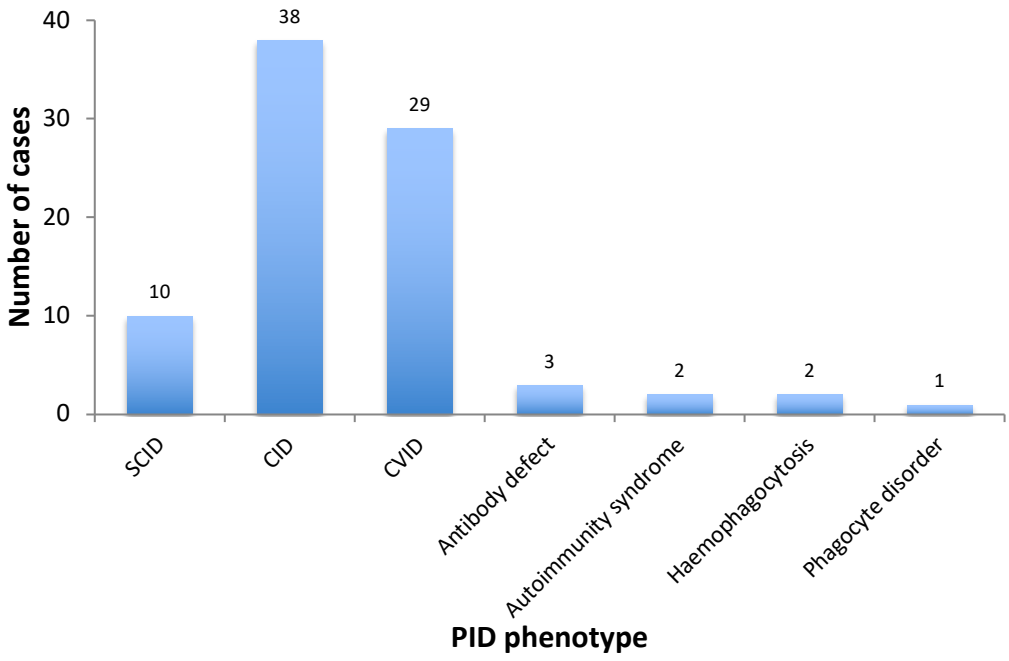
After probing membranes, they were stripped for 10 minutes in ReBlot Plus mild solution (Millipore), blocked for 30 minutes using blocking buffer and re-probed with anti-Tubulin or Vinculin (Cell signalling Technology) at a dilution of 1:1000 for 2 hours at room temperature. The membrane was washed three times with blocking buffer and probed with an anti-mouse HRP secondary antibody at dilution of 1:10,000.

2.4 Novel genetic analysis in PID

2.4.1 Phenotypes of cohort

77 children were recruited to the BRIDGE-PID from Great Ormond Street Hospital NHS Foundation Trust. They all had genetically undiagnosed PID. 19 were subsequently discovered to have a genetic diagnosis by WGS but were included in the association analysis. 9 of the probands were related to at least one other in the cohort (all siblings) and kinship was factored in to the analysis. Phenotype categories of the PID cohort are summarised below (Figure 2.6).

Figure 2.6: PID categories of patient cohort (n= 77 patients)



Controls

320 control subjects were also recruited to the BRIDGE study and had whole genome sequencing done with the same sequencing methods interspersed with other sequencing plates. These were healthy adult individuals, recruited to the BRIDGE study as part of an ethically approved protocol. These were accessed, with kind permission via the high-performance computing domain of BRIDGE study, University of Cambridge. Data quality control for these subjects was conducted by Dr Chris Penkett at University of Cambridge.

2.4.2 Variant annotation

After conducting variant calling and quality control measures, I annotated the sequencing data against several on-line resources, including dbSNP v150, and allele frequencies from Exome Aggregation Consortium (ExAC) [147-149].

Functional annotations were added using Ensembl Variant Effect Predictor (VEP) version 82. Rare variants were defined as those with mean allele frequency (MAF) <1% in 1000G database or absent from the population database. VEP defined non-synonymous variants as those with high or moderate impact. These were transcript ablation, splice acceptor, splice donor, stop gained, frameshift, stop lost, start lost, transcript amplification, inframe insertion, inframe deletion, missense, regulatory region ablation and protein altering variants. Variants within a splice region are classified as thus, rather than synonymous variants or missense variants. This allows prioritisaiton of splice variants which have may have a detrimental effect on the resulting protein.

In-silico tools such as Sorting Intolerant from Tolerant (SIFT) and Polyphen2 were used to predict deleteriousness in non-synonymous missense SNVs [150, 151].

Genomic Evolutionary Rate Profiling (GERP) data was used to assess whether variants affected conserved amino acid sites [152].

2.4.3 Pathway analysis – database and literature search

The intention of this analysis was to find potential disease-causing variants at biologically-meaningful loci. Analysis of the dataset with respect to known disease genes was carried out in Chapter 5. To define the candidate gene list, I collected data for a variety of genetically and biologically relevant sources. Genes known to interact with known PID genes were sought in PID pathways (http://www5.appliedbiosystems.com/tools/pathway), KEGG and GO pathways. Further related genes were obtained from the UCSC genome browser transcriptome database (RNA-seq on UCSC genome browser track search), and the Protein-Protein interaction database (http://string-db.org). The latter analysis uses several databases that curate experimentally validated or algorithmically-predicted protein-protein interactions. In all searches, data was restricted to Homo sapiens data where possible, and moderate associations were sought. Once known PID genes were excluded, a total of 3082 PID-related genes were found (see Appendix B).

WGS data of unsolved cases were analysed for the presence of high impact novel variants. Rare and functional variants were also evaluated for presence in healthy individuals and on genotype consistent with the likely mode of inheritance. The most damaging variants, as predicted by *in-silico* protein models, affecting conserved sequences were prioritised for MDT discussion and analysis.

2.5 Association studies in PID

2.5.1 Quality control

Due to the small sample size of the cohort studied, it is deemed that there is insufficient statistical power to test for single variant association across the genome.

Standard GWAS quality control was followed where possible according to guidelines set out in Anderson et al. (2010) [153]. For rare variant association analysis, the software PLINK version 1.07 was used to carry out the following quality control steps.

Common SNVs of MAF >1% were filtered out, using HapMap v3. If SNVs were followed up, the MAF was checked in 1000G (dbSNP) and ExAC population databases. Missing SNVs throughout the case and control set were reviewed and removed if missingness rate of >0.2. Sex of genome samples were imputed based on the genotype information. Heterozygosity was checked, and subjects with heterozygosity deviating more than 3sd from the mean were removed. To check for relatedness, case and control relationships were checked (PLINK pihat >0.2). Siblings in the case set were removed if one was already represented; the sibling removed was the one found to have the lowest call rate.

To identify novel association implicated in PID, I conducted a case-control analysis for an enrichment of variants in individual genes in PID cases versus controls. Phenotype categories were labelled as binary, rather than continuous. Genetic variants were initially filtered on allele frequency (MAF <0.01) and all variants including those in untranslated regions (UTRs), and essential splice sites, and were given equal weight of being causal. The BURDEN test in PLINK v1.07 was used in SMP mode to correct for population stratification. This allows for co-variation and allows for variants to be grouped per gene. PID cases with at least one variant per

gene were counted and tested for association with disease using a one-tailed Fisher's exact test. To test the significance of each gene in the association study, permutation testing was utilised. This involves repeatedly sampling n number of times with random reassigning case-control labels each time, to sample under the null hypothesis, with permutation of n=1000 equating to an *p*-value of 0.001.

2.5.2 Rare variant aggregation-based analysis using SKAT-O

Rare variant aggregation analyses use multiple regression methods, that do not assume that all tested rare variants act in the same direction, including the C-alpha test and the Sequence Kernel Association Test (SKAT) [154, 155]. Sequence Kernel Association Test -optimised (SKAT-O) was developed shortly after, as the best linear combination of burden and SKAT tests, to maximise power for association, with SKAT-O statistics generally found to be more significant than SKAT [156]. SKAT-O is a variance-component multiple regression test which retains power in settings where neutral variants, or variants in opposing directions could result in loss of power. SKAT-O was applied using the R 'SKAT' package with weight applied to variants less than MAF<0.001 *(Alpha (0.001), Beta(1,20)*, covariates applied for the first 10 principal components to account for population stratification, and all other conditions kept the same as burden test. Coding variants were given equal weight of being as causal as each other, regardless of variant consequence.

Testing SNVs for Hardy-Weinberg equilibrium (HWE) was not carried out for this analysis, particularly given that rare causal variants are more likely to be in HW disequilibrium.

3 ANALYSIS OF NEXT GENERATION SEQUENCING IN DIAGNOSIS OF PID

In this chapter, I present an analysis of genetic sequencing data of a phenotypically heterogenous cohort of children with PID. The studied individuals comprise of cases where no prior causative mutations had been identified. The aim of the study was to analyse the PID diagnostic yield and efficacy in 20 patients who had targeted gene panel sequencing (TIGER gene panel), whole genome sequencing (WGS) and a genotyping SNP array. Whole exome sequencing was also analysed in a separate, unrelated cohort of children with PID where the genetic diagnosis was not previously known. I will analyse variant detection sensitivity and specificity between the different sequencing methods and coverage metrics in targeted gene regions, and known PID genes [7] (Appendix A).

3.1 Introduction

3.1.1 Targeted gene panel studies in primary immunodeficiencies

Targeted gene panels are the main stay of genetic diagnostics in PID. Targeted gene panels sequence specified regions of selected genes, typically at a high depth of coverage. The TIGER panel is utilised in PID diagnostics due to its considerable advantages, particularly for a clinical service: fewer variants generated, short turnaround time and relatively low costs. The diagnostic yield for targeted gene panels varies according to the PID population entered, but can be as high as 70% for a selective severe early-onset PID cohort [157]. However, a considerable disadvantage of targeted gene panel sequencing is that as new genes associated with PID are discovered, panels require updating to remain current. Moreover, the question arises of whether it is then incumbent on the diagnosticians to re-sequence patient DNA through the most recent iteration of the gene panel, particularly where analysis of the previous iteration has not yielded a diagnosis.

For identified variants, robust validation must take place prior to reporting. A major challenge lies in the interpretation of the large quantity of genetic data generated. The vast majority of variants identified in an individual are benign, and over interpretation of a given variant could results in potentially damaging clinical consequences. Currently potentially pathogenic variants are prioritised during variant analysis and brought forward to a clinical multi-disciplinary meeting.

3.1.2 Variant calling

Variant calling is the process of identifying differences from the reference genome within read data. The accuracy of variant calling is dependent on sequencing library and DNA preparation, sequencing platform, alignment and variant calling analysis software. Errors in variant calling fall into two main categories: false positive variants

where an identified variant is not found in the true genomic sequence, and false negative calls where a true variant is not identified through the sequencing process. False positive variant calling is greater in regions of low coverage [158]. For example, where a reference base is covered by only two sequencing reads and one is an alternative allele, a heterozygous variant can be called. However, if the read depth increases with more reference match bases, one variant read is less likely to be called as a variant as the variant read forms a lower proportion of the total. Similarly, false negative errors occur when not enough reads contain a real variant to lead to a correct variant call. This particularly afflicts WES where reference bias can occur as reads are enriched for those which match the reference probes [159, 160]. In PID, the exclusion of a molecular diagnosis is often as important as the identification of a genetic cause. Therefore, there is a requirement for uniformity of sequencing and depth of coverage sufficient to exclude false negative results.

3.1.3 Variant coverage

In NGS, the overall aim is to compare sequenced DNA bases to reference bases. A DNA sequencer has a raw per-base sequencing error rates of 0.5% but by sequencing the base multiple times, the error rate reduces proportionally [161]. Optimal DNA sequencing would entail all reference bases in the genome to be sequenced many times over. Where the assembly size is small e.g. coding region of a gene, it is preferred to map as many reads as possible to the targeted region. However, where the targeted region is very large e.g. WGS, this is very unwieldy and expensive to do, and so a compromise must occur between the breadth of assembly size and the depth of mapped reads. The latter is interchangeably referred to the depth of coverage or read depth and is defined as number of mapped reads aligned at each reference base.

A DNA sequencer have raw per-base sequencing error rates of 0.5% but by sequencing the base multiple times, the error rate reduces proportionally [161]. Current criteria for the quality of a genetic variant set out by the American College of Medical Genetics and Genomics (ACMG) are not definitive, as different sequencing methods and chemistries are used in clinical diagnostics. Many studies have attempted to determine the amount of sequence coverage sufficient to reliably identify single nucleotide variants (SNVs) [145, 162-164] (Table 3.1). However, in each case different sequencing chemistries and variant callers are used which make inter-comparison difficult. The overall lowering of sequence coverage threshold over time, reflect improving technologies. One of the latest studies demonstrate reliable average coverage of >20x in 95% of exons [164].

Table 3.1: Summary of sequencing studies.

			Average coverage of 95% targets		95% SNV detection sensitivity	
Study	**Sequencer**	**Read length**	**WGS**	**WES**	**WGS**	**WES**
Bentley et al (2008)	Illumina HumanHap550 BeadChip	Paired 35 bp read	>33x			
Ajay et al (2011)	Illumina Hi-Seq 2000	Paired 100 bp reads	>50x			
Clark et al (2011)	Illumina Hi-Seq 2000	Paired 101 bp reads		>30x		
Meynert et al (2014)	Illumina Hi-Seq 2000	Paired 98 bp reads	>18x	>34x	>14x	>40x
Lelieveld et al (2015)	Illumina Hi-Seq 2000	Paired 35-101 bp (WGS), 101 bp reads (WES)	>20x	>20x		

3.1.4 Variant sensitivity

Analysis of sensitivity and specificity are often applied to clinical diagnostic tests to measure the test's clinical validity, particularly when measured against the current 'gold standard' testing. These measures inform the accuracy with which the genetic test identifies the patients' clinical phenotype. In relation to sequencing studies, the variant detection sensitivity of the test is the proportion of those with the genetic variant, that have a positive test result. Loss of sensitivity leads to causative genetic variants being undetected by the test. This is mostly due to low read depth of the variant or inadequate breadth of coverage of the targeted region.

The specificity of the test is defined as the proportion of who have a negative rest results and do not have the condition. In genetic tests, the variant detection specificity is the ability to the test to correctly identify true negatives, and is low when false positives are detected. This tends to be very high in next generation sequencing as reads are matched to long segments of reference bases with a high degree of accuracy, and mismatched reads are rejected.

Meynert et al (2013) devised a method of downsampling sequencing files to identify a mean on-target read depth to accurately identify SNVs. By correlating the read depth with the proportion of correctly identified SNVs, they were able to show that with increasing depth of coverage, the variant detection sensitivity also increases. I aim to utilise this method here to define read depth criteria for NGS tests for the genetic diagnosis of PID.

3.2 Results

3.2.1 Variant detection analysis of gene panel and WGS compared to SNP array

For the 20 patients who had both whole genome sequencing and targeted gene panel (TIGER), targeted SNP array genotyping was also carried out. SNVs in SNP array genotyping files were sought within TIGER gene regions for the 20 patients, and the detection of these SNVs were sought in whole genome files and TIGER gene panel sequencing files, as well as in their respective downsampled files.

All genotyped SNVs bases within the respective genes were detected in the TIGER gene panel files and WGS with acceptable average read depths. (Table 3.2)

Table 3.2: Number of genotyped SNVs in different gene regions

	Genotyped SNVs in genome	Genotyped SNVs located within TIGER gene regions	Genotyped SNVs located within IUIS gene regions
Number of SNVs	806,601	1448	5025

All genotyped bases within the respective genes, were detected in the TIGER gene panel files and WGS with acceptable average read depths (Table 3.3).

Table 3.3: Genotyped bases detectable in different sequencing files

	Number of SNVs	Mean read depth (95% CI)
SNVs present within TIGER gene regions using TIGER panel	1448	220.11 (156.32 – 392.2)
SNVs present within TIGER gene regions in WGS	1448	27.64 (21.3 – 35.45)
SNVs present within IUIS gene regions in WGS	5025	25.61 (15.3 – 29.45)

The sensitivity of whole genome sequencing and TIGER gene panel sequencing were sought to detect the same SNVs present in genotyping SNP array, and in the same zygosity, using downsampled files. The sensitivity of WGS and TIGER gene panel variant detection was defined as the ability to correctly detect SNVs present in the SNV array genotyping file.

The variant detection sensitivity of each downsampled file was plotted against the corresponding mean read depth of variants, with best-fit curves (Figure 3.1). The 2rd order polynomial equation used to fit the data is shown below:

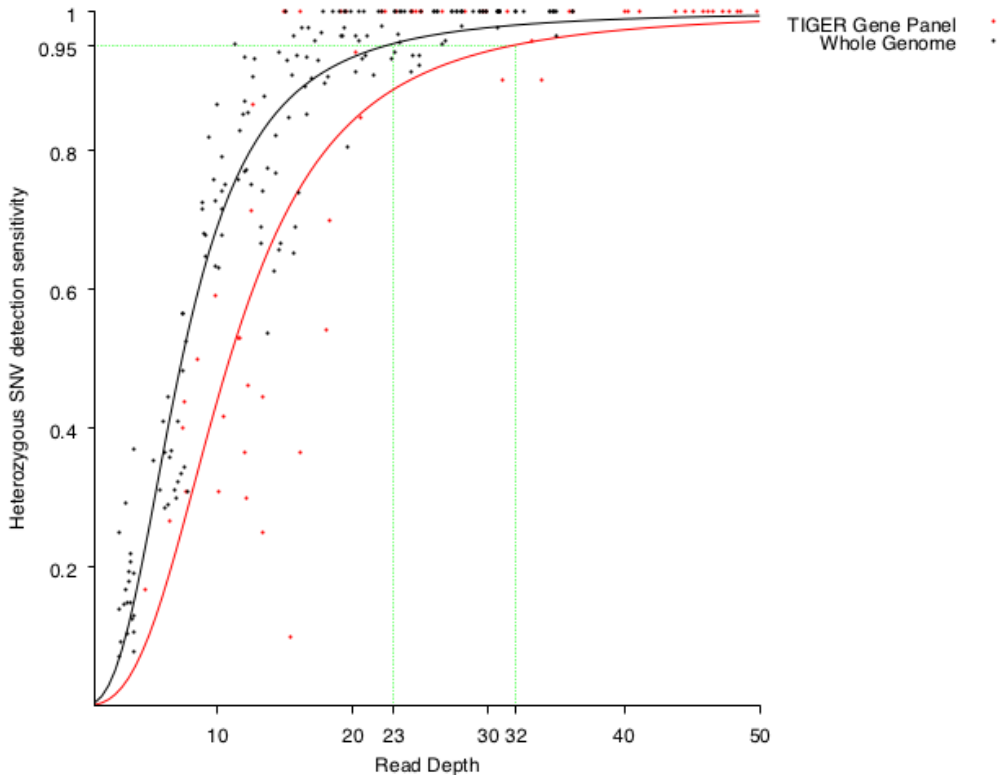
$f(x) = x^n/(b + x^n)$

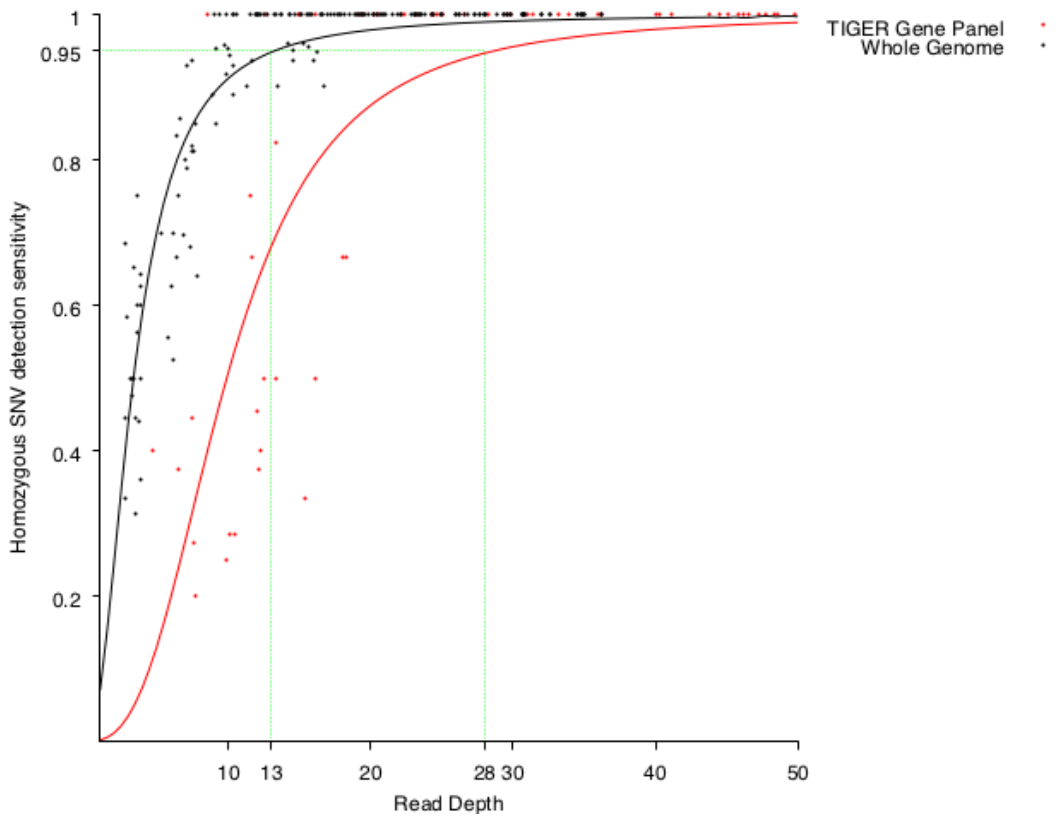
with constants: b = 20, n = 0.95

The graph below shows that at 95% variant detection sensitivity for a heterozygous SNV, the depth of coverage required is 32 for TIGER for the gene panel and 23 for the whole genome sequencing data in the TIGER genes. Figure 3.1 shows that at 95% variant detection sensitivity for a homozygous SNV, the depth of coverage required

is 28 for TIGER for the gene panel and 13 for the whole genome sequencing data in the TIGER genes (Figure 3.1).

Figure 3.1: Heterozygous and homozygous SNV detection sensitivity of each downsampled file (dot) and best-fit curves (line) of the data. Red lines indicate TIGER gene panel files and black indicate WGS data. Green dotted lines show the read depths required at 95% variant sensitivity

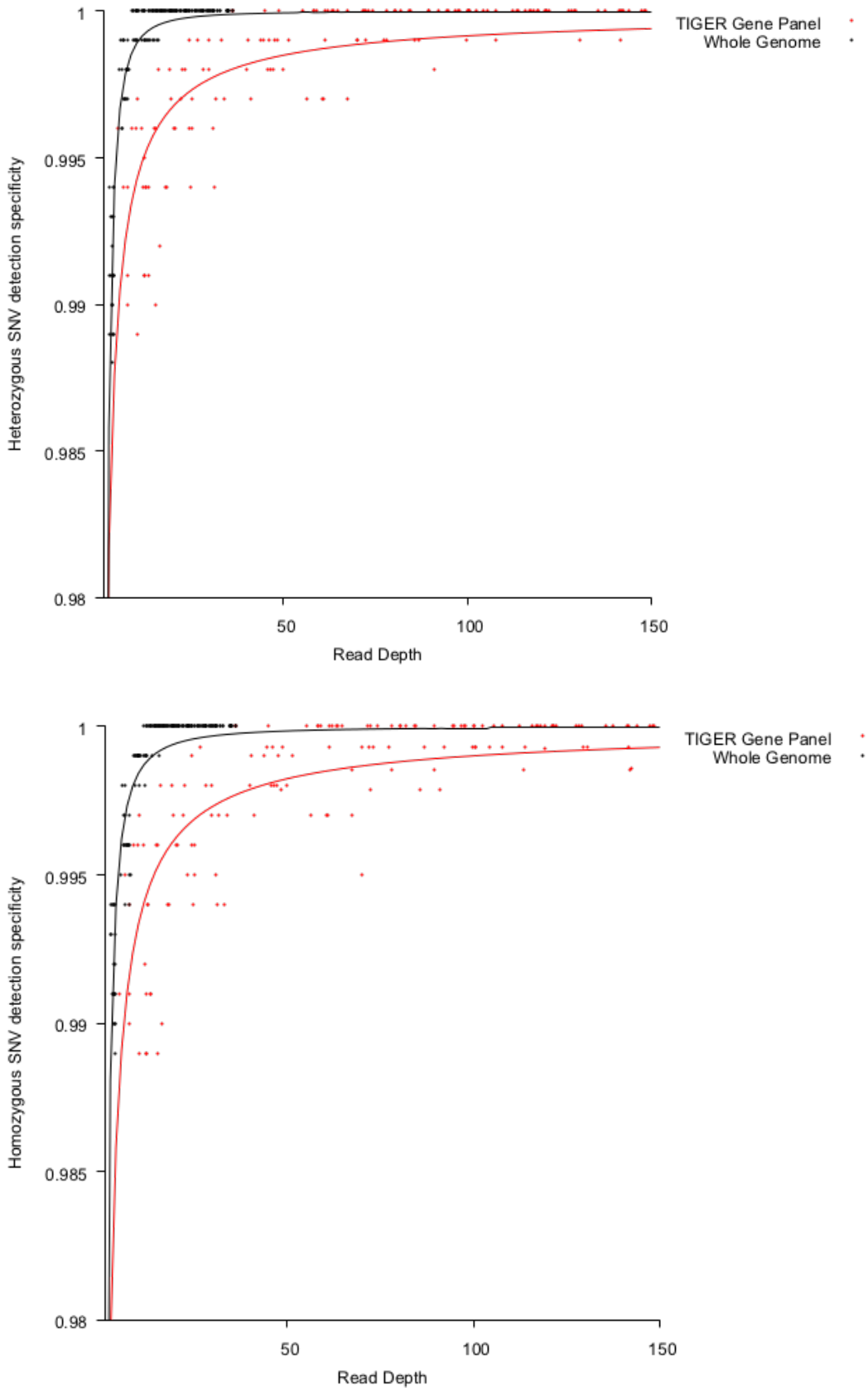




3.2.2 Variant detection specificity

Variant detection specificity was plotted in a similar method used above. The variant detection specificity was found to be high in WGS and TIGER gene panel downsampled files (minimum 98.5% for both for WGS and TIGER) for heterozygous and homozygous SNVs (Figure 3.2).

Figure 3.2: Heterozygous and homozygous SNV detection specificity of each downsampled file (dot) and best-fit curves (line) of the data. Red lines indicate TIGER gene panel files and black indicate WGS data



3.2.3 Variant detection sensitivity within PID gene regions

To assess the performance of whole exome sequencing and whole genome sequencing, variant detection sensitivity and coverage metrics were sought in coding regions of known PID genes.

As genotyping data is not available for whole exome sequencing, variants called from full BAM file were used as a defacto for the genotyping data. This assumption is made as the full representation contained all variants that were present on the genotyping SNP array were present in both targeted panel sequencing and whole genome sequencing. Given that the methodologies are similar in whole exome, we have no reason to assume that this would not be the same for whole exome sequencing.

For patients who had either WGS or WES, variants located within IUIS coding gene regions were sought in the full BAM files. The results are shown in Table 3.4.

Table 3.4: Number of variants found in WGS and WES BAM files within IUIS coding gene regions

	WGS (n=20)		WES (n=20)	
	Total number of variants	**Average number of variants (95% CI)**	**Total number of variants**	**Average number of variants**
	Total	Average	Total	WES
Heterozygous SNVs	6913	345.7 (309.9 – 381.4)	6916	345.8 (317.4 – 374.2)
Homozygous SNVs	3754	187.7 (172.3 – 203)	3982	199.1 (181.2 – 217)
Heterozygous Indels	646	32.3 (27.7 -36.9)	552	27.6 (23.4 – 31.9)
Homozygous Indels	297	14.9 (11.2 – 18.6)	227	11.35 (9.7 -13)

Figure 3.3: Variant detection sensitivity for heterozygous and homozygous SNVs within PID gene regions of each downsampled file (dot) and best-fit curves (line) of the data. Blue lines indicate WES data and black indicate WGS data.

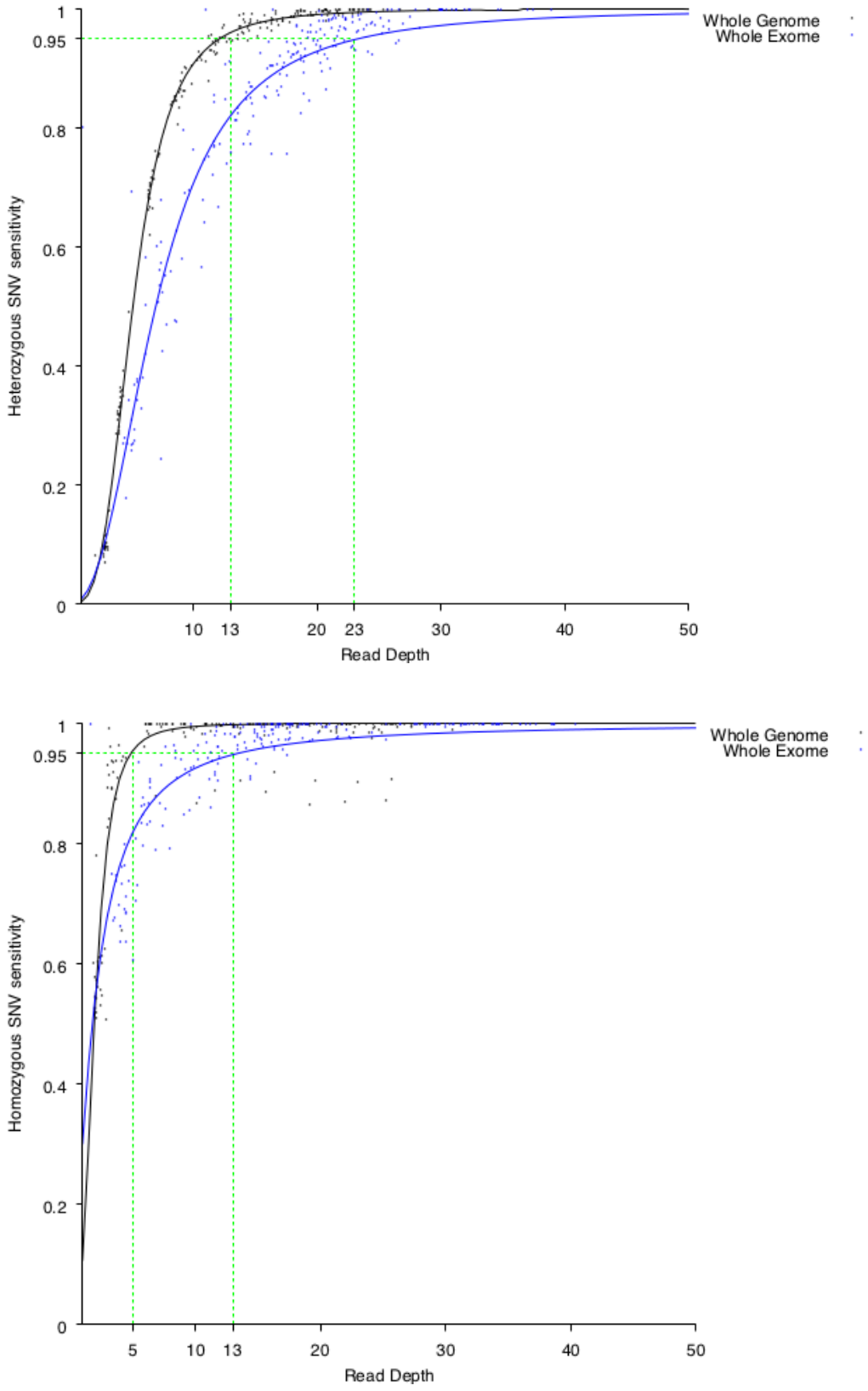


Figure 3.4: Variant detection sensitivity for heterozygous and homozygous Indels within PID gene regions of each downsampled file (dot) and best-fit curves (line) of the data. Blue lines indicate WES data and black indicate WGS data.

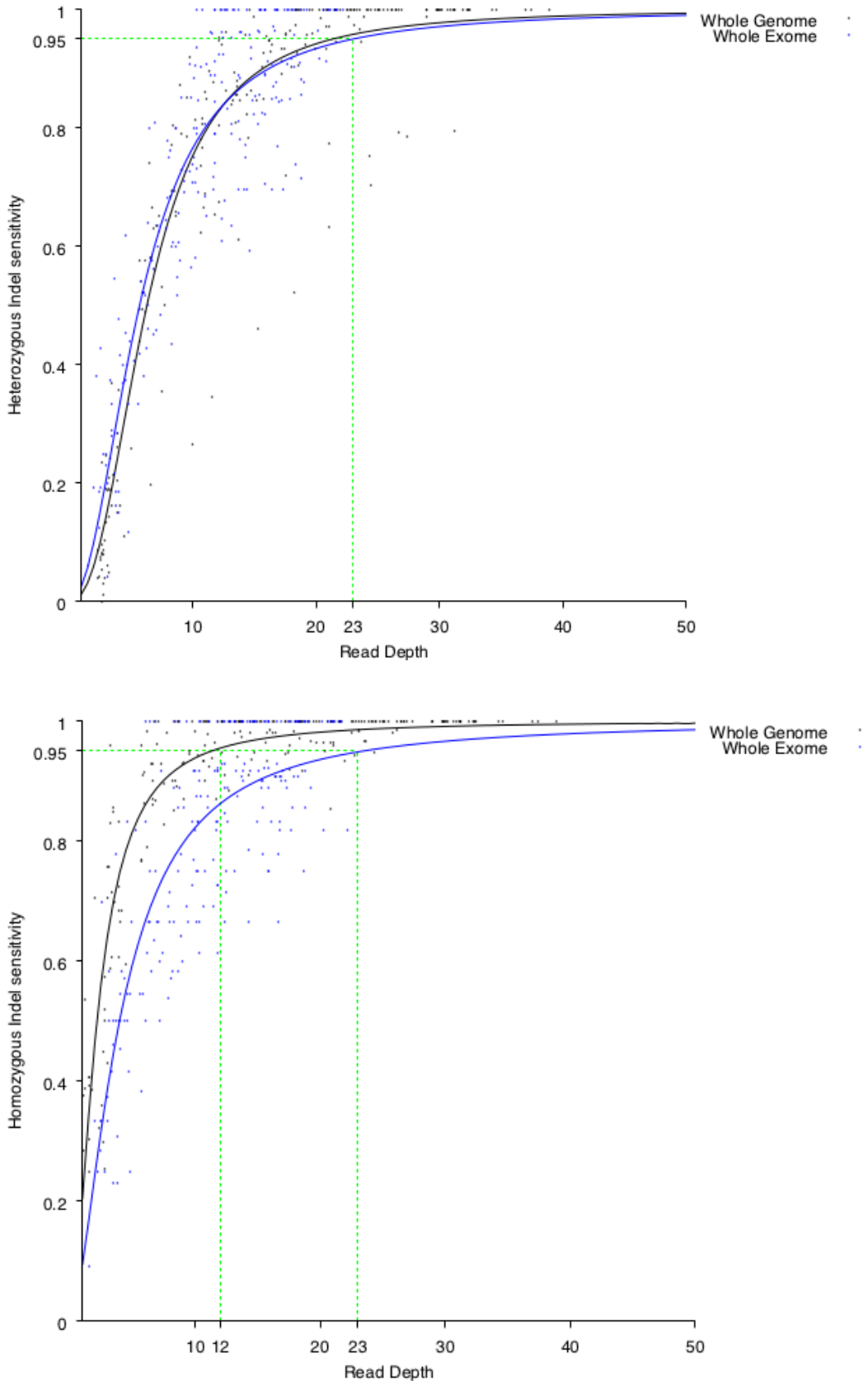


Table 3.5: Estimated read depth to obtain 95% variant sensitivity in PID genes

	Read Depth for 95% variant sensitivity	
	WGS	WES
Het SNV	13	23
Hom SNV	5	13
Het Indel	23	23
Hom Indel	12	23

3.2.4 Depth of coverage

Table 3.6 summarises the coverage metrics of the sequencing methods over known PID genes and the targeted genes. On average, whole genome sequencing had much lower average target coverage of 33.9 reads compared to gene panel sequencing (224.3x) and whole exome sequencing (68.9x). To reach average coverage of 95% of the targeted region, a minimum read depth of 45x was required for TIGER gene panel sequencing, 20x for WGS and 38x for WES.

Table 3.6: Coverage metrics in PID genes

Sequencing method	Average read depth	%>5	%>10	%>20	%>30	%>40	%>50
Gene panel	224.3	99.2	99.1	98.8	98.4	97.9	97.1
WGS	33.9	99.9	99.6	95.2	69.6	25.5	4.0
WES	68.9	89.2	86.1	78.6	70.4	62.4	55.1

3.2.5 Low coverage regions

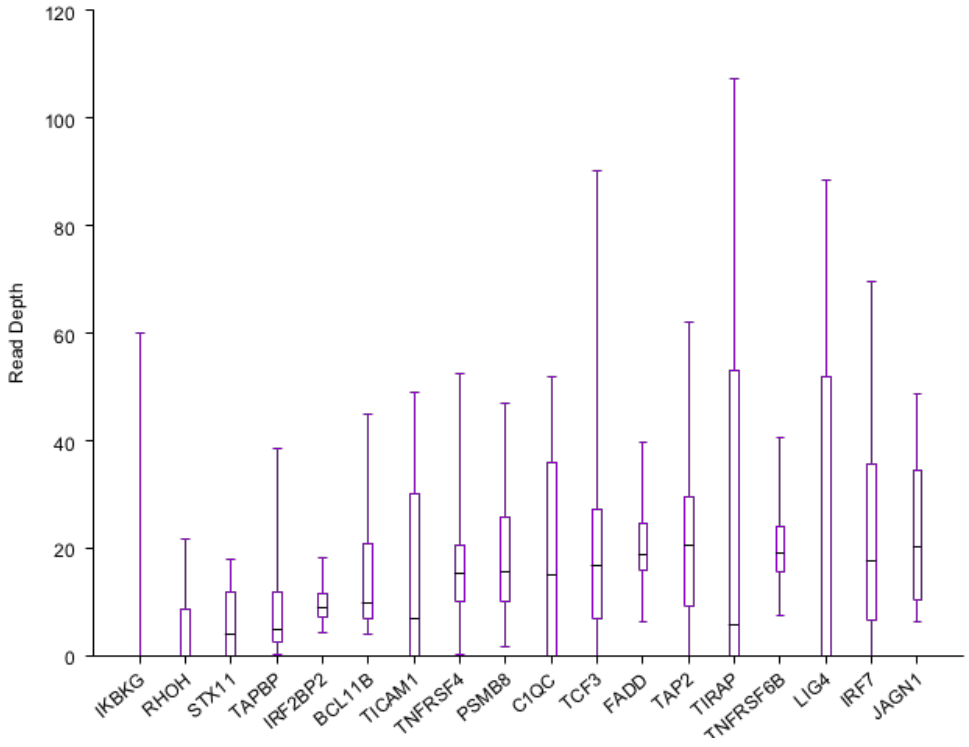
Given the significant discrepancies in depth of coverage between the different sequencing methods studied, I applied the aforementioned 95% variant detection sensitivities.

Table 3.7: Projected coverage at 95% variant detection sensitivity

Sequencing method	Read depth of 95% heterozygous SNV detection sensitivity	Percentage of coding exons covered at 95% variant detection sensitivity (GP – TIGER genes, WGS/WES – IUIS genes)	Percentage of Disease Variants (CLINVAR) detectable at 95% variant detection sensitivity (GP – TIGER genes, WGS/WES – IUIS genes)
Gene Panel	32	99.1% (1153/1164)	99.28% (3171/3194)
WGS	13	99.9% (3998/4001)	99.95% (16443/16452)
WES	23	82.9% (3316/4001)	76.1% (12526/16452)

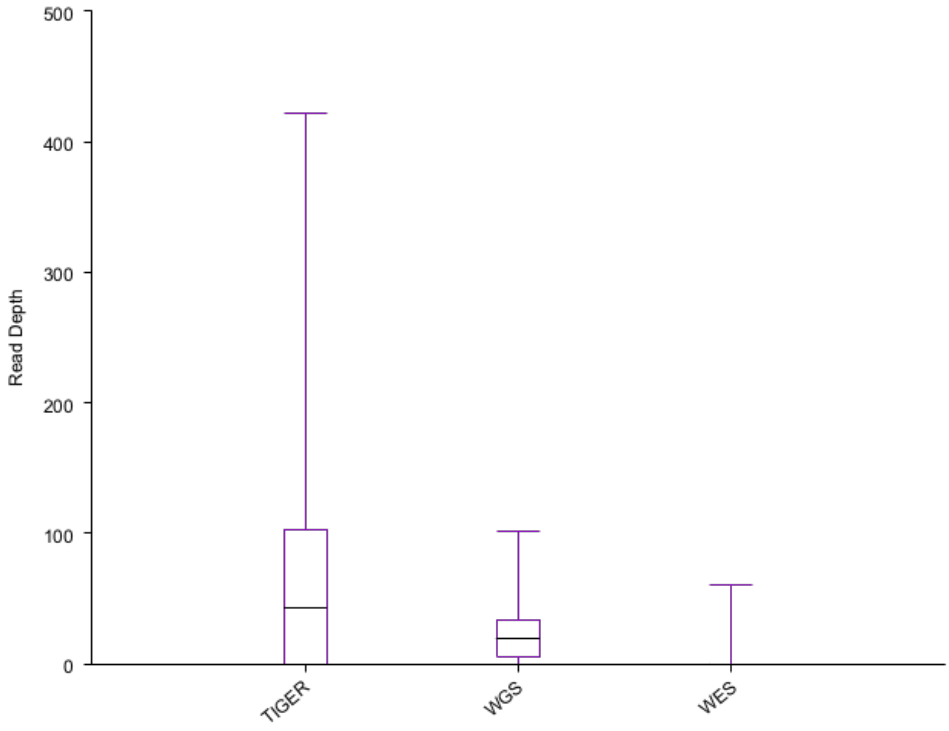
Particular PID genes were noted to have mean low coverage in WES samples (Figure 3.5).

Figure 3.5: Low coverage PID genes in WES with mean low coverage <23 with range depicted as box and whisker plot, (box shows 1st quartile to 3rd quartile and whiskers from minimum to maximum)



IKBKG was noted to be have low mean coverage across all 3 sequencing methods (Figure 3.6).

Figure 3.6: Range of read depths in IKBKG across samples using TIGER gene panel, WGS and WES with range depicted as box and whisker plot, (box shows 1st quartile to 3rd quartile and whiskers from minimum to maximum)



It is important for a diagnostic test that known pathogenic variants are detectable through sequencing methods. To test this, I analysed coverage of known pathogenic variants from the ClinVar database. Known disease variants such as those in the ClinVar database are very informative in variant analysis. I analysed ClinVar variants that were found on average to be below the 95% variant detection sensitivity for the sequencing method, and so have the potential to be disregarded, due to low read depth (TIGER gene panel Table 3.9, WGS Table 3.10). TIGER gene panel had the potential to miss 23 of 3194 ClinVar variants (0.72%) and WGS missing 9 of 16452 ClinVar variants (0.05%). WES had the potential to miss 3925 of 16452 (23.9%), as 23.9% of ClinVar variants were represented at a depth of <23 in WES samples. Some variants in the ClinVar database are benign, or latterly judged to be benign, as is the case with some pathogenic variants. All have been included for the purposes of this analysis, as clinical interpretation of these variants falls outwith the remit of this analysis.

