From THE DEPARTMENT OF LABORATORY MEDICINE
Karolinska Institutet, Stockholm, Sweden

NGS BASED STUDIES ON PRIMARY IMMUNODEFICIENCIES (PIDS): CAUSATIVE GENE IDENTIFICATION, TOOL DEVELOPMENT AND APPLICATION

Mingyan Fang

Karolinska Institutet

Stockholm 2019

Published by Karolinska Institutet.
Printed by Eprint AB, 2019

ISBN 978-91-7831-493-5

NGS based studies on Primary immunodeficiencies (PIDs): causative gene identification, tool development and application
THESIS FOR DOCTORAL DEGREE (Ph.D.)

Venue: Lecture Hall 9Q Månen, Alfred Nobels allé 8, Karolinska Institutet, Huddinge

Date: Wednesday 12th June 2019, 10:00

By

Mingyan Fang

Principal Supervisor:
Professor Lennart Hammarström
Karolinska Institutet
Department of Laboratory Medicine
Division of Clinical Immunology and Transfusion Medicine

Co-supervisor(s):
Professor Xiuqing Zhang
BGI-Shenzhen, China

Opponent:
Professor Thomas Fleisher
NIH Clinical Center
Department of Laboratory Medicine
Bethesda, MD, USA

Examination Board:
Professor Anders Örn
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology (MTC)

Associate Professor Karlis Pauksens
Uppsala University
Department of Medical Sciences

Professor Ola Winqvist
Karolinska University Hospital Huddinge
Klinisk immunologi och Transfusionsmedicin

Never Lose a Holy Curiosity.

-Albert Einstein

君子学以聚之，问以辨之。

-《周易》

To my dearest family

ABSTRACT

Primary immunodeficiency diseases (PIDs) are composed by a group of highly heterogeneous immune system diseases, of which approximately 350 forms of PID have been described so far. The causative gene of around 60% of patients with PIDs has yet unknown. In recent years, Next Generation Sequencing (NGS) has been increasingly adopted for gene identification and molecular diagnosis of rare diseases, including PIDs. An overview of the genetic makeup that underlies PID using NGS has been suggested as a promising approach to elucidate the etiology of PIDs, which could yield diagnostic and, possibly, provide new treatment advances for PID.

To approach this goal, we performed either whole exome sequencing (WES, 454 samples) or targeted region sequencing (TRS, 217 samples) on 602 samples of 500 PID pedigrees. We have summarized the practical suggestions for the interpretation of NGS data and the techniques that can be used to search disease-causative PID genes in **Paper I.** This work aims to improve data annotation, interpretation, and application of NGS data in PIDs, which also facilitates a wide range of application of NGS data analysis in other Mendelian disorders.

The genetic approach together with immunological investigations have identified potential pathogenic variants in 86 primary antibody deficiency (PAD) patients (68.2%), and a correct diagnosis can guide/change treatment plan in around half of the patients with PAD (**Paper II**). We identified potentially disease-causing variants (including variants classified as VUS (variants of unknown clinical significance)) in around 34% of genetically unidentified PID samples, which had been subjected to TRS using a panel of 219 common PID genes. Notably, the genetic diagnosis of a specific atypical ITK deficiency case adds to the growing amount of evidence supporting the importance of genetic investigations initiated at an early stage of the patient´s disease (**Paper III**).

Altogether, around 60% of PID patients have a possible diagnosis via WES/TRS. Copy number variation defects were identified in 16 patients (4 genes were involved, *LRBA*, *ATM*, *DOCK8* and *PMS2*). Beyond the identification of the monogenic causal gene based on pedigree analysis, mutation frequency analysis has been used to identify genes with rare functional variants in the higher proportion of patients in specific patient group compared to control samples, which have discovered several potential novel PID genes (*TNFRSF18*, *PIK3CG*, *LILRB1*, *EPHB2*, *TXNIP*, *CD5* and *NLRP5*). Other possible models beyond the monogenic scenario were also explored, and 16 severe combined immunodeficiency (SCID) or common variable immunodeficiency (CVID) patients might be due to an accumulation of rare amino acid substitution variants in genes related to the same function or pathway (*RAG1* & *RAG2*, *RAG1* & *ATM*, *C3* & *ITGB2*, *PRKDC* & *ATM*, *C5* & *NIPBL, LRBA* & *CR2, CR2* & *NFKB1, UNC93B1* & *NIPBL, PLCG2* & *NOD2 and IGLL1* & *ATM*). These findings indicate that NGS, together with a large sample size, is powerful in decoding the genetic characteristics of PID and provide insight into molecular mechanisms that cause the disease.

Existing variants impact prediction software/algorithms still have a challenge to evaluate the pathological consequences of the prioritized variants or genes. We thus developed a Random

Forest-based discriminator, Variant Impact Predictor for PIDs (VIPPID), to refine the prediction algorithms, which utilized the features of pathogenic variants and benign mutations, integrated with other 24 predictive softwares currently used. Evaluation of VIPPID showed that it had superior performance (AUC=0.95) over existing tools, we also showed the gene-specific model outperformed the non-gene-specific model and provided a possibility to explore the underlying molecular mechanism based on our gene-specific model in **Paper IV**.

Specific mutations of PID causative genes may exert different effects on TCR repertoire diversity and composition, which ultimately lead to heterogeneous phenotypes. DNA damage response/methylation is an essential process during antigen receptor recombination. To investigate the effect of mutations in DNA repair genes on adaptive immunity, 19 patients with DNA repair/methylation defects were selected and subdivided into several groups based on their causative genes, we then performed deep immune repertoire sequencing and comparison with 14 age-matched healthy controls.

Patients with different molecular diagnosis exhibited distinct repertoire diversity, clonality and V-J pairing patterns. Aberrant complementarity-determining region 3 (CDR3) length distribution was observed both in unproductive and productive TCRs in all patients, suggesting that it predominantly arose before thymic selection. Shorter CDR3 lengths in AT patients resulted from a decreased number of insertions, led to an increase in the number of shared clonotypes, whereas patients with *DNMT3B* and *ZBTB24* mutations presented longer CDR3 lengths and reduced specificity for pathogen-associated CDR3 sequences (**Paper V**). This study revealed the role of DNA repair/methylation machinery in patients with ATM, DNMT3B and ZBTB24 deficiency, and shed light on the mechanistic etiology of their T cell dysfunction.

LIST OF SCIENTIFIC PAPERS

I. **M**. **Fang**, H. Abolhassani, C. K. Lim, J. Zhang, L. Hammarstrom, Next Generation Sequencing Data Analysis in Primary Immunodeficiency Disorders - Future Directions. *Journal of clinical immunology* 36 Suppl 1, 68-75 (2016).