Table 3.8: ClinVar variants not covered by TIGER panel at 95% variant detection sensitivity

Gene	Variant position	Variant	Disease category	Average coverage
CASP8	chr2:202134321	c.394T>C p.Leu132= (NM_001228)	Uncertain signficance	0
TAP2	chr6:32790089	NM_018833.2(TAP2):c.1939C>T (p.Leu647Phe)	Benign	0
TAP2	chr6:32790099-32790100	NM_018833.2(TAP2):c.1933-5dupT	Benign	0
RET	chr10:43572705	NM_020975.4(RET):c.-2C>A	Uncertain signficance	17.4
RET	chr10:43572724	NM_020975.5(RET):c.18C>T (p.Ser6=)	Likely benign	16.9
RET	chr10:43572737	NM_020975.5(RET):c.31C>A (p.Leu11Met)	Uncertain significance	17.25
RET	chr10:43572749-43572765	NM_020975.4(RET):c.43_44insTGCTGCTGC (p.Leu19_Pro20insLeuLeuLeu)	Uncertain significance	16.6935
HPS6	chr10:103825454	NM_024747.5(HPS6):c.223C>T (p.Gln75Ter)	Pathogenic	30.15
HPS6	chr10:103825467-103825468	NM_024747.5(HPS6):c.238dupG (p.Asp80Glyfs)	Pathogenic	29.25
HPS6	chr10:103825485	NM_024747.5(HPS6):c.254C>T (p.Pro85Leu)	Uncertain significance	28
HPS6	chr10:103825529	NM_024747.5(HPS6):c.298C>T (p.Leu100=)	Uncertain significance	27.7
HPS6	chr10:103825568	NM_024747.5(HPS6):c.337C>T (p.Arg113Trp)	Uncertain significance	31.9

SLC37A4	chr11:118896407	NM_001164278.1:c.1045C>T	Uncertain significance	2.9
ORAI1	chr12:122064648	NM_032790.3(ORAI1):c.1A>T (p.Met1Leu)	Uncertain significance	19.1
ORAI1	chr12:122064773-122064779	c.132_137=, p.Pro46_Pro47del	Likely benign	11.413
CYBA	chr16:88709869	NM_000101.3(CYBA):c.480G>A (p.Pro160=)	Benign	31.05
CYBA	chr16:88709915-88709963	c.386_433dup, (p.Ile144_Gly145insGluGlnTrpThrProIleGluProLysProArgGluArgProGlnIle)	Uncertain significance	17.774
CYBA	chr16:88709968	NM_000101.3(CYBA):c.381T>C (p.Arg127=)	Benign	16.1
CYBA	chr16:88709984	NM_000101.3(CYBA):c.370-5C>T	Uncertain significance	11.35
WAS	chrX:48547170-48547171	NM_000377.2(WAS):c.1058delC (p.Pro353Hisfs)	Pathogenic	30.3
WAS	chrX:48547197	NM_000377.2(WAS):c.1080A>C (p.Pro360=)	Uncertain significance	25.05
WAS	chrX:48547209-48547210	NM_000377.2(WAS):c.1097delG (p.Gly366Alafs)	Pathogenic	25.575
BTK	chrX:100645467	NM_001287344.1(BTK):c.71A>G (p.Glu24Gly)	Benign	0.1

Table 3.9: ClinVar variants not covered by WGS at 95% variant detection sensitivity

Gene	Variant position	Variant	Disease category	Average read depth
MSH6	2:48030112-48032381	c.3173-433_3556+228del	Pathogenic	11.4325
TERC	3:169481650-169482471	g.169481655_169482475del, NR_001566.1(TERC):n.374_1194del821	Pathogenic	6.1755
TERC	3:169482526-169485504	NR_001566.1(TERC):n.323C>T	Pathogenic	4.262
PMS2	7:6015622-6017501	NM_000535.6(PMS2):c.2276-113_2445+1596del	Pathogenic	7.991
PMS2	7:6028724-6029882	NM_000535.6(PMS2):c.989-296_1144+706del	Pathogenic	10.391
PMS2	7:6038282-6039384	NM_000535.6(PMS2):c.538-478_705+456del	Pathogenic	10.2715
CFTR	7:117249593-117252108	NM_000492.3(CFTR):c.2989-977_3367+248del	Pathogenic	10.561
ORAI1	12:122064773-122064779	NM_032790.3(ORAI1):c.132_137dup (p.Pro47_Ser48insProPro)	Likely benign	8.85
TAZ	X:153640060-153640061	NM_000116.3(TAZ):c.-120delT	Benign	12.725

3.3 Discussion

In this chapter, I analysed the diagnostic capabilities of WGS, WES and the TIGER gene panel in PID. By analysing the variant detection sensitivities, I sought to define a read depth 'cut-off' where sequence data can be scrutinized against. Subsequently, I studied the PID gene regions that fell short of the read depth cut-off to highlight potential deficiencies which invariably occur in all sequencing methods. This also identified PID genes in which there are areas of low coverage that may harbour undetected disease variants.

3.3.1 Variant detection sensitivity and its significance

Generally, thousands of genetic variants are filtered during analysis and only one likely causative pathogenic variant is subsequently confirmed by Sanger re-sequencing. Therefore, the onus lies on inclusion of a potential disease variant within the variant analysis pipeline. Disease-causing variants can be found at a depth of coverage of <10x, and if the bioinformatic pipeline has a stringency cut-off >10x, it is possible that the disease-causing variant will be discarded. Current read depth targets for inclusion are ambiguous, but a consensus read depth >30x has been applied for next generation sequencing data [145, 165-167]. My analysis shows that this cut-off is too stringent particularly for WGS, and individual read depths cut-off should be applied for the different sequencing methods, and to some extent, the type of variant. NGS biotechnologies invariably guarantee sequencing depths of an *average* read depth of >30x over all known genes. However, I have demonstrated that specifically for PID genes, an average gene coverage can mask individual deficits in sequencing particular genes that are clinically significant.

By utilising the genotyping data, I have shown that all genotyping SNVs within the TIGER gene regions, were found in both WGS and TIGER gene panel sequencing,

without exception. This finding validates its use as an accurate genotype marker which can be reliably identified in PID genes regions. The variant detection specificity was also found to be very high (>99.8%), suggesting that false positive variants are not generally found at low read depths.

Within the TIGER gene regions, the variant detection sensitivity of heterozygous sensitivity is high and to achieve a 95% sensitivity, the read depth required is >32 for TIGER gene panel and >23 for WGS. Realistically, the read depths of gene panel variants tend to run to the hundreds to be considered to be robust. However, using this cut-off exposes gene regions and potential missed disease variants within the targeted TIGER gene regions. My results show that caution should be exercised when considering patients with disease phenotypes that may be attributable to deleterious variants in *CASP10* and *CASP8*. Pathogenic genetic variants in these genes cause autoimmune lymphoproliferative syndrome which does have a distinct presentation from typical PID clinical features.

The WGS data is best analysed with comparison to WES in all known PID genes in order to give an indicator of its complete diagnostic performance. WGS has the lowest 95% variant detection sensitivity of heterozygous and homozygous SNVs of >13x and >5x respectively. WES can detect at a heterozygous SNV at >23x and homozygous SNV at >13x for 95% variant detection sensitivity. This demonstrates that variants at lower coverage than the current guidelines of >30x should be included as part of PID genetic diagnostic analysis.

Across all sequencing methods, homozygous variants can be picked up at lower read depth than heterozygous variants. This may seem paradoxical, but as the alternative allele is detected as a higher percentage of the total reads, a homozygous variant can be called accurately at low depth of coverage. This is significant as the majority of

PID is autosomal recessive disease. Therefore, homozygous variants should not be excluded from variant interpretation on read depth alone.

3.3.2 Depth of coverage using different sequencing technologies

The ACMG guidelines suggest a minimum mean coverage of 30x of the gene in which the variant is located [167]. This average measurement allows for deficiencies in the breadth of coverage to be masked by regions of the gene which have a high depth of coverage. This global metric is ultimately unhelpful for clinical diagnostics where a causative genetic variant may lie in any base within the gene, be it covered by 5 reads or 500 reads. My results which are restricted to PID genes, are similar to the recent studies reflecting that the coding regions of PID genes are typical protein coding genes and by extension, typical Mendelian disease causing genes [168]. My findings are also restricted to coding regions of genes that are fewer repetitive regions which can require proportionally greater reads to accurate call variants.

My findings show a higher number of low coverage regions in PID genes in WES (82.9%) compared to WGS (99.9%) despite the increased average depth of coverage; 68.9 compared to 33.9. This is likely due to two main differences in sequencing technologies: namely PCR-free amplification and longer read lengths in WGS compared to WES.

PCR-free WGS has been shown to be much less sensitive to GC rich regions and leads to more uniform coverage [169]. This is due to biases inherent in PCR amplification which result in uneven read coverage and increase duplicate reads present in the library. Long read coverage also allows sequencing through repetitive regions by spanning complex regions. Most alignment software e.g. Burrows Wheeler aligner (BWA), use gapped alignment methods to map reads to as many targets as possible.

BWA Smith-Waterman alignment will split the read in many cases into multiple alignments, and clip paired reads that cannot map [170]. Depth of coverage subsequently is reduced where gaps occur and variant detection, particularly indel variant calling also becomes less accurate as a result. Interestingly, this was not found in WES in PID gene regions, where the same read depth of >23 was required to detect both heterozygous SNVs and indels.

3.3.3 Low coverage genes

Targeted gene panels overcome shortcomings in read length by increasing depth of coverage. All 3 sequencing methods have low coverage regions on the X chromosome, which harbours many PID disease regions including the gene *IKBKG* or NEMO. *IKBKG* is particularly difficult to sequence due to its close homology with its pseudogene *IKBKGP1*. The gene and pseudogene have >99% homology, particularly in the 35kb region from exon 3 to exon 10, which are identical and flanked by low copy repeats [171]. This makes PCR-based amplification and hybridisation capture extremely challenging and has previously led to misattribution of genetic variants to disease [172].

Disease caused by *IKBKG* variants can often be difficult to diagnose due to wide phenotypic range from isolated immunodeficiency to anhidrotic ectodermal dysplasia. [173, 174]. My results show that WGS covers *IKBKG* relatively well, but it is likely that variants found in the gene are of low quality due to low homology mapping of paired-end reads. Interestingly this is only partly reflected in the average read depth and so emphasises the need for cognisant variant analysis prior to ascribing genetic variants to disease. Similarly, caution should also be exercised, in other known PID genes with close proximity of pseudogenes or low copy repeats, namely: *NCF1, C4B, FCGR1A* and *FCGR3A*. Frans et al (2018) have developed a long-range PCR method which utilises specific primers for the region between exons 3 and

10, which the authors suggest be carried out alongside targeted NGS of the first two exons. Given the phenotypic variability of *IKBKG* genetic disease, this long-range PCR method should be applied in all NGS studies in PID.

Another potential solution would be the addition of more specific capture regions of the particular coding exon to cover deficiencies in read depth, however this is not always possible where highly homologous pseudogenes are present. In such cases, the specificity of hybridisation capture during sequencing may need further refining. In cases of ambiguity within these particular regions, candidate gene sequencing should be considered.

3.3.4 Disease variant databases in primary immunodeficiency

The TIGER gene panel detected the vast majority of disease-related variants in the ClinVar database within its targeted regions (3171/3194 - 99.28%). Of the 23 variants with <32x coverage, only 3 variants were not covered at all. Interestingly, only 4 of the 23 with read depth <32 were pathogenic variants with the remainder either likely benign or of uncertain significance. This is likely reflective of the composition of the database which is mostly comprised of single submitter variants, which are either corroborated or refuted with additional data.

There are caveats associated with using such a database for the analysis. Our personal experience has shown that many variants reported to be pathogenic, have latterly been found to be of uncertain clinical significance and this is borne out in the literature [175]. With the large number of genetic variants being generated from large-scale next generation studies now and in the future, such approved databases with fairly rigorous submission criteria will inevitably struggle to remain up-to-date. Genetic variants in databases should not be assigned remotely without

accompanying expert phenotypic analysis, and this interface will continue to be a challenge going into the future.

4 USE OF WHOLE GENOME SEQUENCING IN THE DIAGNOSIS OF PID

In this chapter, I present a whole genome sequencing study to a phenotypically heterogeneous cohort of 77 paediatric PID patients, recruited from Great Ormond Street Hospital NHS Foundation Trust. No causative mutations had been identified prior to enrolment to the study. I applied bioinformatic tools to analyse WGS data to find genetic diagnoses that could account for their clinical phenotype.

4.1 Introduction

As described in Chapter 1, primary immunodeficiency is heterogeneous in onset and severity. Unlike most monogenetic disorders, most PIDs are not recognisable at birth and only manifest on exposure to pathogens. The resulting infection can often be severe and life-threatening, and even after treatment may cause significant complications. Those with severe disease may be referred to specialist centres early in life, and molecular diagnoses are sought as part of an evidential process to direct complex management options such as haematopoietic stem cell transplantation. However, for the majority of patients with PID, access to genomic data is uneven and patients often endure a delay to establishing a molecular diagnosis. In some cases, next-generation sequencing is not warranted as more efficient, cheaper methods of diagnoses are possible e.g. measurement of adenosine deaminase in the severe combined immunodeficiency disorder, ADA-SCID. Similarly, some patients with PID have clear phenotypic handles that allow efficient diagnoses e.g. thrombocytopaenia in Wiskott-Aldrich Syndrome [176]. Treatment decisions currently hinge on clinical severity, rather than molecular diagnosis. However, a molecular diagnosis can indicate disorders that are amenable to early treatment, can provide the patient and family with important prognostic information and indicate family members that may require medical follow-up, treatment and inform decision making regarding pre-implantation genetic diagnosis.

The broad spectrum of PID is reflected in the wide range of cells and tissues, which comprise the immune system. In this thesis, I have used as the International Union of Immunological Societies (2017) classification with 320 genes associated with inborn errors of immunity, collated by the committee of the International Union of Immunological Societies in 2017 (Appendix A) [7]. Since then, a more recent version has been published in 2020 [177].

4.1.1 PID variant analysis

Genetic diagnosis using next generation sequencing methods, has historically been slow, particularly in the analysis of data and prioritisation of the variants. However next generation approaches are useful for establishing a molecular diagnosis for many patients where symptoms are not specific enough for candidate gene sequencing. The identification of causal genetic variants can be challenging and often requires immunophenotypic and genetic expertise within a multidisciplinary team.

Genetic PID is currently a collection of monogenic diseases following Mendelian inheritance. It is generally assumed that such causative variants are protein-coding, rare or private to the family and show complete penetrance [178, 179]. A key challenge of WGS is pinpointing the causative mutation amongst the large number of bystander variants that do not play a role in the disease aetiology. Strategies used to filter out such variants are typically based on systematic elimination of variants according to prior genetic knowledge and available data, and reducing down the number of variants until the causative mutation is isolated.

A common filtering strategy employed is to remove all variants that are less likely to impact protein function i.e. intronic variants outside of splice site positions and synonymous coding variants. The remaining variants are protein truncating variants (PTV), splice variants and missense mutations. Interpretation of missense variants is challenging but guided by specialized programs such as PolyPhen2 and SIFT, which predict the extent of damage a variant causes to the protein using sequence homology and physical amino acid properties [150, 151]. Filtering according to allele frequency makes the assumption that deleterious mutations that cause disease are rare in the population. The use of public SNP databases such as ExAC and dbSNP allow common variants to be identified. However, pathogenicity does exist within 'control' datasets and therefore, allele frequencies are conservative in novel variants [147]. Use of existing disease databases and known genes associated with PID are

also be used to determine a genetic diagnosis in children with PID.

Whole genome sequencing (WGS) has recently been utilised to discover novel disease predisposition genes in PID and in other rare Mendelian disorders [99, 100, 106, 180]. This potential to uncover pathogenic variants in all genes, distinguishes it from gene panel and targeted Sanger sequencing that seek out known disease associations in genes predicted to cause disease.

4.2 Results

4.2.1 Variant analysis

Raw sequencing files were analysed as set out in Chapter 2. Genomic data was analysed to identify potentially disease-causing coding variants within PID genes as set out in by the IUIS [7] (see Appendix A).

Disease-causing variants were prioritised according to Ensembl variant effect predictor (VEP) version 78. For each variant that is mapped to the reference genome, VEP identifies the corresponding Ensembl transcript that overlaps the variant. This canonical transcript is the longest consensus coding sequence (CCDS) of a gene without stop codon. The variant is categorised according to the effect it has to the canonical transcript; the variant consequence with greater protein altering properties were prioritised in analysis. These are listed below in order of variant consequence, with most damaging first (As defined by Ensembl Variation version 78 http://www.ensembl.org/info/genome/variation/predicted_data.html) (Table 4.1).

Table 4.1: Variant annotation and consequences (adapted from Ensembl VEP)

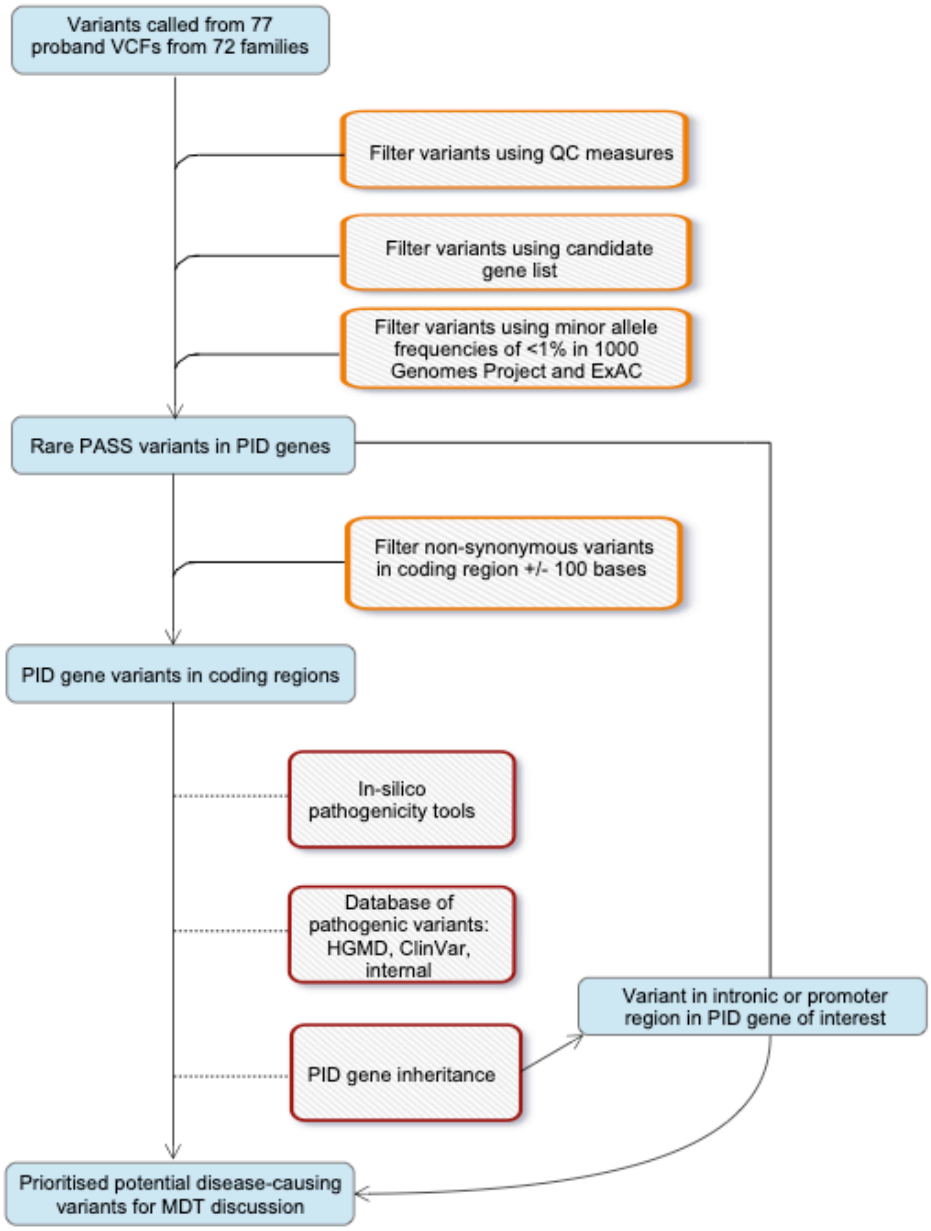
Variant annotations	Variant consequence
Transcript ablation	HIGH
Splice acceptor variant	HIGH
Splice donor variant	HIGH
Stop gained	HIGH
Frameshift variant	HIGH
Stop lost	HIGH
Start lost	HIGH
Transcript amplification	HIGH
Inframe insertion	MODERATE
Inframe deletion	MODERATE
Missense variant	MODERATE
Regulatory region ablation	MODERATE
Protein altering variant	MODERATE

The predicted effects of missense variants on protein function were computed using PolyPhen2 (Ramensky et al., 2002) and SIFT (Ng and Henikoff, 2001). All splice region and intronic variants were analysed for their potential effect on splicing, via the Alamut software interface (Interactive Biosoftware) hosted at HPC.

If a disease-causing variant was found using the pipeline illustrated in Figure 4.1, verification of its presence was sought on the corresponding BAM file. If the variant was represented by less than 5 reads, the variant was excluded from further analysis. Other informative metrics: BRIDGE allele frequencies, nucleotide-specific aggregate quality scores, and nucleotide conservation scores, including Genomic Evolutionary Rate Profiling (GERP) score were all hosted at HPC. Where possible, the ACMG guidelines were followed to assign pathogenicity [181]. Where sufficient evidence was not available to assign pathogenicity to rare non-synonymous variants within

phenotypically-relevant genes, genetic variants were classed as 'variants of uncertain significance' (VUS).

Figure 4.1: Overview of PID variant analysis pipeline

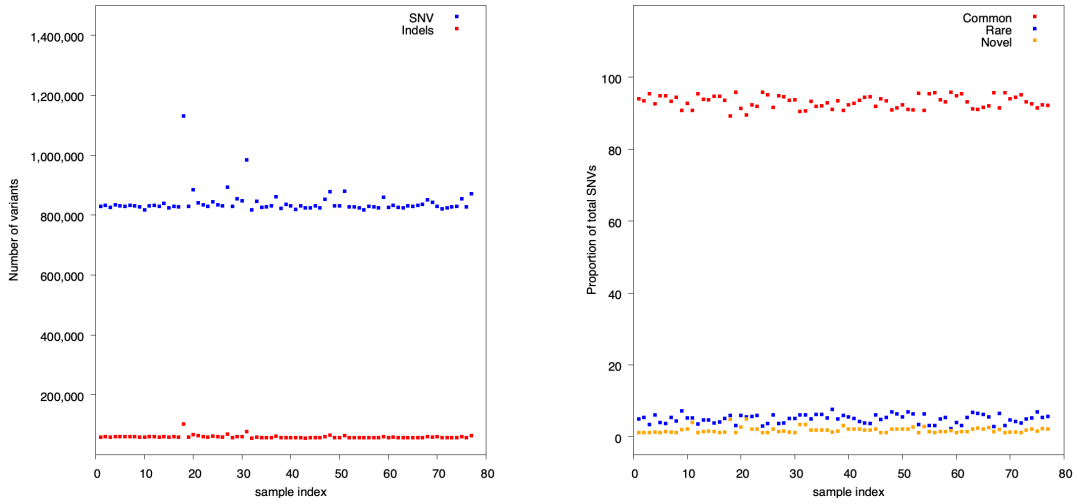


Structural variants were hosted by University of Cambridge high performance computing systems. Data analysis provided by Dr Keren Carss at University of Cambridge using CANVAS and MANTA callers showed that no significant deletions were found in known PID genes in the cohort.

4.2.2 Patients with identified pathogenic variants in PID genes

Variants were analysed according to bioinformatic pipeline as shown in Figure 4.1. The variant metrics were also carried out to assess validity of data (Figure 4.2).

Figure 4.2: Variant metrics for the patient cohort. A) Total number of variants – SNVs and Indels. B) Proportion of total SNVs that are common (>5% of mean allele frequency (MAF) in dbSNP), rare (<1% MAF in dbSNP) and novel (absent from population databases dbSNP and ExAC, and disease databases, ClinVar and HGMD)



Out of the 77 children with PID entered for WGS between June 2014-November 2016, 19 (24.4%) had a known pathogenic or likely pathogenic genetic variant that could explain the full clinical phenotype (Table 4.2). In these cases, *in-silico* evidence (SIFT, Polyphen, Combined Annotation Dependent Depletion (CADD) and

MutationTaster as appropriate) indicated pathogenicity of the genetic variant. Pathogenicity guidance according to the ACMG guidelines was also sought [167, 182]. Variants were classified with the following modifiers: 1) pathogenic, 2) likely pathogenic, 3) uncertain significance, 4) likely benign, or 5) benign. The evidence framework as described in [167] was utilised with information sought on population data, in-silico analysis, functional data and segregation data if this was available. Novel variants were defined as those absent from the Human Gene Mutation Database (HGMD) Professional database, and population databases such as 1000G (dbSNP) and ExAC, at time of writing [147]. The variants were all confirmed by Sanger re-sequencing and further functional evidence was sought, if considered necessary at MDT discussion.

Table 4.2: Confirmed pathogenic variants identified through WGS in patient cohort

Patient ID	PID category	Gene	Nucleotide change	Amino acid change	Zygosity	SIFT	PolyPhen2	PhyloP	CADD	MAF (ExAC) (%)	Reference/ HGMD no.
3	CID	*STK4*	c.349C>T	p.R117X	hom	NA	NA	1	41	0%	Nehme et al. (2012) CM122722
4,5,6	CID	*CD79A*	c.323T>G	p.V108G	hom	Deleterious	Probably damaging	0.573	19.13	0	-
8,9	CID	*DOCK8*	c.3437_3439delinsA	p.M1146Nfs*85	hom	NA	NA	5.243	22.3	0	-
13	CVID	*STAT1*	c.1154C>T	p.T385M	het	Deleterious	Probably damaging	6.083	26.4	0	Soltesz et al. (2013)
16	CID	*CFTR*	c.293A>G	p.Q98R	het	Deleterious	Possibly damaging	5.19	28	0.0000167	CM015351
		CFTR	c.3808G>A	p.D1270N	het	Deleterious	Possibly damaging	5.594	31	0.001463	CD064527
25	CID	*STAT3*	c.1145G>A	p.R382Q	het	Deleterious	Possibly damaging	6.258	34	0	Minegishi et al. (2007)
29	SCID	*TERT*	c.1439G>T	p.V461L	hom	Deleterious	Probably damaging	1	25.9	0	-
32	CVID	*TTC7A*	c.4228T>C	p.R265W	comp. het	Deleterious	Probably damaging	1.477	32	0.00082	-
		TTC7A	c.1802+3G>C	-	comp. het	-	-	1.96	14.19	0	-
34	SCID	*RFXANK*	c.584T>C	p.L195P	hom	Deleterious	Probably damaging	4.517	29.6	0	-

44	CID	*ZAP70*	c.283C>T	p.P95S	hom	Tolerated	Benign	1	21.1	0	-
52	CID	*PRF1*	c.1357G>A	p.V453M	hom	Deleterious	Possibly damaging	0.073	7.08	0.001321	-
55	CID	*NFKBIA*	c.32G>A	p.W11*	het	NA	NA	0.739	37	0	McDonald et al (2007)
59	AD	*BTK*	c.280A>T	p.I94F	hem	Deleterious	Possibly damaging	3.453	26	0	Internal database
63	AIS	*KRAS*	c.38G>A	p.G13D	het	Deleterious	Probably damaging	5.862	24.6	0	Takagi et al (2011)
66	CID	*STAT3*	c.2144C>T	p.P715L	het	Deleterious	Probably damaging	5.372	35	0	-
69	CVID	*TMEM173*	c. 482G>A	p.W161*	het	NA	NA	4.212	38	0	-

4.2.2.1 Patient 3

This patient presented at 9 years old with combined immunodeficiency, persistent Epstein-Barr virus (EBV) viraemia, and Hodgkin's lymphoma. He presented with low CD4+ T cells, low naïve T cells for age and low IgG and IgM levels. He is the eldest sibling of parents of Indian origin.

Analysis of the proband's WGS showed a homozygous stop gained mutation in *STK4* c.349C>T (p.R117*) (NM_006282). The alternative allele variant was not found in 1000G nor ExAC databases, and is predicted to cause nonsense-mediated decay of the resulting protein. This variant has been described previously in patients with similar clinical features as the proband [183] (HGMD Public Dataset CM122722).

STK4 (serine-threonine kinase 4, also alternatively known as MST1) is known to have a pro-apoptotic role in several cellular pathways including HIPPO and via JNK. *STK4* deficiency has been described to cause primary T-cell immunodeficiency syndrome characterized by progressive loss of naive T cells, recurrent infections, autoimmune manifestations, and cardiac malformations, including atrial septal defect in more than 3 separate cases [184] [62]. However, a recent case report of Epstein Barr Virus (EBV)-negative lymphoproliferative disease and immunodeficiency in a patient with *STK4* deficiency may indicate the pathogenic mechanism is not driven specifically by the Epstein-Barr virus [185].

4.2.2.2 Patients 4, 5 and 6 (family 1)

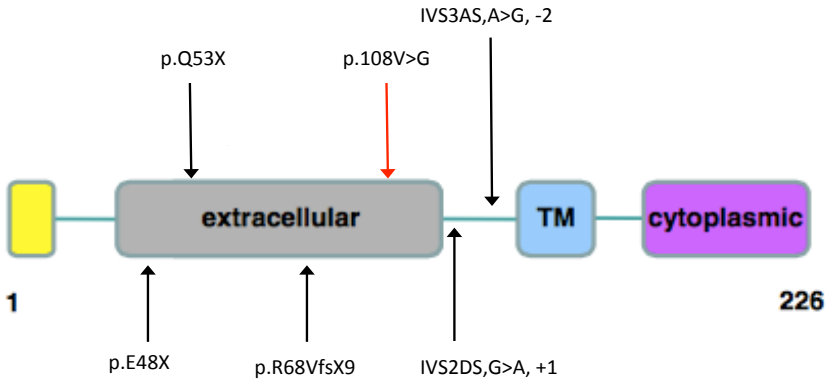
These patients are in a family of 3 affected siblings, born of consanguineous parents of Somali origin. The eldest sibling initially presented with multiple upper respiratory tract infections and hypogammaglobulinaemia at age 5. Her two affected younger sister and brother presented shortly afterwards with lower tract respiratory infections at 1 year, and subsequent investigations found

panhypogammaglobulinaemia. All 3 siblings were found to have less than 1% CD19+ B cells. All clinical courses have improved with replacement immunoglobulin. Given the presence of affected siblings within a consanguineous family, a genetic cause was sought but at the time of the diagnosis, only X-linked genetic causes of agammaglobulinaemia affecting boys were described [186].

PID gene analysis from results of WGS revealed a homozygous missense mutation in *CD79A* c. 323T>G p.V108G (NM_001783). This variant was unique to the 3 affected siblings and was heterozygous in the siblings' mother. The missense variant was predicted to be deleterious and probably damaging in SIFT and Polyphen2 respectively. The variant was not found in the 1000G or ExAC database, and has not been described in the literature.

CD79A codes for a component of the pre-B cell receptor (immunoglobulin alpha), which is essential in B cell differentiation in the bone marrow. Other reported genetic variants have been described in the gene; all being homozygous for their mutations and predicted to cause protein-truncating defects of the final protein. The variant we have identified is novel and is the only missense variant described to cause *CD79A* deficiency. Critical components of the pre-B cell receptor differentiation pathway have been identified that when disrupted, result in a complete block in B-cell differentiation. Deleterious genetic variants in these critical genes have a phenotypically identical presentation of early-onset panhypogammaglobulinaemia, and this novel missense variant appears to be phenotypically identical to other cases caused by the deficiency of the same gene (Figure 4.3).

Figure 4.3: Schematic representation of the protein domains of CD79A (this variant marked in red; others known *CD79A* variants in black) [186-188]



4.2.2.3 Patients 8 and 9 (Family 2)

Patient 8 and her affected sister both presented as infants with recurrent bacterial infections, mouth ulcers, genital warts and eczema. The elder sister also has autoimmune thrombocytopaenia, joint pain and swelling. Both were noted at presentation to have low CD4+ T cell count, and abnormal T cell proliferation to PHA. They also both had raised IgE levels (>2000 mU/l).

WGS of both sisters and their unaffected mother reveal a homozygous frameshift variant in *DOCK8* c.3437_3439delinsA p.M1146Nfs*85 (NM_203447) with their mother carrying the monoallelic version. Given their phenotype and its alignment with that of previously reported cases of DOCK8 deficiency, a diagnosis of DOCK8 deficiency immunodeficiency syndrome was made [189]. Autosomal recessive DOCK8 (Detection of cytokinesis 8) deficiency is a combined CD4 and CD8 T, and B cell deficiency, correlating with recurrent respiratory infections, viral and staphylococcal skin infections with a predisposition to malignancy. A distinguishing feature may be a low serum IgM in DOCK8 deficiency, as was seen in this familial

case. The siblings are now undergoing consideration for haematopoietic stem cell transplantation.

4.2.2.4 Patient 13

This patient presented at age 9 with an 18-month history of recurrent candida infection and oral ulcers. It was also noted that she had evidence of early changes of bronchiectasis. Her blood cell counts and immunoglobulin levels were within normal range. She is treated with prophylactic voriconazole.

Sequencing identified a heterozygous missense mutation in *STAT1* c.1154C>T p.T385M (NM_007315). The variant is predicted to be deleterious and potentially damaging in SIFT and Polyphen2 respectively. This variant has been previously attributed to chronic mucocandidiasis [74]. Her mother did not share the variant and there were no instances of similar disease phenotype within the family, suggesting a *de novo* occurrence. Unfortunately, the father's DNA sample was not available for analysis.

Deleterious heterozygous genetic variants in *STAT1* have been implicated in two distinct PID diseases, one caused by a loss of function of *STAT1* and one caused by gain of function of *STAT1* have been associated with chronic mucocutaneous candidiasis and autoimmunity [75, 190]. This variant in the DNA-binding domain of the protein leads to hyperphosphorylation of *STAT1* in response to the cytokine interferon gamma [191]. *STAT1* gain-of-function variants have also been associated with an increased chance of developing severe autoimmunity in adulthood such as enteropathy with villus atrophy and Type 1 diabetes mellitus [192].

4.2.2.5 Patient 16

Patient 16 was referred to the paediatric immunology department at age 15. He had been in ill-health from age 2, with recurrent lower respiratory tract infections, precipitating chronic lung fibrosis requiring infrequent BiPAP support. He has had several significant respiratory tract infections including Aspergillus, Actinomyces and CMV pneumonitis. He has a history of cerebral fungal infection and abscess with left upper limb paresis, which has mostly resolved. His other co-morbidities include portal venous thrombosis, right nephrectomy, pancreatic insufficiency, and hypertension secondary to renal impairment. He also has severe short stature, delayed puberty and steroid-induced cataracts bilaterally. He was referred to immunology owing to recent development of lymphopenia, particularly low B cells and NK cells.

WGS revealed two heterozygous missense mutations in *CFTR*: c.293A>G, p.Q98R and c.3808G>A p.D1270N (NM_000492), both at conserved bases across species. *In-silico* methods predict that both base changes to be deleterious and probably damaging in SIFT and Polyphen2. The variants are described in the Human Gene Mutation Database (HGMD) as known variants implicated in cystic fibrosis. Genetic variants in *CFTR* are well known to cause the multisystem disease cystic fibrosis. He had a sweat test carried out several years earlier, which appeared to exclude the likely cause of the disease to be cystic fibrosis. However, sequencing of *CFTR* had never been carried out. Genetic databases such as CFTR2 database catalogue *CFTR* variants, and both are present in the database and described in patients with cystic fibrosis and pancreatic insufficiency. Late diagnosis of cystic fibrosis has been described in patients, typically where initial symptoms of recurrent infections have only been picked up later in life [193].

4.2.2.6 Patient 25

Patient 25 presented at 5 months of age with a lung abscess and pneumonia. She has had recurrent staphylococcal infections, including staphylococcal folliculitis, and noted to have a dry skin rash from birth. Her blood results have shown an intermittently raised IgE level. There is no family history of immune deficiency.

Analysis of the proband's WGS showed a heterozygous missense mutation in *STAT3* c.1145G>A p.R382Q (NM_139276). The genetic variant is predicted to be deleterious using *in-silico* testing. This variant has been described previously in relation to a case of autosomal dominant Hyper IgE syndrome (AD-HIES) [194]. Unfortunately, parental DNA was not obtained in this case but both parents and siblings are clinically well. It is likely that the variant occurred as a *de novo* variant, and so may not have implications in the wider pedigree at present. To aid diagnosis and direct genetic studies, a scoring system for AD-HIES was devised and further refined [7, 195, 196]. However, the patient scored at the lower end of scoring (score of 21, between 20-40, suggesting there is only a suspicion of AD-HIES). This score is also different to other unrelated cases described with the same genotype; suggesting that a clear genotype: phenotype correlation is not available for all variants [196] [197].

STAT3 is one of the STAT family of signal transducers involved in many signalling pathways involving immunity activated through cytokines. Autosomal dominant Hyper-IgE syndrome is thought to occur as a result of a dominant negative mechanism of *STAT3*. This is supported by the lack of evidence gained from mice models where a complete heterozygous deletion of one *STAT3* allele do not recapitulate the phenotype [198].

4.2.2.7 Patient 29

Patient 29 is a young girl of Albanian descent who presented at 1 year of age with CMV colitis. She was diagnosed with severe combined immunodeficiency with aplastic anaemia. She had evidence of CMV disease in her blood, stool and on intestinal biopsy, and was found to be neutropaenic. She also had reticular patches of hyperpigmented skin. Her blood results showed low CD4+ T cell, CD8+ T cell, and low B cell counts. Bone marrow aspirate showed maturational arrest of B cells. Her telomere length was measured to be shortened (<10^{th} centile for age). Unfortunately, she became severely ill from enteropathy and died at 10 months old.

WGS of DNA and analysis of known PID genes revealed a homozygous missense mutation in *TERT* c.1381G>T p.V461L (NM_198253). The variant has not been described in databases nor in the literature in association with disease. The same mutation was also found in the proband's mother but as a heterozygous mutation. Heterozygous variants of *TERT* have been described in the literature to cause pulmonary fibrosis but the proband's mother does not have any clinical features of disease [199].

Viewed retrospectively, this patient likely had Hoyeraal-Hreidarsson syndrome, which is a severe variant of dyskeratosis congenita associated with bone marrow failure, severe immunodeficiency, developmental delay and in some cases, cerebellar hypoplasia. *TERT* is one of the genes implicated in Hoyeraal-Hreidarsson syndrome; all currently described have a role in telomere functioning and maintenance. *TERT* encodes the catalytic subunit of telomere reverse transcriptase, the enzyme required to maintain the ends of telomeres post-replication. Mutations in *TERT* result in reduction of the telomere length, particularly in cells with a high telomerase expression such as immune and stem cells in the bone marrow.

4.2.2.8 Patient 32

This proband was referred to the immunology service at age 3 with clinical features consistent with early onset inflammatory bowel disease and low IgG, IgA and IgM levels. She initially presented with features of inflammatory bowel disease at 22 months, confirmed histologically a few months later. Total parenteral nutrition was repeatedly required for several months over the next 2 years.

Analysis of known PID genes showed a compound heterozygous mutation in *TTC7A.* A heterozygous missense mutation was inherited from father c.793C>T p.R265W and a splice donor variant was inherited from mother c.1802+3G>C. The splicing mutation is predicted to be damaging according to Alamut software. These variants are novel to the literature and not described in population databases.

TTC7A is highly expressed in enterocytes and is particularly crucial in directing polarity in development in the foetal intestinal epithelial cells [200, 201]. *TTC7A* is associated with the phenotypes, inflammatory bowel disease, with or without multiple intestinal atresia, and combined T and/or B cell immunodeficiency [201, 202]. In early case reporting, severe phenotypes were described in patients with homozygous protein-truncating genotypes [200, 202]. A mutational analysis study suggests that biallelic missense mutations, and variants not affecting the tetratricopeptide repeat domains appear to have a relatively better prognosis [203]. This finding appears to relate in our patient who presented relatively late and with milder intestinal pathology, compared to other cases.

4.2.2.9 Patient 34

The proband presented at 1 year with repeated Staphylococcus aureus and Pseudomonas infections affecting the lungs, and protracted diarrhoea. He was

panhypogammaglobulinaemic with low CD4+ T cell count, but near normal levels of CD8+ T cells. He also had autoimmune thrombcytopaenia.

The patient had a homozygous missense mutation in *RFXANK* c. 584T>C, p.L195P (NM_003721). This is predicted to disrupt the Pfam domain of the protein. This variant is known to cause disease and is listed on the Human Gene Mutation Database (HGMD CM001330) to cause Bare Lymphocyte Syndrome or MHC class II deficiency, a rare form of SCID. The parents are unrelated, but were both heterozygous for the variant. Genetic counselling and carrier testing has subsequently been offered to the proband's younger sibling, and wider family.

MHC Class II deficiency is caused by biallelic genetic defects in one of 4 genes *RFXANK*, *CIITA*, *RFXAP*, and *RFX5*, with *RFXANK* defects most frequently encountered due to founder mutations [204]. Despite genetic heterogeneity, no clinical correlation has been found between the causative gene and the patient phenotype [205].

4.2.2.10 Patient 44

Patient 44 presented at age 10 with recurrent respiratory infections and occasional episodes of bloody diarrhoea. He was also noted to have persistent high EBV viraemia and had evidence of bronchiectasis. He was diagnosed with CID, with hypogammaglobulinaemia and enterocolitis. He had depletion of CD8+ T lymphocytes, with normal CD4+ T cells and serum immunoglobulin levels.

The proband was found to have a homozygous missense mutation in the *ZAP70* c.283C>T, p.P95S (NM_001079). His parents and younger brother were all found to be heterozygous at this variant. The variant is novel and predicted to be deleterious

using *in-silico* pathogenicity tools. The patient has since undergone a successful haematopoietic stem cell transplantation.

ZAP70 is a cytoplasmic tyrosine kinase that interacts with the zeta-chain of the T-cell receptor with highest levels of expression in T and NK cells [206]. T-cell receptor are rendered unresponsive to stimuli in *ZAP70* deficiency leading to deficiency of CD8+ T cells and lack of functioning in circulating CD4+ T cells. [207].

4.2.2.11 Patient 52

This patient was referred for further assessment and management at 15 months when she developed a barking cough, hepatosplenomegaly and cytopenia. She subsequently developed sepsis, requiring respiratory and inotropic support. Investigations revealed hypocellular bone marrow with infiltration of histiocytic cells. She had absent B cells, NK cells and low IgG and IgM. Her nitroblue tetrozolium (NBT) and vaccination responses were normal, but intracytoplasmic perforin expression and granule release assay were abnormal. Her history of note is of intra-uterine growth restriction, delayed developmental milestones. She is the first daughter of consanguineous Pakistani parents. She later deteriorated and became encephalopathic. CT images revealed multiple brain lesions consistent with fungal septic emboli. Her condition was stabilised, but she continued to show evidence of disseminated multiorgan infection. The pathogen was not clearly identified. Unfortunately, she died at 20 months due to multi-organ failure associated with the infection.

WGS analysis identified a homozygous mutation in *PRF1* c.1357G>A p.V453M (NM_005041), which is a novel variant, predicted *in-silico* to be deleterious and possibly damaging. It is found as a heterozygous variant in the population database ExAC with a minor allele frequency of 0.00132. Homozygous missense mutations in

PRF1 gene cause haemophagocytic lymphohistiocytosis, which does explain her symptoms. For the variant in *PRF1*, the parents were found to have a monoallelic genetic variant each.

PRF1 encodes perforin, a pore-forming cytolytic protein found in granules in CD8+ T cells and NK cells, and is thought to be critical in the creating holes in the plasma membrane of foreign target cells [208]. Absence of perforin has been described as the cause of familial haemophagocytic lymphohistiocytosis type 2, a rare disease of severe hyperinflammation of activated lymphocytes and macrophages usually presenting after a systemic infection in very young children [34, 35, 209, 210]. This variant is novel, and a defect in perforin is indicated in the absence of its expression.

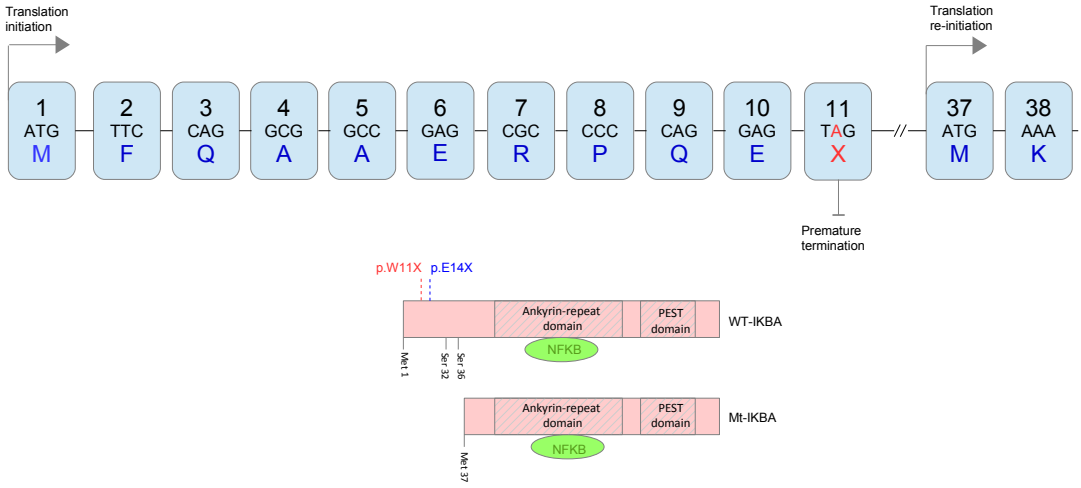
4.2.2.12 Patient 55

This 3-year-old girl presented with recurrent bilateral knee joint swelling, recurrent fevers and history of failure to thrive. Her immune cell counts were within normal range. Synovial biopsy of both knees on multiple occasions yielded inflammatory infiltrates, but did not show evidence of acute infection nor culturing of an organism. Her father has a history of ectodermal dysplasia.

WGS analysis revealed a heterozygous stop gained variant *NFKBIA* c.32G>A p.W11* (NM_020529). The genetic variant was also found in her father's WGS sequence. This change has been described in a female patient with ectodermal dysplasia and immunodeficiency, suggesting that *NFKBIA*-associated disease occurs via haploinsufficiency [211]. After the genetic diagnosis, she subsequently had specialist tests at Cambridge University NHS Trust which showed overall reduced innate response to MYD88/IRAK4 dependent toll-like receptor ligands, low production of interferon-gamma in response to polyclonal T-cell stimulation and reduced LPS induced production of IL-10 as described in McDonald et al (2007).

Our patient appears to have a milder phenotype compared to patients with heterozygous missense variants in *NFKBIA*. It is thought that heterozygous variants in *NFKBIA* cause disease in a dominant negative fashion, such that the mutated allele interferes with the function of the other allele [212]. This may be due to the early mutation in the reading frame, and the presence of an initiation codon located shortly after the mutated base. This may cause continued presence of part of the protein which may preserve some of the function [21, 212] (Figure 4.4).

Figure 4.4: Schematic of the NFKBIA codons. Wildtype (WT) and mutated NFKBIA (Mt) are shown below, with patient's genetic variant in red and a previously described variant in blue, based on figure in [21]



4.2.2.13 Patient 59

This young boy presented at 4 years old with recurrent chest infections and short stature, less than 3 standard deviation from expected height. He was found to have <5% CD19+ B cells and low immunoglobulins (IgG, IgA and IgM). His symptoms have improved since commencing subcutaneous immunoglobulin treatment.

WGS analysis revealed a missense variant in *BTK* c.280A>T p.I94F (NM_000061). This variant is predicted to be deleterious and probably damaging and is not described in population databases, nor in mutation databases. However, it is noted in internal hospital genomic databases associated with another case with similar presenting clinical features. *BTK* (Bruton agammaglobulinemia tyrosine kinase) is a well-described gene implicated in x-linked hypogammaglobulinemia [213, 214]. It plays a critical role is in B lymphocyte development, differentiation and signalling in the myeloid compartment. Disruption of the enzyme causes failure to produce mature B lymphocytes and a subsequent failure of immunoglobulin heavy chain rearrangement [214].

4.2.2.14 Patient 63

This patient presented at 3 years old with joint swelling, hepatosplenomegaly and low platelet counts, and was subsequently diagnosed with systemic lupus erythematous and autoimmunity syndrome. Investigations confirmed the presence of anti-nuclear antibody. Her immune cell counts were within normal range.

Analysis of PID genes revealed a heterozygous variant in *KRAS* c.38G>A p.G13D (NM_004985). The variant has been described in previous 2 cases in patients with symptoms similar to autoimmune lymphoproliferative syndrome and with the somatic variant found exclusively in haematopoietic cells [45]. The variant has not been found in sequencing of the *KRAS* gene from DNA from oral mucosa. Germline *KRAS* mutations have been associated with a distinct cardio-facial-cutaneous syndrome called Noonan syndrome [215]. Significantly, the patient does not have cardiac defects or facial dysmorphism suggestive of Noonan's syndrome. This suggests that the somatic KRAS variant is likely the cause of the patient's clinical features, which is specific to the haematopoietic cell population. As DNA has been derived from peripheral blood sample, it is likely that the variant is specific blood

cells. This has been confirmed in the patient's blood using Sanger sequencing, with salivary DNA awaiting Sanger sequencing to confirm the likely absence of the variant in other cells.

4.2.2.15 Patient 66

The proband presented at 1 year with gross lymphadenopathy and splenomegaly, associated with autoimmune haemolytic anaemia and thrombocytopenia. These symptoms are suggestive of the disease ALPS (Autoimmune lymphoproliferative syndrome). However, prior genetic testing of the *FAS* gene and of the apoptosis pathway related to *FAS* binding were normal.

WGS revealed a novel heterozygous missense mutation in *STAT3* c.2144C>T p.P715L (NM_139276). The variant is absent in both parents suggesting that the inheritance is de novo. As well as causing Hyper IgE syndrome via a loss of STAT3 function, heterozygous *STAT3* variants have been described to cause gain of function effect, leading to clinical features of autoimmune syndrome [216].

4.2.2.16 Patient 69

This patient is a 7-year Caucasian boy with recurrent herpes simplex, with eruptions particularly prominent around the lips. He also persistent verrucae on his feet and hand, thought to be caused by human papilloma virus, isolated on one occasion. Other symptoms include skin photosensitivity and mild joint pain and hypermobility. Investigations show normal lymphocyte and immunoglobulin counts. However, he was noted to have an absent CD3 stimulation response, but normal to PHA.

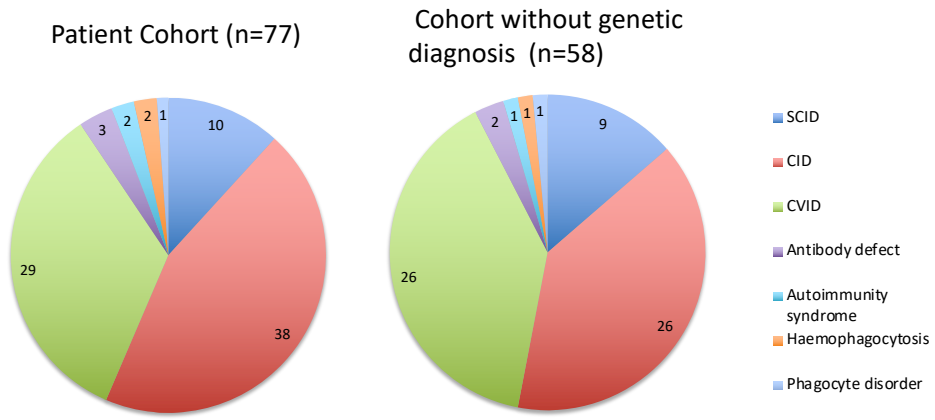
WGS revealed a heterozygous stop gained variant in *TMEM173* c. 482G>A p.W161* (NM_198282). The variant is likely to disrupt to an allele of the resulting protein. The mechanism of disease appears to be via a dominant negative mechanism [217, 218].

Heterozygous mutations in *TMEM173*, otherwise known as STING, have been associated with a syndrome labelled SAVI (stimulator of interferon genes STING-associated vasculopathy with onset in infancy) [218-220]. The clinical features of SAVI include cutaneous vasculitis, fevers, interstitial lung fibrosis, and systemic inflammation. More recently, lupus-type joint pain and photosensitivity features have been described [221]. All of the variants that have been described are gain-of-function heterozygous genetic variants [218, 220]. Interestingly familial heterozygous TMEM173 variants appear to have milder clinical manifestations. Some heterozygous protein-truncating variants in coding genes are not tolerated, and in these circumstances, it is likely that haploinsufficiency may be more common than currently recognised [222]. However, in gain of function immunodeficiencies, a heterozygous protein-truncating variant appears to have a milder phenotype via a mechanism thought to be due to lower levels of the dominant protein that has a noxious role to the cell [212]. Other atypical features in this patient are the recurrent viral infections and reduced CD3 T cell proliferation. It is unclear how significant the clinical features are, given the patient's age, and further follow-up is warranted.

4.2.3 Patients with insufficient PID genotype to meet phenotype

Despite WGS efficacy, 58 of 77 children with PID remain without a firm genetic diagnosis. Disease categories shown in the cohort, and in those without a genetic diagnosis (Figure 4.5). The distribution of disease categories and those unsolved appear broadly similar, suggesting that genetic variants are not being found in both severe disease forms and in milder phenotypes.

Figure 4.5: PID disease categories represented in cohort (n=77) left, and in those without genetic diagnosis (n=58)



In these patients, the sequencing metrics suggest that the depth and accuracy of sequencing was adequate to detect SNVs and indels in genes known to cause PID. Significant deletions encompassing known PID genes were not found in these patients (data provided by Dr K Carss, University of Cambridge).

It was noted that there were instances where single damaging variants were found in phenotypically-relevant genes, but a deleterious variant was not found on the other allele to fulfil the genetic diagnostic criteria. These variants are termed variants of unknown significance (VUS) [181]. Such VUS in 3 different patients in the cohort are described below.

4.2.3.1 Patient 11

Patient 11 was referred to the department at age 5 following recurrent lower respiratory infections with *Streptococcus pneunomiae*. Baseline immune cell counts including complement C3 levels were within normal limits.

A heterozygous stop gained variant was found in the patient: *C9* c.346C>T p.R95* (NM_001737). The protein-truncating variant has been described to be pathogenic in the homozygous state, particularly found in the Japanese population in which it has a carrier frequency of 6.7%, suggesting a founder effect [223, 224]. The variant was also found in the patient's mother who is of Japanese ancestry, but does not have immunodeficiency. C9 deficiency is an autosomal recessive genetic disorder [92]. On further analysis of the non-coding space, a second variant was found in the deep intronic region between the first two exons *C9* g.39347251A>T c.78-4953T>A (NM_001737). The intronic variant was found to have been inherited from the patient's father. However, on assessment with Alamut software, the variant did not appear to have an effect on splicing, and it was concluded that there was insufficient evidence to classify the variant as pathogenic.

C9 is the last complement component in the cylindrical membrane attack complex, specifically causing cell lysis in foreign cells, essential for the functioning of the innate immune system. Deficiency of C9 component predisposes to infections with *Neisseria* with both autosomal dominant and recessive forms described [92, 223]. Our patient has clinical features similar to those with C9 deficiency, but remains without a genetic diagnosis to explain the full phenotype. Further investigations are now indicated such as haemolytic activity of sMAC and antigenic levels of C9.

4.2.3.2 Patient 50

This child initially presented at 3 years old and when entered into the BRIDGE-PID study, he was 8 years with a long history of recurrent lower respiratory infections, bronchiectasis, combined immunodeficiency and lymphocytic interstitial tubular nephritis. He had absent B and NK cells, very low naïve population of CD4+ and CD8+ T cells. He was also noted to have hepatosplenomegaly and short stature. Unfortunately, the patient had a severe episode of sepsis and after an acute deterioration, passed away.

WGS analysis found a potentially pathogenic variant in *RTEL1* c. 334G>A p.A112T (NM_016434). The missense mutation is predicted to be deleterious and probably damaging, with a low MAF in ExAC database (0.00022). These changes were unique to the patient, in the BRIDGE population cohorts, and were both absent in parents.

Given that the typical disease characteristics fit well with the patient's phenotype (B and NK lymphopenia, hepatosplenomegaly), a second variant was sought as only biallelic RTEL1 variants have been described to cause disease [225]. A rare insertion of 35 bases in the deep intronic region was found between exon 14 and 15: *RTEL1* g.62315223_62315224ins35, c.1192-1653insA> AGTTCTGAGGATCCCATATACATATTCTCTCTCTAA (NM_001283009). Using the splicing *in-silico* methods, Alamut Visual software estimated that overall, the insertion did not affect donor sites in the intron using the splicing tools SpliceSiteFinder-like, MaxEntScan, NNSPLICE, and GeneSplicer [146, 226-229]. However, the score in MaxEntScan score from 1.66 to 4.15 (a change of 149.9%). This score can be independently interpreted to mean that the insertion creates a new acceptor splice site in the middle of the intron. This would cause a disruption in RNA production, and may lead to nonsense-mediated disruption of the allele. MaxEntScan uses the maximum entropy model which appears to outperform other models that predict splicing variants [230, 231].

RTEL1 is a DNA helicase, utilised specifically in DNA metabolism and deficiency of the enzyme appears to halt conversion of telomere loops to free telomere circles in replicating cells [232]. Patients with *RTEL1* deficiency have features of B and NK cell lymphopenia, variable onset of bone marrow failure and like other causes of dyskeratosis congenita, result in shortened telomeres in the patient [225]. After discussion at MDT, it was felt that variants in *RTEL1* did fit the patient's phenotype,

but there were no remaining cells from the patient, with which to perform the necessary functional confirmation e.g. telomere measurement.

4.2.3.3 Patient 61

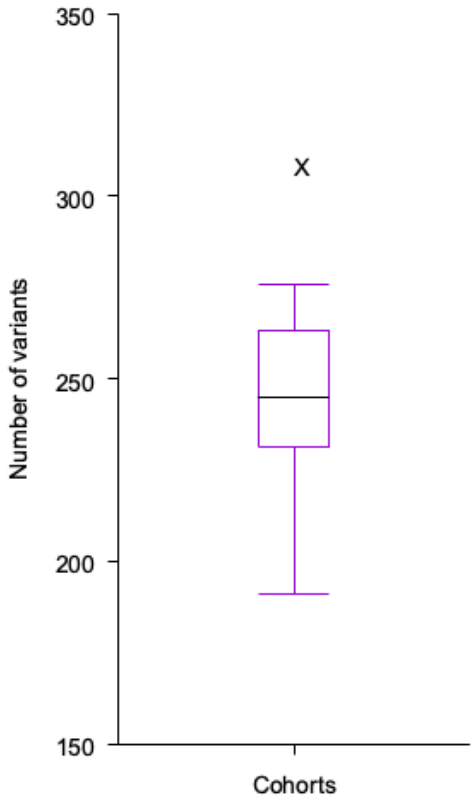
Patient 61 presented to the immunology service with recurrent chest infections, and features of severe combined immunodeficiency: very lower levels of CD3+ and CD4+ T cells, absent gamma-delta T cells, normal B and NK cell count. He is the eldest sibling of parents of Caucasian origin.

Analysis of the proband's WGS revealed a heterozygous missense mutation in *IL7R* c. 908T>C, p.F303S (NM_002185). The variant is predicted to be deleterious and probably damaging *in-silico*, and is at a highly conserved base in the protein. The rare variant is found in the population databases (ExAC MAF = 0.0000083), and is shared with his unaffected mother and younger sibling. *IL7R* deficiency is an autosomal recessive genetic disorder with immunological features similar to the patient's phenotype, but is described in the literature to be exclusively biallelic and completely penetrant [233, 234]. On further analysis, a second damaging genetic variant in *IL7R* exclusive to the patient was sought in the genes non-coding regions, but not found. Functional tests carried out by Dr Kimberly Gilmour and Dr Nadia Shahid (Great Ormond Street Hospital NHS Trust Clinical Immunology Laboratory) using patient and parental blood samples indicate specific IL7R deficiency in the patient only (not shown here as patient consent not obtained). This suggests the presence of another factor that has caused disease in the patient without it also being manifest in his mother.

4.2.4 Variants of unknown significance (VUS)

Given the notable genetic variants found in key genes described above, I sought to find other potential variants of unknown significance (VUS) in PID genes within the cohort. VUS were defined as non-synonymous SNVs and indels within the coding regions of PID genes with a MAF <0.01. Missense variants were included if they were predicted to be deleterious using SIFT and damaging using PolyPhen2. I compared the number of VUS across the cohort with randomly-selected 77 subjects across 11 of the other disease domains in the BRIDGE project. There were collectively more VUS in PID genes in the PID cohort (number of VUS=309), compared to other BRIDGE project cohorts of the same size (mean =244.36, range =191-276) (Figure 4.6). This was found to be statistically significant with a paired t-test(p=0.0212). This suggests enrichment in VUS in patients with PID, either via unrecognised monogenic disease in those in which genetic diagnosis was not made or other oligogenic variants.

Figure 4.6: Box and Whisker plot of range of VUS in BRIDGE cohorts. The range of VUS varying between 191-276, median=245. The BRIDGE-PID cohort's VUS is denoted with a cross (number of VUS=309)



4.3 Discussion

4.3.1 Use of whole genome sequencing to investigate PID

In this study, whole genome sequencing has enabled the genetic diagnosis of 19 children with primary immunodeficiency in a sample of 77 children recruited to the BRIDGE-PID study (24.7%). Pathogenic genetic variants were sought in known genes where the contribution of gene to the phenotype has been previously verified. The diagnostic yield of 24.7% is broadly in-line with similar studies seeking PID genetic diagnoses in affected populations [23, 132, 133]. Patients entered for WGS for the BRIDGE-PID study were pre-selected to those without a prior genetic diagnosis, and were likely those without a readily recognisable PID syndrome. Recruitment restrictions were also not initially imposed. This discouraged 'cherry-picking' of those with severe phenotypes or those more unusual features, allowing the patient cohort described here to be a fair representation of a standard paediatric immunology patient population.

For some PID phenotypes, biochemical or immunological studies may be sufficient to suggest a likely genetic diagnosis via candidate gene sequencing. However, for the majority of PID, a gene-by-gene Sanger sequencing or limited gene panel is not economical or efficient due to genetic and phenotypic heterogeneity involved. However, whole genome sequencing yields many thousand variants and inefficiencies may also occur in the variant analysis. Overall the combination of filtering strategies used has allowed me to reduce a list of many thousands of variants down to a small manageable set of the most compelling variants for further consideration. At this stage, restricting whole genome sequencing data to variants within known PID genes has also facilitated efficient diagnosis of the patient cohort.

A genetic diagnosis of PID can guide treatment, help prevent complications and provide prognostic information. As in Patient 44, the detection of ZAP70 deficiency

has prompted haematopoietic stem cell transplantation to be offered as a potential cure [4]. As gene therapy becomes increasingly available for a broader range of primary immunodeficiencies, an accurate genetic diagnosis is vital [235]. Genetic diagnosis may also inform carrier testing, prenatal or indeed preimplantation diagnosis.

Currently for finding new genetic diagnoses for patients with PID, whole genome sequencing is comparatively expensive but has advantages in SNV accuracy. As more population and disease genetic variants are collated, it is likely that the genetic diagnoses will become further automated, rendering Sanger sequencing unnecessary. To facilitate quick analysis, accurate coded phenotype information is also required, such that genomic diagnoses can be made accurate.

Genetic variants that have previously attributed to disease make the genetic diagnosis more straight-forward. With increasing availability of combined genetic and phenotypic data, and with greater confidence in the sequencing quality, more accurate prognostic information may be delivered to the patient, and often treatment decisions may become better defined. However, there are currently only relatively few PID genes where this may be applicable, with the majority, only recently linked with immunodeficiency. We suspect with greater use of genetic sequencing data, through databases, and increasingly efficient laboratory techniques and likely machine-learning bioinformatic tools to predict pathogenicity.

4.3.2 Challenges in the interpretation of the pathogenicity of novel variants in known PID genes

My results have highlighted the challenges encountered in the interpretation of novel variants. The novel variants found expand the genotypic and phenotypic

spectrum of some primary immunodeficiencies and highlight the complexity of assigning a diagnosis and the advantages of using an untargeted approach. Whole genome sequencing allows an agnostic approach to novel gene associations and exploration of non-coding regions of PID genes.

All patients in the cohort were discussed at the MDT stage, with an expert, individual assessment of the patient's phenotype applied to a filtered, prioritised variant list. As whole genome studies become larger in future studies, automation in phenotyping is increasingly being explored. Current drawbacks include inter-operator variability of assigning phenotypes and the collection of 'snapshot' static phenotypes. The latter is particularly pertinent to immunological diseases, which may not fully manifest until exposure to an organism, or development of an unusual but characteristic complication e.g. HSV-1 encephalitis in *TLR3* and *UNC93B* deficiency [236]. As the patient cohort are children, it is likely that the immunological features will evolve with age. This also emphasises the importance of revisiting the genetic diagnosis assigned to assess whether updated literature can reveal further genotype: phenotype associations.

Variant analysis prioritises PTVs as they have a strong prior probability of causing disease. Any given genome may carry multiple disease-associated variants, and the high frequency in which PTVs occur in apparently healthy individuals indicates that the majority are likely to be benign [237]. Data collected by the 1000 Genomes Project showed most subjects (all recruited as healthy adults) carried between 3 to 24 homozygous disease-causing variants, with one of these classified by HGMD as highly damaging [148]. This emphasises the value of functional analysis in novel genetic discovery to better understand the underlying disease pathogenesis.

Protein-truncating variants, including variants in splice donor and acceptor regions are expected to have significant effects on protein function. *In-silico* prediction tools can give indications of variant pathogenicity, particularly for missense mutations. Indeed, combined tools such as MutationTaster and Combined Annotation Dependent Depletion (CADD) scoring are useful by combining evidence scores of variant type, protein domain information, disease variant databases and base conservation scores, to provide an accurate *in-silico* tool for prediction. Further bioinformatic tools may be used to predict the effect of the variant on protein secondary structure, and with an ever-increasing understanding of 3D protein structures due to sophisticated microscopy, it is likely that this will better inform variant filtering in the future [238, 239].

Bioinformatic tools to predict splicing effects in the non-coding region are also useful, but currently only focus on predicted splice-sites which have an effect on the resulting protein. Indeed, current practice dictates that immunological tests based on the function of the protein are required to confirm pathogenicity, particularly for novel variants. This may be too restrictive to find novel genetic variants as non-coding regions may cause disease by mainly unexplored mechanisms, such as disruption of transcription factor binding, or via non-coding RNA that may regulate expression of target genes [240, 241]. It is clear that inclusion of new methods such as transcriptomics and immune cell -specific sequencing will also allow further elucidation and evaluation of novel genetic variants in rare diseases [242].

I found a greater number of rare deleterious PID genetic variants are harboured by those with immunodeficiency. These variants may represent a potential expansion of the immunophenotypic spectrum of the PID genes, including mechanisms of inheritance. PID genetic variants that are present in the proband are often present in the parents too, and this often negates the likelihood of the variant being the sole cause of the phenotype. However, this phenomenon may be due to variable

expressivity. In Mendelian inheritance, this may be due to lack of environmental trigger, although this is probably unlikely in childhood onset of disease. Splicing variants specific to cell types may also contribute to milder phenotypes. Similarly, somatic mosaicism may also cause a cell/tissue specific effect, such that identical genome sequences in a particular PID gene can cause different phenotypes.

5 NOVEL GENE DISCOVERY AND ASSOCIATIONS IN PID

Within a relatively short space of time, next generation sequencing studies have accelerated our understanding of the genetic mechanisms of immunodeficiency. The discovery of novel genes has allowed researchers to build pathways of interacting genes that share a common immune function. In this chapter, whole genome sequencing (WGS) was used to find novel genetic associations enriched in a cohort of children with primary immunodeficiency.

In the first part of this chapter, a variant analysis pipeline was designed to find potential novel genes in a PID cohort. The second part focusses on association studies within these genes, to find novel genes of interest that may contribute towards disease pathogenicity.

5.1 Introduction

5.1.1 Use of whole genome sequencing for novel gene discovery in rare disease

A full discussion of whole genome sequencing is found in the Introduction. The advent of NGS technologies has made it possible to perform whole genome sequencing (WGS) to identify novel disease predisposition genes in rare Mendelian disorders [32, 99, 100, 180]. Exome sequencing is currently the most commonly used NGS technique to identify novel disease-causing genes in PID but there remains a large number of patients in which exome sequencing has not founda causative gene due to inherent limitations [75, 101, 102]. Exome sequencing covers ~1% of the genome, and does not capture all exons nor other gene regulatory regions [103]. Structural variants such as copy number variants (CNVs) are also not readily detectable in WES; CNVs in exonic areas are very rare and generally extend outwith the targeted regions [104]. WGS is better than WES in finding rare variants within exons [106]. This is, in large part, due to reliance of WES to PCR-amplify targeted regions [107]. Despite the focus of protein-coding regions in novel gene discovery, intronic mutations have been described in PID [101, 109]. The potential to uncover pathogenic variants in unexplored areas within known disease genes as well as novel disease variants, is therefore significantly enhanced by using WGS. It is estimated that the current cost of sequencing a whole genome is ~$1000 [110]. This makes the diagnosis and understanding of the genetic basis of rare diseases through WGS increasingly attainable and cost-effective.

5.1.2 Variant Analysis using filtering strategies

Genetic PID is a currently thought of as a collection of monogenic diseases following Mendelian inheritance. It is generally assumed that such causative variants are protein-coding, rare or private to the family and show complete penetrance [178, 179]. A key challenge of WGS is pinpointing the causative mutation amongst the large number of bystander variants that do not play a role in the disease aetiology.

Strategies used to filter out such variants are typically based on systematic elimination of variants according to prior genetic knowledge and available data, and reducing down the number of variants until the causative mutation is isolated.

A common filtering strategy employed in is to remove all variants that are less likely to impact protein function i.e. intronic variants outside of splice site positions and synonymous coding variants. The remaining variants are protein truncating variants (PTV), splice variants and missense mutations. Interpretation of missense variants is challenging but guided by specialized programs such as PolyPhen2 and SIFT which predict the extent of damage a variant may cause the protein. Filtering according to allele frequency assumes that deleterious disease-causing mutations are rare in the population. The use of public SNP databases such as ExAC and dbSNP allow common variants to be identified. However, pathogenicity does exist within 'control' datasets and therefore, allele frequencies are conservative in novel gene discovery [147].

Filtering based on expected mode of inheritance can be effective, but relies on accurate disease categorization to predict. Genome sequencing of parent-child trios can be used to identify novel gene mutations, in suspected de-novo, and inherited disease models [243]. Where there are two or more affected individuals within a pedigree, concordance filters can be useful to narrow variants to those present in all affected individuals [244]. Sequencing the two most distantly related individuals with the phenotype of interest can substantially reduce the number of variants under investigation. The proportion of bystander variants can be calculated for two individuals related by degree r using the following formula:

$$\text{Proportion of variants excluded} = (2^{r+1} - 2)/(2^{r+1} - 1)$$

Therefore, for cousins, one could expect to exclude 14 out of 15 bystander variants (or 93%). However, the total number of shared variants may still be too great to allow direct identification of the causal mutation in most situations.

Thereafter, remaining variants are often subsequently scrutinized in further detail and prioritised according to deleteriousness to protein and the gene's biological relevance to the phenotype. Putative genetic variants are confirmed experimentally to cause disease. This may include *in-silico* modeling of structural effect on the protein of the variant, *in-vitro* studies using the patient's cells or tissues, or *in-vivo* studies such as animal models, or gene therapy. Finally, genetic homogeneity is sought in other unrelated pedigrees to define the mode of inheritance and full extent of phenotype.

5.1.3 Novel gene associations by pathway analysis

Genetic heterogeneity in PID may be explained by the varied group of immune cells in humans and their diverse functional immune pathways. Immune cell pathways are complex and are tightly regulated, given their critical role in cell and ultimately, organism survival. Many are well characterised and have been utilised to find novel disease-causing genes [186].

Early successes of novel genetic causes in PID were based on family studies and well-defined phenotypic disorders. Absence of proteins in a common pathway which cause an identical phenotype (e.g. absent B cells and agammaglobulinaemia) have allowed identification of genes e.g. *CD79A* (Immunoglobulin α), *CD79B* (Immunoglobulin β), and *BTK* (Bruton tyrosine kinase) [186, 245].

Current efforts to find new gene associations in heterogeneous disorders include use network pathway analysis. *In-silico* analysis using pathway enrichment tools have

been used with success in more complex heterogeneous disorders such as autism and epilepsy [246, 247].

5.1.4 Novel gene associations

Genomic causes underlying rare, highly penetrant monogenic diseases are relatively easier to identify, as the mode of inheritance is readily established and the genotype is noted to be absent or at a very low frequency in health individuals. However, more common, complex disorders are influenced by a number of genomic phenomena including incomplete penetrance, locus heterogeneity and the presence of gene-environment interactions, which render the underlying genetic causes inherently more difficult to isolate. Genome-wide association studies (GWAS) are analyses used to detect associations between genetic variants and traits including diseases in samples from populations [248]. While the underlying genomic architecture of most common diseases is incompletely understood, the 'common disease, rare variant' hypothesis posits that a significant proportion of the missing heritability may be due to multiple rare variants in multiple genes, each acting dominantly, independently, and conferring a moderate increase in relative risk. There is growing interest in defining novel contributory variants by genetic association studies in primary immunodeficiency [39, 249]. A risk allele for primary immunodeficiency has been identified in a known PID gene, namely the heterozygous C104R variant in *TNFRSF13B* [33]. However, novel genes have yet to be identified using these methods.

Rare variants contributing to disease are unlikely to be found by single variant-based association testing, as statistical power of the association test relies upon both frequency of the variant and a large effect size to discriminate cases from controls. A rare variant would require a large effect size, and even with large effects, single rare variants can only be detected in large sample sizes. Instead of testing association

of rare variants individually, an alternative strategy is to group variants likely to have a similar function, into genes. This method, a burden test, allocates a score per individual on presence of rare variants in both cases and controls. Various burden tests exist and differ in the way that they take into account allele frequencies of individual variants and whether they take weighted combination of variants based on a priori information [250]. A limitation of burden tests is the assumption that all rare variants within a gene influence the phenotype in the same direction and unless, weighting is applied, to the same extent. To mitigate the directionality, bidirectional aggregation methods have developed, including the C-alpha test and sequence kernel association test (SKAT) [154, 155]. SKAT is a variant component multiple regression test which retains power in settings where neutral variants or variants with opposite direction of effects could result in loss of power, while the optimised sequence kernel association test (SKAT-O) test is recommended where there is no prior assumption of number or directionality of causal rare variants [126]. The SKAT-O has been shown to outperform burden tests, particularly where both protective and deleterious variants are present within a gene, and the underlying genetic architecture is largely unknown [127].

5.1.5 Association studies in PID

Primary immunodeficiency in children is thought to be caused by disruption of rare, Mendelian-inherited genetic variants causing early-onset disease. However, milder symptomatology in immunodeficiency, particularly common variable immunodeficiency (CVID) has identical clinical features for those who present as children, as those who develop symptoms in adulthood. CVID make up the largest group of PID, but collectively have fewer number of genes related to its cause.

Genome-wide association studies (GWAS) can be useful for identifying genetic variants that appear to increase or decrease susceptibility to various diseases, with

common variants identified by GWAS shown to be informative to identify novel biological processes [248, 251, 252].

In PID, three GWAS studies have been conducted: two conducted to discover genetic variation in CVID specifically, and the other to detect common variants responsible for IgA deficiency, the most common form of immunodeficiency [253-255]. These were all based on individuals of European ancestry. The largest study report only one significant loci using 363 CVID cases and total of 3319 samples: the major histocompatibility (MHC) locus [254]. The combined suggestive loci from all 3 studies are shown in Table 5.1.

Table 5.1: GWAS regions with significant SNPs associated with CVID [254, 255]

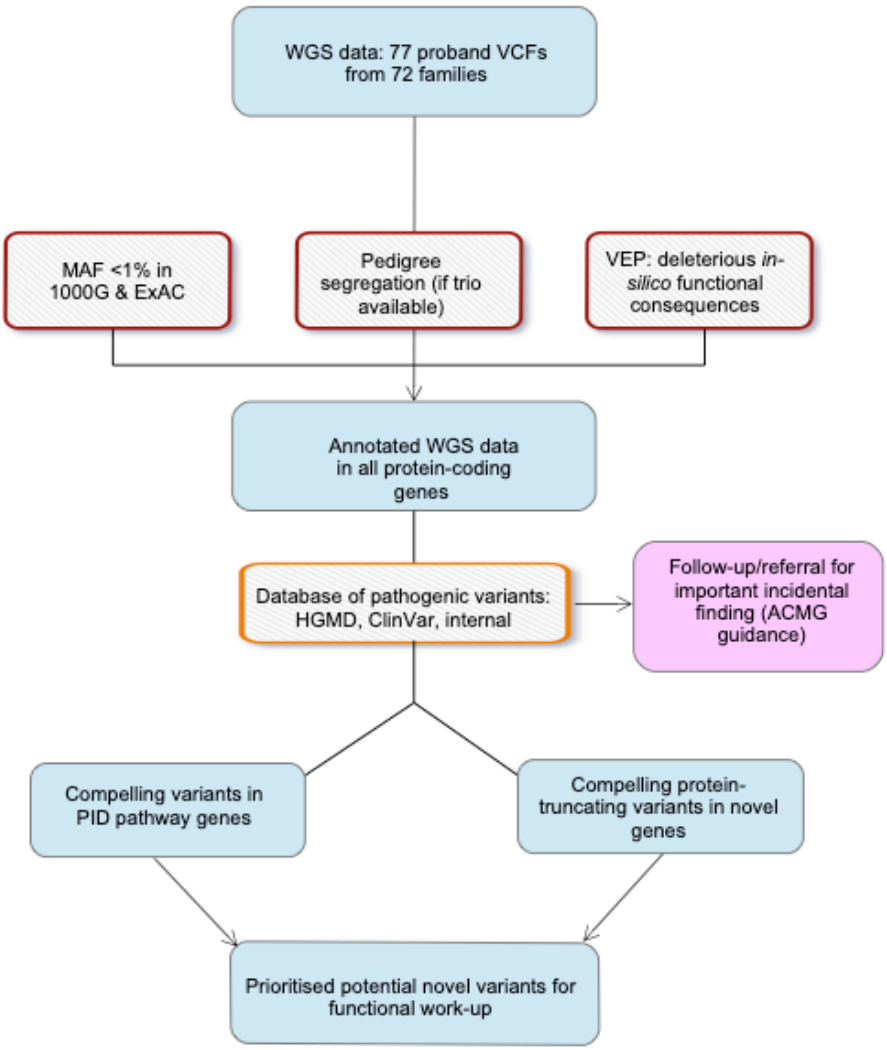
Locus/Position	Gene	Significant SNPs	p-value	Trait	Study
Chr 6: 27236785-32521295	*MHC*	rs3117426	8.6×10^{-10}	CVID	Orange et al (2011), Maggadottir et al (2015)
Chr 8: 23746576-24681608	*ADAM7*	rs4872262	6.24×10^{-6}	CVID	Orange et al (2011)
Chr 7: 4270929-4287047	*SDK1*	rs895710	4.7×10^{-5}	CVID	Orange et al (2011)
Chr 1: 20369369-20490791	*UBXN10*	rs7514144	9.09×10^{-5}	CVID	Orange et al (2011)
Chr 1: 59987116-59993733	*FGGY*	rs11207520	1.52×10^{-4}	CVID	Orange et al (2011)
Chr 2: 65518304-65520560	*FLJ16124*	rs1194849	5.2×10^{-4}	CVID	Orange et al (2011)
Chr 16: 31191482	*FUS*	rs929867	5.49×10^{-10}	CVID	Maggadottir et al (2015)
Chr 16: 31094089	*ZNF646*	rs750953	4.25×10^{-8}	CVID	Maggadottir et al (2015)

5.2 Results

5.2.1 Identification of candidate genes for further investigation within PID cohort

Analysis of the whole genome sequence for pathogenic variants capable of causing immunodeficiency, required filtering of the 174,738 variants within protein-coding genes, present in the cohort. Filtering the variants to just those that are predicted to be protein-truncating, generated 3,228 variants. During in-depth exploration of the list of genetic variants, it was noted that the majority of proteins were either not related to the immune system, or not expressed by immune cells. To ameliorate this, I identified critical immune pathways, often already featuring known PID genes to find PID candidate genes that function in immune cells or contribute to immune cell functioning (Figure 5.1).

Figure 5.1: Variant analysis pipeline to find novel genetic variants in PID candidate genes



Using the variant analysis pipeline to find novel variants in PID candidate genes, I identified a total of 2179 rare non-synonymous coding variants in PID candidate genes in 58 probands. The majority of variants were missense and splice region variants, with 47 protein-truncating variants remaining. None were homozygous or compound heterozygous, that would predict complete absence of expression of the protein. Of the heterozygous variants remaining, many did not segregate with disease. When only novel variants were sought (within family or population databases), only 10 remained. However, none were considered to be a good functional match for the patient phenotype.

As this approach was too stringent for a small cohort, I next focused on trios or informative pedigrees, where bystander variants could be readily excluded. Complete trios were present in only 9 of the remaining probands without PID genetic diagnosis. Heterozygous variants of likely pathogenic consequence were also considered which could cause disease by haploinsufficiency or gain-of-function mechanisms.

The cohort was also examined for known pathogenic variants in Human Genome Mutation Database. All of these were classified variants of uncertain significance, particularly as relevant symptoms were not present, and there was deemed inconsistent evidence of the pathogenicity of the variant in question. Thus, no further specific follow-up was warranted for the paediatric cohort.

Similar analysis was carried out for the adult patient cohort based at Royal Free NHS Foundation Trust. Two variants were selected for further investigation in two separate patients, one in a known PID gene (*RAC2*) and one in a PID candidate gene (*EP300*). Further exploration of these variants is set out in Chapter 6 of this thesis.

5.2.2 Gene variant-based association analysis

Due to the small sample size of the PID cohort, there is no statistical power to test for single variant association across the genome, or at known PID-risk loci [256, 257]. Gene-based association analyses are used to detect rare genetic variants that affect a disease, as it takes into consideration the correlations among SNVs within a single gene. This is done by comparing presence of rare variants in PID cases compared to population-matched controls. The method is similar to a SNV-based genome-wide association studies, but tends to have higher power as individual rare SNVs are likely

to have larger clinical effect, with cumulative clustering within a gene associated with disease. Multiple testing corrections are applied to control the type-I error rate.

The cohort examined was restricted to those with European or South Asian ancestry (67 cases, 320 controls), as they were the most plentiful in the PID cohort. By excluding more genetically diverse samples, I attempted to minimise the effect of population stratification.

Instead, I focussed on rare variants: firstly, analysis of variants with mean allele frequency (MAF) <0.01 in PID candidate genes, and then those MAF <0.1. I also used different burden tests and filtered further to functional variants, to identify noteworthy genetic variants.

5.2.2.1 Burden test

Association analysis was conducted across all protein-coding genes with MAF <0.01, where mean gene coverage was >80% at 20x (number of genes=18,674). The analysis covered 971,112 SNVs, with 79,765 in the case cohort. The quantile-quantile (QQ) plot for both analyses show the per-gene Z-statistic following a null distribution (Figure 5.2). This suggests that samples met strict quality control criteria, and that the population stratification of both cases and controls were appropriately matched. However, there is inflation at the start of the QQ plot, deviating away from the normal distribution. This is likely to occur when there is an imbalance between the number of case and control samples.

Figure 5.2: Quantile-quantile plot of burden test showing null distribution in red, and observed spread of p-values in black.