II. H. Abolhassani, A. Aghamohammadi, **M. Fang**, N. Rezaei, C. Jiang, X. Liu, Q. Pan-Hammarstrom, L. Hammarstrom, Clinical implications of systematic phenotyping and exome sequencing in patients with primary antibody deficiency. *Genetics in medicine* 21, 243-251 (2019).

III. **M. Fang**, H. Abolhassani, Q. Pan-Hammarstrom, E. Sandholm, X. Liu, L. Hammarstrom, Compound Heterozygous Mutations of IL2-Inducible T cell Kinase in a Swedish Patient: the Importance of Early Genetic Diagnosis. *Journal of clinical immunology* 39, 131-134 (2019).

IV. **M. Fang**, H. Abolhassani, Z. Su, Y. Itan, L. Hammarström. VIPPID: a gene specific single nucleotide variant pathogenicity prediction tool for Primary Immunodeficiency Diseases. Preliminary Manuscript.

V. **M. Fang**, Z. Su, H. Abolhassani, W. Zhang, C. Jiang, B. Cheng, L. Lin, X. Wang, S. Wang, L. Wang, L. Luo, A. Aghamohammadi, T. Li, J. Wu, X. Zhang, L. Hammarström, X. Liu. Abnormality in the T cell repertoire contributes to immunodeficiency in patients with DNA repair/methylation defects. Manuscript.

PUBLICATIONS NOT INCLUDED IN THE THESIS (2015-2019)

(Equal contributions)*

1. C. Jespersgaard*, **M. Fang***, M. Bertelsen, X. Dang, H. Jensen, Y. Chen, N. Bech, L. Dai, T. Rosenberg, J. Zhang, L. B. Moller, Z. Tumer, K. Brondum-Nielsen, K. Gronskov, Molecular genetic analysis using targeted NGS analysis of 677 individuals with retinal dystrophy. *Scientific reports* 9, 1219 (2019).

2. S. Bigoni, M. Neri, C. Scotton, R. Farina, P. Sabatelli, C. Jiang, J. Zhang, M. S. Falzarano, R. Rossi, D. Ognibene, R. Selvatici, F. Gualandi, D. Bosshardt, P. Perri, C. Campa, F. Brancati, M. Salvatore, M. C. De Stefano, D. Taruscio, L. Trombelli, **M. Fang**, A. Ferlini, Homozygous recessive versican missense variation is associated with early teeth loss in a Pakistani family. *Frontiers in genetics* 9, 723 (2018).

3. V. Yahalom, N. Pillar, Y. Zhao, S. Modan, **M. Fang**, L. Yosephi, O. Asher, E. Shinar, G. Celniker, H. Resnik-Wolf, Y. Brantz, H. Hauschner, N. Rosenberg, L. Cheng, N. Shomron, E. Pras, SMYD1 is the underlying gene for the AnWj negative blood group phenotype. *European journal of haematology* 101, 496-501 (2018).

4. N. Pillar*, O. Pleniceanu*, **M. Fang***, L. Ziv, E. Lahav, S. Botchan, L. Cheng, B. Dekel, N. Shomron, A rare variant in the FHL1 gene associated with X-linked recessive hypoparathyroidism. *Human genetics* 136, 835-845 (2017).

5. F. Gualandi, F. Zaraket, M. Malagu, G. Parmeggiani, C. Trabanelli, S. Fini, X. Dang, X. Wei, **M. Fang**, M. Bertini, R. Ferrari, A. Ferlini, Mutation load of multiple ion channel gene mutations in Brugada Syndrome. *Cardiology* 137, 256-260 (2017).

6. H. Abolhassani, E. S. Edwards, A. Ikinciogullari, H. Jing, S. Borte, M. Buggert, L. Du, M. Matsuda-Lennikov, R. Romano, R. Caridha, S. Bade, Y. Zhang, J. Frederiksen, **M. Fang**, S. K. Bal, S. Haskologlu, F. Dogu, N. Tacyildiz, H. F. Matthews, J. J. McElwee, E. Gostick, D. A. Price, U. Palendira, A. Aghamohammadi, B. Boisson, N. Rezaei, A. C. Karlsson, M. J. Lenardo, J. L. Casanova, L. Hammarstrom, S. G. Tangye, H. C. Su, Q. Pan-Hammarstrom, Combined immunodeficiency and Epstein-Barr virus-induced B cell malignancy in humans with inherited CD70 deficiency. *The Journal of experimental medicine* 214, 91-106 (2017).

7. R. F. Schindler, C. Scotton, J. Zhang, C. Passarelli, B. Ortiz-Bonnin, S. Simrick, T. Schwerte, K. L. Poon, **M. Fang**, S. Rinne, A. Froese, V. O. Nikolaev, C. Grunert, T. Muller, G. Tasca, P. Sarathchandra, F. Drago, B. Dallapiccola, C. Rapezzi, E. Arbustini, F. R. Di Raimo, M. Neri, R. Selvatici, F. Gualandi, F. Fattori, A. Pietrangelo, W. Li, H. Jiang, X. Xu, E. Bertini, N. Decher, J. Wang, T. Brand, A. Ferlini,

POPDC1(S201F) causes muscular dystrophy and arrhythmia by affecting protein trafficking. *The Journal of clinical investigation* 126, 239-253 (2016).

8. S. Olgiati*, M. Quadri*, **M. Fang***, J. P. Rood, J. A. Saute, H. F. Chien, C. G. Bouwkamp, J. Graafland, M. Minneboo, G. J. Breedveld, J. Zhang, F. W. Verheijen, A. J. Boon, A. J. Kievit, L. B. Jardim, W. Mandemakers, E. R. Barbosa, C. R. Rieder, K. L. Leenders, J. Wang, V. Bonifati, DNAJC6 mutations associated with early-onset Parkinson's disease. *Annals of neurology* 79, 244-256 (2016).

9. I. Masuho*, **M. Fang***, C. Geng, J. Zhang, H. Jiang, R. K. Ozgul, D. Y. Yilmaz, D. Yalnizoglu, D. Yuksel, A. Yarrow, A. Myers, S. C. Burn, P. L. Crotwell, S. Padilla-Lopez, A. Dursun, K. A. Martemyanov, M. C. Kruer, Homozygous GNAL mutation associated with familial childhood-onset generalized dystonia. *Neurology. Genetics* 2, e78 (2016).

10. E. Gregianin, G. Pallafacchina, S. Zanin, V. Crippa, P. Rusmini, A. Poletti, **M. Fang**, Z. Li, L. Diano, A. Petrucci, L. Lispi, T. Cavallaro, G. M. Fabrizi, M. Muglia, F. Boaretto, A. Vettori, R. Rizzuto, M. L. Mostacciuolo, G. Vazza, Loss-of-function mutations in the SIGMAR1 gene cause distal hereditary motor neuropathy by impairing ER-mitochondria tethering and Ca2+ signalling. *Human molecular genetics* 25, 3741-3753 (2016).

11. L. Gao, X. Dang, L. Huang, L. Zhu, **M. Fang**, J. Zhang, X. Xu, L. Zhu, T. Li, L. Zhao, J. Wei, J. Zhou, Search for the potential "second-hit" mechanism underlying the onset of familial hemophagocytic lymphohistiocytosis type 2 by whole-exome sequencing analysis. *Translational research* 170, 26-39 (2016).

12. C. Dallabona, T. E. Abbink, R. Carrozzo, A. Torraco, A. Legati, C. G. van Berkel, M. Niceta, T. Langella, D. Verrigni, T. Rizza, D. Diodato, F. Piemonte, E. Lamantea, **M. Fang**, J. Zhang, D. Martinelli, E. Bevivino, C. Dionisi-Vici, A. Vanderver, S. G. Philip, M. A. Kurian, I. C. Verma, S. Bijarnia-Mahay, S. Jacinto, F. Furtado, P. Accorsi, A. Ardissone, I. Moroni, I. Ferrero, M. Tartaglia, P. Goffrini, D. Ghezzi, M. S. van der Knaap, E. Bertini, LYRM7 mutations cause a multifocal cavitating leukoencephalopathy with distinct MRI appearance. *Brain* 139, 782-794 (2016).

13. O. K. Alkhairy, H. Abolhassani, N. Rezaei, **M. Fang**, K. K. Andersen, Z. Chavoshzadeh, I. Mohammadzadeh, M. A. El-Rajab, M. Massaad, J. Chou, A. Aghamohammadi, R. S. Geha, L. Hammarstrom, Spectrum of phenotypes associated with mutations in LRBA. *Journal of clinical immunology* 36, 33-45 (2016).

14. A. Reyes, L. Melchionda, A. Nasca, F. Carrara, E. Lamantea, A. Zanolini, C. Lamperti, **M. Fang**, J. Zhang, D. Ronchi, S. Bonato, G. Fagiolari, M. Moggio, D. Ghezzi, M. Zeviani, RNASEH1 mutations impair mtDNA replication and cause Adult-Onset Mitochondrial Encephalomyopathy. *American journal of human genetics* 97, 186-193 (2015).

15. A. E. Pen, M. Nyegaard, **M. Fang**, H. Jiang, R. Christensen, H. Molgaard, H. Andersen, B. P. Ulhoi, J. R. Ostergaard, S. Vaeth, M. Sommerlund, A. P. de Brouwer, X. Zhang, U. B. Jensen, A novel single nucleotide splice site mutation in FHL1 confirms an Emery-Dreifuss plus phenotype with pulmonary artery hypoplasia and facial dysmorphology. *European journal of medical genetics* 58, 222-229 (2015).

16. A. Masotti, P. Uva, L. Davis-Keppen, L. Basel-Vanagaite, L. Cohen, E. Pisaneschi, A. Celluzzi, P. Bencivenga, **M. Fang**, M. Tian, X. Xu, M. Cappa, B. Dallapiccola, Keppen-Lubinsky syndrome is caused by mutations in the inwardly rectifying K+ channel encoded by KCNJ6. *American journal of human genetics* 96, 295-300 (2015).

17. O. K. Alkhairy, R. Perez-Becker, G. J. Driessen, H. Abolhassani, J. van Montfrans, S. Borte, S. Choo, N. Wang, K. Tesselaar, **M. Fang**, K. Bienemann, K. Boztug, A. Daneva, F. Mechinaud, T. Wiesel, C. Becker, G. Duckers, K. Siepermann, M. C. van Zelm, N. Rezaei, M. van der Burg, A. Aghamohammadi, M. G. Seidel, T. Niehues, L. Hammarstrom, Novel mutations in TNFRSF7/CD27: Clinical, immunologic, and genetic characterization of human CD27 deficiency. *The Journal of allergy and clinical immunology* 136, 703-712.e710 (2015).

18. M. Robusto, **M. Fang***, R. Asselta, P. Castorina, S. C. Previtali, S. Caccia, E. Benzoni, R. De Cristofaro, C. Yu, A. Cesarani, X. Liu, W. Li, P. Primignani, U. Ambrosetti, X. Xu, S. Duga, G. Solda, The expanding spectrum of PRPS1-associated phenotypes: three novel mutations segregating with X-linked hearing loss and mild peripheral neuropathy. *European journal of human genetics* 23, 766-773 (2015)

CONTENTS

LIST OF ABBREVIATIONS

ACMG	American College of Medical Genetics and Genomics
AD	Autosomal Dominant
AID	Autoimmune disease
AR	Autosomal Recessive
AT	Ataxia-Telangiectasia
ATM	Ataxia-telangiectasia mutated kinase
AUC	Area under the curve
BCR	B cell receptors
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation-Dependent Depletion
CD	Coverage depth
CDR3	Complementarity-determining region 3
ChIP	Chromatin immunoprecipitation
CID	Combined Immunodeficiency
CNV	Copy number variation
CVID	Common variable immunodeficiency
DDR	DNA-damage response
DSBs	DNA double-strand breaks
EBV	Epstein-Barr virus
ELISA	Multiplex enzyme-linked immunosorbent assay
EMSA	Electrophoretic mobility shift assay
ENCODE	The Encyclopedia of DNA Elements Project
ESID	European Society for Immunodeficiencies
ESP	Exome Variant Server
ExAC	Exome Aggregation Consortium
FDR	False discovery rate
GATK	Genome Analysis Toolkit
GC	Guanine-Cytosine
GDI	Gene damage index
GTEx	Genotype-Tissue Expression Project
gnomAD	The Genome Aggregation Database
GOF	Gain-of-function
GQ	Genotyping quality
GWAS	Genome-wide association study
HGMD	Human Gene Mutation Database
HLH	Hemophagocytic Lymphohistiocytosis
HR	Homologous recombination
ICF	The Immunodeficiency, Centromeric instability, and Facial anomalies
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IgM	Immunoglobulin M
InDels	Insertion/deletion
KG	The 1000 Genomes Project
KRECs	κ-deleting recombination excision circles
LOF	Loss-of-function
mRNA	Messenger RNA
MSC	Mutation Significance Cutoff

ncRNA	Non-coding RNA
NGS	Next-generation sequencing
NHEJ	Non-homologous DNA end joining
nsSNVs	Non-synonymous single nucleotide variation
PAD	Primary antibody deficiency
paTCR	Pathology-associated TCR
PCA	Principal components analysis
PCR	Polymerase chain reaction
PID	Primary immunodeficiency
PolyPhen2	Polymorphism Phenotyping v2
PTM	Post-translational modification
RAPID	The Resource of Asian Primary Immunodeficiency Diseases database
ROC	Receiver Operating Characteristics
RSSs	The recombination signal sequences
RT-PCR	Real-time polymerase chain reaction
SCID	Severe combined immunodeficiency
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
SnpEff	Single nucleotide polymorphism effect
SNV	Single-nucleotide variant
SOAP	Short oligonucleotide alignment program
stLFR	Single tube long fragment read
TCR	T cell receptors
TdT	Terminal deoxynucleotidyl transferase
TRECs	T-cell receptor excision circles
TRS	Targeted region sequencing
UTR	Untranslated Regions
VEP	Variant Effect Predictor
VIPPID	Variant Impact Predictor for Primary immunodeficiency
VUS	Variants of unknown significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
XR	X-linked recessive