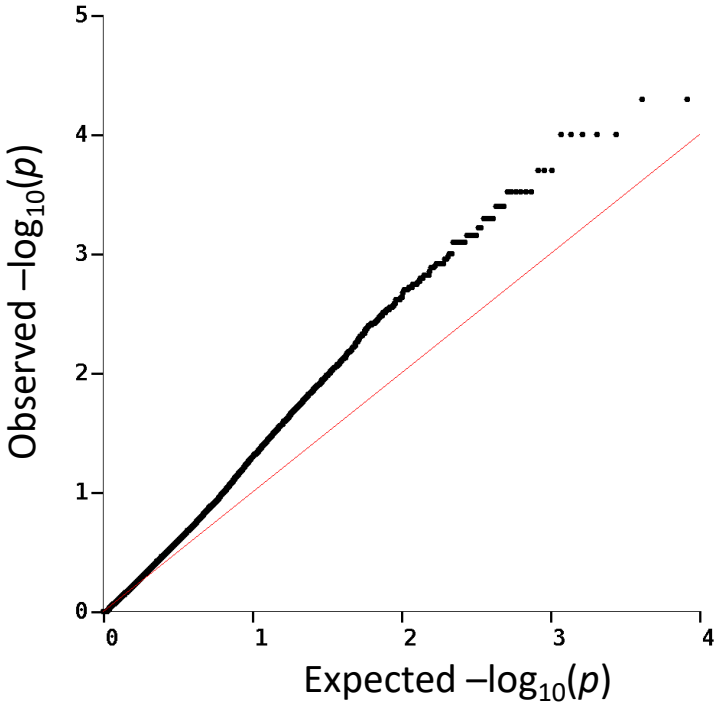
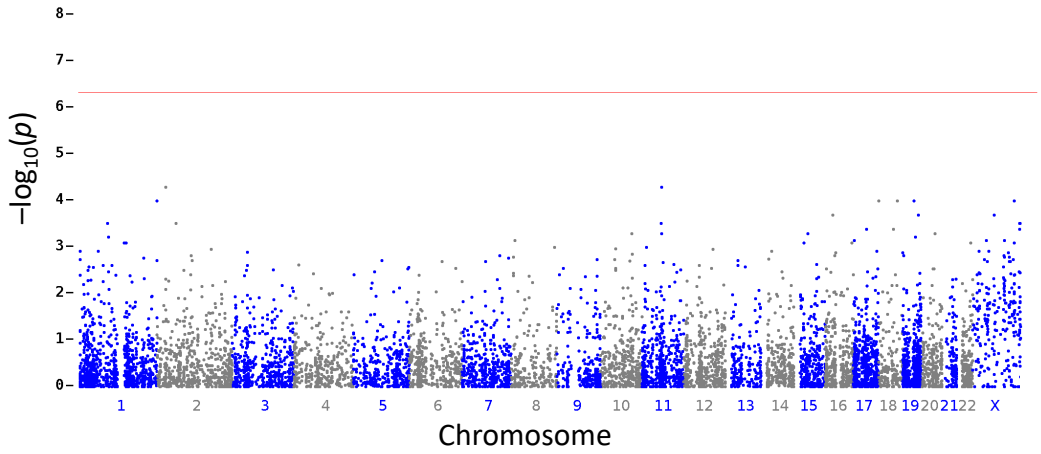


Figure 5.3: Manhattan plot of gene-based association analysis. Genome-wide association line in red (1.6×10^{-6})



The Manhattan plot show p-values obtained from gene burden tests (Figure 5.3). No gene reached genome-wide gene significance using the Bonferroni multiple correction (1.6×10^{-6}).

In order to pick out significant genes in burden tests, I used more stringent criteria by analysing burden in variants with MAF <0.01 and variants restricted to PID candidate genes (Appendix B). Again, no genes were noted to be statistically significant, but two genes were closer to significance than the others: *PPM1G* on Chromosome 2, and *TAF6L* on Chromosome 11.

PPM1G (Protein phosphatase 1G) encodes a manganese and magnesium-dependent Ser/Thr protein phosphatases which is required for spliceosome formation, and promotion of gene-specific transcription [258, 259]. It also has a role in the DNA damage response by downregulating hypoxia-inducible factors [260]. It is widely expressed, most abundantly in the skeletal muscle and heart [261]. It also interacts with an anti-apoptotic protein Nol3 which acts as a mediator of FAS and FADD receptors [262].

TAF6L (TATA-box binding protein associated factor 6-like RNA polymerase II) is a gene implicated in the initiation of transcription by RNA polymerase II. The gene product is a component of the p300/CBP-associated factor (PCAF) histone acetylase complex which is a key transcriptional coactivator of nuclear factor-κB (NF-κB) and p53, themselves key transcription factors. The PCAF complex has been shown to regulate the expression of NF-κB target genes in response to TNF-α stimulation [263, 264]. The gene product has high expression in plasma, liver and CD8 T cells.

The burden test collapses variants into genes and assume that all rare variants influence the phenotype in the same direction, and are equal in magnitude. This can create noise into the combined index, and may cause false positive associations.

5.2.2.2 SKAT-O test

The optimised sequence kernel association test (SKAT-O) is similar to a burden test for association of disease variants in cases compared to controls, but uses a multiple regression model to directly regress the dichotomous phenotype on different variants, allowing for effects of different directions and magnitude [156].

No gene achieved significance after multiple testing corrections were applied (Figure 5.4 and Figure 5.5). Interestingly, different genes were highlighted in the analysis: *FFAR3* (chromosome 19), *GBP2* (chromosome 1), *LAT2* (chromosome 7), and *FAM122B* (chromosome X) with p-values between 0.000123 and 0.000238 (Table 5.2).

Figure 5.4: Quantile-quantile plot of SKAT-0 test showing null distribution in red, and observed spread of p-values in black.

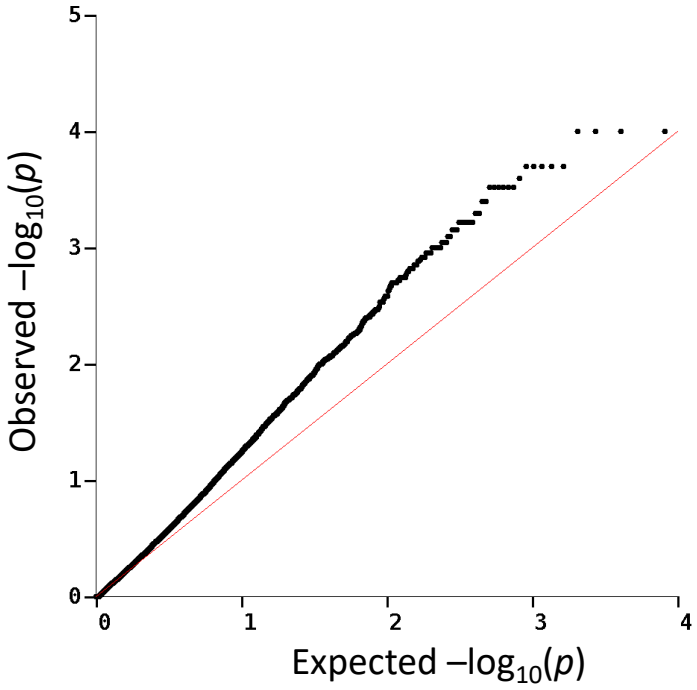


Figure 5.5: Manhattan plot of gene-based association analysis. Genome-wide association line in red (1.6×10^{-6})

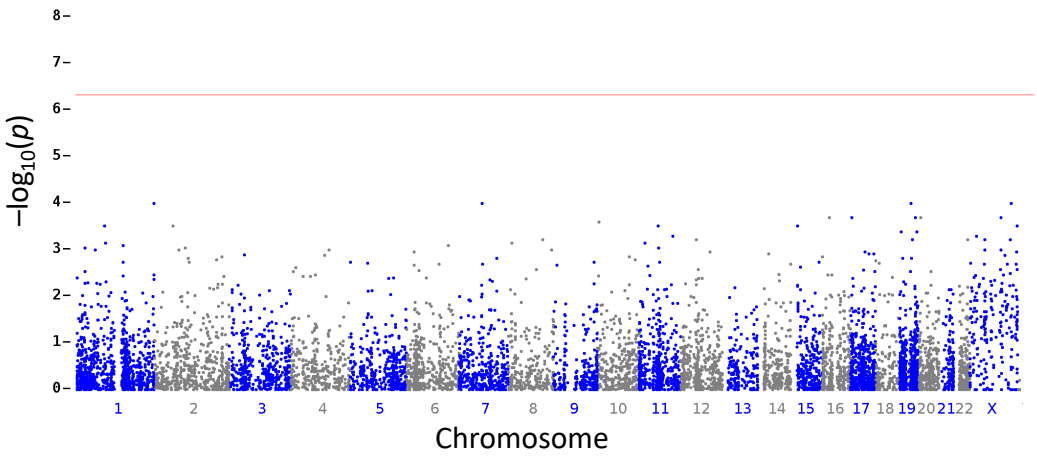


Table 5.2: Top 10 genes and adjusted p-values identified by case-control comparison using SKAT-O

Gene	Locus/Position	SKAT-O p-value
FFAR3	Chr 7: 4270929- 4287047	0.000123
GBP2	Chr 6: 27236785-32521295	0.000187
LAT2	Chr 8: 23746576- 24681608	0.000207
FAM122B	Chr 1: 20369369- 20490791	0.000238
HCFC1	Chr X:153213007-153236819	0.008964
CDH13	Chr 16:82660398- 83830215	0.002946
AKAP10	Chr 17:19807764-19881169	0.002946
BSG	Chr 19:571276- 583493	0.002946
ZNF334	Chr 20:45128268- 45142198	0.006765
FAM69A	Chr 1:93298285- 93427079	0.006765
BRD3	Chr 9:136895445-136933141	0.006765

Genes highlighted in the association analysis

GBP2 (Guanylate-binding protein 2) lies on chromosome 1 and is implicated in immune cells and pathways. The GTPase protein is interferon γ-inducible and primarily hydrolyses GTP to GDP in macrophages, thereby facilitating oxidative killing via transport against bacteria [265]. There is high expression in immune derived cells including lymph nodes, spleen and PBMCs. Although implicated in critical immune pathways, with interactions with NF-κB and Rac proteins, the gene itself is yet to be solely implicated in a primary immune disorder [266].

LAT2 (Linker For Activation Of T Cells Family Member 2) lies on chromosome 7. It is commonly one of a number of genes deleted in one allele in Williams syndrome, a multisystem developmental disorder; itself not primarily associated with immunodeficiency. While the *LAT* gene appears critical for T-cell function with loss of protein causing severe combined immunodeficiency, *LAT2* may have a role in

activation of B cells [267, 268]. It is highly expressed in immune cells such as B cells, NK cells, monocytes and PBMCs, and notably absent in T-cells.

FFAR3 (Free Fatty Acid Receptor 3) is a G-protein coupled receptor gene found on chromosome 19. Its primary role is in glucose homeostasis, activated by short chain fatty acids in the gut [269]. It appears to have a role in intestinal immunity by regulating cytokine response in intestinal epithelial cells [270, 271]. It is expressed primarily in blood cells and adipose tissue.

FAM122B (Family with Sequence Similarity 122B) lies on chromosome X and encodes an intracellular phosphopeptide. The function of the protein is not well defined, but interacts with protein phosphatase 2, regulatory subunit B, gamma (*PPP2R2C*) and implicated on PI3K-akt signalling pathway (Kegg pathway 04151). However, there is low protein expression of the protein in immune cells.

To focus on variants more likely to be impactful, I ran the SKAT-O analysis considering only non-synonymous variants within the coding region (missense, frameshift, stop, splice-acceptor and splice-donor SNVs), with rare variants weight applied (MAF<0.001). From my analysis, no gene surpassed the *p*-value significance threshold (1.6×10^{-6}). Considerable inflation was also noted in the QQ plot, due to the lack of rare coding variants per gene in the analysis (Figure 5.6).

Figure 5.6: Quantile-quantile plot of non-synonymous SKAT-0 test showing null distribution in red, and observed spread of p-values in black.

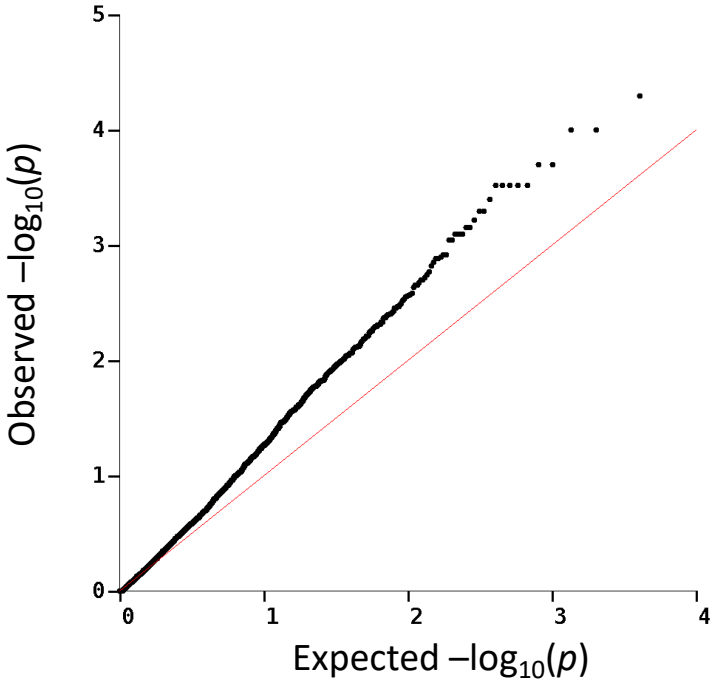
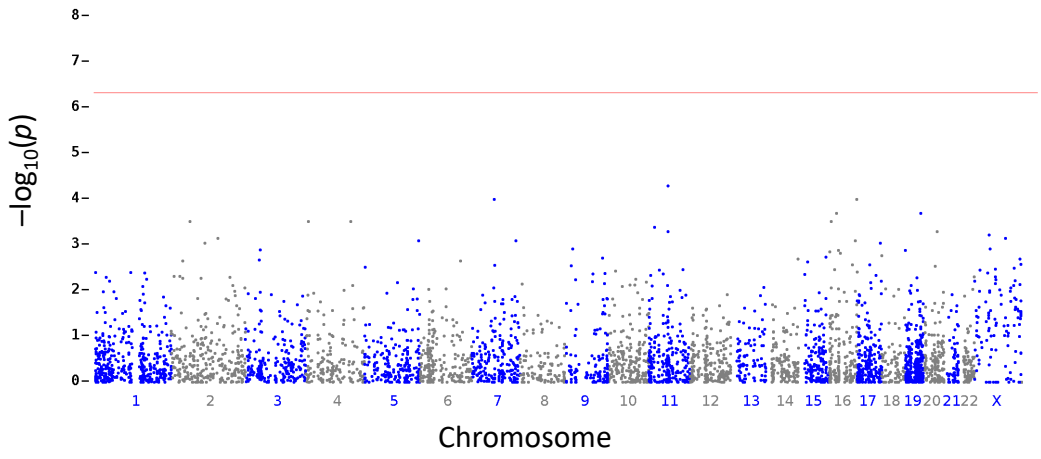


Figure 5.7: Manhattan plot of gene-based association analysis. Genome-wide association line in red (1.6×10^{-6})



The analysis found more coding variants present in *TAF6L* gene, in the cases more than the controls (p=0.00042) (Table 5.3). Other coding variants with a p-value <0.001 were found in the *LAT2* and *CBFB* genes are also shown in Table 5.3.

The functions of *TAF6L* and *LAT2* genes have been highlighted above.

CBFB (Core-Binding Factor Subunit Beta) gene on chromosome 16, encodes a component of the heterodimeric core-binding factor complex with the transcription regulatory RUNX proteins. The beta component has specifically been shown to enhance the DNA-binding capacity of T-cell enhancers and is critical to haematopoietic stem and progenitor cell formation [272-274]. There is high expression of *CBFB* in PBMCs, and T cells.

Table 5.3: Rare variants aggregation tests of SKAT-O for PID cohort: coding variants of significant genes

Gene	Chr:position	Variant	Consequence	No. of variants in cases (n=67)	No. of variants in controls (n=320)	ExAC MAF	BURDEN p-value	SKAT-O p-value	SKAT-O coding variants p-value
TAF6L	11:62538774 - 62554813						0.00005	0.0018	0.00042
	11:62543326	c.71C>T, p.T24M	missense	5	0	0.008484			
	11:62545490	c.275G>A, p.R92H	missense	1	0	0.000083			
	11:62545525	c.310G>C, pD104H	missense	2	0	0			
	11:62549397	c.563A>G, p.K188R	missense	1	0	0			
	11:62549768	c.790C>T, p.264W	missense	2	0	0.0000083			
	11:62554113	c.1214C>G, p.S405W	missense	4	1	0.001937			
	11:62554194	c.1295C>T, p.S432F	missense	2	0	0.0002201			
LAT2	7:73624086 - 73644164						0.001	0.00021	0.0007
	7:73631155	c.95G>T, p.G32V	missense & splice region	2	0	0.0004979			
	7:73634076	c.137G>A, p.R46H	missense & splice region	1	0	0.0002571			
	7:73634295	c.184G>A, p.G62R	missense	2	0	0.00004205			
CBFB	16:67063049 - 67134958						0.001	0.028	0.0009
	16:67116140	c. 424G>T, p.A142S	missense	1	0	0			
	16:67116156	c.440G>A, p.R147Q	missense	1	0	0			

5.3 Discussion

To the best of my knowledge, this is the first time that WGS has been used to find novel genes using both pathway analysis, or rare variant association studies in a cohort of children affected by PID. By considering the whole genome, and focussed examination of key regions, this bioinformatic project represents a proof-of-concept study that can be built upon on a larger scale in future work.

5.3.1 PID-pathway gene screening to identify genetic causes of PID

Using bioinformatic tools, I identified 3082 PID candidate genes known to interact with previously-described PID genes (Appendix B). Deleterious variants harboured in these genes allow for a further variant filtering tool, which can be utilised in WGS, where deleterious variants are often plentiful, but may be biologically harmless.

WGS is a valid strategy to undertake in the identification of genes likely to cause PID. I developed a strategy to use a targeted screening of potential candidate genes. I used this strategy to identify a gene associated with the genetic neurodevelopmental syndrome Rubinstein-Taybi syndrome *EP300*. The novel variant identified, fit the clinical manifestations and immuno-phenotype of the patient, and is explored further in Chapter 6. This finding adds further strength to a growing body of recent evidence suggesting that defects in genes responsible for orphan monogenic conditions, can underlie PID phenotypes, because many syndromes encompass immune pathology or present with susceptibility to infections [178, 275, 276]. My finding of a variant in *EP300* is consistent with very recent case reports which report immunodeficiency with and without co-existing Rubinstein-Taybi syndrome.

Routine use of NGS technologies for genetic diagnosis has not yet become established practice, but these results suggest it is likely to be fruitful to screen as wide a gene-list as possible, especially given the non-specific clinical phenotypes of PID patients.

5.3.2 Rare variant case-control analyses to find novel associations in PID

Association studies are a useful approach particularly where the variant's contribution to disease is undefined. Genome-wide association studies to detect common single-variant indicative of disease require large sample sizes for adequate statistical power; sample numbers which were not available in this study [256, 257]. The alternative approach to aggregate rare variants within a functional unit such as a gene, which can be interrogated for increased burden of rare variation. Using this method, I have identified three PID candidate genes: *TAF6L*, *LAT2* and *CBFB* in which rare coding SNVs are enriched in our paediatric cohort of patients with PID. However, the difference in enrichment between our cohort and the control group, was not statistically significant upon multiple correction methods.

This finding suggests that there is no overall difference in frequency of gene variants in PID cases compared to controls. However, my case cohort may be limited for several reasons. Firstly, it is likely that my analysis is underpowered, according to power calculations generated by The et al (2015). Secondly, burden tests are best utilised when there is clear phenotyping for a particular trait. Given the large number of immune cells implicated, and the diverse phenotypes exhibited, it is increasingly difficult to treat PID as one unifying phenotype. Thirdly, only genes with adequate coverage were included for assessment, and there is a chance of a true association signal being missed due to low depth of coverage. Use of WGS data instead of genotyping arrays, was initially not to be used for association studies particularly as there is relatively low coverage per individual. However, given the wide ranging

capability of WGS, it is perhaps the optimal modality for conducting genome-wide association studies. Indeed, it is now likely to be more sensitive and more amenable to fine mapping regions of interest.

Burden tests are limited in that the rare variants analysed, are assumed to influence the phenotype in the same direction and with equal magnitude [127]. They have also been shown to have low statistical power for rare disease genetic models, particularly where common and rare variants within a region have little, no or unknown effect [126].

There is currently a lack of independent WGS cohorts for case and control replication studies. Although it is plausible to conduct a genome -wide association study if large enough sample size is collected, this approach is agnostic to function and therefore, has less power to detect true signals than those with reliable prior knowledge of genomic function. However, it has the potential to uncover groups of putatively functionally correlated rare variants within regulatory regions. The variants highlighted by these studies could be causal, but are more likely to be associated indirectly by being correlated other genomic variants. GWAS are mostly conducted via gene-microarray chips, and may not completely accommodate the particular type of variant, e.g. insertion/deletion variant (indels) or copy number variants (CNVs). It is also likely that the genetic effects are highly likely to be probabilistic and variable with external factors, rather than causative as observed in Mendelian disorders.

5.3.3 Clinical interpretation of rare variant association studies

Novel PID loci have been identified through association studies including *FUS* and *ITGAM* [254, 255]. The MHC (major histocompatibility complex) locus is known to be highly polymorphic with high levels of linkage disequilibrium [277]. As more data

becomes available via large-scale genomic studies, it is expected that more loci will be identified through a variety of approaches focussed on common and low-frequency variation, and collaborative meta-analyses in multi-ethnic populations.

Despite progress in methods of identification, the task of determining causal variants among association loci remain challenging. This work will undoubtedly be enhanced by improved understanding of regulatory regions of DNA at a genome-wide level.

The genes that were discovered in my analysis were not significant on SKAT-O analysis. However, it is interesting to note that two of the genes enriched in this analysis are *TAF6L* and *CBFB* which both appear to have a role in transcription and downstream regulation.

TAF6L (TAF6-like RNA polymerase II) is a component of the p300/CBP-associated factor (PCAF) histone acetylase complex which is a key transcriptional coactivator of nuclear factor-κB (NF-κB) and p53, both implicated in immunodeficiency [278]. Given the role in histone acetylation, it likely has many targets, and acts in several pathways [279, 280]. As a gene, there is little known about its function specific in immune cells.

CBF-β (core binding factor-beta) subunit is also part of a complex regulating transcription via immune-specific RUNX (runt-related transcription factor) transcription factors [272, 281]. These proteins act in the transcriptional control of lineage decision during T-cell development, and there is evidence that they support differentiation of naïve CD4+ T cells into different subsets of helper and regulatory T cell, and in extrathymic differentiation of T-cell progenitors [274, 282, 283]. Given this finding in a mixed cohort of children, some of whom have absent CD4+ or CD8+ T cell populations, it is not unreasonable to hypothesise that there is a role for *CBFB*

in lineage-specific fate of T-cell populations. With the potential to disrupt T-cell differentiation, *CBFB* is an interesting candidate gene for further investigation.

5.3.4 Novel gene discovery requires functional validation

Current validation of novel genetic variants would usually require functional analysis as evidence of disease pathogenicity. Statistical causation and replication are also important considerations, particularly in analysis of rare diseases. The focus of these analyses remains on coding variants which can be analysed in protein studies. However, the reality is undoubtedly more complex with interaction of regulatory aspects of the genome.

Early whole genome studies revealed that diverse specifies from yeast, mouse and primates share the common functional genes [284]. Animal models of disease can provide valuable insight into a gene's functional effect, with the assumption that disruption of the orthologous gene in humans will have a similar phenotype. They can guide the investigator as to the likelihood of the gene's potential pathogenicity, and in variant analysis, can be used as a late filtering tools to narrow candidate genes. The International Knockout Mouse Consortium aims to mutate all protein-coding genes in mice, but applicability to human disease, would only be relevant in loss of function disease [285, 286].

The main driver of rare variant heritability is natural selection with most rare variants that impact function expected to have deleterious impact. In rare diseases such as PID, rare variant analysis will likely remain the focus of continuing PID research. However, the genetic architecture is incompletely understood, with some evidence of genetic risk factors suggesting a polygenic basis, or incomplete penetrance [33]. Genome wide association studies were set up to discover master genes behind the

phenotype of interest, for example, genes that drive transcriptional responses that may act as therapeutic targets [287]. However, in PID, many master genes are well-defined, with loss of function or absence causing early-onset completely penetrant severe disease [7, 157].

5.3.5 Future work

There are still challenges in applying statistical methods to rare variant association tests, especially when the sample size is small. Future work should focus on burden testing of larger groups of immunophenotyped individuals, where clear genic associations may become statistically significant.

Use of WGS for association studies is growing, with initial discoveries of rare informative alleles suggesting highly penetrant rare variants are likely to contribute to phenotypic variance in commonly encountered diseases such as coronary heart disease [288]. Many different burden, component and combined rare-variant association tests are available, but their respective use in different diseases is not yet defined. Establishing a genetic architecture in both common and rare disease through use of association studies should provide a clearer picture as to the scale of rare and more common genetic variation in diseases such as PID and guide future analyses.

Future studies of enriched rare coding SNVs could assess their individual function through gene reporter assays and binding affinity for haematopoetic-specific transcription factors. Another approach is to overlay larger disease-association regions with chromatin marks or regions of DNase I hypersensitivity which suggest open chromatin and active transcription [289].

Environmental influences including exposure to infections have been associated with PID [136]. New methods of association study such as epigenome-wide association study (EWAS) allow epigenomic interrogation in PID cases [290, 291]. As this has yet to be explored in PID, it would certainly yield noteworthy results that may be functionally and clinically relevant.

6 NOVEL GENETIC VARIANTS IN PATIENTS WITH PRIMARY IMMUNODEFICIENCY

In this chapter, I will describe two different patients with novel genetic variants in disease genes. I have investigated the genetic defect in closer detail, to build evidence of the genotype's pathogenicity.

The patients are unrelated young adults who were recruited to the BRIDGE-PID project from Royal Free Hospital NHS Foundation Trust. They both had disease symptoms in childhood. WGS data was accessed with kind permission via the HPC at University of Cambridge. Quality control measures were undertaken as described previously, with quality control metrics held with the patient genetic data. I have not included a formal WGS analysis of the whole adult cohort with immunodeficiency as this is beyond the scope of this thesis and is on-going work. However, I chose to functionally analyse these clinical variants in more detail, and present the work in this thesis.

All the work presented is my own work, unless otherwise stated. This research was carried out at the UCL Institute of Immunity and Transplantation at Royal Free Hospital under the supervision and guidance of Dr Siobhan Burns.

Some additional functional data is presented in this chapter, undertaken by other colleagues and it is explicitly stated where this has occurred.

6.1 Introduction

6.1.1 Use of whole genome sequencing in discovery of novel genetic variants

Whole genome sequencing (WGS) has recently been utilised to discover novel disease predisposition genes in PID and in other rare Mendelian disorders [99, 100, 106, 180]. This potential to uncover pathogenic variants in all genes, distinguishes it from gene panel and targeted Sanger sequencing that seek out known disease associations in genes predicted to cause disease.

Genetic variants that have previously been attributed to disease make the genetic diagnosis more straight-forward. With increasing availability of combined genetic and phenotypic data, and with greater confidence in the sequencing quality, more accurate, prognostic information may be delivered to the patient, and potentially impact treatment decisions. Presently, there are relatively few PID genes where functional diagnostic tests are available in accredited laboratories. This has the potential to hamper future investigations of variants of unknown significance within known PID genes.

6.1.2 RAC2

RAC2 (Rac Family Small GTPase 2) gene is located at chromosome 22q13, and is approximately 19.1kb in size. There are 7 alternative transcripts ranging from 86 to 192 amino acids long. The consensus sequence (CCDS) consists of 7 exons, and is expressed at highest levels in immune cells, specifically T cells, NK cells, neutrophils, and monocytes [292-294].

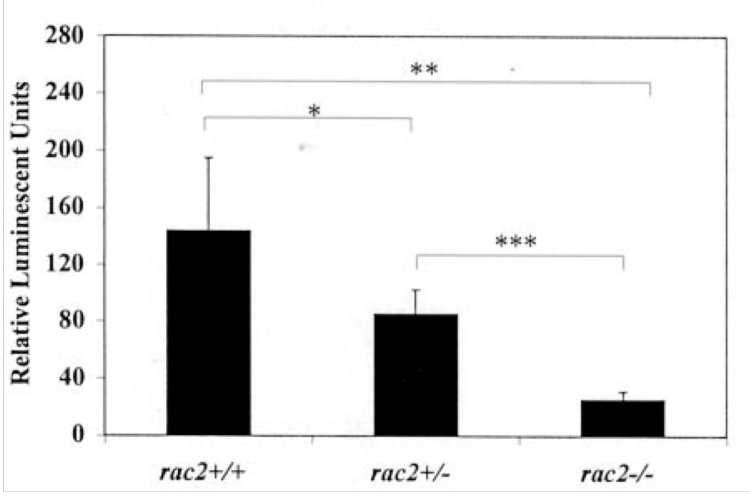
RAC2 is a 192-amino acid protein belonging to the Rho family of guanosine triphosphatases (GTPases) [295, 296]. Secondary structure prediction indicates that

the protein is composed of 8 alpha-helices connected by beta strands, to form a compact twisted beta sheet. Protein domains include two switch regions and a differentiating C-terminal region [297, 298].

Like other Rho proteins, Rac2 functions as a GTPase signal transducer, becoming active and interacting with downstream effector molecules when bound to GTP [299]. The protein cycles between GTP and GDP bound states, regulated by guanine nucleotide exchange factors. Rac2 has been shown to have non-redundant functions, particularly as it is the predominant Rac protein in immune cells [293, 300-302]. Rac1 is expressed ubiquitously and Rac3 expressed in the brain and nervous system [298, 303, 304]. Rac1 and Rac2 appear to share some functions, but have differential protein expression with Rac1 having greater expression in neurons [300, 305]. Rac1 missense mutations have recently been implicated in neurodevelopmental delay [306].

Both Rac1 and Rac2 have been shown to be critical in cell migration, particularly at the leading edge of leucocytes, where lamellipodia and filopodia are extended by actin polymerisation to drive forward the plasma membrane [307, 308]. Neutrophils deficient in Rac2 have been shown to have undirected migration caused by defective filamentous (F-) actin assembly [296, 309]. In response to fMLP (N-Formylmethionyl-leucyl-phenylalanine) stimulation, $Rac2^{-/-}$ neutrophils in mice are impaired in the onset and extent of actin polymerization responses to FMLP-induced polymerised F-actin [299, 309]. The level of activated Rac2 is titratable for fMLP-induced F-actin polymerisation and NAPDH-oxidase activity of neutrophils [296] (Figure 6.1).

Figure 6.1 Graph taken from Li et al (2002) [296]. The authors have demonstrated significant reductions in f-MLP-activated NAPDH oxidase activity (measured in relative luminescent units) in Rac2 normal (+/+), haploinsufficient (+/-) and null (-/-) mouse neutrophils.



Rac2 is the main GTPase implicated in the production of reactive oxygen species (ROS) by NADPH oxidase in neutrophils predominantly [302, 310]. In addition, Rac2 is required for formation of neutrophil extracellular traps (NETs) which are thought to contribute to pathogen control [311-313].

Genetic variants in *RAC2* have been shown to cause distinct immunodeficiency syndromes. Neutrophil Immunodeficiency syndrome (OMIM 608203) is a rare, autosomal dominant syndrome characterised by early onset severe bacterial infections, poor wound healing and functional neutrophil defect. To date, only 2 cases have been described in the literature: each with the same genotype *RAC2* c.169G>A p.D57N (NM_002872) and with equivalent clinical presentation [54, 314]. The variant acts as a dominant negative mutation; the resulting protein causes disease by interfering with the GTP binding, to inactivate the protein [314]. A differing distinct phenotype is described associated with a homozgyous rare stop gained Rac2 mutation (*RAC2* c.168G>A p.W36X (NM_002872)). These patients were

in a consanguineous kindred and both have early-onset common variable immunodeficiency (CVID) [315]. This variant caused complete absence of the protein, but appears to cause different clinical features including falling B cell count, immunoglobulins, moderately-reduced neutrophil chemotaxis and normal oxidative burst using phorbol myristate acetate (PMA) [315]. More recently, a *RAC2* variant (*RAC2* c.184G>A p.E62K (NM_002872)) has been associated with combined immunodeficiency, cytoskeletal defects and progressive loss of white cell count, particularly B cells. Functional data in this case suggests a gain-of-function variant with excessive superoxide production and impaired migration to fMLP (N-formylmethionyl-leucyl-phenylalanine) [316].

6.1.3 EP300

EP300 (E1A-associated cellular p300 transcriptional co-activator) is a gene located on chromosome 22q13. Three protein-generating transcripts have been found with the canonical sequence being 9587 bases long. The resulting protein is 2414 amino acids long, with expression greatest in bone marrow stromal cells. This protein's structure contains several conserved protein motifs including a bromodomain, and CREB domain [317]. Bromodomains contain lysine residues which are differentially acetylated to guide protein-histone docking, while the CREB domain is specific to p300 and CBP (CREB binding protein) as an activation site between the protein and specific transcription factors.

The resulting protein p300 functions as a histone acetyltransferase that regulates transcription via chromatin remodelling [318, 319]. Histone acetyltransferases (HAT) are enzymes responsible for transfer of acetyl group to conserved lysine amino acids on histone proteins, to allow transcription factors to bind more easily to DNA [320]. In general, HATs increase DNA transcription and gene expression. P300 also interacts with Rel-A (p65) component of NFκB, with p300 depletion studies showing decreased

Rel-A-associated histone kinase activity by 75% [264]. Rel-A haploinsufficiency is known to cause associated with increase in $CD21^{low}CD38^{low}$ B cells, suggesting an impact on B cell homeostasis [32].

Rubinstein-Taybi syndrome (RTS) (OMIM 180849) is a rare well-described syndrome, predominantly of intellectual disability, postnatal growth restriction, characteristic facial features and broad thumbs and halluces [321, 322]. Pathogenic heterozygous variants in two genes are known to cause RTS, *CREBBP* causing RTS1 and *EP300* implicated in RTS2 [323, 324]. The mutational spectrum in both genes range from whole gene deletions to missense mutations with no particular correlation found between the genotype and the severity of phenotype. Variants in *EP300* account for the minority of individuals with RSTS (<10%). Both types are inherited in an autosomal dominant manner and typically arise de novo in a pedigree with no family history of the disorder. Individuals with RTS2 appear to have milder features than RSTS1 and have also been shown to have an increased incidence of tumours, with risk of malignant tumours rising with age [325, 326]. RTS has been inconsistently associated with variable B-cell defects [327, 328]. Specific *EP300* variants relating to co-existing RSTS and B-cell deficiency have only recently been reported; a heterozygous missense variant in the HAT region of the protein [329].

6.2 Results and Discussion of *RAC2* case

By screening genes associated with known PID disorders, I identified a novel likely-pathogenic variant in *RAC2*. Variant analysis and pipeline used are described in Chapter 5.2.

6.2.1 Clinical summary

Patient 1053 is a 40-year old male who was referred to the adult immunology service at the age of 20 with low immunoglobulins. He reported numerous chest infections as a child and teenager and was found to have epithelioid granulomas on liver biopsy and splenomegaly at presentation. He was diagnosed with granulomatous-type common variable immunodeficiency (CVID) and has since developed cutaneous, ocular and cerebral granulomas, the latter implicated in the patient's seizure disorder. Ocular granulomas have caused diplopia and he has been intermittently treated with steroids for this. Over the last 10 years, he has mainly had upper and lower respiratory tract infections, and frequent sinusitis. He currently has bronchiectasis with *Pseudomonas aeruginosa* colonisation, and is treated with nebulised Colomycin for this. Notably, he had persistent *Campylobacter jejuni* infection with prolonged episode of bacteraemia in 2016.

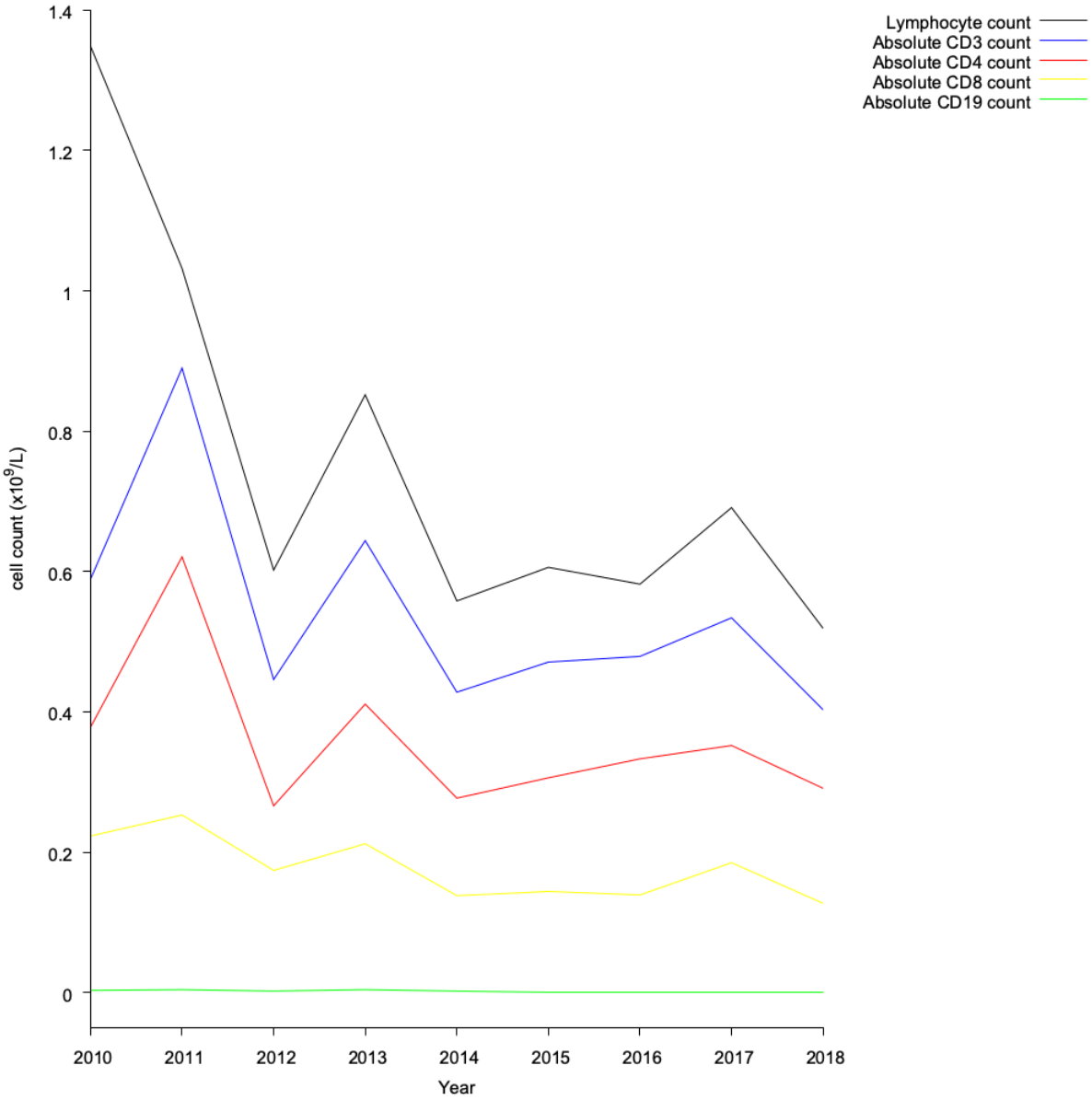
He has associated auto-immune phenomena, a finding not uncommon in adults with CVID. He was initially diagnosed with neutropaenia and thrombocytopaenia at age 19, with identification of anti-platelet and neutrophil IgG antibodies, confirming the diagnosis.

He had low lymphocyte cell count at diagnosis (0.9 x10^9/L, normal range 1-2.8 x 10^9/L) which has steadily fallen in the last 9 years (Figure 6.2). He also now has virtually

absent B cells with a related antibody defect. He also has T cell-lymphopenia and borderline low $CD4^+$ T cell count (Figure 6.3). Immunoglobulin levels were low at presentation with IgG level 0.1g/L (normal range, 7-16g/L), IgA <0.1g/L (normal range, 0.7-4g/L), except IgM which has remained at normal levels throughout, latest count 2.2g/L. He also has autoimmune neutropenia, latest counts 0.73 x 10^9/L. (normal range 1.7-7.5 x10^9/L) and thrombocytopaenia 113 x 10^9/L (normal range 140-400 x10^9/L). His other standard blood results are within normal range.

He is treated with intravenous immunoglobulin replacement which is administered 3-weekly, sodium valproate and antibiotics as required. Sodium valproate is well recognised to cause secondary hypogammaglobulinaemia and B cell lymphopenia. However, given the other clinical features, it is unlikely to be responsible for the full spectrum of immunodeficiency seen in the patient.

Figure 6.2: Lymphocyte counts of the patient over the last 9 years. Normal ranges ($x10^9$/L): Lymphocyte: 1-2.8, Absolute CD3 count: 0.7-2.1, Absolute CD4 count: 0.3-1.4, Absolute CD8 count: 0.2-0.9, Absolute CD19 0.1-0.5



6.2.2 Genetic diagnosis of *RAC2* variant

WGS analysis of coding regions of known PID genes found 3 rare variants (MAF <0.01) (Table 6.1). No significant large deletions nor copy number variants (CNV) were found in this patient.

Table 6.1: Rare coding variants found in patient in known PID genes. Canonical transcripts were used to derive in-silico predictions (Het – heterozygous)

Gene	Variant type	Nucleotide change	Amino acid change	Zygosity	SIFT	PolyPhen 2	CADD	ExAC MAF
PTPRC	Missense	c.982A>G	I328V	het	Benign	Tolerated	0.086	0.0004
RAC2	Missense	c.106G>A	V36M	het	Deleterious	Probably damaging	34	0
CSF2RA	Missense	c.191A>G	D64G	hemi	Tolerated	Benign	11.1	0.0005

Both *PTPRC* and *CSF2RA* genes cause disease in autosomal recessive and in X-linked recessive inheritance patterns respectively. The *RAC2* variant was selected for further investigation, given the greater predicted pathogenicity scoring, and was not found in normal (ExAC, 1000G), disease databases (HGMD, ClinVar), nor in BRIDGE data (data not shown).

PTPRC (Protein tyrosine phosphatase, receptor type C, also called CD45) gene variants cause autosomal recessive severe combined B+, T-, NK+ immunodeficiency, with only homozygous or compound heterozygous variants (OMIM: 608971) [330, 331]. The *CSF2RA* gene is located in the pseudo-autosomal region of the X and Y chromosomes. Genetic variants found in the pseudo-autosomal region are effectively handled by the variant annotation software, and a rare Y chromosome variant was

not found. Biallelic *CSF2RA* variants can cause pulmonary alveolar proteinosis (OMIM: 300770). This condition is characterised by early-onset diffuse interstitial lung disease, due to impaired surfactant homeostasis [57]. There is often co-existing hypogammaglobulinaemia, but otherwise, the clinical features described in the literature do not fit our patient's phenotype.

The variant *RAC2* c.106G>A, p.V36M NM_002872 (het) was found only in the patient within the BRIDGE database. Unfortunately, other family member data was not available for segregation studies. The *RAC2* variant was also not found in other genetic databases ExAC and 1000G. In-silico pathogenicity predictor tools suggest that the *RAC2* missense variant is expected to be cause disruption to the protein, including a CADD score >30, suggesting high pathogenicity. The valine base is a conserved site across several species including mammals such as chimpanzees and in zebrafish (Ensembl data). Other rare *RAC2* variants were sought in the non-coding space of the gene, but none were found.

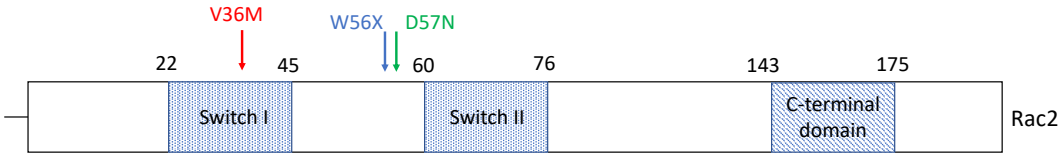
The variant is located in a splice region, at the start of Exon 3 (Figure 6.3). To explore the impact of the splice region variant further, I used MaxEntScan to compare the predicted splicing scores from the reference to the variant found in the patient [146]. MaxEntScan uses a maximum entropy model and scores 5' splice sites on modelling of short sequence motifs. My analysis showed the missense variant shift from guanine to adenine, increased the maximum entropy score from 7.42 to 7.83. In a maximum entropy model as utilised by MaxEntScan, a score deviation of 30% caused by a base substitution is likely to disrupt the splice site. This result suggests the original splice donor site is preserved and the mutation in this case does not affect the splicing of the mRNA.