1 INTRODUCTION

1.1 PRIMARY IMMUNODEFICIENCY

Primary immunodeficiency diseases (PIDs) consist of a class of immune deficient diseases with high heterogeneity, predisposing individuals to an increased frequency and severity of infections, immune dysregulation, autoimmune manifestations and malignancy (*1*). The incidence of PID varies in different populations and there are an expected 6 million affected individuals worldwide (*2*). Around 400 PID genes have been identified to date (*3, 4*). It has been estimated that around 3,100 genes are involved in the connectome of cells within the immune system (*5*), suggesting that the genetic etiology of many PIDs have not identified to date. Most PIDs described until now are monogenic in origin but it is increasingly recognized that some of these diseases are polygenic in origin and may involve multiple cell types or organs.

Some of the PIDs are life-threatening and diagnosed during childhood, resulting in serious symptoms, requiring a high level of care (e.g. gammaglobulin substitution, cytokine treatment and recently developed cellular and gene therapies (*6*)) and for the vast majority, no effective treatment presently exists. There are contraindications in the use of vaccines for PID patients, as significant vaccine-related adverse events may occur when live attenuated vaccines are given (*7*). Besides, severe PIDs such as severe combined immunodeficiency (SCID) requires early identification in order to initiate prompt treatment and improve survival (*8*).

The identification of the causative gene in PID patients serves as the starting point for intervention which makes a more rapid and accurate diagnosis or test of pediatrics diseases and new therapeutic interventions possible. All of those effective precautionary tests, such as Premarital test, Antenatal test and Newborn Screening Test can not only promote the genetic testing progress of the PIDs and reduce the number of children with severe forms of immunodeficiency but also prevent transmission of hereditary diseases.

1.2 NEXT GENERATION SEQUENCING TECHNOLOGY

The rapid development of sequencing technologies in the past decade has dramatically changed the strategy and increased the pace of identification of causative genes of human diseases, especially the rare disorders (*9*). Considering the sequencing cost (*10*) and the technical effort involved, NGS is also rapidly becoming a routine for detecting SNVs and small InDels (*11*) (see **Figure** 1), which has resulted in a surge of discoveries on the cause of inherited diseases. Identification of pathogenic mutations and understanding of the underlying molecular mechanism serve as starting points for genetic diagnosis and counseling (*12*), while diseases such as familial forms of PID provides an opportunity to uncover mutations that may have implications both in PID and other immune-related diseases and may shed light on new therapeutic approaches for the disease. Over the past few years, researchers have investigated the etiology of many subtypes of PID on a molecular level and their clinical outcome by using Whole Exome Sequencing (WES) (*13-18*).

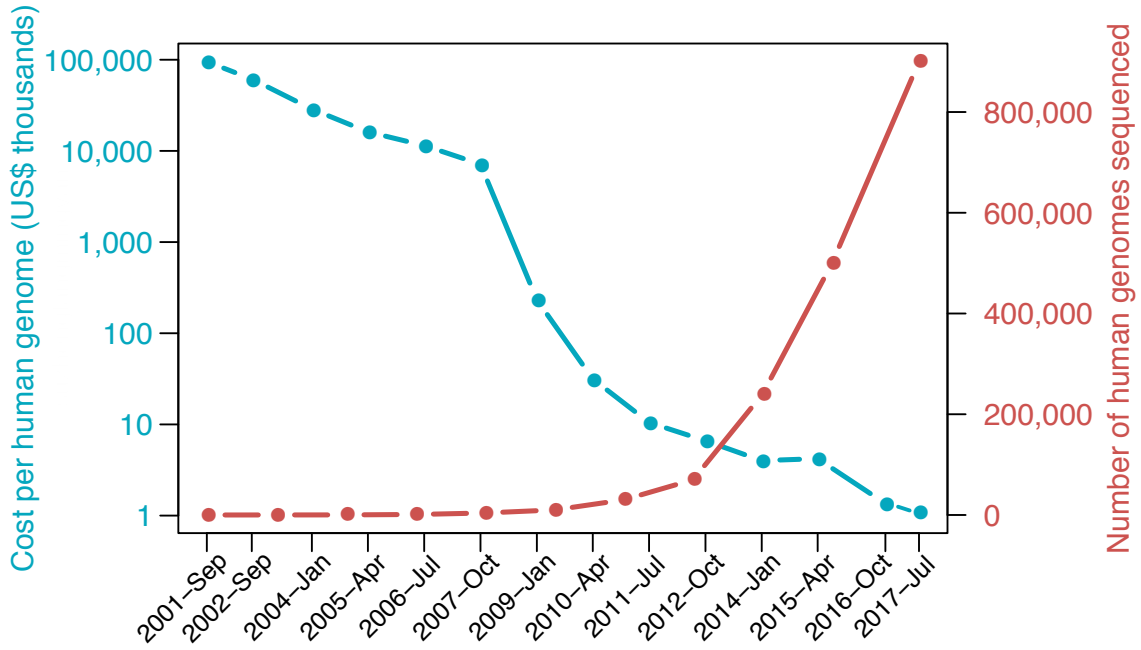

Figure 1. An exponential decay curve shows the cost of DNA sequencing has dropped dramatically (blue line), and the exponential growth curve shows sequenced human genomes number (red line) increased significantly over the past decade.

Although WES focuses on the coding region of the genes which only constitutes approximately 1% of the whole genome, these regions are considered to contain 85% of the mutations underlying Mendelian disorders (*19*). Its use has dramatically increased research efficiency and has also significantly reduced the cost of research and the computation burden as well as save computer storage. Therefore, WES has been widely used for the causative gene identification of various types of inherited diseases (*20-29*). Several capture kits (**Figure 2**) have been designed to capture all human exons, ranging from approximately 34 Mb to 71 Mb in covered region size.

The main limitation of WES after employing a sequence capture technology is the variability and noise due to capture and hybridization. Besides, WES is neither readily able to detect structural DNA changes (SVs) or copy number changes (CNVs), although these two kinds of variations cause a number of diseases, nor to detect variants located in the non-coding regions or coding regions where it is challenging to design the capture probes. Whole Genome Sequencing (WGS), which is more complete, thus tends to be increasingly attractive as an alternative for WES (*30*). Despite WGS being more reliable to detect all kinds of variation, SNVs, InDels, CNVs and SVs, our understanding of other genomic elements, such as in non-coding regions, restricts the efficiency of annotation of pathogenic sequence variations (*31*). The Encyclopedia of DNA Elements (ENCODE) Project and Genotype-Tissue Expression (GTEx) Project (*32-34*) provide possibilities to accurately and completely annotate the various functional elements of the genome.

Targeted region sequencing (TRS) is a more customized approach, which is a single test to evaluate many of the genes related to a particular disease. TRS has also been used to study diseases based on linkage analysis or homozygosity mapping. It usually uses a high depth sequencing result and has excellent performance in evaluating large genes (*35*) and detecting mosaic mutations. This method also allows an update of the panel regularly.

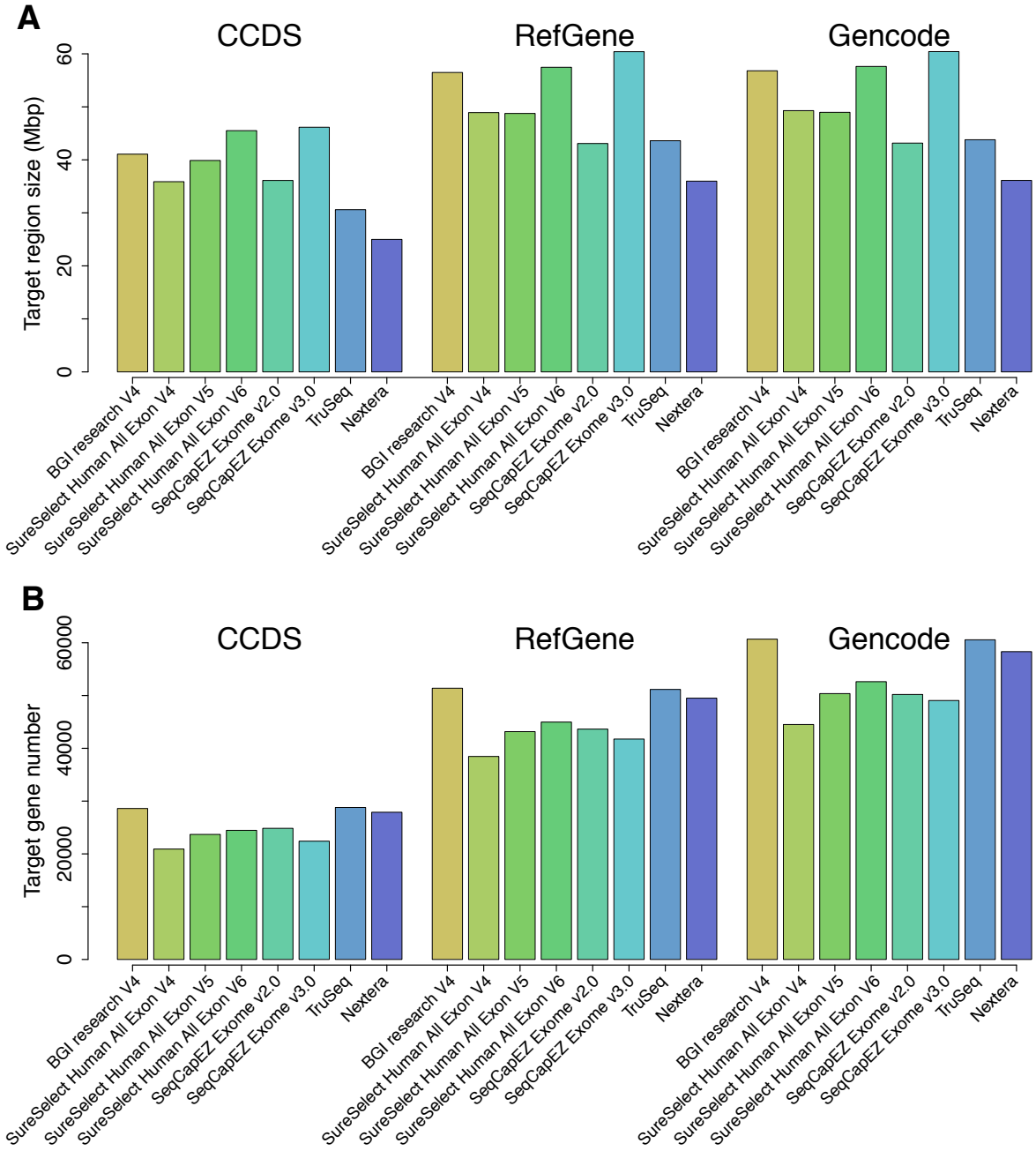

Figure 2. Covered region sizes (**A**) and gene number (including protein-coding genes, long/small non-coding RNA genes, pseudogenes and Immunoglobulin/T-cell receptor gene segments) (**B**) of the capture target genes in different annotation databases for various exome enrichment designs.

1.3 BIOINFORMATIC ANALYSIS

Data analysis can be divided into two distinct parts: 1) Standard analysis which includes sequencing alignment and variant calling; 2) Advanced analysis which aims to determine the pathogenic variant.

Burrows-Wheeler Aligner (BWA) (*36*) and the Short Oligonucleotide Alignment Program (SOAP) package SOAPaligner (SOAP2.21) (*37*) can be used to align reads onto the human reference genome (GRCh37/hg19 or GRCh38/hg38) and only mapped reads are kept for subsequent analysis. The calculation of coverage is based on all mapped reads in the target region. The fraction of target positions where we called a high-confidence consensus

genotype is comparable to the regions covered by probes in the capture kit. The consensus genotypes can be called by the Genome Analysis Toolkit (GATK) (*38*) or SOAPsnp (*39*).

The genotypes different from the human genome reference were extracted as candidate variants, then the unreliable proportion was filtered out and finally obtained high-confidence variants for the investigated samples. These SNPs and InDels would be further annotated and categorized by an automated pipeline or annotation tools, such as Ensembl's VEP (Variant Effect Predictor) (*40*), single nucleotide polymorphism effect (SnpEff) (*41*), ANNOVAR (*42*) and the Variant Annotation, Analysis and Search Tool (VAAST) suite's Variant Annotation Tool (VAT) (*43*). Both canonical transcripts and the ones with the most severe consequence should be annotated to reduce the number of false negatives due to the incomplete annotation of transcripts (*44*).