Figure 6.3: The full length Ensembl sequence for human RAC2 encoded amino acids for transcript ENST00000249071 (https://grch37.ensembl.org/index.html). The blue and black font indicate alternating exons. Amino acids highlighted in red font indicate a residue overlap splice site

MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVMVDS
KPVNLGLWDTAGQEDYDRLRPLSYPQTDVFLICFSLVSPASYENVRAK
WFPEVRHHCPSTPIILVGTKLDLRDDKDTIEKLKEKKLAPITYPQGLA
LAKEIDSVKYLECSALTQRGLKTVFDEAIRAVLCPQPTRQQKRACSLL

The variant lies central within the Switch I domain, which contrasts to other pathogenic variants previously described in *RAC2* disease (Figure 6.4). The Switch I domain is thought to play a key role in GTP-binding activity of all RhoGTPases including Rac2, by undergoing conformational change and allowing binding of downstream signal transducers to effect key neutrophil functions such as chemotaxis and reactive oxygen species granule release [332, 333].

Figure 6.4: Representative schematic of pathogenic variants identified in RAC2. The variant in red font is seen in Patient 1053. The variant in blue is the loss of function homozygous stop gained mutation and the mutation in green is the heterozygous missense variant associated with neutrophil immunodeficiency syndrome

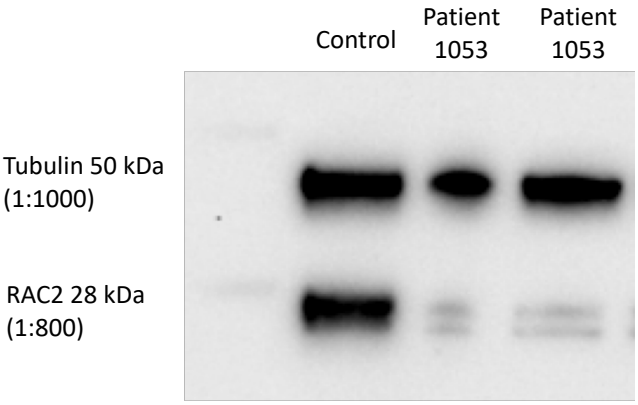


6.2.3 Rac2 expression analysis

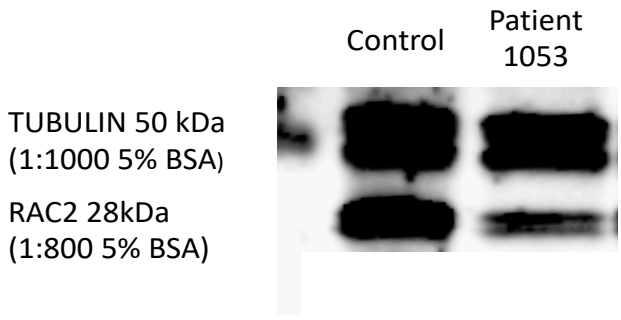
The protein expression of Rac2 was measured in control and patient peripheral blood mononuclear cells (PBMCs) on an immunoblot, and probed with both anti-Rac2 and the loading control anti-tubulin antibodies (Figure 6.6). It shows relative decrease in Rac2 protein expression, compared to a healthy control. This was repeated a further two times at separate blood lets using separate controls (Figure 6.5). Densitometry showed an average Rac2 expression of 32% (95% CI: 14.57-49.4%) in the patient compared to control samples, relative to the loading control (Table 6.2).

Figure 6.5: Rac2 expression in patient PBMCs, compared to a loading dose control

Immunoblot 1:



Immunoblot 2:



Immunoblot 3:

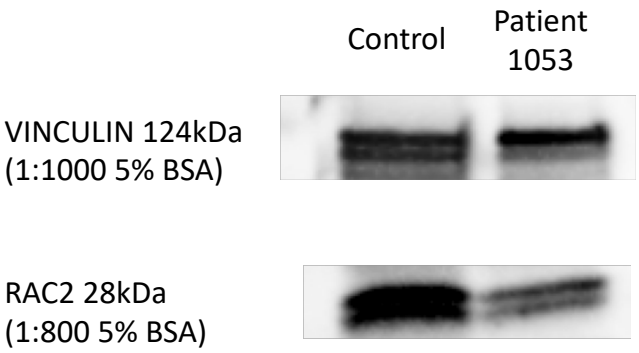


Table 6.2: Densitometry analysis (computed using Bio-rad Image Lab software)

Immunoblot	Loading Control (%)	Patient 1053 (%)
1	100	12
2	100	44
3	100	40
Mean optical density compared to control (%)		**32 (95% CI – 14.57-49.4)**

6.2.4 Further experimental analysis

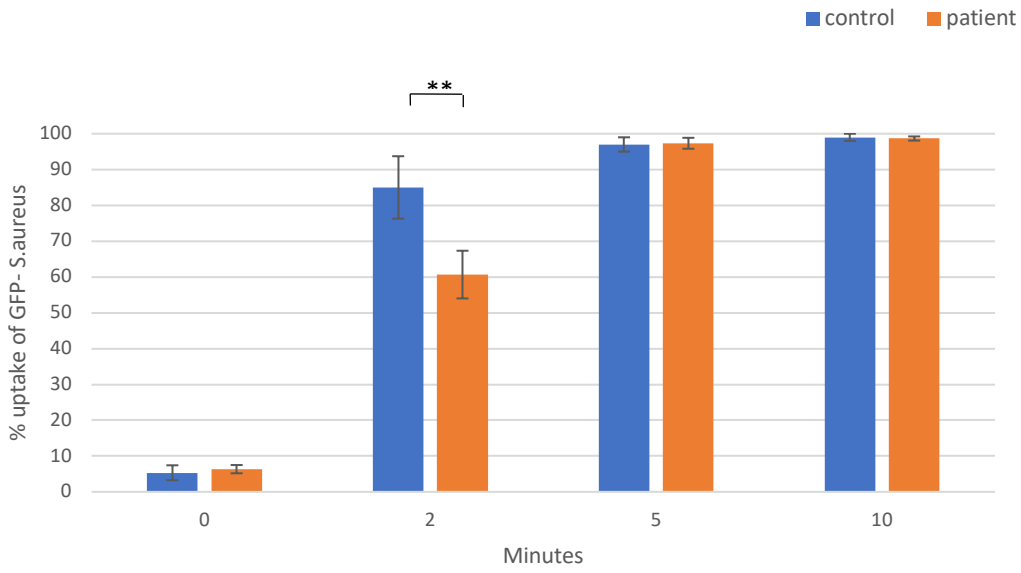
Neutrophil function was analysed using a number of functional tests, undertaken at different centres. The first two tests were performed at Department of Immunology molecular laboratory, Great Ormond Street Hospitals NHS Foundation Trust, by Dr Kimberly Gilmour. The Hydrogen Peroxide assay (Amplex-red) was performed at Sanquin diagnostic laboratory, Amsterdam by Dr Anton Tool, under the supervision of Dr Taco Kuijpers. The neutrophil chemotaxis assay was performed at Institute of Immunity and Transplantation, Royal Free Hospitals NHS Foundation Trust by Dr Adriana Albuquerque.

1. Neutrophil phagocytosis flow cytometric assay
2. Dihydrorhodamine (DHR) flow cytometric assay
3. Amplex-red test of hydrogen peroxide release assay
4. Neturophil chemotaxis

Neutrophil phagocytosis assay

The test was carried out using flow cytometric analysis of neutrophils suspended with tagged *Staphylococcus aureus*- green fluorescent protein (GFP) with serial measurements taken at 0, 2, 5 and 10 mins, from 3 individual assays performed over 8 months (Figure 6.6). At 2 mins, the patient neutrophils had not taken up the GFP-tagged bacteria as readily as the control neutrophils, with mean reduction in patient of 24.33% compared to control (95% CI: 11.59-37.08, *p*-value = 0.0145). However, bacterial uptake was equivalent to control at 5 and 10 minutes. This data suggests a small but reproducible kinetic defect in patient neutrophil phagocytosis.

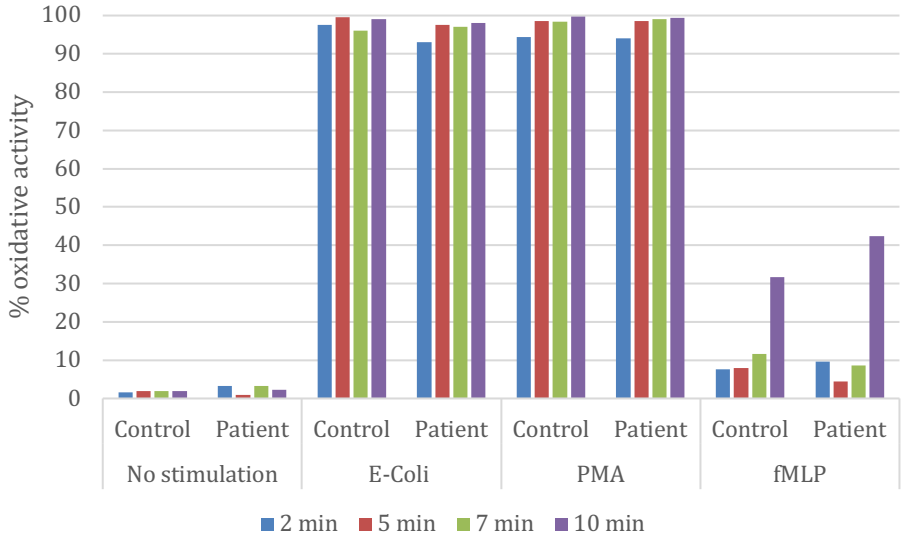
Figure 6.6: Graph showing Neutrophil Phagocytosis Assay results in patient and control. The mean and range of 3 individual assays over 8 months are depicted. ** denotes statistical signficance (p<0.05).



Dihydrorhodamine (DHR) flow cytometric assay

This assay measures extracellular levels of reactive oxygen species (ROS), specifically hydrogen peroxide released by neutrophils. This assay replaces the more subjective nitroblue tetrazolium test (NBT). ROS is typically released to chemotactic factors such as phorbol myristate acetate (PMA) and N-formyl-methionyl-leucyl-phenylalanine (fMLP). Initial results of repeated assays suggest a reduction of oxidative production of neutrophils to fMLP at 5-minute time point compared to healthy control samples but this is not statistically significant (Figure 6.7). The oxidative production index appears equivalent for other timepoints and chemotactic factors and there were no significant differences. A repeat assay is required to confirm these results conclusively.

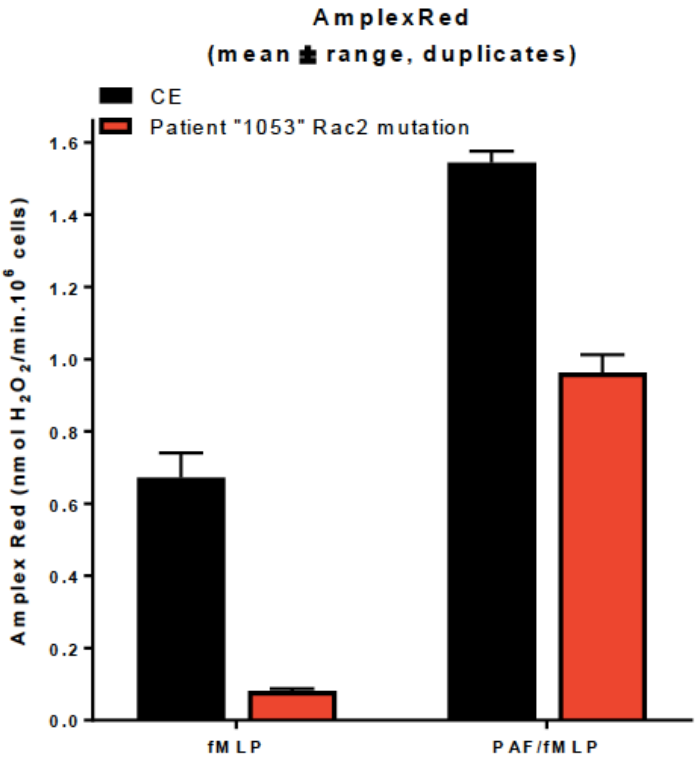
Figure 6.7: Graph showing oxidative production index of dihydrorhodamine (DHR) assay using patient and control neutrophils measured at 2,5,7, and 10 minute timepoint. PMA - phorbol myristate acetate, fMLP - N-formyl-methionyl-leucyl-phenylalanine.

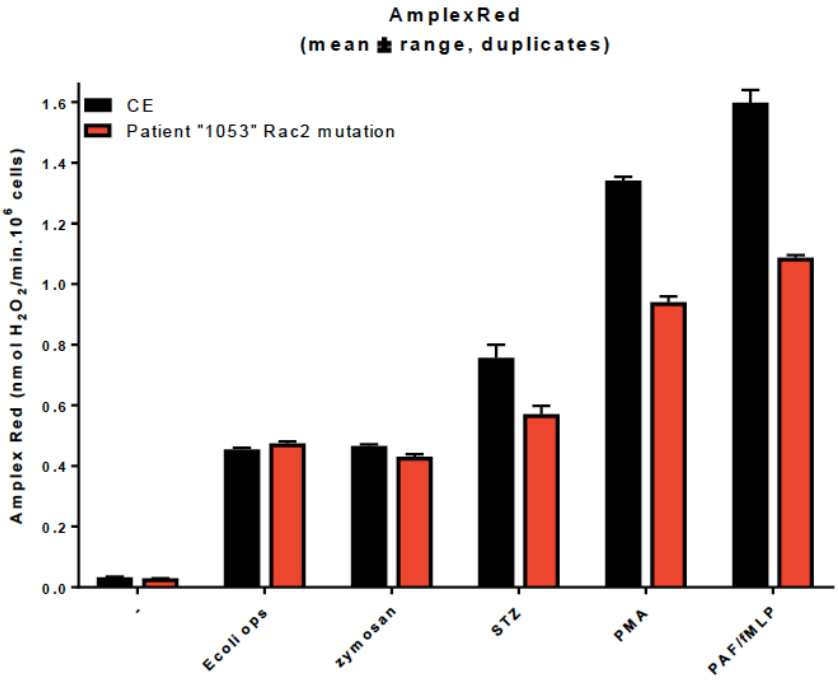


Amplex-red Hydrogen peroxide assay

This assay also measures extracellular release of ROS of neutrophils using a direct oxidation product to detect the presence of the probe. Therefore, this method is more specific and sensitive than the DHR assay [334]. Initial results suggest a partial defect in ROS production in patient neutrophils, however we require further assays to verify this result (Figure 6.8). The Amplex-red detection is not at healthy control levels, but oxidation products are detected at low/intermediate levels.

Figure 6.8: Amplex-red assay showing reduced patient neutrophil production of oxidative products compared to control (above) and using different phagocytic stimulants (below). CE- control, PAF – platelet activating factor, fMLP - N-formyl-methionyl-leucyl-phenylalanine, STZ – serum treated zymosan





Neutrophil chemotaxis

Neutrophil chemotaxis, the directional migration of neutrophils to a chemical stimulus, was measured by tracking cells across a cell chamber slide (Ibidi chamber) using confocal microscopy. Initial results indicate a deficit in chemotaxis as patient cells moved <10μm, both in direction and velocity, and repeated analysis is underway.

In summary, these functional tests suggest a partial defect of neutrophil function, both in ROS production for bacterial killing, and chemotaxis. However, all results are preliminary and further work is required to report conclusively.

6.2.5 *RAC2* haploinsufficiency may cause combined immunodeficiency and partial neutrophil dysfunction

My results demonstrate the presence of a novel pathogenic heterozygous missense mutation in *RAC2* c.106G>A, p.V36M (NM_002872) which may be contribute to combined immune deficiency and partial neutrophil dysfunction. Further analysis is required to show that this variant is causative for the patient's phenotype. Haploinsufficiency is the mechanism where loss of a single allele causes a functional phenotype causing disease. In this case, the patient has a loss of Rac2 protein expression, with disease characteristics suggesting a partial loss of Rac2 function, resulting in neutrophil GTP activity and progressive loss of B cells. Extended family genetic analysis is not available in this patient and so it is difficult to confirm the penetrance of this genetic variant.

Previous case reports share aspects in common with our patient, both in neutrophil dysfunction and common variable immunodeficiency. Our patient's clinical features align more closely with Rac2 loss of function W56X variant, than the dominant

negative D57N phenotype, particularly as progressive B- and T-cell lymphopenia was also observed in W56X [315]. The clinical phenotype of the E62K variant is similar to our patient's clinical features, however functional data from that case suggest a gain of function variant, rather than reduced function [316]. This suggests that Rac2 has a role in B-cell development and differentiation, as shown in knockout Rac2 murine studies [300, 301, 335]. This has yet to be clearly delineated in humans.

It is also not understood why germline haploinsufficiency may cause a delay in phenotype expression. It is plausible that haploinsufficient proteins may have retention of some functional protein that may contribute to a milder phenotype and later presentation in life. Although it is not clear whether the splice region variant in our patient causing a splicing defect, mutations that disrupt mRNA splicing may allow some production of normal mRNA as well as abnormal transcripts. This is likely to be investigated in further work.

It is noteworthy that the W56X loss of function variant also has a later-onset presentation, with progressive immunodeficiency over years [315]. As the patient has complete absence of protein, there is a degree of functional redundancy of Rac proteins in the immune system, particularly as Rac1 has ubiquitous expression, and shares much of the same functionality in murine studies [300, 305].

Investigation of the heterozygous Rac2 D57N variant showed loss of protein expression similar to our patient, but not complete absence as demonstrated in the homozygous Rac2 W56X [54, 315]. However, a direct comparison cannot be made as our western immunoblot was done in PBMCs rather than patient neutrophils as seen in the original Ambruso et al (2000) paper. While heterozygous missense mutations are not expected to impact protein expression, it has been shown that disruption of the GTPase activity, a core function of the protein, is enough to disrupt

protein expression [314]. The neutrophil immunodeficiency caused by heterozygous Rac2 D57N variant, causes disease by dominant negative mechanism; a protein product that acts antagonistically to the wild-type protein. This mechanism may cause more severe effects than null alleles as demonstrated in other PIDs and other genetic disorders [194, 336, 337]. Lack of Rac2 activity has been shown to cause decreased chemotactic motility, inhibits oxidase activation and ROS production by neutrophils, by disrupting the Rac2-GTP-binding site. Our patient also showed reduced reactive oxygen species (ROS) production and slower phagocytosis by neutrophils suggesting a partial defect, but this alone is unlikely to instigate an immunodeficiency phenotype.

Our patient's variant V36M lies in the Switch I domain. This region lies close to GTP-binding sites, and undergoes conformational change of the protein when GTP is bound, hence the name 'switch'. This activates the GTPase and enables interaction with downstream effector proteins such as PAK (p21 activated kinase) [332, 338]. While this pathway is yet to be clearly elucidated, the defect may disrupt interaction of transducer molecules, thereby blocking full functionality of Rac2.

6.2.6 Future Work

Future work will concentrate on consolidating further evidence for RAC2 haploinsufficiency. This includes analysing Rac2 protein expression in the patient's neutrophils. It is clear that Rac2 haploinsufficiency appear to cause a neutrophil migration defect but the nature is not yet defined. This will include the extent of fMLP-induced F-actin polymerisation in neutrophils.

The novel variant affects the Switch domain which is implicated in downstream interactions with effector proteins. By using Flag-tagged patient cells, expression of

potential interacting proteins such as PAK may be studied to further delineate the molecular mechanism of disease.

6.3 Results and Discussion of *EP300* case

By screening genes associated with known PID disorders (PID candidate genes) which have an indistinguishable clinical presentation, I identified a novel likely-pathogenic variant in *EP300*. Variant analysis and pipeline used are described in Chapter 5.

6.3.1 Clinical summary

Patient 1901 is a 24-year-old male who was referred from paediatric immunology services at the age of 16 years with ongoing hypogammaglobulinaemia and recurrent ear and sinus infections. He also has recurrent chronic skin warts. Polycythaemia was also noted on blood investigations.

He was born at 28 weeks gestation and had bronchopulmonary dysplasia with oxygen requirement in early infancy. He was noted to have dysmorphic facial features, specifically: low hairline, thick eyebrows, downslanting palpebral fissures and low-set ears, and mild autism and learning difficulties. He was also noted to have short, wide fingers and toes. Karyotype and array-comparative genomic hybridisation (aCGH) investigations at the time did not reveal a unifying diagnosis.

In the first year of life, he had surgical intervention for hiatus hernia and poor swallow and continues to have oropharyngeal dyscoordination with poor swallow. He had recurrent upper respiratory tract infections throughout childhood, particularly otitis media and tonsillitis. He is his parents' only child and there is no significant family history.

Blood investigations showed raised haemoglobin count of 189g/L (normal range: 133-170 g/L), with raised total red cell count 6.28 $x10^{12}$/L (normal range: 4.3-5.6 x 10^9/L), and raised haematocrit of 0.5 (normal range: 0.34-0.45). Normal total white cell counts were found: 10.42 $x10^9$/L with 69.6% neutrophil, 21.4% lymphocytes and 7.3% monocytes. Serum immunoglobulin levels were low: IgG 3.3 g/L, (normal range: 5.5-15.8 g/L), IgA < 0.1g/L (normal range: 0.67-3.48 g/L) and IgM 0.4 g/L (normal range: 0.23-2.59 g/L). There were absent vaccine responses, indicating poor specific antibody function. Serum erythropoietin levels were within normal range.

He has commenced immunoglobulin replacement and has regular venesection for polycythaemia.

6.3.1.1 Genetic diagnosis

The patient had undergone previous microdeletion assay testing as a child at Great Ormond Street Hospital after dysmorphism was detected, but no genetic cause was found. The patient was entered into BRIDGE-PID research study of whole genome sequencing for patients with immunodeficiency. Sequencing data was filtered and analysed to look for variants in genes known to cause primary immunodeficiency as described in Method chapter 2. Analysis of the coding regions of genes known to cause primary immunodeficiency and polcythaemia (*JAK2, TET2, EPOR, NFE2, SH2B3, VHL, EGLN1, EPAS1, HBB, HBA,* and *BPGM*) revealed a potentially disease-causing variant in *GATA2* (Table 6.3). The variant is found at low frequency (MAF = 0.008), but is present in normal population databases. Monoallelic mutations in *GATA2* have been shown to cause combined immunodeficiency with particular susceptibility to mycobacterial, viral and opportunistic fungal infections [55, 339]. The disease is also dubbed MonoMAC, due to preponderance of a mono-cytopaenia (usually B, natural killer or dendritic cell) and infection with *Mycobacterium avium* complex (MAC) [340]. Other *GATA2* disease manifestations have been described including hypoplastic bone marrow, and association with lymphoedema, acute myeloid leukaemia and sensorineural hearing loss (Emberger syndrome) [55]. However, our

patient does not have these clinical features so we could not attribute clinical disease to this variant.

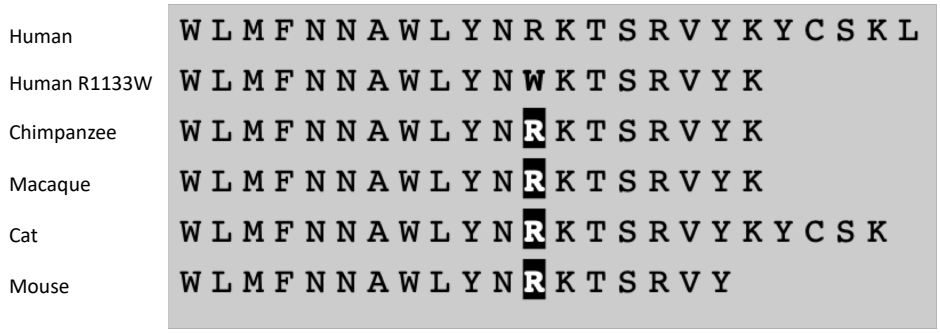
As no other compelling variant was found, the data was interrogated within genes in a known PID gene pathway. Common variants with a minor allele frequency >1% in the 1000 Genomes project were excluded. Variants with a low read depth of <5 were also excluded. Synonymous and non-coding variants were filtered out, and those considered to have a high/moderate impact on the protein were prioritised. Missense variants were also classified using in-silico methods to retain those deleterious to the protein.

Table 6.3: Variants found in patient within PID candidate genes post variant-filtering

Gene	Variant type	Nucleotide change	AA change	Zygosity	SIFT	PolyPhen2	CADD	ExAC MAF
GATA2	Missense	c.121C>G	P41A	het	Tolerated	Probably damaging	23.2	0.0008
RECQL4	Missense	c.2344G>A	D782N	het	Deleterious	Probably damaging	30	0.00009
SPEN	Missense	c.207G>A	M69I	het	Deleterious	Probably damaging	26.8	-
SPINK5	Missense	c.2678C>G	A893G	het	Deleterious	Probably damaging	27.5	0.0005
DOCK2	Missense	c.640C>T	R214W	het	Tolerated	Probably damaging	28.1	0.005
TWIST1	Missense	c.454G>A	A152T	het	Tolerated	Probably damaging	33	0.00008
PTPRZ1	Missense	c.6143A>T	N2048S	het	Deleterious	Probably damaging	27.1	0.00003
ACTN1	Missense	c.2605G>T	G891C	het	Deleterious	Probably damaging	33	0.00023
RAB26	Missense	c.659C>T	A220V	het	Deleterious	Probably damaging	28.8	0.0001
EP300	Missense	c.3397C>T	R1133W	het	Deleterious	Probably damaging	29.5	-

A novel heterozygous missense variant *EP300* c.3397C>T, p.R1133W (het) (NM_001429) was found, and was deemed the most likely causative variant, particularly given the patient's facial gestalt. The variant was predicted to be deleterious using in-silico analysis tools SIFT and PolyPhen-2. It is predicted to be disease-causing in MutationTaster, and has a CADD score of 29.5 (score >20 indicate 1% of most deleterious SNVs). Furthermore, the variant affects an amino acid is conserved among mammals and primates (Figure 6.9).

Figure 6.9: Schematic diagram of EP300 bromodomain showing conservation in mammals and in primates



The variant is not present in the 1000 Genomes or ExAC databases, nor in clinical mutation databases such as ClinVar or Leiden Open Variation Database (LOVD). The variant was confirmed by Sanger re-sequencing at Great Ormond Street Molecular Genetics Laboratory.

6.3.2 A novel variant in EP300 causes a complex syndromic phenotype with combined immunodeficiency and polycythaemia

Heterozygous genetic variants in *EP300* are known to cause Rubinstein-Taybi Type 2 (RSTS2) [323]. The genetic variant is thought to be have arisen sporadically as there is no family history of Rubinstein-Taybi. This variant is interesting as *EP300* variants have not consistently been associated with immunodeficiency and polycythaemia previously.

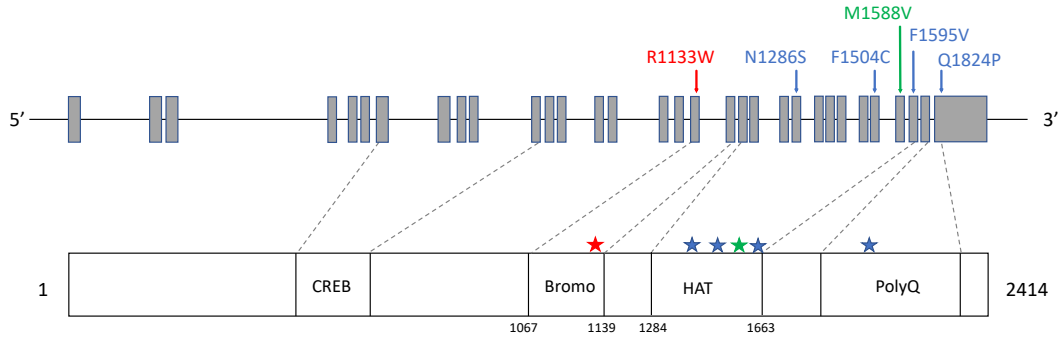
Rubinstein-Taybi syndrome (RSTS) is characterised by key dysmorphic features including broad thumbs and great toes, down-slanting palpebral fissures, short stature and moderate learning difficulties common in affected individuals [323, 341]. It is caused by heterozygous pathogenic genetic variants in 2 genes: *CREBBP* causing

RSTS1 and *EP300* causing RSTS2 [323]. Variants in *EP300* account for the minority of individuals with RSTS (<10%) [341]. Both types are inherited in an autosomal dominant manner and typically arise de *novo* in a pedigree with no family history of the disorder. Clinical features that are shared in this case with RSTS2 are prematurity, intra-uterine growth retardation, mild intellectual difficulty, distinctive facial features such as down-slanting palpebral fissures, high palate, broad hallux and thumb (Figure 6.10).

Figure 6:10: Photographs of our patient showing characteristic facial features including down-slanting palpebral fissures, beaked nose and broad halluces.

The mutational spectrum in both genes range from whole gene deletions to missense mutations with no particular correlation found between the genotype and the severity of phenotype. A recent review found missense mutations to account for only 3.6% of *EP300*-related RSTS2 [342]. Pathogenic missense mutations are shown in Figure 6.11.

Figure 6.11: Schematic diagram showing previously described missense mutations of EP300 known to cause RSTS2 (blue), except variant in shown in green which is only associated with dysmorphic features. Variant in red is found in patient 1901.



The mutational spectrum of *EP300* does not appear to correlate with particular clinical features of RSTS2 nor severity of features. Missense mutations have only been found in 3 of the 86 patients described in the literature with RSTS2, with the majority lying in the HAT region (Figure 6.11). A pathogenic variant in the HAT domain has also been described without RSTS2 syndromic features, but with combined immunodeficiency phenotype [329]. Given this variant and that of our patient's variants, it may be that the location of variant may not be critical in the development of both RSTS2 or immunodeficiency. Specific pathogenesis may arise from downstream effect of protein-protein interaction.

Predisposition to respiratory infections is an infrequent feature of Rubinstein-Taybi syndrome, but with increasing availability to sequencing data, more specific data of *EP300*-related immunodeficiency is being gathered. Immunophenotyping studies are limited, but generally show normal levels of IgG and IgA, but high IgM levels [343]. Most patients with both types of Rubinstein-Taybi have normal vaccine responses to protein and carbohydrate antigens, and normal lymphocyte numbers [343]. Immune defects in RSTS1 (caused by monoallelic mutation in *CREBBP*) have also been

described in patients with severe B-cell dysregulation, hypogammaglobulinaemia and autoimmune cytopenia [344].

Both *EP300* and *CREBBP* code for histone acetyltransferases p300 and CBP respectively, that have homologous, non-redundant function to relax euchromatin structure to allow transcription factors to bind more easily to DNA. Activation and regulation of the histone acetyltransferases is thought to mediated by binding sites on the bromodomain. A key transcription factor to interact with p300 is nuclear factor-kappa B (NFkB), specifically the RelA (p65) component. Haploinsufficiency of Rel-A is associated with increase in $CD21^{low}CD38^{low}$ B cells, suggesting pathogenic deregulation in B cell homeostasis [32]. This is a similar finding in cases of more severe immune defects that have been studied in RSTS, suggesting pathogenesis of immunodeficiency may occur by disruption of NF-kB pathway [329, 344].

6.3.3 The role of p300 in polycythaemia

A novel feature of our patient is primary polycythaemia. Primary polycythaemia is a myeloproliferative disorder in which there is an overproduction of erythrocytes from myeloid progenitor cells in the bone marrow [345]. As a primary disorder of myeloid cells, it is a neoplastic proliferation with accompanying maturation of other cells in the myeloid compartment called panmyelosis [345, 346]. Clinical features are variable and range from asymptomatic, skin rashes to life threatening thrombotic events due to increased blood viscosity.

Rubinstein-Taybi syndrome can predispose to malignancy, but has yet to be linked with myeloproliferative disorders or specifically polycythaemia [325]. The function of the *CREBBP*/*EP300* bromodomain is not yet fully understood, but there is increasing evidence that the bromodomain is necessary for upregulation of haematopoetic

proliferative transcription factors such as Myc and Tet. Specific inhibitors that block bromodomains have been shown to displace p300 from *GATA1* and *MYC* binding sites at chromatin enhancers, resulting in downregulation of oncogene-driven cell proliferation [347-349]. Bromodomain activity has also been associated with cell proliferation in leukaemia, with both bromodomains of p300 and CBP present in somatic chromosomal translocations of mixed lineage leukaemia and myelodysplastic syndromes [350]. Given the current knowledge, it can be postulated that the monoallelic variant *EP300* c.3397C>T p.R1133W found in our patient with erythroid-proliferation, may confer a gain of function of p300 bromodomain. However, this is at odds with the consensus of haploinsufficiency of CBP and p300, and that the loss of HAT activity causes Rubinstein-Taybi syndrome [351]. It is likely the mechanism of pathogenesis of this point mutation to cause the particular constellation of syndromic features is complex, given p300's role as a transcriptional factor and histone modifier. This suggests that an epigenetic role in pathogenesis, which has yet to be fully explored in primary immunodeficiencies.

The HAT region of p300 is known to interact with the potent erythropoetic-specific Gata1 transcription factor which itself activates the Epo receptors on developing red blood cell progenitors [320, 352]. Recent studies using CRISPR/Cas9-engineered transcriptional regulator of the HAT domain of p300, have demonstrated that globin gene expression is activated by epigenomic p300-mediated acetylation, even in the absence of the Gata1 transcription factor, further highlighting the epigenomic role of p300 [353].

Most RSTS2 cases were discovered from screening of RSTS2-affected cases, without prior molecular diagnosis. The main pathogenic effect of *EP300* mutations is RSTS2, a milder phenotype of Rubinstein-Taybi syndrome, but given the pleiotropy of *EP300*, it should be considered as a genetic cause of immunodeficiency. Given the agnostic nature of current large-scale sequencing studies, *EP300* should be included as a gene known to cause immunodeficiency, as there is likely further scope for genetic

diagnosis of patients with *EP300*-associated immunodeficiency with, or without mild syndromic features. Genetic diagnosis may also prompt greater cognisance of RSTS2/*EP300*-associated malignancy, which may allow quicker diagnosis and improve treatment outcomes.

6.3.4 Future work

Given the few cases described in the literature of *EP300*-assoicated immunodeficiency, more work needs to be done to identify further cases, with and without RSTS symptoms. Functional analysis of these cases may elucidate further pathways that cause immunodeficiency, and thereby improve knowledge, and provide new management options to improve prognosis in patients. Further analysis should also include DNA methylation analysis, or chromatin immunoprecipitation sequencing in immune cells which may elucidate the mechanism of hypogammaglobulinaemia.

The data presented in this chapter highlight the importance of a genetic diagnosis in PID and the challenges encountered. Further work is required to refine the genetic analysis to determine more genetic diagnoses in patients with PID. Patients without a firm genetic diagnosis were considered for further investigation of the rest of the sequenced genome, to find potential novel disease-causing variants as described in Chapter 4 and 5. This work is on-going.

7 FINAL DISCUSSION

This thesis describes separate projects in which whole genome sequencing (WGS) was used to identify genetic determinants of immunodeficiency that have not been fully studied thus far. Each research chapter used WGS data from paediatric and adult cohort recruited to the NIHR BioResource Rare Diseases BRIDGE study.

In Chapter 3, I conducted an analysis of current next generation sequencing methods to detect genetic causes of primary immunodeficiency in children. I compared the variant detection sensitivity and gene coverage the three most commonly used methods of NGS sequencing, gene-panel, whole-exome sequencing, and WGS.

In Chapter 4, I conducted a comprehensive screening of known causative genes in 77 children with primary immunodeficiency using whole genome sequencing. Genetic screening had previously been limited to Sanger sequencing of individual genes, or panel sequencing of selected genes such that that the cohort did not have a genetic diagnosis prior to my analysis. By combining a stringent variant filtering pipeline with in-silico predictions of pathogenicity for candidate variants, I was able to ascribe a genetic cause in 24.4% of the patients.

In Chapter 5 I set out to discover novel predisposition genes in a cohort of paediatric patients with PID through analysis of WGS. I collated a list of PID candidate genes that are known to interact directly with known PID genes, or function within immune pathways. I sought deleterious variants within these biologically-relevant genes. I also conducted a small study using rare variant association studies to identify variants in genes which may be enriched in the case population, compared to the control. Although a small study, I found some interesting candidate genes that may be pursued in further research.

In Chapter 6, I described two cases building on work set out in the previous chapter. The cases were both adults: one with combined immunodeficiency caused by a novel heterozygous *RAC2* haploinsufficiency. The second described a novel heterozygous variant in *EP300* in a patient with syndromic features, immunodeficiency and polycythaemia rubra vera.

7.1 Clinical insights of genetic disease diagnosis

This project explored for the first time the utility of WGS for genetic diagnosis in a cohort of children and my analysis has informed current genetic diagnostic methods at Great Ormond Street Hospital NHS Foundation Trust. Moreover, for the patients that we have managed to provide a genetic diagnosis, we have been able to better inform the patients to their condition, and inform management choices for the patients.

A conclusive molecular diagnosis was reached in 24.4% of the cases discussed in Chapter 3. For some patients, medical management significantly changed based on the underlying molecular defect. Even for patients whose medical management remained unchanged, a genetic diagnosis has allowed patients and families to receive prognostic information, genetic counselling, discussion of recurrence risk with families and the identification of asymptomatic mutation carriers at risk of developing PID.

Genetic diagnosis is valuable as early identification of cases in these children may help to avoid detrimental complications of immunodeficiency e.g. overwhelming infections and bronchiectasis. For patients with severe immunodeficiency who may require haematopoetic stem cell transplantation (HSCT), a definitive molecular diagnosis is important to find, or exclude, given the patient risk and cost of transplantation is considerable. Transplantation outcomes improve where a genetic diagnosis is known [354, 355]. Prior detection of familial carriers (who may otherwise have been identified as potential stem cell donors) will also impact upon HSCT outcomes.

Drugs have been developed to regulate particular immune pathways by targeting specific receptors, but these are currently few. Abatacept, a CTLA4 agonist has been developed to bind to downstream receptors, and block T-cell activation, which may be used in CTLA-4 deficiency [29]. However, there have been mixed results to its efficacy in patients' with *CTLA4* mutations, attributed mainly to delayed treatment intervention, or late diagnosis [30]. This suggests that providing a genetic diagnosis as early as possible would be beneficial for treatment outcomes. Given the interconnectedness of immune pathways, it is likely that drugs such as abatacept may be used in individuals with other related genetic mutations as with *LRBA* mutations which has been shown to mediate the effects of autoimmunity [356].

For some genetic diseases, detection of deleterious variants is important as it can impact upon early treatment or screening, these are called actionable variants. Estimates of up to 1% of the population harbouring actionable variants, may be exaggerated, particularly given many deleterious variants have been found in 'normal' population databases. Another anticipated application of WGS in asymptomatic individuals is the identification of carrier status for many autosomal recessive disorders, which could be useful for family planning and for interventions such as preimplantation genetic diagnosis.

Newborn screening for SCID by detection of impaired T cell development will allow patients to be identified before infective complications arise. It follows that better outcomes from treatments such as haematopoietic stem cell transplantation or gene therapy will result. As the cost of gene sequencing falls, all newborns may receive genetic screening of a selection of critical genes such as those that cause SCID. Identification of a previously worked-up genetic mutation allows investigators to make a more confident genetic diagnosis, provided the phenotypes between the two cases overlap sufficiently. Where there is doubt, variants are classified as those of unknown significance (VUS). Until further functional evidence is sought, or found by further sequencing studies, such variants remain difficult to interpret, and act upon.

Genetic testing may pick up false positives (genetic variants that do not translate to *in-vivo* immunodeficiency) or false negatives, where failure to identify children with a genetically undefined form of SCID. Therefore, clinical and immunoprotein analysis will continue to play an important role in diagnosis of genetic conditions. Furthermore, the investigation of false positives and false negatives is likely to better illuminate mechanisms of primary immunodeficiency.

7.2 Translation from genetic variant to patient phenotype

As PID is genetically heterogeneous, selection of appropriate genes to test is challenging, when based only on clinical features. WGS allows for more scope for genetic diagnosis, as well as novel genetic associations in PID and in other clinical settings. In pathology for which preventive measures are available, detection of disease-causing variants before the onset of disease in asymptomatic patients will be beneficial in PID, such as newborn screening. WGS screening is currently too expensive as a screening tool, but as the cost of sequencing falls, it is likely that genetic diagnoses will be made earlier, and perhaps at first medical intervention in the future.

Due to the hypothesis-free method of WGS, it is likely that variants of unknown significance (VUS) will continue to confound researchers and clinicians. Undoubtedly, more detailed phenotypic information is needed to assist laboratory functional testing, in analysing and interpreting the variant. As more sequencing data becomes available, our ability to assign pathogenicity to a genetic variant may improve, particularly as in parallel, *in-silico* mechanisms and mutation modelling are becoming more sophisticated [286, 357-359]. However, given the natural evolution of PID symptoms and features, it is likely that frequent re-analysis of sequencing data will be required, to arrive upon a diagnosis.

As more genetic data becomes available, we may have to functionally validate more potential genetic variants. Reports suggest up to 27% of mutations labelled as pathogenic in medical literature are in fact, not disease-causing [360]. False assignments of pathogenicity can have detrimental consequences on patient care as incorrect prognostic, therapeutic and reproductive advice may be given [361]. Due to this, functional demonstration of the defect in patient-derived samples is currently the only method to demonstrate pathogenicity.

However, functional tests to assess the pathogenicity of genetic variants have been established in <10% of known PID genes. This is set to hamper progress and applicability of genetic data to the field in the near future, particularly as large-scale genome sequencing projects are capable of identifying many promising variants in novel or known genes in faster turnaround times.

Functional assays are often time-consuming and expensive to set up and run. Due to this, a large number of potential candidates are required to test for validity and for cost-effectively. Functional assays may also differ in their ability to recapitulate the *in-vivo* environment within each patient, so-called genetic background, and therefore, may be limited in their ability to determine functional impairment. Novel biochemical methods are required to support and corroborate large-scale DNA sequencing findings, particularly where the yield of interesting variants is high, despite accurate filtering mechanisms.

Multiple mouse models based on suspected genetic aberrations may correlate with the immunodeficiency phenotype. They may also uncover spatial and temporal resolution of *in-vivo* mechanisms, which *in-silico* methods of gene pathway curation may not identify. However, these studies are time-consuming and may not be directly applicable to the human phenotype.

Emerging genome-editing tools such as the CRISPR-Cas9 system, allowing genome modifications at single-base resolution, will play an important role in future functional studies, that aim to understand variant effects, gene function and disease mechanisms in animal studies. In addition, gene editing tools can be used in a variety of biological models from cell lines, induced pluripotent stem cells (iPSC) and more complex animal studies.