1.3.1 Candidate gene prioritization

Artificial signals caused by various steps involved in the target region capture and sequencing process are generally thought of as noise. Neutral variants that were enriched in the sequencing platform, false positive variants due to low quality (quality score < 20 (Q20) or read depth < 4 folds) and variants with frequency >5% in geographical ancestry matched general population, were thus filtered out. The estimated copy number is no more than 2 and a distance between two SNPs shorter than 5 were also considered to be excluded in further analysis.

Massive sequencing efforts such as the 1000 Genomes Project (KG), Exome Variant Server (ESP), Exome Aggregation Consortium (ExAC) and The Genome Aggregation Database (gnomAD) and data from large scale private sequencing projects provide a powerful foundation for candidate prioritization. An allele frequency higher than 1% in the normal population could be a reasonable threshold for filtering out polymorphisms for the most of rare diseases, pedigree and genetic inheritance should also be taken into account.

The majority of variants cause disease by changing the sequence of amino acids which may further affect the function of the protein. Such amino-acids-sequencing-changing includes non-synonymous substitutions, donor and acceptor splice sites mutations, and insertions/deletions as well as truncation of proteins due to a premature stop codon. Thus, variants are prioritized based on the extent to which they can change the sequence of the protein product.

A number of prediction software tools based on evolutionary conservation or protein structure have previously been developed, providing helpful information on the impact of variants. However, these tools should be used with caution since each of them has a certain percentage of false positives and false negatives. A uniform cutoff (the Mutation significance cutoff (MSC)) (*45*) is preferred in order to improve genome-wide accuracy. Genes that have a high gene damage index (GDI) (*46*) value are prone to enrich phenotype-irrelevant mutations as they are relatively less evolutionarily conserved, including sensory perception genes and long coding sequences. Besides, many false positive results come from these genes.

The mutations identified in the studied families represent all modes of Mendelian inheritance, autosomal dominant, recessive and X-linked. For many of the families, however, the exact mode of inheritance is unknown, especially for the families where only the proband is sequenced or the proband is the only affected case in the family. Therefore, different inheritance patterns should be considered to identify potential disease-causing variants (heterozygous for autosomal dominant, homozygous or compound heterozygous for autosomal recessive and hemizygous for X-linked), and identification of the causative gene in these cases would allow determining the inheritance pattern.

Consanguineous families and families with multiple affected and unaffected individuals provide an advantage in disease gene identification by NGS. Since the former's causal mutations are assumed to be inherited from a common ancestor, homozygosity mapping using the data from NGS will help to further narrow down the candidate genes to those within the homozygous regions. The latter will allow the evaluation of the co-segregation of candidate variants with the disease phenotype among family members. Thus, it is able to exclude all variants that do not show linkage to the disease.

1.3.2 CNV detection

The NGS field is evolving rapidly and novel cutting-edge methods are frequently made available for structural variant analysis (*47-49*), including paired-end mapping (PEM), read-depth analysis, split-read strategies, and sequence assembly comparisons. By allowing the detection of point variants and inversion, NGS can estimate the exact location of a breakpoint since it is not based on hybridization, allowing a better estimation of high copy numbers (*50*).

Nevertheless, CNV analysis based on WES data is still a challenge (*51, 52*). Abnormal GC content (*53*), PCR amplification bias and pre-existing InDels/CNVs (*51, 52*) can introduce bias. The latter causes bias of the sequence read quality score which needs to be re-calibrated for those known InDels in nearby regions. In addition, noise also exists due to the non-uniform gene coverages, where longer genes and complex regions tend to have a better relative coverage. Statistical models have been introduced to solve this problem. Modeling the read-depth of each position, applying multiple algorithms (*51*) to estimate the noise (*53*) is a promising strategy to reduce false positives (*54*). Another option is to detect rare CNVs by using a complicated algorithm on WES data and infer using copy number polymorphic genotypes (*55*). However, each statistical model only serves a specific situation.

In contrast to the extensively studied WGS workflow, identification of CNVs from exome data still can be erroneous when mistaking using packages or tools, even though some of them are customized for dealing with variabilities in WES studies. The usage of each tool/package can be case sensitive. By categorizing the existing algorithms using certain practical criteria, like reference dependent or independent, the way of setting the reference, the way of normalization, read-depth/split reads, etc., to get a better vision of the current status of existing algorithms, it is possible to add novel or complement ideas efficiently.

1.4 VARIANT INTERPRETATION AND CLASSIFICATION

According to the joint consensus recommendation of variants interpretation suggested by the American College of Medical Genetics and Genomics (ACMG) (*56*), we found that most detected SNVs are classified as variants of unknown significance (VUS) which are private in many cases, and the more samples sequenced, the more VUSs are observed. To single out the pathogenic variants from the benign variants is critical for clinical pathogenicity. The recurrent observation of VUS after sequencing more samples and data sharing serves as an approach to knowing the potential consequence of VUSs. Functional assessment is another comprehensive approach but might be a labor-intensive way to analysis the molecular function of VUS. Site-directed mutagenesis and massively parallel functional assays had been used to distinguish pathogenic variants from polymorphisms in *MSH2* (*57*) and *BRAC1* (*58*).

In addition, penetrance is not considered in the existing classification system and it should be reported in the clinical screening reports. We learned that large-scale genome studies and experimental function can be informative for computational prediction development, and these results were also extremely useful for clinical practice, such as providing proper carrier counseling and allowing precision medicine.

Although synonymous mutations are thought to be harmless as they do not alter the amino acid sequence are thus generally excluded in the further causative gene prioritization, synonymous mutations may alter protein expression, post-translational modification (PTM), stability and function. Rare codons reduce protein expression and affect rates up to various fold (*59*). Species-specific codon bias analysis might help to accurately interpret the consequence of gene variants and the effects of codon usage on gene expression, which would be beneficial to improve design principles for gene therapy.

1.5 OLIGOGENIC ANALYSIS

The above-described approaches have been most effectively utilized in familial and cohort studies for causative gene identification, mirroring the classic strategies of Mendelian disease studies. However, earlier findings suggest that a majority of patients with different forms of PID remain undiagnosed. In order to understand the underlying pathogenic mechanisms in undiagnosed patients, in particular in newly recognized polygenic forms of PID and diseases caused by a combination of heterozygous mutations in several PID associated genes, a broader and more comprehensive analytic approach might be necessary.

The identification of the pathophysiological processes involved in polygenic forms of PID has been illuminated in recent years where combinations of heterozygous mutations in several PID associated genes have been recognized. More than one gene involved in a specific pathway (eg. the degranulation pathway, the IL7 signaling pathway and the DNA repair pathway) can contribute to a disease phenotype in a given family (*60*) or patients (*61, 62*), where each of the variants may only explain a part of the phenotype (*60, 61*). Our previous work has suggested a potential ‘‘second-hit’’ mechanism (*63*) which also provides clues to the different phenotypes in family members although they carry the same causative variants. An increasing number of studies indicate that PID may be caused by

haploinsufficiency of genes, including *CTLA4* (*64, 65*), *NFAT5* (*66*), *IKZF1* (*67*), *GATA2* (*68*), *SERPING1* (*69*), *CHD7* (*70*), *FAS* (*71*), *IFNGR2* (*72*), *TNFRSF13B* (*73*), *NFKB1* (*74*) and *NFKB2* (*75*). A combination of heterozygous mutations in two of them could result in an increased disease susceptibility that deserves special attention.

Even so, since PID known genes and immunologically essential genes are spread throughout the genome, and NGS has unraveled a large number of rare variants. It is still a significant challenge to find a single variant in autosomal recessive traits genes that are compatible with the affected subject's phenotype. In order to understand the underlying biological mechanisms in undiagnosed patients, in particular for polygenic forms of PID, the development of novel analytical approaches to filtering potential pathogenic variants and rapidly correlate them with patient phenotypes is in great need.

1.6 VALIDATION AND REPLICATION

Except for previously well-characterized variants and pathogenic variants in well-characterized genes, validation is indispensable to eliminate false positive findings. First, these studies allow validation of the original variant in the studied pedigrees and geographical ancestry-matched healthy individuals, which is challenging by only using *in silico* analysis pipeline or predictor. Replication in additional samples provides convincing evidence for any novel causative genes. Confirmed data of genetic validation and segregation analyses from earlier studies show that all of these *in silico* methods have a 10-20% false negative and 10-20% false positive rate. These methods provide an essential prerequisite but are not sufficient for determining whether a variant is pathogenic or disease-causing and should not be relied entirely on for identification. *In vitro* and *in vivo* studies can both be employed to explore the biological effect of a variant and no other method can replace a functional analysis for previously uncharacterized variants.

Standard functional tests can be used to determine how causative genes are involved in immunity. These tests include molecular analyses (e.g., qPCR, microarrays, chromatin immunoprecipitation (ChIP) and RNA seq), biochemical analyses (e.g., western blot, electrophoretic mobility shift assay (EMSA), luciferase assays), gene transfer experiments, CRISPR/Cas9 knock-out/in trials and cellular studies. The cellular studies include immunophenotyping by flow cytometry (to test the appropriate CD markers), *in vitro* activation (Fascia) and differentiation tests, cytokine measurement by flow cytometry or multiplex enzyme-linked immunosorbent assay (ELISA).

1.7 MULTIOMICS APPROACHES

Despite tremendous advances in analyzing NGS data and efforts to create massive databases for mutations/variants, the observed mutations are often not possible to relate to the clinical phenotype. Moreover, a majority of patients with different forms of PID remain undiagnosed at the molecular genetic level. Given these difficulties, it is widely accepted that additional experimental methods, including epigenetics, transcriptomics, proteomics, metagenomics and determination of immune repertoires should also be employed in order to diagnose the patients.

1.7.1 Epigenetics

Epigenetic factors, including DNA methylation, histone modifications and non-coding RNAs, play fundamental roles in cell development, differentiation and function (*76, 77*). It is well known that both during immune cell differentiation and B or T cell activation, DNA methylation combined with the transcription factors play a vital role in the regulation of gene expression (*78*), which through DNA methylation of cytosines in the context of cytosine-guanine dinucleotides (CpG), often located in clusters (CpG islands) within regulatory regions such as promoters and enhancers. Epigenetics is also thought to be essential for V(D)J rearrangement, and immunological memory (*79, 80*) and recent studies have identified DNA methylation alterations in common variable immunodeficiency (CVID) (*78*), the Immunodeficiency, Centromere instability and Facial anomalies (ICF) (*81, 82*) and ATM deficiency (*83*).

1.7.2 Transcriptomics

The transcriptome is a direct link between genomic information and the proteome, and it is a starting point for studying the structure and function of a given gene. With transcriptome sequencing, variations in RNA sequence (SNPs and InDels), and downstream information including isoform expression, exon expression, gene structure refinement, alternative splicing, novel transcripts, and gene fusion can serve to complement the genomic sequencing. Transcriptomics can be directly interpreted to prioritize candidate disease-causing genes of Mendelian disorders, by detection of the aberrantly expressed genes, reveal mono-allelic expression and direct probing of splice isoforms to discover splicing defects. The aberrantly expressed genes can be due to the variants in the promoter, enhancer or intronic regions; these variants are tricky during the interpretation of genomics studies. It is also difficult to predict the consequence of splice site variant from the WES or WGS data by using current algorithms (*84*). Besides, we usually neglected the mono-allelic expression which due to heterozygous mutation of recessive genes, but it is essential for some specific patients. Transcriptome studies have proven to be a powerful strategy to the molecular diagnosis of mitochondriopathy patients (10% (5 of 48)) and to identify potential causative genes of the remaining patients (*84*).

Comparative analysis of transcriptomic and methylome data will potentially be able to reveal changes in gene expression as a consequence of alterations in the methylation of relevant CpG sites. Further increase of the use of RNA sequencing may ultimately identify non-coding RNAs, alternatively spliced transcripts as well as allele-specific expression underlying immune dysfunction.

1.7.3 Proteomics

Proteome data represent the combined effect of the genetic, epigenetic, transcriptional and translational regulation and proteomics applications will thus provide valuable information about the "real-time" dynamic molecular phenotype (*85*). The quantitative proteomic analysis will be essential to understand underlying disease mechanisms which in turn will shed light on disease diagnosis, prevention, intervention, monitoring, and prediction of the treatment response (*86*). By acquiring both deep cellular and humoral proteome data, combined with available genomic data, incorporating sequence information on online proteomics database and analyzing the pathway activation status, it could be used to define PIDs associated

immune dysfunctions. It may thus be possible to understand the heterogeneity of the immune system and immune dynamics and has the potential to unravel diagnostic biomarkers.

1.7.4 Metagenomics

Over the past decades, emerging studies have suggested that the human gut microbiome plays a crucial role in the development of a range of diseases, including multiple sclerosis (*87, 88*), rheumatic diseases (*89*), amyotrophic lateral sclerosis (*90*) and atherosclerotic cardiovascular disease (*91*). Specific bacteria play a role in affecting or regulating the immune system (*87, 88*). The composition changes of the gut microbiota have recently been observed in several forms of immunodeficiency in mice (*92-94*). Investigation of the population structure and taxonomic or functional features of the microbiota may unfold how gut bacteria regulate the immune system both in patients with PID and healthy individuals.