7.3 Disease mechanisms

Use of WGS to further develop our understanding of novel genetic mechanisms of immunodeficiency have led to numerous discoveries of disease-causing mutations, and novel gene associations. However, it is likely that this represents a first step to understanding the biological mechanisms by which mutations and disease-susceptibility alleles contribute to disease. These are likely to be further elucidated for patient benefit via greater understanding of in-vivo pathways.

The umbrella term of PID is currently thought of as a collection of monogenic disorders. However, as a large proportion remain undiagnosed it may have a more complex genetic architecture with some more common variants impacting on disease pathology. Large scale genome wide association studies are currently few and are hampered by limited power, but may be feasible with more widespread use of sequencing. An issue of relevance exists for many non-coding regions implicated in association studies. Further analysis of the genome is required to map and assign regulatory activity to non-coding regions, and enhance our understanding of their mechanistic effects in disease development.

An ever-expanding range of structural variation exists in individuals including copy number variants (CNVs), insertions, deletions, and microsatellite repeat expansions among others. Their impact in PID is largely unknown, and may represent a proportion of genetic heritability. Epigenetic changes which modify the genome to affect gene expression or cellular phenotype without affecting the underlying DNA sequence, are also likely to contribute to disease.

It is likely that future attempts to study genetic causes of immunodeficiency will amalgamate data from genomics, but also serum and plasma for proteomics and metabolomics, RNA for transcriptomics, lymphocyte DNA for epigenomics and cells for expression quantitative trait locus (eQTL) studies. The clinical phenotype may be a product of a specific mutated allele, epigenetic and posttranscriptional changes, epistatic variation, or the age of each patient and this represents a rich area of future genotype-phenotype correlation studies in PID.

Enabling a multi-omic intergrated approach will be challenging, particularly in analysis of large quantities of data, and to be usefully interpreted will require algorithmic machine-learning approaches. Gene expression databases and use of transcriptomic data have been used to find potential novel genes that cause PID [362, 363]. Expansion into single cell transcriptomics may also yield novel genetic variants and pathogenic mechanisms of disease.

As costs decrease, WGS is set to become more available as a diagnostic and public health tool, allowing us to more rapidly diagnose disease and identify or predict onset of disease. Currently, the greatest impact of WGS is in diagnosis of patients with early onset disorder caused by single gene mutations of high impact. The increasing deployment of WGS in a variety of clinical settings will facilitate the assessment of its overall benefits and limitations, and the diagnostic yield in different genetic disorders is likely to increase as WGS becomes more widespread in clinical use.

Appendix A

Table of IUIS diseases and genes with proposed inheritance (2017)[7]:

Disease	Genetic defect	Inheritance
γ chain deficiency (common gamma chain SCID, CD132 deficiency)	IL2RG	XL
JAK3 deficiency	JAK3	AR
IL7Ra deficiency	IL7R	AR
CD45 deficiency	PTPRC	AR
CD3d deficiency	CD3D	AR
CD3e deficiency	CD3E	AR
CD3z deficiency	CD247	AR
Coronin-1A deficiency	CORO1A	AR
LAT deficiency	LAT	AR
RAG1 deficiency	RAG1	AR
RAG2 deficiency	RAG2	AR
DCLRE1C (Artemis) deficiency	DCLRE1C	AR
DNA PKcs deficiency	PRKDC	AR
Cernunnos/XLF deficiency	NHEJ1	AR
DNA ligase IV deficiency	LIG4	AR
Reticular dysgenesis	AK2	AR
Adenosine deaminase (ADA) deficiency	ADA	AR
DOCK2 deficiency	DOCK2	AR
CD40 ligand deficiency (CD154)	CD40LG (TNFSF5)	XL
CD40 deficiency	CD40 (TNFRSF5)	AR
ICOS deficiency	ICOS	AR
CD3g deficiency	CD3G	AR
CD8 deficiency	CD8A	AR

ZAP-70 deficiency (ZAP70 LOF)	ZAP70	AR
MHC class I deficiency	TAP1	AR
MHC class I deficiency	TAP2	AR
MHC class I deficiency	TAPBP	AR
MHC class I deficiency	B2M	AR
MHC class II deficiency group A	CIITA	AR
MHC class II deficiency group B	*RFXANK*	AR
MHC class II deficiency group C	*RFX5*	AR
MHC class II deficiency group D	*RFXAP*	AR
DOCK8 deficiency	*DOCK8*	AR
Rhoh Deficiency	*RHOH*	AR
MST1 deficiency	*STK4*	AR
TCR? deficiency	*TRAC*	AR
LCK deficiency	*LCK*	AR
MALT1 deficiency	*MALT1*	AR
CARD11 deficiency (LOF)	*CARD11*	AR
BCL10 deficiency	BCL10	AR
BCL11B deficiency	BCL11B	AD
IL-21 deficiency	IL21	AR
IL-21R deficiency	IL21R	AR
OX40 deficiency	TNFRSF4	AR
IKBKB deficiency	IKBKB	AR
NIK deficiency	MAP3K14	AR
RelB deficiency	RELB	AR
Moesin deficiency	MSN	XL
TFRC deficiency	TFRC	AR
Wiskott-Aldrich syndrome (WAS LOF)	WAS	XL

WIP deficiency	WIPF1	AR
ARPC1B deficiency	ARPC1B	AR
Ataxia-telangiectasia	ATM	AR
Nijmegen breakage syndrome	NBS1	AR
Bloom Syndrome	RECQL3	AR
Immunodeficiency with centromeric instability and facial anomalies, ICF1	DNMT3B	AR
Immunodeficiency with centromeric instability and facial anomalies, ICF2	ZBTB24	AR
Immunodeficiency with centromeric instability and facial anomalies, ICF3	CDCA7	AR
Immunodeficiency with centromeric instability and facial anomalies, ICF4	HELLS	AR
PMS2 Deficiency	PMS2	AR
RNF168 deficiency (Radiosensitivity, Immune Deficiency, Dysmorphic features, Learning difficulties [RIDDLE] Syndrome)	RNF168	AR
MCM4 deficiency	MCM4	AR
POLE1 (Polymerase subunit 1) deficiency (FILS syndrome)	POLE	AR
POLE2 (Polymerase subunit 2) deficiency	POLE2	AR
Ligase I deficiency	LIG1	AR
NSMCE3 deficiency	NSMCE3	AR
ERCC6L2 (Hebo deficiency)	ERCC6L2	AR
GINS1 deficiency	GINS1	AR
Chromosome 22q11.2 deletion (22q11.2DS) (AKA DiGeorge/velocardiofacial syndrome)	Large (3Mb) deletion	AD
DiGeorge/velocardiofacial syndrome	Unknown	Sporadic/toxin exposure
TBX1 deficiency	TBX1	AD
CHARGE syndrome due to CHD7 deficiency	CHD7	AD

CHARGE syndrome due to SEMA3E deficiency	SEMA3E	AD
CHARGE syndrome	Unknown	
Winged helix nudem FOXN1 deficiency	FOXN1	AR
Chromosome 10p13-p14 deletion Syndrome (10p13-p14DS)	Del10p13-p14	AD
Cartilage hair hypoplasia (CHH)	RMRP	AR
Schimke Immuno-osseous Dysplasia	SMARCAL1	AR
MYSM1 deficiency	MYSM1	AR
MOPD1 deficiency	RNU4ATAC	AR
EXTL3 deficiency	EXTL3	AR
AD-HIES Job syndrome	STAT3	AD LOF
Comel-Netherton syndrome	SPINK5	AR
PGM3 deficiency	PGM3	AR
XL-DKC due to Dyskerin deficiency	DKC1	XL
AR-DKC due to nucleolar protein family A member 2 (NHP2) deficiency	NHP2	AR
AR-DKC due to nucleolar protein family A member 3 (NHP3) or NOP10 deficiency	NOP10	AR
AD/AR-DKC due to regulator of telomere elongation (RTEL1) deficiency	RTEL1	AD or AR
AD-DKC due to TERC deficiency	TERC	AD
AD/AR-DKC due to TERT deficiency	TERT	AD or AR
AD-DKC due to TINF2 deficiency	TINF2	AD
AD/AR -DKC due to TPP1 deficiency	TPP1	AD or AR
AR-DKC due to DCLRE1B deficiency	DCLRE1B/ SNM1/APOLLO:	AR
AR-DKC due to PARN deficiency	PARN	AR (AD?)
AR-DKC due to WRAP53 deficiency	WRAP53	AR
Coats plus syndrome due to STN1 deficiency	STN1	AR

Coats plus syndrome due to CTC1 deficiency	CTC1	AR
SAMD9	SAMD9	AD (GOF)
SAMD9L	SAMD9L	AD (GOF)
Transcobalamin 2 deficiency	TCN2	AR
SLC46A1/PCFT deficiency causing hereditary folate malabsorbtion	SLC46A1	AR
Methylene-tetrahydrofolate dehydrogenase 1 (MTHFD1) deficiency	MTHFD1	AR
EDA-ID due to NEMO /IKBKG deficiency (ectodermal dysplasia, immune deficiency)	NEMO (IKBKG)	XL
EDA-ID due to IKBA GOF mutation	IKBA (NFKBIA)	AD GOF
ORAI-1 deficiency	ORAI1	AR
STIM1 deficiency	STIM1	AR
Purine nucleoside phosphorylase (PNP) deficiency	PNP	AR
Immunodeficiency with multiple intestinal atresias	TTC7A	AR
Hepatic veno-occlusive disease with immunodeficiency (VODI)	SP110	AR
Vici syndrome due to EPG5 deficiency	EPG5	AR
HOIL1 deficiency	HOIL1 (RBCK1)	AR
HOIP deficiency	HOIP1 (RNF31)	AR
Hennekam-lymphangiectasia-lymphedema syndrome due to CCBE1 deficiency	CCBE1	AR
Hennekam-lymphangiectasia-lymphedema syndrome due to FAT4 deficiency	FAT4	AR
STAT5b deficiency	STAT5B	AR
Kabuki Syndrome 1 due to KMT2D deficiency	KMT2D (MLL2)	AD
Kabuki Syndrome 2 due to KDM6A deficiency	KDM6A	XL (females may be affected)

BTK deficiency, X-linked agammaglobulinemia (XLA)	BTK	XL
m heavy chain deficiency	IGHM	AR
l5 deficiency	IGLL1	AR
Iga deficiency	CD79A	AR
Igb deficiency	CD79B	AR
BLNK deficiency	BLNK	AR
PIK3R1 deficiency	PIK3R1	AR
E47 transcription factor deficiency	TCF3	AD
Common variable immune deficiency with no gene defect specified (CVID)	Unknown	Variable
PIK3CD mutation (GOF)	PIK3CD GOF	AD
PIK3R1 deficiency (LOF)	PIK3R1	AD
PTEN Deficiency (LOF)	PTEN	AD
CD19 deficiency	CD19	AR
CD81 deficiency	CD81	AR
CD20 deficiency	MS4A1	AR
CD21 deficiency	CR2	AR
TACI deficiency	TNFRSF13B (TACI)	AD or AR
BAFF receptor deficiency	TNFRSF13C (BAFF-R)	AR
TWEAK deficiency	TNFSF12	AD
Mannosyl-oligosaccharide glucosidase deficiency (MOGS)	MOGS (GCS1)	AR
TRNT1 deficiency	TRNT1	AR
TTC37 deficiency	TTC37	AR
NFKB1 deficiency	NFKB1	AD
NFKB2 deficiency	NFKB2	AD
IKAROS deficiency	IKZF1	AD
IRF2BP2 deficiency	IRF2BP2	AD

ATP6AP1 deficiency	ATP6AP1	XL
AID deficiency	AICDA	AR
UNG deficiency	UNG	AR
INO80	INO80	AR
MSH6	MSH6	AR
Ig heavy chain mutations and deletions	Mutation or chromosomal deletion at 14q32	AR
Kappa chain deficiency	IGKC	AR
Isolated IgG subclass deficiency	Unknown	?
IgG subclass deficiency with IgA deficiency	Unknown	?
Selective IgA deficiency	Unknown	?
Specific antibody deficiency with normal Ig levels and normal B cells	Unknown	?
Transient hypogammaglobuliemia of infancy	Unknown	?
CARD11 GOF	CARD11	AD GOF
Selective IgM deficiency	Unknown	?
Perforin deficiency (FHL2)	PRF1	AR
UNC13D / Munc13-4 deficiency (FHL3)	UNC13D	AR
Syntaxin 11 deficiency (FHL4)	STX11	AR
STXBP2 / Munc18-2 deficiency (FHL5)	STXBP2	AR or AD
FAAP24 deficiency	FAAP24	AR
Chediak-Higashi syndrome	LYST	AR
Griscelli syndrome, type 2	RAB27A	AR
Hermansky-Pudlak syndrome, type 2	AP3B1	AR
Hermansky-Pudlak syndrome, type 10	AP3D1	AR
IPEX, immune dysregulation, polyendocrinopathy, enteropathy X-linked	FOXP3	XL
CD25 deficiency	IL2RA	AR

CTLA4 deficiency (ALPSV)	CTLA4	AD
LRBA deficiency	LRBA	AR
STAT3 GOF mutation	STAT3	AD (GOF)
BACH2 deficiency	BACH2	AD
APECED (APS-1), autoimmune polyendocrinopathy with candidiasis and ectodermal dystrophy	AIRE	AR or AD
ITCH deficiency	ITCH	AR
ZAP-70 combined hypomorphic and activation mutations	ZAP70	AR (LOF/GOF)
Tripeptidyl-Peptidase II Deficiency	TPP2	AR
JAK1 GOF	JAK1	AD GOF
Prolidase deficiency	PEPD	AR
ALPS-FAS	TNFRSF6	AD or AR
ALPS-FASLG	FASLG	AR
ALPS-Caspase10	CASP10	AD
ALPS-Caspase 8	CASP8	AR
FADD deficiency	FADD	AR
IL-10 deficiency	IL10	AR
IL-10Ra deficiency	IL10RA	AR
IL-10Rb deficiency	IL10RB	AR
NFAT5 haploinsufficiency	NFAT5	AD
SH2D1A deficiency (XLP1)	SH2D1A	XL
XIAP deficiency (XLP2)	XIAP	XL
CD27 deficiency	CD27	AR
CTPS1 deficiency	CTPS1	AR
RASGRP1 deficiency	RASGRP1	AR
CD70 deficiency	CD70 (TNFSF7)	AR
RLTPR (CARMIL2) deficiency	RLTPR	AR

ITK deficiency	ITK	AR
MAGT1 deficiency (XMEN)	MAGT1	XL
PRKCD deficiency	PRKCD	AR
Elastase deficiency (SCN1)	ELANE	AD
GFI 1 deficiency (SCN2)	GFI1	AD
HAX1 deficiency (Kostmann Disease) (SCN3)	HAX1	AR
G6PC3 deficiency (SCN4)	G6PC3	AR
VPS45 deficiency (SCN5)	VPS45	AR
Glycogen storage disease type 1b	G6PT1	AR
X-linked neutropenia/ myelodysplasia WAS GOF	WAS	XL
P14/LAMTOR2 deficiency	LAMTOR2	AR
Barth Syndrome, (3-Methylglutaconic aciduria type II)	TAZ	XL
Cohen syndrome	VPS13B	AR
Clericuzio syndrome (Poikiloderma with neutropenia)	USB1	AR
JAGN1 deficiency	JAGN1	AR
3-Methylglutaconic aciduria	CLPB	AR
G-CSF receptor deficiency	CSF3R	AR
SMARCD2 deficiency	SMARCD2	AR
HYOU1 deficiency	HYOU1	AR
Leukocyte adhesion deficiency type 1 (LAD1)	ITGB2	AR
Leukocyte adhesion deficiency type 2 (LAD2)	SLC35C1	AR
Leukocyte adhesion deficiency type 3 (LAD3)	FERMT3	AR
Rac 2 deficiency	RAC2	AD
b actin deficiency	ACTB	AD

Localized juvenile periodontitis	FPR1	AR
Papillon-Lefèvre Syndrome	CTSC	AR
Specific granule deficiency	CEBPE	AR
Shwachman-Diamond Syndrome	SBDS	AR
WDR1 deficiency	WDR1	AR
Cystic fibrosis	CFTR	AR
Schwachman Diamond syndrome due to DNAJC21 deficiency	DNAJC21	AR
Neutropenia with combined immune deficiency due to MKL1 deficiency	MKL1	AR
X-linked chronic granulomatous disease (CGD), gp91phox	CYBB	XL
Autosomal recessive CGD p22phox	CYBA	AR
Autosomal recessive CGD p47phox	NCF1	AR
Autosomal recessive CGD p67phox	NCF2	AR
Autosomal recessive CGD p40phox	NCF4	AR
G6PD deficiency Class I	G6PD	XL
GATA2 deficiency (MonoMac syndrome)	GATA2	AD
Congenital pulmonary alveolar proteinosis due to CSF2RB mutations	CSF2RB	AR
Congenital pulmonary alveolar proteinosis due to CSF2RA mutations	CSF2RA	XL (Pseudoautosomal)
IL-12 and IL-23 receptor b1 chain deficiency	IL12RB1	AR
IL-12p40 (IL-12 and IL-23) deficiency	IL12B	AR
IFN-g receptor 1 deficiency	IFNGR1	AR /AD
IFN-g receptor 2 deficiency	IFNGR2	AR
STAT1 deficiency (AD LOF)	STAT1	AD
Macrophage gp91 phox deficiency	CYBB	XL
IRF8 deficiency (AD)	IRF8	AD

IRF8 deficiency (AR)	IRF8	AR
Tyk2 deficiency	TYK2	AR
ISG15 deficiency	ISG15	AR
RORc deficiency	RORC	AR
JAK1 (LOF)	JAK1	AR
EVER1 deficiency	TMC6	AR
EVER2 deficiency	TMC8	AR
WHIM (Warts, Hypogammaglobulinemia, infections, Myelokathexis) syndrome	CXCR4	AD GOF
STAT1 deficiency (AR LOF)	STAT1	AR
STAT2 deficiency	STAT2	AR
IRF7 deficiency	IRF7	AR
IFNAR2 deficiency	IFNAR2	AR
CD16 deficiency	FCGR3A	AR
MDA5 deficiency (LOF)	IFIH1	AR
TLR3 deficiency	TLR3	AD or AR
UNC93B1 deficiency	UNC93B1	AR
TRAF3 deficiency	TRAF3	AD
TRIF deficiency	TICAM1	AD or AR
TBK1 deficiency	TBK1	AD
IRF3 deficiency	IRF3	AD
CARD9 deficiency	CARD9	AR
IL-17RA deficiency	IL17RA	AR
IL-17RC deficiency	IL17RC	AR
IL-17F deficiency	IL17F	AD
STAT1 GOF	STAT1	AD GOF
ACT1 deficiency	TRAF3IP2	AR
IRAK-4 deficiency	IRAK4	AR

MyD88 deficiency	MYD88	AR
IRAK1 deficiency	IRAK1	XL
TIRAP deficiency	TIRAP	AR
Isolated congenital asplenia (ICA) due to RPSA deficiency	RPSA	AD
Isolated congenital asplenia (ICA) due to HMOX deficiency	HMOX	AR
Trypanosomiasis	APOL1	AD
Acute liver failure due to NBAS deficiency	NBAS	AR
Acute necrotizing encephalopathy	RANBP2	AD
CLCN7 deficiency associated osteopetrosis	CLCN7	AR
SNX10 deficiency associated osteopetrosis	SNX10	AR
OSTM1 deficiency associated osteopetrosis	OSTM1	AR
PLEKHM1 deficiency associated osteopetrosis	PLEKHM1	AR
TCIRG1 deficiency associated osteopetrosis	TCIRG1	AR
TNFRSF11A deficiency associated osteopetrosis	TNFRSF11A	AR
TNFSF11 deficiency associated osteopetrosis	TNFSF11	AR
NCSTN deficiency hidradenitis suppurativa	NCSTN	AD
PSEN deficiency hidradenitis suppurativa	PSEN	AD
PSENEN deficiency hidradenitis suppurativa	PSENEN	AD
TREX1 deficiency, Aicardi-Goutieres syndrome 1 (AGS1)	TREX1	AR or AD
RNASEH2B deficiency, AGS2	RNASEH2B	AR
RNASEH2C deficiency, AGS3	RNASEH2C	AR
RNASEH2A deficiency, AGS4	RNASEH2A	AR
SAMHD1 deficiency, AGS5	SAMHD1	AR
ADAR1 deficiency, AGS6	ADAR1	AR
Aicardi-Goutieres syndrome 7 (AGS7)	IFIH1	AD

Spondyloenchondro-dysplasia with immune dysregulation (SPENCD)	ACP5	AR
STING--associated vasculopathy, infantile-onset	TMEM173	AR
X-linked reticulate pigmentary disorder	POLA1	XL
USP18 deficiency	USP18	AR
CANDLE (chronic atypical neutrophilic dermatitis with lipodystrophy)	PSMB8	AR and AD
Familial Mediterranean fever	MEFV	AR or AD
Mevalonate kinase deficiency (Hyper IgD syndrome)	MVK	AR
Muckle-Wells syndrome	NLRP3	AD GOF
Familial cold autoinflammatory syndrome 1	NLRP3	AD GOF
Familial cold autoinflammatory syndrome 2	NLRP12	AD GOF
Neonatal onset multisystem inflammatory disease (NOMID) or chronic infantile neurologic cutaneous and articular syndrome (CINCA)	NLRP3	AD GOF
NLRC4-MAS (macrophage activating syndrome) or familial cold autoinflammatory syndrome 4	NLRC4	AD GOF
PLAID (PLCg2 associated antibody deficiency and immune dysregulation) or familial cold autoinflammatory syndrome 3 or APLAID (c2120A>C)	PLCG2	AD GOF
NLRP1 deficiency	NLRP1	AR
TNF receptor-associated periodic syndrome (TRAPS)	TNFRSF1A	AD
Pyogenic sterile arthritis, pyoderma gangrenosum, acne (PAPA) syndrome, hyperzincemia and hypercalprotectinemia	PSTPIP1	AD
Blau syndrome	NOD2	AD
ADAM17 deficiency	ADAM17	AR
Chronic recurrent multifocal osteomyelitis and congenital dyserythropoietic anemia (Majeed syndrome)	LPIN2	AR

DIRA (Deficiency of the Interleukin 1 Receptor Antagonist)	IL1RN	AR
DITRA (Deficiency of IL-36 receptor antagonist)	IL36RN	AR
SLC29A3 mutation	SLC29A3	AR
CAMPS (CARD14 mediated psoriasis)	CARD14	AD
Cherubism	SH3BP2	AD
COPA defect	COPA	AD
Otulipenia/ORAS	OTULIN	AR
A20 deficiency	TNFAIP3	AD LOF
ADA2 deficiency	CECR1	AR
AP1S3 deficiency	AP1S3	AR
C1q deficiency due to defects in C1QA	C1QA	AR
C1q deficiency due to defects in C1QB	C1QB	AR
C1q deficiency due to defects in C1QC	C1QC	AR
C1r deficiency	C1R	AR
C1s deficiency	C1S	AR
Complete C4 deficiency	C4A+C4B	AR
C2 deficiency	C2	AR
C3 deficiency (LOF)	C3	AR
C3 GOF	C3	AD
C5 deficiency	C5	AR
C6 deficiency	C6	AR
C7 deficiency	C7	AR
C8a deficiency	C8A	AR
C8g deficiency	C8G	AR
C8b?deficiency	C8B:	AR
C9 deficiency	C9	AR

MASP2 deficiency	MASP2	AR
Ficolin 3 deficiency	FCN3	AR
C1 inhibitor deficiency	SERPING1	AD
Factor B GOF	CFB	AD
Factor B LOF	CFB	AR
Factor D deficiency	CFD	AR
Properdin deficiency	CFP	XL
Factor I deficiency	CFI	AR
Factor H deficiency	CFH	AR or AD
Factor H –related protein deficiencies	CFHR1-5	AR or AD
Thrombomodulin deficiency	THBD	AD
Membrane Cofactor Protein (CD46) deficiency	CD46	AD
Membrane Attack Complex Inhibitor (CD59) deficiency	CD59	AR
CD55 deficiency (CHAPEL disease)	CD55	AR

Appendix B

Table of genes mined for pathway analysis

ADD2	DNAJB6	ARHGAP4	BUB1B	ETS1	NF1	PGR	S100A6	CBY1	EP300	CNR1	ITGA4	RPS6KA1	IL22RA2	RND2	MAOA
TNNT2	ADORA2A	PTPN7	ZNF593	KLHL12	ARHGEF12	NTSR1	FBXO4	MAP3K8	RPL11	PTK2	UBE2C	ADRB2	APEX1	AGPAT1	SHC2
MAPRE2	CSNK2A1P	MED13L	CSE1L	PTHLH	NFKBIE	BLMH	RING1	MARK3	L1CAM	PTAFR	TRAF6	KIF2C	SSH1	BRAF	CAPN10
ZBTB33	SPTAN1	GSN	SNRPF	NRP2	ABLIM2	DDX5	CCAR1	HES3	FBXW11	MAD2L1	CLASP2	DOK2	IFNA2	S100A11	PTGES2
CABC1	GPS2	ELP3	SAP30	USP33	XRCC6	SIK3	ALB	RGMB	TARBP2	MAP2K4	SOCS2	IL2RB	PAWR	FLT4	APAF1
CCNO	AKAP8	FPR3	TRAIP	RNF40	PPP2R4	SHH	ACP1	XRN2	STX1B	ESR1	ESCO2	SFPQ	PSMA6	SUMO4	CAMK2B
RAB13	DPP4	AAAS	NCOA1	CDC123	TSPO	KIF18A	HSPA4	MED17	CD97	FBXO6	APOA2	GDI2	HOXC6	ULBP3	TGFBR2
BCL6	FAF1	ZBTB7B	AFF1	PSME1	FANCE	DDIT4	CD1A	FANCA	HGS	CDC26	TRPV6	MORF4L1	RAB8B	TARDBP	MAPK14
IGH2	STAG3	CD8BP	F2	GTF2E1	CD2	YWHAE	GPR77	SIK2	SOD1	SEC61B	TERF2IP	ITPR3	RNF5	IL18RAP	PAPD5
ULBP2	POLR3C	RINT1	GSK3A	IK	NFYB	MCM7	ELK3	NGDN	USP12	ATF6	DHX9	STK38	LAT2	HIST1H4A	PIK3R4
P4HB	SNW1	KCNA5	RBX1	ADAMTS4	DR1	BID	LNX1	SLC18A2	ID4	CD38	SNRPC	DOT1L	INHBA	CTNND2	CDK19
GSPT1	RBPJ	MYO9A	ARID2	SMAD5	UBE2S	MAP2K7	LCAT	SMARCB1	HIST2H3A	CNBP	DCTN1	SELL	TCEB2	DLG3	VCP
MDK	CAV1	YY1	HIRA	SCARB1	SPEN	S100A8	KIAA0368	OPRM1	SCGB1A1	TEC	ONECUT1	PEBP1	SLC11A1	SPG20	ACTN2
RYBP	ADAM12	TUBB4	GADD45G	G3BP2	IKZF3	LPXN	POLR2C	PPP1R9B	ANAPC7	CYR61	DYNLL1	PPM1F	CRX	NSF	APOC2
CD86	MEOX2	GPS1	TAX1BP1	HOXB7	CSF1	RAF1	ACTC1	TRAT1	DRD2	BAG1	IGHG1	UBE2D1	KLK2	OTUB2	GNA12
KDM1A	PTPN6	VEZF1	DLG4	NEURL	TUBA1A	RNF4	MAP3K7	LGALS7	DCUN1D1	SLC9A3	IRAK2	SOX6	DNAJC10	BUB3	VEGFA
MYO1F	ARNT	F11R	MAP2K1	AKT1S1	ERBB4	BHLHE40	DUT	RIPK2	STEAP3	MEF2A	ITSN1	TSC1	ARPC1B	MYOCD	ADAM9
SCFD1	ITGAD	EPHA5	MAP1S	MAPK8IP2	SPTBN4	TOP2B	OTUB1	TACC3	TIAF1	YOD1	LIG3	SYK	GRIK2	TRPC1	RAB19
EPHB1	FCN2	ZFPM2	CDC27	RHOV	CDC42SE1	NOP2	VASP	RPS15	DOCK1	MMP2	PSMD5	EIF2C2	MNAT1	ILKAP	IQGAP1
PMS1	IL22	SMARCC2	UCHL5	FSCN1	VPREB1	PYGL	APOA1	LHX9	CDX1	WDR77	NRAS	FKBP1C	VAMP7	SLC2A4	WDR12
SYT1	PROCR	EPHA1	CPT1A	CABP4	NOX3	DNAJB5	GATA6	PIK3C3	LLGL1	TNFSF10	IL6R	BDKRB2	KDR	GADD45B	USP9X

LST1	ATG7	NCAM1	BAZ2A	CCL5	HES5	CDK8	AGTR1	SOX7	FOSL1	E2F2	NFRKB	GULP1	DSCAM	CBX3	EEF1D
MAP3K13	TF	LACRT	PIK3C2A	ECM1	MED11	RDX	ESRRA	RAP1A	ANXA8L2	RHOF	POLR2L	CAPN1	POLR2F	CSRP3	ZWINT
SERTAD1	ZC3H12A	YWHAQ	CTNNBIP1	NLK	APEH	RAB27B	PAK1	PIM1	KLF11	GINS2	FCER1G	CACNB4	TGS1	G3BP1	GABPA
KLF1	XPO6	NIP7	CNTN2	TLR2	BMP1	SAA2	IGSF8	ACTR3	SON	MED4	AIPL1	RAB1B	CD80	MEAF6	FCER2
DDR1	UBE2L6	NCF1C	ANAPC1	TSG101	REV3L	RPS6KB1	TRIM5	SF3B3	HEY1	SLAMF1	PARD6B	ALKBH2	TNFAIP1	GATA5	CTCF
PMS2L2	PRKCG	CDKN1B	MDFI	NR2E1	HIST1H4I	NGEF	STAMBP	ANK3	VANGL2	GLMN	PPP3R1	TNIK	MYH6	SMG1	RNF111
IFNA21	IGHG3	ATP1A1	ZBTB7A	CSK	NFX1	RABGEF1	TRADD	ARHGAP6	KPNA1	PLK3	IGFBP3	GNG2	COBRA1	SNIP1	DPYSL5
MYOD1	PTPN3	CDC42SE2	PAFAH1B1	SHPRH	RNF20	NFIC	HDAC6	IFI16	PSMD9	RAB9A	GTF2A1L	SMARCD1	BAX	CTNNA2	UBE3A
BMF	CCL22	RUVBL1	TPX2	BIN1	ABLIM3	KRT8	SF1	FCN1	MAPRE1	SMC1A	CD70	DGKD	PSMD14	IL18RAP	RAPGEF1
GNAI2	MPG	MYOF	SPAG9	CAMLG	IGF1	CITED1	POLR2K	NFASC	ARRB1	PRDX3	XRCC6BP1	CLK2	SPP1	HOXD10	EPOR
PSEN1	CD160	SRP54	PLG	NMI	GCKR	EXOC4	PDGFRA	PSMC3IP	CHAF1B	INPPL1	EPO	SELPLG	IFNG	CD47	IFNW1
TBX3	MAPK15	SNCB	HHEX	MCL1	MAP4K4	GIYD1	GATAD2A	USP47	FEM1B	BCL2L2	FBXL5	SNCA	ORC6L	ORC2L	SIRPG
S100A9	CD34	TRIM23	RFC4	USP2	DDX54	TSR1	BTBD12	DNMT3A	PPP2R5A	TNFRSF12A	PAK2	SHFM1	HDAC3	HIST1H4E	TPM4
MARK4	RPH3AL	TOP3A	NPHP1	RASSF1	FBXL2	RAB3D	KLF10	PRKAB1	MED7	JMY	OTUD5	GRB10	CASP3	PDIA3	IDO1
IL7R	CHRM1	RAD51L3	IFNA6	ZNF346	PLXNA3	SSH2	FZR1	KIR2DS1	RYR2	DLK1	HIST2H2AA4	APOE	ADRA1B	SMCHD1	ARIH2
NUMBL	GDF2	THOC5	NCAPH	MET	RFX1	SMPD1	NR4A2	TLE1	TRIM63	PDP2	USP22	SUPT3H	TXLNA	SUPT16H	NOG
NR6A1	MED16	TUBGCP4	BECN1	PROS1	MAGI3	DBC1	RCOR1	CACNA1G	E2F1	ACVR2A	CAPNS1	EZH2	MIS12	TNFRSF17	ST14
HIST2H4B	CALCR	EPHA7	PBX1	PRMT1	BAZ1B	ATP5B	LEF1	LMO4	CAPN3	POLL	PHF17	BAG5	QKI	SYCP1	FASTKD5
CYCS	NAP1L4	SAE1	INPPL1	PSMD12	LMTK2	REL	CHUK	EXOSC9	PPARD	FSHB	PIK3CB	EIF5B	PHF12	TAF9	TDP1
MST1	FOSL2	BRWD1	MAP4K3	TXNL1	POU2F2	NPLOC4	RHOJ	PPIA	MERTK	ING5	NHLH2	IGLC2	TUBG2	PIK3R3	PRDX6
PAK6	FEN1	FGG	CD9	RAB7A	SRP72	SAA1	KCTD7	CDKN2A	E2F4	LANCL2	CDC42EP5	DSP	USP13	CARD6	PSMG2
AKT2	GRK5	EIF4G1	FGR	HIST2H2AC	KSR1	CARM1	MKNK1	LIPC	IL13	RAB8A	PDLIM7	AP2B1	LIF	CTCFL	WDR48
MEF2C	SH2B1	MKNK2	IGF2R	CHEK1	TTBK2	CARD16	HLA-DRA	NAPG	NONO	UBA2	CHAF1A	SMN2	TFAP2A	TCEB1	FGA
DAP3	RUNX3	ICAM5	ING2	HPN	GSTP1	RALB	C4BPB	UBE3C	FANCL	SLK	LAIR1	TNFRSF10C	CAV3	MED24	ITGB1BP1
LCOR	CD14	H2AFJ	RBBP7	G2E3	VEGFB	NAMPT	ITM2B	GCG	HCK	ARF1	GEMIN4	CLP1	SH3RF1	CCNE2	RPS6KA5

SIK1	IFNA16	PLK4	DAPK1	CENPJ	LCP1	ID2	RAB3B	ASCL3	USP8	LRP6	CHD1	MAP3K3	BSG	NKRF	AMPH
IPO5	AP2A1	SH3GLB1	EDNRA	ENO1	ARID1A	MED26	GRAP2	HIST1H2BC	MACF1	ILK	TRIM28	TCF7L1	CBX8	USP31	COL18A1
ARFIP2	CSTA	COL2A1	RAD23A	RND1	RAB4A	BCR	CST3	TOP2A	PLSCR1	IRS2	SGK2	SH2B2	GADD45A	MAPKAPK3	STK11
BIRC7	FBXO7	PES1	FRS2	EIF3E	ADAMTS3	ICOSLG	MED22	TGFBI	ARFIP1	RAC3	BHLHE41	NDUFV2	SMURF1	FCER1A	SP100
PICK1	IL1B	RAB5A	CLN6	LRPAP1	CARD10	ATR	CD164	BMS1	SMARCAD1	CDKN3	TNF	IFT88	KIR2DL4	CCNA2	DVL1L1
TRERF1	PPP2R3A	MID1	CC2D1A	RASGRP3	FKBP1A	UBE2R2	CDK5R1	LYN	CDCA8	CCDC47	MAPK7	RPL24	MAML2	FCGRT	TFCP2
FYN	COG4	CREBBP	BTC	MCM8	EDF1	XRCC1	TREM2	PRSS3	DIABLO	UBE2E3	MLH3	FOXO4	IFNA14	ATN1	ANGPTL4
ATP1A2	SIRT2	POLR2H	NR2C2	TGFBR1	CKS1B	ORC5L	CAPN2	PSMC2	FIS1	HDAC11	PURA	DYSF	PDX1	SLC6A1	SIAH2
TRIM54	GSC	BAK1	FIGF	CENPH	APOC3	CDC25A	CHTF8	ARHGDIG	ENG	YWHAZ	PRND	MAP2K2	BRAP	CINP	CAND1
EIF4A3	TTK	PSMD8	SPTBN2	RAD18	STK24	SHC1	PSMB9	SIAH1	PARK2	CD28	ERP44	RIMS1	MAP3K11	IL27RA	RAD54B
NKX2-1	SP3	EIF2AK2	FAIM2	LDB1	AUP1	LINGO1	BDNF	PPP2R2B	LTA	GMNN	TRAF2	AZGP1	ILF2	VAPA	DHCR7
EPHA6	SSTR2	GATA3	VSIG4	TMSB4X	FANCC	MTPN	TAF6L	PRPF6	INSR	ARHGDIA	ARIH1	SYVN1	DPF1	STAT6	IGF2BP1
TAF5L	ADRM1	DNAJC16	BANP	PSMD11	DEDD2	PTPRA	FOXO1	SERPINF2	ARPC5	RBM5	KLF5	CIB1	AXIN2	USP16	CCNH
CDKN1C	RNF2	EPS15	TLR4	RBM8A	BRIP1	PPARA	MYST2	KCTD13	KRT14	HMGCS1	CD7	PRKAR2B	APLP1	RORA	CGB8
ARF6	IGF1R	RHOBTB2	TYROBP	CTTNBP2	DUSP6	APRT	GDF5	VLDLR	TNFRSF11A	PTGDS	RPS14	USP19	DVL2	CDC6	SNAP25
HLA-DRB4	DTL	PTPRD	NEK8	MADD	HIST1H1A	PRKCA	XPO5	CAMP	PPP2CB	TCF7L2	SENP2	ATPAF1	AOX1	TNFRSF18	JUP
MXD1	POLK	TEK	CRYAB	RFC5	CLDN1	MED23	PIAS2	FBXO18	AP2M1	ADRBK1	KAL1	RBM9	KLF6	CUL3	ESRRG
RUVBL2	CDKN1A	DCTN2	MBD1	TERF1	DUSP22	SRP68	NR2F1	NRIP1	SREBF2	ARHGEF4	PARD6A	HSPA2	CCR5	UBC	SMG5
ETV4	CREM	SUMO2	ZIM2	CTSB	IL32	FAM175A	CBFB	ABL1	DBNL	RIN2	RIPK4	SORT1	NTRK1	GATA4	SMARCC1
ITGA8	GDI1	HES1	MAP3K12	TFPI	NDUFB4	BIRC3	PSMB5	PPEF1	HERC2	LRP8	WRN	ERCC2	DDA1	KLKB1	VTI1B
F13A1	PFN1	UBR5	HERPUD1	SAG	DLL1	VDR	THOC1	KDM5B	SEMA3C	RAB37	MYCN	SOS1	EHMT2	YBX1	PPP1CA
HP1BP3	VPRBP	NOTCH3	YES1	SGSM3	HOXD11	SMC4	TRIM32	CTGF	B4GALT1	GP1BA	TPR	IRAK3	SCNN1B	CUL4A	PRKCQ
PNKP	DAG1	PAX5	TOP1	NOX4	SERPINE1	SFTPD	NAE1	ARHGEF3	CAV2	MSX1	PPARG	TMPRSS11A	NR2F2	RORB	RARA
SPG21	CD4	OS9	CASR	AHCTF1	PDE4A	KPNA3	A2M	UCHL1	RIPK1	ITGA5	ULK2	KNG1	SAP18	GAP43	SMARCE1
ZBTB32	RUNX1	RHOA	CASP4	BTLA	ACD	EGFR	SLIT2	PLK1	TNPO1	RAB1A	BCL2L1	AHSG	PSMB10	ACVR2B	EPHB6

RBBP8	KDM5C	BCL3	STAC	SNRPG	ARHGEF6	PRIM2	RFC2	FGD1	STK35	PROX1	CAPZB	CBX4	STMN3	TGFB1I1	TBL1X
TXNL4A	AIF1	E2F3	CD36	NAP1L2	BMPR2	ANXA4	IRF6	PMF1	NEK9	RIPK3	APBB3	IL9	PLA2G4A	STMN2	HLTF
USP7	NOS3	KDM6B	MSTN	CCDC99	IBTK	SENP3	CCS	MED10	TBP	MAP3K6	IL3	EFNA5	DDX58	SEMA7A	SERPINA1
IL4	ALPL	MRPS26	ING4	MGMT	MAPK8IP3	CSTB	VAV2	UNC13B	REM2	RHOG	NAA16	MED13	RXRG	GNMT	RELA
HSPA1A	RRM1	HOXD8	POLE3	ARHGEF2	LBP	RICTOR	HIST2H4A	UIMC1	TRIM33	POLR3D	TLK1	NR1I2	ELAVL1	IL7	RNASE2
ACVR1	UBE2D3	IL18	EED	PVR	HIST1H4H	DPP3	HOXD3	MSH3	DNAJB2	BMP7	KRT18	NRP1	CETP	BMP4	MMP14
PACSIN3	PLP2	HDAC2	RPS3	NR0B1	ATRIP	PRKAR1A	CECR2	DET1	LRRK2	SF3A2	TP53INP1	PLAU	PRCP	AZI2	AVPR2
SCG2	ALDOA	DNAJA1	CD226	NFATC2IP	PRKCH	MCM6	ELF3	CDK1	NPPB	PYCARD	PACSIN2	PGLYRP1	CTNNBL1	HDAC8	GTF2A1
SUDS3	MCM3	CDC42BPG	FKBP8	SEMA3A	MSH5	FN1	RHO	PRPF31	TBL1XR1	ATP6V1H	NOTCH2	HFE	UQCRC2	HOXA3	UBASH3A
EDNRB	GHR	PCBP4	FLI1	TNFRSF10B	POLR3F	IFNA8	IL13RA1	CCNT1	FBXO30	HSP90AA2	ACTL6A	PSMB6	MED14	PTPN11	APOBEC3G
TUBB	NUP85	ERC1	IL5RA	ERCC8	AFG3L2	ADAM10	CPB2	STXBP3	ICAM1	PCNA	ILF3	TAF4	ARPC3	DDX1	ZFAND5
TRIP12	TNFRSF10A	PCBD2	KIF15	DBF4	HUWE1	HMGA1	CACNA1E	PTPN1	ANGPT2	PLD1	SETD7	DDI2	ABR	KDM4A	SATB1
RPS5	IL23A	ALS2	NTRK2	TRPC3	USP4	NCAPG	NOS1	HCLS1	RGS3	CHD1L	EIF6	KNTC1	RGS4	EPAS1	CCND2
HSD17B10	MAPK8IP1	PSMC6	SLC2A1	CCM2	SDC1	THOC4	FAU	SRI	MAP4K2	CASP8AP2	GRIN2B	RPS15A	IGHG2	MST1R	NKX2-5
ZRANB1	ADRBK2	PIK3CG	PRDM4	CRBN	DNAJA3	INSIG1	TSC2	REG1A	ENY2	PSMA7	CACNB3	AMHR2	TAF9B	TUBA1B	USP28
IFNA13	NOD1	ACVR1B	PCSK9	APC2	CUL4B	MAPKAPK2	PSMD10	CDC25B	SUPT7L	ITGA6	BMP2	SFN	UBE2O	APEX2	SDCBP
SOCS1	TRPC6	FLT1	RAGE	HIST4H4	ECE1	RGS1	HIST1H4D	TOB1	IQGAP2	CBX1	MDM4	PTK6	EPC1	NCK2	SIRT7
FKBP1B	KRT1	HMOX1	CDK7	CPLX2	NCR3	C5AR1	KLF16	RRM2B	DLX2	EFNB1	SERPINA5	CXCR2	TNNC1	MCM5	RNF8
FYB	IREB2	MED20	RMI1	CDH2	BCAR1	CHEK2	ATF4	HIVEP3	SETDB1	CNKSR1	PSMA3	ERCC3	KIAA1967	TOPORS	XAF1
SIN3B	MCM3AP	GRM1	SH3KBP1	FBXW8	RPS17	GAS7	CHRM5	IGBP1	EGLN1	UBE2I	TLR6	MYBPC3	IL1R1	BUB1	TBKBP1
EPS8	POLD3	TLR5	EFS	RAB22A	TAF12	DNAJC5	PLAUR	CSNK1D	STRBP	APBA3	EPB41	CHD4	USP5	TRIM24	SPRY2
TNNI3K	OTC	PRLR	KIR3DL2	ZFP57	FMOD	HLA-G	FST	F2RL2	TIA1	ROBO1	UBA1	CISH	PRDM1	CUL5	ROCK1
DPT	MMP1	CLASP1	NRF1	BLK	TAF10	CRKL	IDE	MPO	EDEM1	APBA1	SHANK2	IGKV4-1	PA2G4	PPP1R8	FANCB
PARP1	GNG12	ANAPC5	EPHA2	CCNB1	PEX5	BRD8	APITD1	ORC1L	SUV420H2	ASF1A	RASL10B	HSPA9	XRCC4	PLXNB1	ATG16L1
PTPN12	NEDD4L	UBE2J1	ESR2	NOX1	EREG	ABCC1	TNFRSF25	HOXD9	FGFR1OP	RALA	YAP1	RCAN1	AR	HRK	MAML1

DULLARD	RAD1	ZFP36L1	COL1A2	HIST1H3E	NEDD4	NCBP1	PPP1R2P3	FBXO31	CEL	ABCB1	EP400	STXBP4	KIT	ATRX	KRIT1
ARL2	AIFM1	CAPZA1	GNB1	PPP5C	HRAS	CATSPER3	REST	BPTF	RHOU	IFITM1	FSHR	MAP4K1	UHRF2	APCS	IL12A
EIF4EBP1	ZNF496	TAPBPL	RECQL5	APOH	HCFC1	FARP2	GRB2	XRCC5	VPS72	HPR	AMFR	BCAP31	NFYC	IFNA5	LY96
SHMT1	PTPN18	UBOX5	IER3	PSMB11	PSMA1	CREB5	VAMP8	RAD9A	MEF2D	F8	EDARADD	FBXO44	ARID4A	MED21	KDM5A
PPM1J	AATF	BMX	NF2	GTPBP4	RAD17	UBE2T	NFIB	PKN2	TEAD3	GINS4	BCL2	GSK3B	MYST4	ZW10	BCL2L11
HIST1H4L	CACNA1F	SIVA1	ACHE	BAT1	MAGED1	SET	TM2D1	RBBP4	DDX20	TNPO2	ACTN1	ZER1	GCH1	UBXN4	PHPT1
SH3BGR	MIA	ATPIF1	HP	UBE2V1	PF4	CD93	CITED2	CLU	INTS3	LPL	NR1H4	ELN	FOS	XPA	MEIS2
PDZK1	RB1CC1	POLR2E	TCF21	CASP1	ZNF830	PAIP2	PIF1	SMAD9	DAD1	CGA	MYOG	USP14	MYO6	RAB5C	FGF2
PDIA2	ADORA2B	SELE	ITGA11	MLF1	PAK7	RALGDS	DIRAS3	MED8	RPS9	CASC3	STK3	KAT2A	HIST1H2BM	SYT7	MECOM
PPP2R3B	BNIP1	EIF5A	UBB	FCAR	SNCAIP	REV1	ELMO1	LMX1A	AGT	IL15RA	ABI1	TRPV4	TNFRSF9	TAF5	GAS1
APBB1	BIRC2	IL6ST	SREBF1	RARG	SRP14	PLXNB2	TXN2	SLC9A3R1	ARHGAP1	GP5	USP34	GAB2	JAK1	ERCC4	POLM
VTN	TRIM16	STK19	IGF2	RAB10	BDKRB1	PMPCA	PSEN2	AGER	UBE2A	DNTT	RASA1	C1D	SIGIRR	MBL2	TIMP1
TNKS2	MVD	FGD6	ANAPC10	XRCC3	DDB1	POLD2	PARK7	GATA1	ARF4	FGB	AHR	TAF6	HIST1H1C	PSMB3	DNMT1
PRPF4B	SUFU	EFNA1	TNFRSF6B	ITGA1	LDLR	MYST1	HSP90AB1	SOCS3	GRIN2A	NUMB	HTRA2	SCAP	FGF13	TXNIP	TNFSF13
POLR2J	SH2B3	DEPDC1	FLNA	TERF2	KLRK1	PARD3	MED9	SKIV2L	PLA2G6	TLK2	LTB	CPD	ROCK2	FEZF1	NUDC
CUL1	ANXA1	IRF1	CBLC	CDC42EP1	MBD2	CBLB	PDGFB	ARL1	UBE2E1	SMAD3	TSC22D3	JUNB	DYRK1A	FLT3	FGFR3
DERL2	PWP2	CDC37	MIER1	IL1R2	RPTOR	DUSP16	NDE1	FBXO27	RSF1	CTDP1	TGFA	XPC	ANP32A	TNFRSF21	PMVK
CELA3B	NUSAP1	LEP	PRKCE	ANAPC13	NCR1	NR5A1	MECP2	NR4A1	KLF13	NCOA2	PDGFRB	CAMKK1	IGFBP2	PSMD2	DLC1
HIST1H4F	NAP1L3	IL1RAP	TBX2	STRADB	HIP1	GRIP1	TAF1	TUBA3D	EIF4E	PLN	FGF6	NEDD8	TNFSF15	KCNJ11	IL6
TWIST1	SMC2	ARFGEF2	CLK1	UNC5C	UBE2L3	SPG7	TUBGCP2	TGM2	CSF2	MT2A	MLF2	MCF2L	MAP3K1	PLAT	PRKCB
NEUROD1	NFKBIB	ACTL6B	ARHGEF7	IRS1	RBL1	EZH1	SPTBN1	S100A1	GALNT2	PSMC4	ABL2	FOXP1	PRNP	SNRNP200	HIST1H2AE
CSNK1A1	ITGAL	RET	CDC20	UBE2B	TCF19	POLA2	HLA-F	SMAP1	SIN3A	HIST2H3C	JUB	CAST	FKBPL	FGF1	F9
LDOC1	PEA15	CACYBP	DERL1	APP	GLI2	MGEA5	TOR1A	SIRT6	SS18	F2RL3	PLCB2	CTNNB1	CDK2AP1	ADAM15	GTF2F1
TUBA3C	POU2AF1	RUNX2	IRF9	CASP9	HIST1H4C	LAMC1	MAGI2	POLR2A	NUP98	PSMG1	EYA2	MTNR1A	CASP14	OPTN	UBE2H

ATG12	HTR2A	PTMS	PDPK1	PAG1	PTPN22	BMPR1B	JMJD7-PLA2G4B	SPRY1	TBL3	APH1A	TGFB2	IGFBP5	PIN1	CYP1A1	TLR9
MTSS1	SMAD4	CLUL1	ZBTB16	NFATC2	PPP1CB	XRCC2	TGIF2	ASF1B	JMJD6	FOXP2	RAB3C	PRMT5	GCGR	GFI1B	NCOR2
STX4	RAN	ADM	EDN1	ARHGEF11	RNF11	ACAP1	CNPY3	NIPA1	ARHGAP29	RAB14	RAB33A	RNF128	BCORL1	PDCD10	HIST1H4B
NOL3	RHOB	LRRC29	SAP130	CLOCK	MAPK12	BCCIP	KEAP1	ERN1	ULK1	TEAD4	SKP1	NDEL1	PDCD6	SIRT3	BOC
PELO	SKIL	CAMK2G	NASP	STX1A	SHC3	SLIT3	IHH	SERPIND1	TADA2A	PSMD4	JAG1	CD84	GRLF1	VIM	UBE2D2
TUBB2B	PSMB1	ACTR2	IL8	NFE2	VHL	IVNS1ABP	SNRPD2	MAML3	RBP3	NCBP2	LOX	REG3A	RPL22	KLK5	CDC7
SOCS6	TLR1	ATP2A1	FANCM	ERBB3	DMAP1	LMAN1	SERPINH1	PSMD1	COPS3	SNX6	CALR	CDH5	POT1	UBE2N	CSNK1E
PKN3	UNC5A	ITGB6	ATXN1	CDK2	LEPR	TIPIN	RHOD	PIAS4	ATXN7	SYCP2	MARK2	GPAM	ATP6V0A1	TCF12	SARM1
PKN1	CUL2	DNMT3L	BIRC5	SMARCD3	CHD3	GAB1	KRT19	AZGP1P1	NR0B2	EPHA4	KLF4	TJP1	PPP1R13B	ARNT2	RABEP1
COL1A1	SPOP	LRDD	NCOA3	RIT2	CDK11B	MYO1E	MITF	VCAM1	PPP1R15B	SUV39H1	APBB2	LAMB1	TAB2	TRIP6	SH3PXD2B
TNXB	NEK2	TSC22D4	MSX2	ADAMTS1	UBA52	GTSE1	PIAS3	PSMA8	ZIC2	MAPK11	NFYA	HTATIP2	STAT4	HDAC4	INS
ABCB4	UCN3	PDP1	AXL	INTS1	RELN	FCGR2B	RAB40AL	GTF2I	PRKCSH	KRT5	RUNX1T1	CD209	PANX1	GRN	LILRB3
MTA2	KLHL13	UBXN6	AURKB	CASP5	ANAPC2	AIMP2	DUSP1	NOSTRIN	IL29	PSMB2	PAX8	SPDYA	NR3C1	ITGAV	H3F3B
PAK3	RIF1	NCOA6	MAFG	HIPK2	CENPK	NUP62	FKBP4	BAT5	YEATS4	MR1	CDH1	DDOST	MAD2L2	PTPRM	TAOK1
SMAD1	STRN	VPS4B	POLR2B	SUGT1	HDAC5	SQLE	DCLRE1A	PPM1L	MYH10	SORD	MAGI1	PADI4	FUS	SRP19	WHSC1
CAMK2D	BRE	IGHE	ID3	IKBKE	PMS2L1	FPR2	ARHGDIB	MBD4	CDC16	BRCA2	ACTN4	ADAM8	CD5	PINK1	PPP6C
SP1	BMPR1A	TNFSF14	IFNA17	NCDN	CHRAC1	RAD9B	NR2E3	TRRAP	CREB1	ATXN7L3	FGFR2	SIP1	SH3BP5	HIST1H3A	SHARPIN
HIST1H2BB	WASF3	MSH4	POLD4	NTN1	RPA1	JPH3	PIK3CA	CD1B	RAD52	PPARGC1A	CRHR2	NCKAP1	NEIL1	HMGB1	CTBP1
GTF3C4	MTMR15	VCL	SF3A3	YWHAH	TLR7	KLK13	HBEGF	DAB1	STXBP1	CD300LG	ATF3	CUEDC2	HMGCR	POLR2G	ASB2
MYF6	PTPN23	FBXW7	ANAPC11	FOXH1	PRKCI	GNAI1	PSMA2	TUSC2	MLL	C1QBP	PSMA4	CCNK	ATG5	MAFF	DFFA
HIST2H2BE	CDC23	CACNA1C	PSME2	POLR2I	GP1BB	SND1	HNRNPUL1	IGHD	DSCC1	LIME1	MAPK8	AP2S1	HSPA8	CDC42	HIST1H4J
ADRB3	EPHA3	BMI1	LDLRAP1	CHD8	BAIAP2	CXCL12	DDX11	ID1	IL17A	MSH2	POLR2D	PSMF1	CAMK1	ZFP36	MAEA
AMBP	APTX	SETD8	TFAM	TPT1	ITGB1	RNF41	ALK	POMP	PARVA	TINAGL1	GNAQ	POLR3G	SLC6A3	AFAP1L2	RACGAP1P
RPS6KA3	IL34	MYO9B	ANAPC4	CEBPB	PCSK2	BANF1	IL9R	AVP	SHKBP1	RFC3	NPM1	FCGR3B	RAC1	CMTM3	CAP1
SHOC2	SPINT1	JAK2	MAPK3	WWP1	THEG	GH1	DLX5	ITGAM	CCL3	CFLAR	IL5	PROC	NLRX1	DAXX	LCN2

EDA2R	EGF	POLD1	STUB1	F12	BAD	HAT1	CATSPER1	CHD9	NISCH	FANCG	RAET1G	PTPN2	ZMYND11	CD48	USP25
EGR1	EPHB2	TP73	MMP9	LIMK1	ACVR1C	DECR1	TIMP2	SMARCA1	TFDP1	NR3C2	HLA-DRB3	ANXA2	CLPTM1	PRPF19	OSM
RHEBL1	PRKG1	USP21	SRC	TLN1	ARID1B	SDC4	IL28RA	SENP1	TIAM1	NR1H3	KCNA2	RB1	TRPC4AP	F7	SGK3
HNF4A	PIKFYVE	CCNG1	PPP3CA	ESRRB	MAPK1	RHOBTB1	IFT57	CLK3	MYF5	MYC	HIST3H3	RANBP9	PFDN6	ZNF354A	IL2
SETD1A	CCNE1	DHX58	NGF	MAF	RBM4	TNFRSF10D	MED1	SELP	UBE2G2	PDCD1	PACSIN1	CTSD	TNIP2	MAS1	SLC9A6
SSRP1	AMH	CD44	FER	HNF4G	MAP1LC3A	MXI1	KHDRBS1	ANXA5	CDK5	CYLD	TRAF5	KIR2DS3	EIF3B	TAB1	GTF2E2
ARL3	PSMD3	TP53BP2	FOSB	RAD21	ABI2	VNN2	PPP2R2A	COPS2	TNIP1	ERCC5	POLR3A	PSMA5	TNFRSF8	CUX1	ELK1
DOK3	BHLHA15	PCGF2	STAR	DNAJA2	TP63	SRF	MED18	CARD8	UBE2E2	LHX4	ANPEP	POLI	NDFIP2	CDC34	TLR8
KLK3	JUND	ERCC1	ZMIZ2	ACVRL1	PPM1E	SUMO3	ATP2A2	TRIB3	TNFRSF14	MTOR	AMOT	PSMB7	DCAF11	SMAD6	RPS19
PPP2R5C	SSBP1	ZNF148	ASB3	ABCF1	CALM2	IL1A	GRAP	MAP3K5	PAK4	SMC3	HIST1H3B	DAPP1	PHF21A	NPC1	CFL1
GPC3	IL18R1	NFATC3	RIOK3	MSL3	SCRIB	HLA-E	BRD4	SRPK2	DCX	MXD3	UBA7	SPHK2	NFE2L2	BMP6	YWHAB
ITPR1	NOTCH1	PHB2	TSHR	UBXN1	GAS2	RBL2	THRB	CALM3	APPL1	PPM1B	GTF2H3	NUP205	ULBP1	FHL2	FANCD2
GLRX	BCOR	MX1	HIF1A	TNFSF4	SMG7	PI4KA	AP2A2	PSMC1	IRF4	PTPRCAP	CAPZA2	DVL3	HERC5	GNAL	CDKN2D
YAF2	RASA3	ODC1	STK25	TRIOBP	BRD7	CABIN1	APBA2	EEF1A2	ZHX2	NRD1	IGHG4	PRKD3	CD2AP	EIF4E2	ERLEC1
HSPE1	CSNK2A2	IFNA1	ITGB3	LNX2	IP6K2	MEN1	USP11	GTF2F2	THY1	COL3A1	VIP	SMAD7	SETMAR	NOS2	SERTAD3
HOMER1	RHOQ	HDAC1	PLA2G2A	LHCGR	NME1	NCK1	PYDC1	PTGES3	CAMTA2	SLA2	SKAP1	FBXL15	LAMA2	SPN	TOLLIP
LTF	IGFBP7	SLC9A3R2	RNF19A	RAB3A	NGFRAP1	PDCD4	GNA15	SPON2	HBZ	HGFAC	IL1RL2	MED12	SAPS3	HSPA5	SEMA4D
TNFSF9	HSPD1	UBA5	RAB26	PRKD2	ING3	UBE2K	MLH1	SRCAP	HDAC10	VEGFC	RARB	PPM1D	S100B	ARR3	UBE4B
SKI	LMO2	RAB11A	SIRT1	DMC1	PLCG1	ELMO2	ARHGEF1	DNA2	CD63	YWHAG	KIR3DS1	IGSF11	CRK	SNUPN	CSF1R
SQSTM1	OCA2	DLG1	MUSK	CLNS1A	HIST1H4K	TXK	USF1	STIM2	KLHL21	RPA4	MIF4GD	TIFA	XPO1	MUTYH	DCC
TAOK3	UNC5B	RTN4	H2AFX	C4BPA	MAP4K5	RAB6A	CENPA	ADCY8	HACE1	PPP3CB	GTF2H1	HOXB3	MAP3K15	F2RL1	ACTG1
PRKRIR	PPP1R13L	TRIM27	KPNB1	BAG4	APC	DTNBP1	AES	TDG	TRBC1	TIMM50	EBI3	TAF11	IGHA2	NFATC1	BRCC3
TNNI3	PSMD6	PTCRA	SGOL2	RNF216	SMARCA4	TXN	NOL8	FAP	NFKBIL2	APOC1	RNGTT	SF3A1	FTH1	APPL2	SPARC
ESPN	MLLT1	XAB2	PRSS2	RPA3	PSMC5	RAB2A	CACNG2	GP9	SLC25A4	MORF4L2	GOLT1B	CD8B	HAND2	ATP5A1	NCOA4
DNM2	DOK1	SLC26A9	IGLC3	DDB2	IGHA1	CUL7	RXRA	DDIT3	F11	SHB	CDK6	PLAA	BLOC1S2	ZHX1	EGFL8

PPP2CA	DFFB	CXCL10	ACTA1	CD109	TXNRD1	ING1	AP1M1	CLN3	NCAPD3	THBS1	KAT2B	FIG4	NFIX	GOPC	CD24L4
FBXW2	TLX1	AURKA	MIB2	ETF1	NGFR	OCLN	CBL	F2R	TRIB1	IL18BP	MXD4	APLP2	MED27	TPM1	MAP3K4
RAET1E	TNKS	PKD2	VAPB	MMP25	POU1F1	C3AR1	H3F3A	RAD23B	MCM2	HUS1	PGF	IL12RB2	EPHB3	SLAMF7	SERPINB9
HIPK3	TRAF4	PRKACG	CREG1	TNFRSF1B	MAPK9	TFF3	DHH	TNK2	TOPBP1	ATP6V1B1	RAET1L	PPM1K	RGS6	HOXB1	MED6
FTL	PRSS29P	CENPE	WIBG	HSF1	CRADD	OLR1	EIF4B	NDC80	MAP3K2	HSPH1	DUSP10	RPS24	APLF	STAT5A	MAP3K14
CASP2	VAV3	BMP15	CDC73	RPA2	SMPD2	TNFSF13B	SOX2	SMAD2	LTBR	ARHGAP5	SCNN1G	RHEB	CBFA2T3	BAG3	SLC1A2
RAD51	WDTC1	CCND3	FBXO2	PML	IGHV3-23	F3	DEDD	IPO13	TAL1	RRAS2	IRAK1BP1	PPM1G	KPNA2	MLL5	BUD31
MIB1	HSPB1	DDX23	TJP2	TGFB1	CLPTM1L	PARP2	CDT1	CAMK4	CUL9	SUZ12	TUBG1	PSMC3	SNCG	IRF2	HIST1H1B
NTN3	UFSP2	KAT5	FGFR1	NR2F6	LSP1	HSP90B1	VAMP3	AKT1	IL23R	IFNAR1	TRHDE	RAB12	NFKBIZ	HMG20B	FES
TSPAN4	GIPC1	CIRBP	CER1	PRL	CBX5	FANCF	NR2C1	H2AFY	PSMD13	HMGB2	HSPB2	RGMA	ERLIN2	GREM1	CALD1
SPI1	IFNB1	AGAP2	GML	E2F6	NDRG1	DAB2	TAF7	AXIN1	MDM2	SMURF2	SSR1	LGALS1	ZEB2	CD1D	LTK
NLRP2	TCP1	CDC45	KCNQ1	CALM1	TADA1	TP53	FBP1	WARS	CEP63	HBB	MAP1LC3B	PTPN13	ARAF	DOM3Z	ORC3L
UBE2W	PDF	SH2D2A	NR1H2	MBD3	CRP	CADM1	PDCD5	SHROOM2	DGKG	SYCP3	PRKCZ	ASH2L	SIT1	LY9	NRBP1
EGR4	REN	CCND1	SMARCA5	ATXN3	CD74	RPS13	EBF1	DUSP9	CAPN6	CA2	UPF1	DEPDC6	ETS2	ANKHD1-EIF4EBP3	HDAC7
MCM10	CPEB1	NDN	SYNJ2	NCOR1	MED30	PURB	CXCR3	SMN1	SERPINF1	CDK4	HIST2H2AA3	PARP3	RAB34	DPF2	TUBA4A
PTGS2	ZFYVE9	HSPA1B	IGLC1	ERO1L	MNT	SKP2	GRB14	SGK1	TCF4	RFC1	SERPINC1	FBXW5	PMS2L5	XBP1	SLC30A1
RASL10A	PPP2R1A	CCK	GDF6	PLEKHB1	FZD1	IGHV2-70	IKZF2	CDK9	ITGA2	FCGR1A	TYRO3	MPL	CHRM3	TANK	MED15
SUMO1P3	ACE2	RAD50	CDC42EP4	GTF2B	PI4KB	ARF5	KDM2B	CASK	JUN	TOM1L1	PSMB4	RALBP1	DNMBP	PIAS1	NMT1
BACE1	AQP2	GJA1	PSME3	POLH	GAA	HOXC8	EIF2C1	GLI3	BTRC	DPF3	NFKBIL1	PRKAA1	LNPEP	THRA	TRIM21
VCPIP1	PRKD1	PRDX1	TGFB3	SNRPD3	ORC4L	PXN	RPS6	SOS2	PTK2B	MED19	POU2F1	PPM1A	AIP	HDAC9	APOA4
IL15	KLK4	NCAPD2	F10	TAF2	ANGPT1	FBXL3	PPP1CC	SORBS1	ERG	POLB	CDC25C	PTTG1	JPH2	SMARCA2	KIFAP3
SPHK1	PHB	ATMIN	CSNK2B	VAV1	WASL	UFD1L	PBRM1	SPAG5	TRIO	LCP2	CYP1B1	DSTN	FOXO3	TADA3	TRIM69
MLST8	PRIM1	RFWD2	COL11A1	MAP2	TRAF1	PSMD7	ADCY2	PTPRZ1							

References

1. Bousfiha, A.A., et al., *Primary immunodeficiency diseases worldwide: more common than generally thought.* J Clin Immunol, 2013. **33**(1): p. 1-7.

2. Edgar, J.D., et al., *The United Kingdom Primary Immune Deficiency (UKPID) Registry: report of the first 4 years' activity 2008-2012.* Clin Exp Immunol, 2014. **175**(1): p. 68-78.

3. Resnick, E.S., et al., *Morbidity and mortality in common variable immune deficiency over 4 decades.* Blood, 2012. **119**(7): p. 1650-7.

4. Cuvelier, G.D., et al., *Long-Term Outcomes of Hematopoietic Stem Cell Transplantation for ZAP70 Deficiency.* J Clin Immunol, 2016. **36**(7): p. 713-24.

5. Shillitoe, B., et al., *The United Kingdom Primary Immune Deficiency (UKPID) registry 2012 to 2017.* Clin Exp Immunol, 2018. **192**(3): p. 284-291.

6. Laberko, A. and A.R. Gennery, *Clinical considerations in the hematopoietic stem cell transplant management of primary immunodeficiencies.* Expert Rev Clin Immunol, 2018. **14**(4): p. 297-306.

7. Bousfiha, A., et al., *The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies.* J Clin Immunol, 2018. **38**(1): p. 129-143.

8. Notarangelo, L.D., *Primary immunodeficiencies.* Journal of Allergy and Clinical Immunology, 2010. **125**(2): p. S182-S194.

9. Balduini, C.L. and A. Savoia, *Genetics of familial forms of thrombocytopenia.* Hum Genet, 2012. **131**(12): p. 1821-32.

10. Morio, T., *Recent advances in the study of immunodeficiency and DNA damage response.* Int J Hematol, 2017. **106**(3): p. 357-365.

11. Staines Boone, A.T., et al., *Failing to Make Ends Meet: The Broad Clinical Spectrum of DNA Ligase IV Deficiency. Case Series and Review of the Literature.* Front Pediatr, 2018. **6**: p. 426.

12. Chrzanowska, K.H., et al., *Nijmegen breakage syndrome (NBS).* Orphanet J Rare Dis, 2012. **7**: p. 13.

13. Martinez, A.R., et al., *Differential requirements for DNA repair proteins in immortalized cell lines using alternative lengthening of telomere mechanisms.* Genes Chromosomes Cancer, 2017. **56**(8): p. 617-631.

14. Cavalieri, S., et al., *Deep-intronic ATM mutation detected by genomic resequencing and corrected in vitro by antisense morpholino oligonucleotide (AMO).* Eur J Hum Genet, 2013. **21**(7): p. 774-8.

15. Schatorje, E., et al., *Primary immunodeficiency associated with chromosomal aberration - an ESID survey.* Orphanet J Rare Dis, 2016. **11**(1): p. 110.

16. Sanyal, M., et al., *Lack of IL7Ralpha expression in T cells is a hallmark of T-cell immunodeficiency in Schimke immuno-osseous dysplasia (SIOD).* Clin Immunol, 2015. **161**(2): p. 355-65.

17. Llorens-Agost, M., et al., *Analysis of novel missense ATR mutations reveals new splicing defects underlying Seckel syndrome.* Hum Mutat, 2018. **39**(12): p. 1847-1853.

18. Battaglia, A., J.C. Carey, and S.T. South, *Wolf-Hirschhorn syndrome: A review and update.* Am J Med Genet C Semin Med Genet, 2015. **169**(3): p. 216-23.

19. Ramirez-Alejo, N., et al., *Novel hypomorphic mutation in IKBKG impairs NEMO-ubiquitylation causing ectodermal dysplasia, immunodeficiency, incontinentia pigmenti, and immune thrombocytopenic purpura.* Clin Immunol, 2015. **160**(2): p. 163-71.

20. Ohnishi, H., et al., *Immunodeficiency in Two Female Patients with Incontinentia Pigmenti with Heterozygous NEMO Mutation Diagnosed by LPS Unresponsiveness.* J Clin Immunol, 2017. **37**(6): p. 529-538.

21. Lopez-Granados, E., et al., *A novel mutation in NFKBIA/IKBA results in a degradation-resistant N-truncated protein and is associated with ectodermal dysplasia with immunodeficiency.* Hum Mutat, 2008. **29**(6): p. 861-8.

22. Sanchez, L.A., et al., *Two Sides of the Same Coin: Pediatric-Onset and Adult-Onset Common Variable Immune Deficiency.* J Clin Immunol, 2017. **37**(6): p. 592-602.

23. Maffucci, P., et al., *Genetic Diagnosis Using Whole Exome Sequencing in Common Variable Immunodeficiency.* Front Immunol, 2016. **7**: p. 220.

24. Berglund, L.J., S.W. Wong, and D.A. Fulcher, *B-cell maturation defects in common variable immunodeficiency and association with clinical features.* Pathology, 2008. **40**(3): p. 288-94.

25. Warnatz, K., et al., *Severe deficiency of switched memory B cells (CD27(+)IgM(-)IgD(-)) in subgroups of patients with common variable immunodeficiency: a new approach to classify a heterogeneous disease.* Blood, 2002. **99**(5): p. 1544-51.

26. Yazdani, R., et al., *Selective IgA Deficiency: Epidemiology, Pathogenesis, Clinical Phenotype, Diagnosis, Prognosis and Management.* Scand J Immunol, 2017. **85**(1): p. 3-12.

27. Abolhassani, H., et al., *Clinical, immunologic, and genetic spectrum of 696 patients with combined immunodeficiency.* J Allergy Clin Immunol, 2017.

28. Bogaert, D.J., et al., *Genes associated with common variable immunodeficiency: one diagnosis to rule them all?* J Med Genet, 2016. **53**(9): p. 575-90.

29. Hou, T.Z., et al., *Identifying functional defects in patients with immune dysregulation due to LRBA and CTLA-4 mutations.* Blood, 2017. **129**(11): p. 1458-1468.

30. Schwab, C., et al., *Phenotype, penetrance, and treatment of 133 cytotoxic T-lymphocyte antigen 4-insufficient subjects.* J Allergy Clin Immunol, 2018. **142**(6): p. 1932-1946.

31. Volanakis, J.E., et al., *Major histocompatibility complex class III genes and susceptibility to immunoglobulin A deficiency and common variable immunodeficiency.* J Clin Invest, 1992. **89**(6): p. 1914-22.

32. Tuijnenburg, P., et al., *Loss-of-function nuclear factor kappaB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans.* J Allergy Clin Immunol, 2018.

33. Salzer, U., et al., *Relevance of biallelic versus monoallelic TNFRSF13B mutations in distinguishing disease-causing from risk-increasing TNFRSF13B variants in antibody deficiency syndromes.* Blood, 2009. **113**(9): p. 1967-76.

34. Zur Stadt, U., et al., *Mutation spectrum in children with primary hemophagocytic lymphohistiocytosis: molecular and functional analyses of PRF1, UNC13D, STX11, and RAB27A.* Hum Mutat, 2006. **27**(1): p. 62-8.

35. Mukda, E., et al., *Exome sequencing for simultaneous mutation screening in children with hemophagocytic lymphohistiocytosis.* International Journal of Hematology, 2017. **106**(2): p. 282-290.

36. Zhang, K., et al., *Hypomorphic mutations in PRF1, MUNC13-4, and STXBP2 are associated with adult-onset familial HLH.* Blood, 2011. **118**(22): p. 5794.

37. Nijman, I.J., et al., *Targeted next-generation sequencing: a novel diagnostic tool for primary immunodeficiencies.* J Allergy Clin Immunol, 2014. **133**(2): p. 529-34.

38. Rigaud, S., et al., *XIAP deficiency in humans causes an X-linked lymphoproliferative syndrome.* Nature, 2006. **444**(7115): p. 110-4.

39. Fodil, N., D. Langlais, and P. Gros, *Primary Immunodeficiencies and Inflammatory Disease: A Growing Genetic Intersection.* Trends in immunology, 2016. **37**(2): p. 126-140.

40. Uhlig, H.H., et al., *The diagnostic approach to monogenic very early onset inflammatory bowel disease.* Gastroenterology, 2014. **147**(5): p. 990-1007.e3.

41. Zhu, L., et al., *IL-10 and IL-10 Receptor Mutations in Very Early Onset Inflammatory Bowel Disease.* Gastroenterology research, 2017. **10**(2): p. 65-69.

42. Chun, H.J., et al., *Pleiotropic defects in lymphocyte activation caused by caspase-8 mutations lead to human immunodeficiency.* Nature, 2002. **419**(6905): p. 395-9.

43. Rieux-Laucat, F., A. Magerus-Chatinet, and B. Neven, *The Autoimmune Lymphoproliferative Syndrome with Defective FAS or FAS-Ligand Functions.* J Clin Immunol, 2018.

44. Li, P., et al., *Updated Understanding of Autoimmune Lymphoproliferative Syndrome (ALPS).* Clin Rev Allergy Immunol, 2016. **50**(1): p. 55-63.

45. Takagi, M., et al., *Autoimmune lymphoproliferative syndrome-like disease with somatic KRAS mutation.* Blood, 2011. **117**(10): p. 2887-90.

46. Oliveira, J.B., et al., *NRAS mutation causes a human autoimmune lymphoproliferative syndrome.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8953-8.

47. Humbert, L., et al., *Chronic Mucocutaneous Candidiasis in Autoimmune Polyendocrine Syndrome Type 1.* Front Immunol, 2018. **9**: p. 2570.

48. Gambineri, E., et al., *Clinical, Immunological, and Molecular Heterogeneity of 173 Patients With the Phenotype of Immune Dysregulation, Polyendocrinopathy, Enteropathy, X-Linked (IPEX) Syndrome.* Front Immunol, 2018. **9**: p. 2411.

49. Cepika, A.M., et al., *Tregopathies: Monogenic diseases resulting in regulatory T-cell deficiency.* J Allergy Clin Immunol, 2018. **142**(6): p. 1679-1695.

50. Kuehn, H.S., et al., *Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4.* Science, 2014. **345**(6204): p. 1623-7.

51. Chiriaco, M., et al., *Chronic granulomatous disease: Clinical, molecular, and therapeutic aspects.* Pediatr Allergy Immunol, 2016. **27**(3): p. 242-53.

52. Thomas, D.C., *The phagocyte respiratory burst: Historical perspectives and recent advances.* Immunol Lett, 2017. **192**: p. 88-96.

53. Arnadottir, G.A., et al., *A homozygous loss-of-function mutation leading to CYBC1 deficiency causes chronic granulomatous disease.* Nat Commun, 2018. **9**(1): p. 4447.

54. Ambruso, D.R., et al., *Human neutrophil immunodeficiency syndrome is associated with an inhibitory Rac2 mutation.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4654-9.

55. Kazenwadel, J., et al., *Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature.* Blood, 2012. **119**(5): p. 1283-91.

56. Serwas, N.K., et al., *CEBPE-Mutant Specific Granule Deficiency Correlates With Aberrant Granule Organization and Substantial Proteome Alterations in Neutrophils.* Front Immunol, 2018. **9**: p. 588.

57. Hildebrandt, J., et al., *Characterization of CSF2RA mutation related juvenile pulmonary alveolar proteinosis.* Orphanet journal of rare diseases, 2014. **9**: p. 171-171.

58. Suzuki, T., et al., *Hereditary pulmonary alveolar proteinosis caused by recessive CSF2RB mutations.* Eur Respir J, 2011. **37**(1): p. 201-4.

59. Siler, U., et al., *Severe glucose-6-phosphate dehydrogenase deficiency leads to susceptibility to infection and absent NETosis.* J Allergy Clin Immunol, 2017. **139**(1): p. 212-219.e3.

60. Klein, C., et al., *HAX1 deficiency causes autosomal recessive severe congenital neutropenia (Kostmann disease).* Nat Genet, 2007. **39**(1): p. 86-92.

61. Skokowa, J., et al., *Severe congenital neutropenias.* Nat Rev Dis Primers, 2017. **3**: p. 17032.

62. Dang, T.S., et al., *Defective Leukocyte Adhesion and Chemotaxis Contributes to Combined Immunodeficiency in Humans with Autosomal Recessive MST1 Deficiency.* J Clin Immunol, 2016. **36**(2): p. 117-22.

63. Etzioni, A., *Genetic etiologies of leukocyte adhesion defects.* Curr Opin Immunol, 2009. **21**(5): p. 481-6.

64. Harris, E.S., A.S. Weyrich, and G.A. Zimmerman, *Lessons from rare maladies: leukocyte adhesion deficiency syndromes.* Curr Opin Hematol, 2013. **20**(1): p. 16-25.

65. Robert, P., et al., *A novel leukocyte adhesion deficiency III variant: kindlin-3 deficiency results in integrin- and nonintegrin-related defects in different steps of leukocyte adhesion.* J Immunol, 2011. **186**(9): p. 5273-83.

66. Hall, G.W., P. Dale, and J.A. Dodge, *Shwachman-Diamond syndrome: UK perspective.* Archives of disease in childhood, 2006. **91**(6): p. 521-524.

67. Cohen, T.S. and A. Prince, *Cystic fibrosis: a mucosal immunodeficiency syndrome.* Nat Med, 2012. **18**(4): p. 509-19.

68. Balikova, I., et al., *Deletions in the VPS13B (COH1) gene as a cause of Cohen syndrome.* Hum Mutat, 2009. **30**(9): p. E845-54.

69. Picard, C., et al., *Clinical features and outcome of patients with IRAK-4 and MyD88 deficiency.* Medicine (Baltimore), 2010. **89**(6): p. 403-25.

70. Maglione, P.J., N. Simchoni, and C. Cunningham-Rundles, *Toll-like receptor signaling in primary immune deficiencies.* Ann N Y Acad Sci, 2015. **1356**: p. 1-21.

71. Chapgier, A., et al., *Human complete Stat-1 deficiency is associated with defective type I and II IFN responses in vitro but immunity to some low virulence viruses in vivo.* J Immunol, 2006. **176**(8): p. 5078-83.

72. Shahni, R., et al., *Signal transducer and activator of transcription 2 deficiency is a novel disorder of mitochondrial fission.* Brain, 2015. **138**(Pt 10): p. 2834-46.

73. Hernandez, P.A., et al., *Mutations in the chemokine receptor gene CXCR4 are associated with WHIM syndrome, a combined immunodeficiency disease.* Nat Genet, 2003. **34**(1): p. 70-4.

74. Takezaki, S., et al., *Chronic mucocutaneous candidiasis caused by a gain-of-function mutation in the STAT1 DNA-binding domain.* J Immunol, 2012. **189**(3): p. 1521-6.

75. Liu, L., et al., *Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis.* J Exp Med, 2011. **208**(8): p. 1635-48.

76. Lanternier, F., et al., *Inherited CARD9 deficiency in otherwise healthy children and adults with Candida species-induced meningoencephalitis, colitis, or both.* J Allergy Clin Immunol, 2015. **135**(6): p. 1558-68.e2.

77. Bogunovic, D., et al., *Mycobacterial disease and impaired IFN-gamma immunity in humans with inherited ISG15 deficiency.* Science, 2012. **337**(6102): p. 1684-8.

78. Okada, S., et al., *IMMUNODEFICIENCIES. Impairment of immunity to Candida and Mycobacterium in humans with bi-allelic RORC mutations.* Science, 2015. **349**(6248): p. 606-13.

79. Ramirez-Alejo, N. and L. Santos-Argumedo, *Innate defects of the IL-12/IFN-gamma axis in susceptibility to infections by mycobacteria and salmonella.* J Interferon Cytokine Res, 2014. **34**(5): p. 307-17.

80. Bustamante, J., et al., *Mendelian susceptibility to mycobacterial disease: genetic, immunological, and clinical features of inborn errors of IFN-gamma immunity.* Semin Immunol, 2014. **26**(6): p. 454-70.

81. Granel, B., et al., *Overlap syndrome between FMF and TRAPS in a patient carrying MEFV and TNFRSF1A mutations.* Clin Exp Rheumatol, 2007. **25**(4 Suppl 45): p. S93-5.

82. Dode, C., et al., *The enlarging clinical, genetic, and population spectrum of tumor necrosis factor receptor-associated periodic syndrome.* Arthritis Rheum, 2002. **46**(8): p. 2181-8.

83. Marzano, A.V., et al., *Autoinflammation in pyoderma gangrenosum and its syndromic form (pyoderma gangrenosum, acne and suppurative hidradenitis).* Br J Dermatol, 2017. **176**(6): p. 1588-1598.

84. Arostegui, J.I., et al., *NOD2 gene-associated pediatric granulomatous arthritis: clinical diversity, novel and recurrent mutations, and evidence of clinical improvement with interleukin-1 blockade in a Spanish cohort.* Arthritis Rheum, 2007. **56**(11): p. 3805-13.

85. Liu, Y., et al., *Activated STING in a vascular and pulmonary syndrome.* N Engl J Med, 2014. **371**(6): p. 507-18.

86. Rodero, M.P. and Y.J. Crow, *Type I interferon-mediated monogenic autoinflammation: The type I interferonopathies, a conceptual overview.* J Exp Med, 2016. **213**(12): p. 2527-2538.

87. Genel, F., et al., *Properdin deficiency in a boy with fulminant meningococcal septic shock.* Acta Paediatr, 2006. **95**(11): p. 1498-1500.

88. Ross, S.C. and P. Densen, *Complement deficiency states and infection: epidemiology, pathogenesis and consequences of neisserial and other infections in an immune deficiency.* Medicine (Baltimore), 1984. **63**(5): p. 243-73.

89. Nishizaka, H., et al., *Genetic bases of human complement C7 deficiency.* J Immunol, 1996. **157**(9): p. 4239-43.

90. Kojima, T., et al., *Genetic basis of human complement C8 alpha-gamma deficiency.* J Immunol, 1998. **161**(7): p. 3762-6.

91. Kaufmann, T., et al., *Genetic basis of human complement C8 beta deficiency.* J Immunol, 1993. **150**(11): p. 4943-7.

92. Witzel-Schlomp, K., et al., *The human complement C9 gene: identification of two mutations causing deficiency and revision of the gene structure.* J Immunol, 1997. **158**(10): p. 5043-9.

93. Kirschfink, M., et al., *Complete functional C1q deficiency associated with systemic lupus erythematosus (SLE).* Clin Exp Immunol, 1993. **94**(2): p. 267-72.

94. McAdam, R.A., D. Goundis, and K.B. Reid, *A homozygous point mutation results in a stop codon in the C1q B-chain of a C1q-deficient individual.* Immunogenetics, 1988. **27**(4): p. 259-64.

95. Petry, F., et al., *Non-sense and missense mutations in the structural genes of complement component C1q A and C chains are linked with two different types of complete selective C1q deficiencies.* J Immunol, 1995. **155**(10): p. 4734-8.

96. Zipfel, P.F., et al., *Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome.* PLoS Genet, 2007. **3**(3): p. e41.

97. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-351.

98. Huptas, C., S. Scherer, and M. Wenning, *Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly.* BMC Res Notes, 2016. **9**: p. 269.

99. Sobreira, N.L., et al., *Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene.* PLoS Genet, 2010. **6**(6): p. e1000991.

100. Lupski, J.R., et al., *Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.* N Engl J Med, 2010. **362**(13): p. 1181-91.

101. Conley, M.E., et al., *Agammaglobulinemia and absent B lineage cells in a patient lacking the p85alpha subunit of PI3K.* J Exp Med, 2012. **209**(3): p. 463-70.

102. Engelhardt, K.R., et al., *Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by Whole Exome Sequencing.* J Clin Immunol, 2017. **37**(1): p. 42-50.

103. Marian, A.J., *Sequencing Your Genome: What Does It Mean?* Methodist DeBakey Cardiovascular Journal, 2014. **10**(1): p. 3-6.

104. Danielsson, K., et al., *Next-generation sequencing applied to rare diseases genomics.* Expert Rev Mol Diagn, 2014. **14**(4): p. 469-87.

105. Samarakoon, P.S., et al., *Identification of copy number variants from exome sequence data.* BMC Genomics, 2014. **15**: p. 661.

106. Belkadi, A., et al., *Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants.* Proc Natl Acad Sci U S A, 2015. **112**(17): p. 5473-8.

107. Kebschull, J.M. and A.M. Zador, *Sources of PCR-induced distortions in high-throughput sequencing data sets.* Nucleic Acids Res, 2015. **43**(21): p. e143.

108. Ekblom, R. and J.B. Wolf, *A field guide to whole-genome sequencing, assembly and annotation.* Evol Appl, 2014. **7**(9): p. 1026-42.

109. Stepensky, P., et al., *Deep intronic mis-splicing mutation in JAK3 gene underlies T-B+NK- severe combined immunodeficiency phenotype.* Clin Immunol, 2016. **163**: p. 91-5.

110. DeFrancesco, L., *Life Technologies promises $1,000 genome.* Nat Biotechnol, 2012. **30**(2): p. 126.

111. Ma, C.S., et al., *Functional STAT3 deficiency compromises the generation of human T follicular helper cells.* Blood, 2012. **119**(17): p. 3997-4008.

112. Boztug, H., et al., *NF-kappaB1 Haploinsufficiency Causing Immunodeficiency and EBV-Driven Lymphoproliferation.* J Clin Immunol, 2016. **36**(6): p. 533-40.

113. Takagi, M., et al., *Autoimmune lymphoproliferative syndrome–like disease with somatic KRAS mutation.* Blood, 2011. **117**(10): p. 2887-2890.

114. Omoyinmi, E., et al., *Brief Report: whole-exome sequencing revealing somatic NLRP3 mosaicism in a patient with chronic infantile neurologic, cutaneous, articular syndrome.* Arthritis Rheumatol, 2014. **66**(1): p. 197-202.

115. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.

116. Whiffin, N., J.S. Ware, and A. O'Donnell-Luria, *Improving the Understanding of Genetic Variants in Rare Disease With Large-scale Reference Populations.* Jama, 2019.

117. Meyts, I., et al., *Exome and genome sequencing for inborn errors of immunity.* The Journal of allergy and clinical immunology, 2016. **138**(4): p. 957-969.

118. Merkle, Florian T. and K. Eggan, *Modeling Human Disease with Pluripotent Stem Cells: from Genome Association to Function.* Cell Stem Cell, 2013. **12**(6): p. 656-668.

119. Deshpande, A., et al., *Cellular Phenotypes in Human iPSC-Derived Neurons from a Genetic Model of Autism Spectrum Disorder.* Cell Rep, 2017. **21**(10): p. 2678-2687.

120. Hemminki, K., A. Försti, and J.L. Bermejo, *The 'common disease-common variant' hypothesis and familial risks.* PloS one, 2008. **3**(6): p. e2504-e2504.

121. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.

122. Lander, E.S., *The new genomics: global views of biology.* Science, 1996. **274**(5287): p. 536-9.

123. Lee, J.C., et al., *Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease.* Nature genetics, 2017. **49**(2): p. 262-268.

124. Verstockt, B., K.G. Smith, and J.C. Lee, *Genome-wide association studies in Crohn's disease: Past, present and future.* Clinical & translational immunology, 2018. **7**(1): p. e1001-e1001.

125. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

126. Lee, S., M.C. Wu, and X. Lin, *Optimal tests for rare variant effects in sequencing association studies.* Biostatistics, 2012. **13**(4): p. 762-75.

127. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* Am J Hum Genet, 2014. **95**(1): p. 5-23.

128. Horstman, B., J. Li, and Y. Chen, *Detecting epistatic effects in association studies at a genomic level based on an ensemble approach.* Bioinformatics, 2011. **27**(13): p. i222-i229.

129. Bolze, A., et al., *Whole-exome-sequencing-based discovery of human FADD deficiency.* Am J Hum Genet, 2010. **87**(6): p. 873-81.

130. Stray-Pedersen, A., et al., *Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders.* Journal of Allergy and Clinical Immunology, 2017. **139**(1): p. 232-245.

131. Gallo, V., et al., *Diagnostics of Primary Immunodeficiencies through Next-Generation Sequencing.* Frontiers in immunology, 2016. **7**: p. 466-466.

132. Moens, L.N., et al., *Diagnostics of primary immunodeficiency diseases: a sequencing capture approach.* PLoS One, 2014. **9**(12): p. e114901.

133. Stray-Pedersen, A., et al., *Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders.* J Allergy Clin Immunol, 2017. **139**(1): p. 232-245.

134. Al-Mousa, H., et al., *Unbiased targeted next-generation sequencing molecular approach for primary immunodeficiency diseases.* Journal of Allergy and Clinical Immunology, 2016. **137**(6): p. 1780-1787.

135. Stoddard, J.L., et al., *Targeted NGS: A Cost-Effective Approach to Molecular Diagnosis of PIDs.* Frontiers in Immunology, 2014. **5**(531).

136. Shabani, M., K.E. Nichols, and N. Rezaei, *Primary immunodeficiencies associated with EBV-Induced lymphoproliferative disorders.* Crit Rev Oncol Hematol, 2016. **108**: p. 109-127.

137. Hoeger, B., N.K. Serwas, and K. Boztug, *Human NF-kappaB1 Haploinsufficiency and Epstein-Barr Virus-Induced Disease-Molecular Mechanisms and Consequences.* Front Immunol, 2017. **8**: p. 1978.

138. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 2006. **38**: p. 904.

139. Zhu, W., et al., *Concurrent nucleotide substitution mutations in the human genome are characterized by a significantly decreased transition/transversion ratio.* Hum Mutat, 2015. **36**(3): p. 333-41.

140. Wang, J., et al., *Genome measures used for quality control are dependent on gene function and ancestry.* Bioinformatics (Oxford, England), 2015. **31**(3): p. 318-323.

141. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.

142. Guo, Y., et al., *Three-stage quality control strategies for DNA re-sequencing data.* Briefings in bioinformatics, 2014. **15**(6): p. 879-889.

143. Poplin, R., et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2018.

144. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

145. Meynert, A.M., et al., *Variant detection sensitivity and biases in whole genome and exome sequencing.* BMC Bioinformatics, 2014. **15**: p. 247.

146. Desmet, F.O., et al., *Human Splicing Finder: an online bioinformatics tool to predict splicing signals.* Nucleic Acids Res, 2009. **37**(9): p. e67.

147. Kobayashi, Y., et al., *Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation.* Genome Med, 2017. **9**(1): p. 13.

148. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

149. Karczewski, K.J., et al., *The ExAC browser: displaying reference data information from over 60 000 exomes.* Nucleic Acids Res, 2017. **45**(D1): p. D840-d845.

150. Sim, N.L., et al., *SIFT web server: predicting effects of amino acid substitutions on proteins.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W452-7.

151. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

152. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.

153. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.

154. Neale, B.M., et al., *Testing for an Unusual Distribution of Rare Variants.* PLOS Genetics, 2011. **7**(3): p. e1001322.

155. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.

156. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* Am J Hum Genet, 2012. **91**(2): p. 224-37.

157. Yu, H., et al., *Rapid molecular diagnostics of severe primary immunodeficiency determined by using targeted next-generation sequencing.* J Allergy Clin Immunol, 2016. **138**(4): p. 1142-1151.e2.

158. Ribeiro, A., et al., *An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome.* BMC Bioinformatics, 2015. **16**: p. 382.

159. Ross, M.G., et al., *Characterizing and measuring bias in sequence data.* Genome Biology, 2013. **14**(5): p. R51.

160. Muyas, F., et al., *Allele balance bias identifies systematic genotyping errors and false disease associations.* Hum Mutat, 2018.

161. Mardis, E.R., *Next-generation sequencing platforms.* Annu Rev Anal Chem (Palo Alto Calif), 2013. **6**: p. 287-303.

162. Ajay, S.S., et al., *Accurate and comprehensive sequencing of personal genomes.* Genome Res, 2011. **21**(9): p. 1498-505.

163. Clark, M.J., et al., *Performance comparison of exome DNA sequencing technologies.* Nat Biotechnol, 2011. **29**(10): p. 908-14.

164. Lelieveld, S.H., et al., *Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions.* Hum Mutat, 2015. **36**(8): p. 815-22.

165. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

166. Gargis, A.S., et al., *Good laboratory practice for clinical next-generation sequencing informatics pipelines.* Nature Biotechnology, 2015. **33**: p. 689.

167. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.

168. Petrovski, S., et al., *Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes.* PLoS Genet, 2013. **9**(8).

169. Meienberg, J., et al., *Clinical sequencing: is WGS the better WES?* Human Genetics, 2016. **135**: p. 359-362.

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

171. Frans, G., et al., *Conventional and Single-Molecule Targeted Sequencing Method for Specific Variant Detection in IKBKG while Bypassing the IKBKGP1 Pseudogene.* J Mol Diagn, 2018. **20**(2): p. 195-202.

172. Bardaro, T., et al., *Two cases of misinterpretation of molecular results in incontinentia pigmenti, and a PCR-based method to discriminate NEMO/IKKgamma dene deletion.* Hum Mutat, 2003. **21**(1): p. 8-11.

173. Smahi, A., et al., *Genomic rearrangement in NEMO impairs NF-kappaB activation and is a cause of incontinentia pigmenti. The International Incontinentia Pigmenti (IP) Consortium.* Nature, 2000. **405**(6785): p. 466-72.

174. Doffinger, R., et al., *X-linked anhidrotic ectodermal dysplasia with immunodeficiency is caused by impaired NF-kappaB signaling.* Nat Genet, 2001. **27**(3): p. 277-85.

175. Dorschner, M.O., et al., *Actionable, pathogenic incidental findings in 1,000 participants' exomes.* Am J Hum Genet, 2013. **93**(4): p. 631-40.

176. Ochs, H.D. and A.J. Thrasher, *The Wiskott-Aldrich syndrome.* J Allergy Clin Immunol, 2006. **117**(4): p. 725-38; quiz 739.

177. Bousfiha, A., et al., *Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification.* J Clin Immunol, 2020. **40**(1): p. 66-81.

178. Ku, C.S., N. Naidoo, and Y. Pawitan, *Revisiting Mendelian disorders through exome sequencing.* Hum Genet, 2011. **129**(4): p. 351-70.

179. Rabbani, B., et al., *Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders.* J Hum Genet, 2012. **57**(10): p. 621-32.

180. Jensson, B.O., et al., *COPA syndrome in an Icelandic family caused by a recurrent missense mutation in COPA.* BMC Med Genet, 2017. **18**(1): p. 129.

181. Kalia, S.S., et al., *Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics.* Genet Med, 2017. **19**(2): p. 249-255.

182. Kleinberger, J., et al., *An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants.* Genet Med, 2016. **18**(11): p. 1165-1165.

183. Nehme, N.T., et al., *MST1 mutations in autosomal recessive primary immunodeficiency characterized by defective naive T-cell survival.* Blood, 2012. **119**(15): p. 3458-68.

184. Halacli, S.O., et al., *STK4 (MST1) deficiency in two siblings with autoimmune cytopenias: A novel mutation.* Clin Immunol, 2015. **161**(2): p. 316-23.

185. Schipp, C., et al., *EBV Negative Lymphoma and Autoimmune Lymphoproliferative Syndrome Like Phenotype Extend the Clinical Spectrum of Primary Immunodeficiency Caused by STK4 Deficiency.* Frontiers in Immunology, 2018. **9**(2400).

186. Minegishi, Y., et al., *Mutations in Igalpha (CD79a) result in a complete block in B-cell development.* J Clin Invest, 1999. **104**(8): p. 1115-21.

187. Wang, Y., et al., *Novel Igalpha (CD79a) gene mutation in a Turkish patient with B cell-deficient agammaglobulinemia.* Am J Med Genet, 2002. **108**(4): p. 333-6.

188. Khalili, A., et al., *Autosomal recessive agammaglobulinemia: a novel non-sense mutation in CD79a.* J Clin Immunol, 2014. **34**(2): p. 138-41.

189. Zhang, Q., et al., *Combined immunodeficiency associated with DOCK8 mutations.* N Engl J Med, 2009. **361**(21): p. 2046-55.

190. van de Veerdonk, F.L., et al., *STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis.* N Engl J Med, 2011. **365**(1): p. 54-61.

191. Soltesz, B., et al., *New and recurrent gain-of-function STAT1 mutations in patients with chronic mucocutaneous candidiasis from Eastern and Central Europe.* J Med Genet, 2013. **50**(9): p. 567-78.

192. Uzel, G., et al., *Dominant gain-of-function STAT1 mutations in FOXP3 wild-type immune dysregulation-polyendocrinopathy-enteropathy-X-linked-like syndrome.* J Allergy Clin Immunol, 2013. **131**(6): p. 1611-23.

193. Nick, J.A. and D.P. Nichols, *Diagnosis of Adult Patients with Cystic Fibrosis.* Clin Chest Med, 2016. **37**(1): p. 47-57.

194. Minegishi, Y., et al., *Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome.* Nature, 2007. **448**(7157): p. 1058-62.

195. Grimbacher, B., et al., *Genetic linkage of hyper-IgE syndrome to chromosome 4.* Am J Hum Genet, 1999. **65**(3): p. 735-44.

196. Woellner, C., et al., *Mutations in STAT3 and diagnostic guidelines for hyper-IgE syndrome.* J Allergy Clin Immunol, 2010. **125**(2): p. 424-432.e8.

197. Kumanovics, A., et al., *Diffuse large B cell lymphoma in hyper-IgE syndrome due to STAT3 mutation.* J Clin Immunol, 2010. **30**(6): p. 886-93.

198. Takeda, K., et al., *Targeted disruption of the mouse Stat3 gene leads to early embryonic lethality.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3801-4.

199. Diaz de Leon, A., et al., *Telomere lengths, pulmonary fibrosis and telomerase (TERT) mutations.* PLoS One, 2010. **5**(5): p. e10680.

200. Avitzur, Y., et al., *Mutations in tetratricopeptide repeat domain 7A result in a severe form of very early onset inflammatory bowel disease.* Gastroenterology, 2014. **146**(4): p. 1028-39.

201. Bigorgne, A.E., et al., *TTC7A mutations disrupt intestinal epithelial apicobasal polarity.* J Clin Invest, 2014. **124**(1): p. 328-37.

202. Samuels, M.E., et al., *Exome sequencing identifies mutations in the gene TTC7A in French-Canadian cases with hereditary multiple intestinal atresia.* J Med Genet, 2013. **50**(5): p. 324-9.

203. Lien, R., et al., *Novel Mutations of the Tetratricopeptide Repeat Domain 7A Gene and Phenotype/Genotype Comparison.* Front Immunol, 2017. **8**: p. 1066.

204. Ouederni, M., et al., *Major histocompatibility complex class II expression deficiency caused by a RFXANK founder mutation: a survey of 35 patients.* Blood, 2011. **118**(19): p. 5108-18.

205. Hanna, S. and A. Etzioni, *MHC class I and II deficiencies.* Journal of Allergy and Clinical Immunology, 2014. **134**(2): p. 269-275.

206. Chan, A.Y., et al., *A novel human autoimmune syndrome caused by combined hypomorphic and activating mutations in ZAP-70.* J Exp Med, 2016. **213**(2): p. 155-65.

207. Arpaia, E., et al., *Defective T cell receptor signaling and CD8+ thymic selection in humans lacking zap-70 kinase.* Cell, 1994. **76**(5): p. 947-58.

208. Thiery, J., et al., *Perforin pores in the endosomal membrane trigger the release of endocytosed granzyme B into the cytosol of target cells.* Nat Immunol, 2011. **12**(8): p. 770-7.

209. Bordbar, M.R., et al., *A case report of novel mutation in PRF1 gene, which causes familial autosomal recessive hemophagocytic lymphohistiocytosis.* BMC Med Genet, 2017. **18**(1): p. 49.

210. Kim, M.S., et al., *Familial Hemophagocytic Lymphohistiocytosis Type 2 in a Korean Infant With Compound Heterozygous PRF1 Defects Involving a PRF1 Mutation, c.1091T>G.* Ann Lab Med, 2017. **37**(2): p. 162-165.

211. McDonald, D.R., et al., *Heterozygous N-terminal deletion of IkappaBalpha results in functional nuclear factor kappaB haploinsufficiency, ectodermal dysplasia, and immune deficiency.* J Allergy Clin Immunol, 2007. **120**(4): p. 900-7.

212. Boisson, B., et al., *Human IkappaBalpha Gain of Function: a Severe and Syndromic Immunodeficiency.* J Clin Immunol, 2017. **37**(5): p. 397-412.

213. Vihinen, M., P.T. Mattsson, and C.I. Smith, *Bruton tyrosine kinase (BTK) in X-linked agammaglobulinemia (XLA).* Front Biosci, 2000. **5**: p. D917-28.

214. de Weers, M., et al., *B-cell antigen receptor stimulation activates the human Bruton's tyrosine kinase, which is deficient in X-linked agammaglobulinemia.* J Biol Chem, 1994. **269**(39): p. 23857-60.

215. Schubbert, S., et al., *Germline KRAS mutations cause Noonan syndrome.* Nat Genet, 2006. **38**(3): p. 331-6.

216. Milner, J.D., et al., *Early-onset lymphoproliferation and autoimmunity caused by germline STAT3 gain-of-function mutations.* Blood, 2015. **125**(4): p. 591-9.

217. Yi, G., et al., *Single nucleotide polymorphisms of human STING can affect innate immune response to cyclic dinucleotides.* PloS one, 2013. **8**(10): p. e77846-e77846.

218. Jeremiah, N., et al., *Inherited STING-activating mutation underlies a familial inflammatory syndrome with lupus-like manifestations.* J Clin Invest, 2014. **124**(12): p. 5516-20.

219. Liu, Y., et al., *Activated STING in a Vascular and Pulmonary Syndrome.* The New England journal of medicine, 2014. **371**(6): p. 507-518.

220. Picard, C., et al., *Severe Pulmonary Fibrosis as the First Manifestation of Interferonopathy (TMEM173 Mutation).* Chest, 2016. **150**(3): p. e65-e71.

221. Melki, I., et al., *Disease-associated mutations identify a novel region in human STING necessary for the control of type I interferon signaling.* J Allergy Clin Immunol, 2017. **140**(2): p. 543-552.e5.

222. Bartha, I., et al., *The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome.* PLoS Comput Biol, 2015. **11**(12): p. e1004647.

223. Horiuchi, T., et al., *A non-sense mutation at Arg95 is predominant in complement 9 deficiency in Japanese.* J Immunol, 1998. **160**(3): p. 1509-13.

224. Kira, R., et al., *Molecular epidemiology of C9 deficiency heterozygotes with an Arg95Stop mutation of the C9 gene in Japan.* J Hum Genet, 1999. **44**(2): p. 109-11.

225. Speckmann, C., et al., *Clinical and Molecular Heterogeneity of RTEL1 Deficiency.* Frontiers in Immunology, 2017. **8**(449).

226. Cartegni, L., et al., *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res, 2003. **31**(13): p. 3568-71.

227. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.

228. Stapleton, M., J.W. Carlson, and S.E. Celniker, *RNA editing in Drosophila melanogaster: New targets and functional consequences.* Rna, 2006. **12**(11): p. 1922-32.

229. Pertea, M., X. Lin, and S.L. Salzberg, *GeneSplicer: a new computational method for splice site prediction.* Nucleic Acids Res, 2001. **29**(5): p. 1185-90.

230. Houdayer, C., et al., *Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants.* Hum Mutat, 2012. **33**(8): p. 1228-38.

231. Ohno, K., J.I. Takeda, and A. Masuda, *Rules and tools to predict the splicing effects of exonic and intronic mutations.* Wiley Interdiscip Rev RNA, 2018. **9**(1).

232. Walne, A.J., et al., *Constitutional mutations in RTEL1 cause severe dyskeratosis congenita.* Am J Hum Genet, 2013. **92**(3): p. 448-53.

233. Puel, A., et al., *Defective IL7R expression in T(-)B(+)NK(+) severe combined immunodeficiency.* Nat Genet, 1998. **20**(4): p. 394-7.

234. Roifman, C.M., et al., *A partial deficiency of interleukin-7R alpha is sufficient to abrogate T-cell development and cause severe combined immunodeficiency.* Blood, 2000. **96**(8): p. 2803-7.

235. Rivat, C., et al., *Gene therapy for primary immunodeficiencies.* Hum Gene Ther, 2012. **23**(7): p. 668-75.

236. Lafaille, F.G., et al., *Impaired intrinsic immunity to HSV-1 in human iPSC-derived TLR3-deficient CNS cells.* Nature, 2012. **491**(7426): p. 769-73.

237. Macarthur, D., *Methods: Face up to false positives.* Nature, 2012. **487**(7408): p. 427-8.

238. Achoch, M., et al., *Protein structural robustness to mutations: an in silico investigation.* Phys Chem Chem Phys, 2016. **18**(20): p. 13770-80.

239. Borguesan, B., M. Inostroza-Ponta, and M. Dorn, *NIAS-Server: Neighbors Influence of Amino acids and Secondary Structures in Proteins.* J Comput Biol, 2017. **24**(3): p. 255-265.

240. Brajic, A., et al., *The Long Non-coding RNA Flatr Anticipates Foxp3 Expression in Regulatory T Cells.* Front Immunol, 2018. **9**: p. 1989.

241. Ogawa, C., et al., *Blimp-1 Functions as a Molecular Switch to Prevent Inflammatory Activity in Foxp3(+)RORgammat(+) Regulatory T Cells.* Cell Rep, 2018. **25**(1): p. 19-28.e5.

242. Stubbington, M.J.T., et al., *Single-cell transcriptomics to explore the immune system in health and disease.* Science, 2017. **358**(6359): p. 58-63.

243. The Deciphering Developmental Disorders, S., *Large-scale discovery of novel genetic causes of developmental disorders.* Nature, 2015. **519**(7542): p. 223-228.

244. Chong, J.X., et al., *The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.* Am J Hum Genet, 2015. **97**(2): p. 199-215.

245. Ferrari, S., et al., *Mutations of the Igbeta gene cause agammaglobulinemia in man.* J Exp Med, 2007. **204**(9): p. 2047-51.

246. Guo, W., et al., *Identifying and Analyzing Novel Epilepsy-Related Genes Using Random Walk with Restart Algorithm.* BioMed research international, 2017. **2017**: p. 6132436-6132436.

247. Ayhan, F. and G. Konopka, *Regulatory genes and pathways disrupted in autism spectrum disorders.* Prog Neuropsychopharmacol Biol Psychiatry, 2019. **89**: p. 57-64.

248. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* American journal of human genetics, 2017. **101**(1): p. 5-22.

249. Stranger, B.E. and P.L. De Jager, *Coordinating GWAS results with gene expression in a systems immunologic paradigm in autoimmunity.* Curr Opin Immunol, 2012. **24**(5): p. 544-51.

250. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* American journal of human genetics, 2014. **95**(1): p. 5-23.

251. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

252. Li, M.J., et al., *GWASdb v2: an update database for human genetic variants identified by genome-wide association studies.* Nucleic Acids Res, 2016. **44**(D1): p. D869-76.

253. Bronson, P.G., et al., *Common variants at PVT1, ATG13-AMBRA1, AHI1 and CLEC16A are associated with selective IgA deficiency.* Nat Genet, 2016. **48**(11): p. 1425-1429.

254. Orange, J.S., et al., *Genome-wide association identifies diverse causes of common variable immunodeficiency.* J Allergy Clin Immunol, 2011. **127**(6): p. 1360-7.e6.

255. Maggadottir, S.M., et al., *Rare variants at 16p11.2 are associated with common variable immunodeficiency.* J Allergy Clin Immunol, 2015. **135**(6): p. 1569-77.

256. The, U.K.K.C., et al., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**: p. 82.

257. Asimit, J. and E. Zeggini, *Rare variant association analysis methods for complex traits.* Annu Rev Genet, 2010. **44**: p. 293-308.

258. Khoronenkova, S.V., et al., *ATM-dependent downregulation of USP7/HAUSP by PPM1G activates p53 response to DNA damage.* Mol Cell, 2012. **45**(6): p. 801-13.

259. Gudipaty, S.A., et al., *PPM1G Binds 7SK RNA and Hexim1 To Block P-TEFb Assembly into the 7SK snRNP and Sustain Transcription Elongation.* Mol Cell Biol, 2015. **35**(22): p. 3810-28.

260. Pyo, J., et al., *The Protein Phosphatase PPM1G Destabilizes HIF-1alpha Expression.* Int J Mol Sci, 2018. **19**(8).

261. Foster, W.H., et al., *Nuclear phosphatase PPM1G in cellular survival and neural development.* Developmental dynamics : an official publication of the American Association of Anatomists, 2013. **242**(9): p. 1101-1109.

262. Bell, B.D., et al., *FADD and caspase-8 control the outcome of autophagic signaling in proliferating T cells.* Proc Natl Acad Sci U S A, 2008. **105**(43): p. 16677-82.

263. Huang, B., et al., *Posttranslational modifications of NF-kappaB: another layer of regulation for NF-kappaB signaling pathway.* Cellular signalling, 2010. **22**(9): p. 1282-1290.

264. Rothgiesser, K.M., M. Fey, and M.O. Hottiger, *Acetylation of p65 at lysine 314 is important for late NF-kappaB-dependent gene expression.* BMC Genomics, 2010. **11**: p. 22.

265. Kim, B.H., et al., *A family of IFN-gamma-inducible 65-kD GTPases protects against bacterial infection.* Science, 2011. **332**(6030): p. 717-21.

266. Balasubramanian, S., et al., *The interferon-gamma-induced GTPase, mGBP-2, inhibits tumor necrosis factor alpha (TNF-alpha) induction of matrix metalloproteinase-9 (MMP-9) by inhibiting NF-kappaB and Rac protein.* J Biol Chem, 2011. **286**(22): p. 20054-64.

267. Bacchelli, C., et al., *Mutations in linker for activation of T cells (LAT) lead to a novel form of severe combined immunodeficiency.* J Allergy Clin Immunol, 2017. **139**(2): p. 634-642.e5.

268. Fuller, D.M. and W. Zhang, *Regulation of lymphocyte development and activation by the LAT family of adapter proteins.* Immunol Rev, 2009. **232**(1): p. 72-83.

269. Hudson, B.D., et al., *Extracellular ionic locks determine variation in constitutive activity and ligand potency between species orthologs of the free fatty acid receptors FFA2 and FFA3.* J Biol Chem, 2012. **287**(49): p. 41195-209.

270. Brown, A.J., et al., *The Orphan G protein-coupled receptors GPR41 and GPR43 are activated by propionate and other short chain carboxylic acids.* J Biol Chem, 2003. **278**(13): p. 11312-9.

271. Kim, M.H., et al., *Short-chain fatty acids activate GPR41 and GPR43 on intestinal epithelial cells to promote inflammatory responses in mice.* Gastroenterology, 2013. **145**(2): p. 396-406.e1-10.

272. Bravo, J., et al., *The leukemia-associated AML1 (Runx1)--CBF beta complex functions as a DNA-induced molecular clamp.* Nat Struct Biol, 2001. **8**(4): p. 371-8.

273. Oo, Z.M., et al., *A tool compound targeting the core binding factor Runt domain to disrupt binding to CBFbeta in leukemic cells.* Leuk Lymphoma, 2018. **59**(9): p. 2188-2200.

274. Setoguchi, R., et al., *Repression of the transcription factor Th-POK by Runx complexes in cytotoxic T cell development.* Science, 2008. **319**(5864): p. 822-5.

275. Hoffman, J.D., et al., *Immune abnormalities are a frequent manifestation of Kabuki syndrome.* Am J Med Genet A, 2005. **135**(3): p. 278-81.

276. Walsh, M.A., et al., *A multicentre study of patients with Timothy syndrome.* Europace, 2018. **20**(2): p. 377-385.

277. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project.* Immunogenetics, 2008. **60**(1): p. 1-18.

278. Ogryzko, V.V., et al., *Histone-like TAFs within the PCAF histone acetylase complex.* Cell, 1998. **94**(1): p. 35-44.

279. Hahn, S., *The role of TAFs in RNA polymerase II transcription.* Cell, 1998. **95**(5): p. 579-82.

280. Struhl, K. and Z. Moqtaderi, *The TAFs in the HAT.* Cell, 1998. **94**(1): p. 1-4.

281. Roudaia, L., et al., *CBFbeta is critical for AML1-ETO and TEL-AML1 activity.* Blood, 2009. **113**(13): p. 3070-9.

282. Collins, A., D.R. Littman, and I. Taniuchi, *RUNX proteins in transcription factor networks that regulate T-cell lineage choice.* Nature Reviews Immunology, 2009. **9**: p. 106.

283. Tenno, M., et al., *Cbfbeta2 controls differentiation of and confers homing capacity to prethymic progenitors.* J Exp Med, 2018. **215**(2): p. 595-610.

284. Dolinski, K. and D. Botstein, *Orthology and functional conservation in eukaryotes.* Annu Rev Genet, 2007. **41**: p. 465-507.

285. Rosen, B., J. Schick, and W. Wurst, *Beyond knockouts: the International Knockout Mouse Consortium delivers modular and evolving tools for investigating mammalian genes.* Mamm Genome, 2015. **26**(9-10): p. 456-66.

286. Dickinson, M.E., et al., *High-throughput discovery of novel developmental phenotypes.* Nature, 2016. **537**(7621): p. 508-514.

287. Zhu, X. and M. Stephens, *Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes.* Nature communications, 2018. **9**(1): p. 4361-4361.

288. Timpson, N.J., et al., *A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans.* Nature Communications, 2014. **5**: p. 4871.

289. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.

290. Tsai, P.C., T.D. Spector, and J.T. Bell, *Using epigenome-wide association scans of DNA methylation in age-related complex human traits.* Epigenomics, 2012. **4**(5): p. 511-26.

291. Liu, D., et al., *EWASdb: epigenome-wide association study database.* Nucleic Acids Res, 2019. **47**(D1): p. D989-d993.

292. Kasper, B., et al., *Differential expression and regulation of GTPases (RhoA and Rac2) and GDIs (LyGDI and RhoGDI) in neutrophils from patients with severe congenital neutropenia.* Blood, 2000. **95**(9): p. 2947-53.

293. Yu, H., et al., *Deficiency of small GTPase Rac2 affects T cell activation.* The Journal of experimental medicine, 2001. **194**(7): p. 915-926.

294. Morillon, Y.M., 2nd, et al., *Antibody Binding to CD4 Induces Rac GTPase Activation and Alters T Cell Migration.* Journal of immunology (Baltimore, Md. : 1950), 2016. **197**(9): p. 3504-3511.

295. Gu, Y., et al., *Biochemical and biological characterization of a human Rac2 GTPase mutant associated with phagocytic immunodeficiency.* J Biol Chem, 2001. **276**(19): p. 15929-38.

296. Li, S., et al., *Chemoattractant-stimulated Rac activation in wild-type and Rac2-deficient murine neutrophils: preferential activation of Rac2 and Rac2 gene dosage effect on neutrophil functions.* J Immunol, 2002. **169**(9): p. 5043-51.

297. Filippi, M.D., et al., *Localization of Rac2 via the C terminus and aspartic acid 150 specifies superoxide generation, actin polarity and chemotaxis in neutrophils.* Nat Immunol, 2004. **5**(7): p. 744-51.

298. Moll, J., et al., *The murine rac1 gene: cDNA cloning, tissue distribution and regulated expression of rac1 mRNA by disassembly of actin microfilaments.* Oncogene, 1991. **6**(5): p. 863-6.

299. Akasaki, T., H. Koga, and H. Sumimoto, *Phosphoinositide 3-kinase-dependent and -independent activation of the small GTPase Rac2 in human neutrophils.* J Biol Chem, 1999. **274**(25): p. 18055-9.

300. Walmsley, M.J., et al., *Critical roles for Rac1 and Rac2 GTPases in B cell development and signaling.* Science, 2003. **302**(5644): p. 459-62.

301. Croker, B.A., et al., *The Rac2 Guanosine Triphosphatase Regulates B Lymphocyte Antigen Receptor Responses and Chemotaxis and Is Required for Establishment of B-1a and Marginal Zone B Lymphocytes.* The Journal of Immunology, 2002. **168**(7): p. 3376.

302. Abdel-Latif, D., et al., *Rac2 is critical for neutrophil primary granule exocytosis.* Blood, 2004. **104**(3): p. 832-9.

303. Shirsat, N.V., et al., *A member of the ras gene superfamily is expressed specifically in T, B and myeloid hemopoietic cells.* Oncogene, 1990. **5**(5): p. 769-72.

304. Corbetta, S., et al., *Generation and characterization of Rac3 knockout mice.* Mol Cell Biol, 2005. **25**(13): p. 5763-76.

305. Marcos-Ramiro, B., et al., *RhoB controls endothelial barrier recovery by inhibiting Rac1 trafficking to the cell border.* J Cell Biol, 2016. **213**(3): p. 385-402.

306. Reijnders, M.R.F., et al., *RAC1 Missense Mutations in Developmental Disorders with Diverse Phenotypes.* The American Journal of Human Genetics, 2017. **101**(3): p. 466-477.

307. Ridley, A.J., *Life at the leading edge.* Cell, 2011. **145**(7): p. 1012-22.

308. Tybulewicz, V.L. and R.B. Henderson, *Rho family GTPases and their regulators in lymphocytes.* Nat Rev Immunol, 2009. **9**(9): p. 630-44.

309. Roberts, A.W., et al., *Deficiency of the hematopoietic cell-specific Rho family GTPase Rac2 is characterized by abnormalities in neutrophil function and host defense.* Immunity, 1999. **10**(2): p. 183-96.

310. Nguyen, G.T., E.R. Green, and J. Mecsas, *Neutrophils to the ROScue: Mechanisms of NADPH Oxidase Activation and Bacterial Resistance.* Frontiers in cellular and infection microbiology, 2017. **7**: p. 373-373.

311. Brinkmann, V., et al., *Neutrophil extracellular traps kill bacteria.* Science, 2004. **303**(5663): p. 1532-5.

312. Branzk, N., et al., *Neutrophils sense microbe size and selectively release neutrophil extracellular traps in response to large pathogens.* Nat Immunol, 2014. **15**(11): p. 1017-25.

313. Lim, M.B., et al., *Rac2 is required for the formation of neutrophil extracellular traps.* J Leukoc Biol, 2011. **90**(4): p. 771-6.

314. Williams, D.A., et al., *Dominant negative mutation of the hematopoietic-specific Rho GTPase, Rac2, is associated with a human phagocyte immunodeficiency.* Blood, 2000. **96**(5): p. 1646-54.

315. Alkhairy, O.K., et al., *RAC2 loss-of-function mutation in 2 siblings with characteristics of common variable immunodeficiency.* J Allergy Clin Immunol, 2015. **135**(5): p. 1380-4.e1-5.

316. Hsu, A.P., et al., *Dominant activating RAC2 mutation with lymphopenia, immunodeficiency, and cytoskeletal defects.* Blood, 2019. **133**(18): p. 1977-1988.

317. Arany, Z., et al., *E1A-associated p300 and CREB-associated CBP belong to a conserved family of coactivators.* Cell, 1994. **77**(6): p. 799-800.

318. Tropberger, P., et al., *Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer.* Cell, 2013. **152**(4): p. 859-72.

319. Ogryzko, V.V., et al., *The transcriptional coactivators p300 and CBP are histone acetyltransferases.* Cell, 1996. **87**(5): p. 953-9.

320. Sterner, D.E. and S.L. Berger, *Acetylation of histones and transcription-related factors.* Microbiol Mol Biol Rev, 2000. **64**(2): p. 435-59.

321. Baraitser, M. and M.A. Preece, *The Rubinstein-Taybi syndrome: occurrence in two sets of identical twins.* Clin Genet, 1983. **23**(4): p. 318-20.

322. Berry, A.C., *Rubinstein-Taybi syndrome.* J Med Genet, 1987. **24**(9): p. 562-6.

323. Roelfsema, J.H., et al., *Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease.* Am J Hum Genet, 2005. **76**(4): p. 572-80.

324. Petrij, F., et al., *Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP.* Nature, 1995. **376**(6538): p. 348-51.

325. Boot, M.V., et al., *Benign and malignant tumors in Rubinstein-Taybi syndrome.* Am J Med Genet A, 2018. **176**(3): p. 597-608.

326. Hamilton, M.J., et al., *Rubinstein-Taybi syndrome type 2: report of nine new cases that extend the phenotypic and genotypic spectrum.* Clin Dysmorphol, 2016. **25**(4): p. 135-45.

327. Villella, A., et al., *Rubinstein-Taybi syndrome with humoral and cellular defects: a case report.* Arch Dis Child, 2000. **83**(4): p. 360-1.

328. Herriot, R. and Z. Miedzybrodzka, *Antibody deficiency in Rubinstein-Taybi syndrome.* Clin Genet, 2016. **89**(3): p. 355-8.

329. Saettini, F., et al., *A novel EP300 mutation associated with Rubinstein-Taybi syndrome type 2 presenting as combined immunodeficiency.* Pediatr Allergy Immunol, 2018. **29**(7): p. 776-781.

330. Cale, C.M., et al., *Severe combined immunodeficiency with abnormalities in expression of the common leucocyte antigen, CD45.* Arch Dis Child, 1997. **76**(2): p. 163-4.

331. Tchilian, E.Z., et al., *A deletion in the gene encoding the CD45 antigen in a patient with SCID.* J Immunol, 2001. **166**(2): p. 1308-13.

332. Carstanjen, D., et al., *Rac2 Regulates Neutrophil Chemotaxis, Superoxide Production, and Myeloid Colony Formation through Multiple Distinct Effector Pathways.* The Journal of Immunology, 2005. **174**(8): p. 4613-4620.

333. Hanna, S. and M. El-Sibai, *Signaling networks of Rho GTPases in cell motility.* Cellular Signalling, 2013. **25**(10): p. 1955-1961.

334. Zielonka, J., et al., *Global profiling of reactive oxygen and nitrogen species in biological systems: high-throughput real-time analyses.* The Journal of biological chemistry, 2012. **287**(5): p. 2984-2995.

335. Arana, E., et al., *Activation of the small GTPase Rac2 via the B cell receptor regulates B cell adhesion and immunological-synapse formation.* Immunity, 2008. **28**(1): p. 88-99.

336. Dorjbal, B., et al., *Hypomorphic caspase activation and recruitment domain 11 (CARD11) mutations associated with diverse immunologic phenotypes with or without atopic disease.* J Allergy Clin Immunol, 2018.

337. Liu, W., et al., *Mutant fibrillin-1 monomers lacking EGF-like domains disrupt microfibril assembly and cause severe marfan syndrome.* Hum Mol Genet, 1996. **5**(10): p. 1581-7.

338. Dorrance, A.M., et al., *The Rac GTPase effector p21-activated kinase is essential for hematopoietic stem/progenitor cell migration and engraftment.* Blood, 2013. **121**(13): p. 2474.

339. Bigley, V., et al., *The human syndrome of dendritic cell, monocyte, B and NK lymphoid deficiency.* J Exp Med, 2011. **208**(2): p. 227-34.

340. Camargo, J.F., et al., *MonoMAC syndrome in a patient with a GATA2 mutation: case report and review of the literature.* Clin Infect Dis, 2013. **57**(5): p. 697-9.

341. Negri, G., et al., *From Whole Gene Deletion to Point Mutations of EP300-Positive Rubinstein-Taybi Patients: New Insights into the Mutational Spectrum and Peculiar Clinical Hallmarks.* Hum Mutat, 2016. **37**(2): p. 175-83.

342. Spena, S., C. Gervasini, and D. Milani, *Ultra-Rare Syndromes: The Example of Rubinstein-Taybi Syndrome.* J Pediatr Genet, 2015. **4**(3): p. 177-86.

343. Torres, L.C., et al., *Evaluation of the immune humoral response of Brazilian patients with Rubinstein-Taybi syndrome.* Braz J Med Biol Res, 2010. **43**(12): p. 1215-24.

344. Lougaris, V., et al., *Progressive severe B cell deficiency in pediatric Rubinstein-Taybi syndrome.* Clin Immunol, 2016. **173**: p. 181-183.

345. Raedler, L.A., *Diagnosis and Management of Polycythemia Vera: Proceedings from a Multidisciplinary Roundtable.* American health & drug benefits, 2014. **7**(7 Suppl 3): p. S36-S47.

346. Bilgrami, S. and B.R. Greenberg, *Polycythemia rubra vera.* Semin Oncol, 1995. **22**(4): p. 307-26.

347. Garcia-Carpizo, V., et al., *CREBBP/EP300 bromodomains are critical to sustain the GATA1/MYC regulatory axis in proliferation.* Epigenetics & Chromatin, 2018. **11**(1): p. 30.

348. Loven, J., et al., *Selective inhibition of tumor oncogenes by disruption of super-enhancers.* Cell, 2013. **153**(2): p. 320-34.

349. Jin, L., et al., *Therapeutic Targeting of the CBP/p300 Bromodomain Blocks the Growth of Castration-Resistant Prostate Cancer.* Cancer Research, 2017. **77**(20): p. 5564.

350. Zhang, Y., et al., *Characterization of genomic breakpoints in MLL and CBP in leukemia patients with t(11;16).* Genes Chromosomes Cancer, 2004. **41**(3): p. 257-65.

351. Kalkhoven, E., et al., *Loss of CBP acetyltransferase activity by PHD finger mutations in Rubinstein-Taybi syndrome.* Hum Mol Genet, 2003. **12**(4): p. 441-50.

352. Zon, L.I., et al., *Activation of the erythropoietin receptor promoter by transcription factor GATA-1.* Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(23): p. 10638-10641.

353. Hilton, I.B., et al., *Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.* Nat Biotechnol, 2015. **33**(5): p. 510-7.

354. Slatter, M.A. and A.R. Gennery, *Hematopoietic cell transplantation in primary immunodeficiency - conventional and emerging indications.* Expert Rev Clin Immunol, 2018. **14**(2): p. 103-114.

355. Slatter, M.A. and A.J. Cant, *Hematopoietic stem cell transplantation for primary immunodeficiency diseases.* Ann N Y Acad Sci, 2011. **1238**: p. 122-31.

356. Lo, B., et al., *AUTOIMMUNE DISEASE. Patients with LRBA deficiency show CTLA4 loss and immune dysregulation responsive to abatacept therapy.* Science, 2015. **349**(6246): p. 436-40.

357. Calvanese, L., et al., *Structural Basis for Mutations of Human Aquaporins Associated to Genetic Diseases.* Int J Mol Sci, 2018. **19**(6).

358. N, N., et al., *Analysing the Effect of Mutation on Protein Function and Discovering Potential Inhibitors of CDK4: Molecular Modelling and Dynamics Studies.* PLOS ONE, 2015. **10**(8): p. e0133969.

359. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic Acids Res, 2019. **47**(D1): p. D886-d894.

360. Bell, C.J., et al., *Carrier testing for severe childhood recessive diseases by next-generation sequencing.* Sci Transl Med, 2011. **3**(65): p. 65ra4.

361. MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease.* Nature, 2014. **508**(7497): p. 469-476.

362. Greenberg, D.E., et al., *Simultaneous Host-Pathogen Transcriptome Analysis during Granulibacter bethesdensis Infection of Neutrophils from Healthy Subjects and Patients with Chronic Granulomatous Disease.* Infect Immun, 2015. **83**(11): p. 4277-92.

363. van Schouwenburg, P.A., et al., *Application of whole genome and RNA sequencing to investigate the genomic landscape of common variable immunodeficiency disorders.* Clin Immunol, 2015. **160**(2): p. 301-14.