1.7.5 Immune repertoires

The diversity of antigen receptor repertoires is generated by recombination of the variable (V), diversity (D) and joining (J) gene segments, and is further augmented by junctional diversity (non-templated (N) nucleotide additions in the V-D and D-J junctions inserted by terminal deoxynucleotidyl transferase (TdT) and random deletion of nucleotides at the recombining edges as a consequence of asymmetric hairpin opening by ARTEMIS) (*95-97*). The productive T-cell receptor (TCR) repertoire further undergoes maturation in the thymus and through interaction with antigens.

DNA double-strand breaks (DSBs) are cytotoxic lesions, which can be caused by pathogenic (ionizing radiation and radio-mimetic chemicals) or physiologic (introduced during V(D)J recombination and class switch recombination in developing lymphoid cells) conditions (*98*). Different forms of PID are associated with cancer and can develop due to genomic instability when DSBs are improperly repaired, such as in patients with Fanconi's anemia, SCID, Ataxia-Telangiectasia (AT) and Nijmegen breakage syndrome (*99*). In addition, epigenetic modifications play a crucial role in the regulation of the active chromatin state in order to make the recombination signal sequences (RSSs) accessible to the recombinase (*80*), recruitment and stable binding of the recombination complex (*100*). The epigenetic plasticity involved in lymphocyte development and activation is a risk factor for human diseases associated with the immune system (*100*), such as in patients with mutations in *DNMT3B* (*101-105*) and *ZBTB24* (*106, 107*) who develop ICF syndrome type 1 and type 2, respectively.

It is now possible to analyze specific B or T cell antigen receptor (BCR or TCR) repertoires by using high-throughput sequencing technology and to characterize the repertoire clonality and investigate of the mechanisms of immune surveillance in PIDs in a single experiment. Immune repertoires as distinctive feature of antigen receptor (VDJ) rearrangement have been explored in a number of immunodeficiency diseases, including RAG deficiency (TCRβ and BCR) (*108*), ATM deficiency (TCRα and BCR) (*109, 110*), Cernunnos deficiency (TCRβ, TCRδ and BCR) (*96*), Wiskott-Aldrich syndrome (TCRβ) (*111*) and CVID (TCRβ, BCR) (*112, 113*). The presence of immune repertoire diversity in patients with different immunodeficiency gene mutations provide opportunities to learn the effects of specific

causative genes which will deepen our understanding of the biology of these disorders and will provide insight into the clinical heterogeneity observed (*97*).

1.7.6 Pathogen detection

PIDs can lead to life-threatening infections, and therefore the timely and accurate microbial diagnostics is critical for implementing the individualized treatment plan. Current diagnoses include bacterial culture, qPCR and immunological essays, which are commonly used in clinical settings. However, these methods are either time-consuming or labor-intensive, and they can only detect a narrow range of pathogens in a single test. In contrast, NGS technology has compelling advantages over traditional methods. By unbiased sequencing nucleotide that extracted from the sample, NGS technology can comprehensively characterize underlining pathogens without prior knowledge. In addition, this approach can also detect some of the unculturable pathogens, rare pathogens and mixed infections, which often cannot be identified by traditional methods. Collectively, NGS technology enables rapidly and comprehensively microbial diagnostics of infection in PIDs patients and thus can accelerate the diagnostic process.

2 AIMS

2.1 GENERAL AIM

Using NGS technology to identify the genetic variations that are involved in the pathogenesis of PID and to investigate the immune repertoire characteristics and the monogenic and polygenic etiologies of PID, and to develop tools that can be used to facilitate the process of the genetic investigation.

2.2 SPECIFIC AIMS

2.2.1 We aimed to optimize the NGS bioinformatics analysis pipeline by integrating different analytical strategies, and to employ the pipeline in the investigations of the studied samples.

2.2.2 To identify pathogenic variants in selected PIDs samples by using different NGS approaches.

2.2.3 To improve the accuracy of molecular diagnosis and provide insights into genotype-based care for patients with an atypical clinical presentation.

2.2.4 To evaluate the feasibility and utility of the proposed sequencing approaches in the identification of genetic etiology of PIDs and beyond.

2.2.5 To refine the prediction algorithms by using a Random Forest-based discriminator, which was established by utilizing the features of pathogenic variants and benign mutations and the integration of other existing prediction software.

2.2.6 To characterize the T-cell receptor repertoire in selected PIDs patients, and to elucidate the role of different components of the DNA repair and methylation machinery in TCR repertoire profile and diverse phenotypes of immunodeficiency.

3 MATERIALS AND METHODS

3.1 SUBJECTS

Existing PID patient cohorts were collected at the Karolinska Hospital, the Children's Medical Center (Pediatrics Center of Excellence affiliated to the Tehran University of Medical Sciences, Tehran, Iran) and outside collaborators. More than 600 samples have been subjected to NGS in the purpose of identifying underlying genetic causes. To exclude false-positive findings, healthy members of enrolled pedigrees were also collected for further co-segregation validation. Pre-test counseling was also carried out before genetic testing. The pre-test counseling includes a basic medical examination, medical history, family history, laboratory and molecular data, which are followed by a medical check-up for all participants. Informed consent was obtained from all patients or their legal guardians. Ethical permissions were obtained from the ethics committees of the participating centers.

3.2 GENOMICS DATA GENERATION AND PREPROCESSING

3.2.1 Whole exome sequencing (WES)

To enrich exonic region of DNA, the Agilent SureSelect 50Mb exome capture kit (50Mb, Agilent Technologies, Santa Clara, CA) or the BGI exon capture kit for WES were used. Enriched DNA was loaded on Illumina (Illumina, San Diego, CA, USA) platform or the BGIseq500 platform (BGI, Shenzhen, China); 90 bases paired-end sequencing (Illumina) or 100 bases paired-end sequencing (BGIseq) was performed for exon library sequencing.

3.2.2 Targeted region sequencing (TRS)

A custom capture kit was designed to capture 219 PID known genes (**Table S1**). TRS was used for hybridization, and the amplified library was sequenced on BGIseq500 platform with paired-end 100 bp reads.

3.2.3 Sequence alignment, variants calling and annotation

The standard bioinformatics pipeline of WES/TRS data was performed as previously described (*44*). BWA, GATK and VEP were carried out for reads alignment to the GRCh37/hg19 human reference genome, variants calling and annotation, respectively. Public databases including KG, ESP, ExAC and gnomAD were used to identify and remove polymorphisms. CNVs were detected by ExomeDepth.

3.2.4 Candidate prioritization

A comprehensive pipeline for the annotation and clinical interpretation of the sequencing data relevant to PIDs was used to identify known and novel disease-causing variants in a high throughput fashion. The pipeline prioritizes variants for follow-up, relate them to existing clinical data, and provide a user-friendly interface for decision support. It was implemented a several-step process, which was described in our previously published paper (**Figure 3**) (*44*).

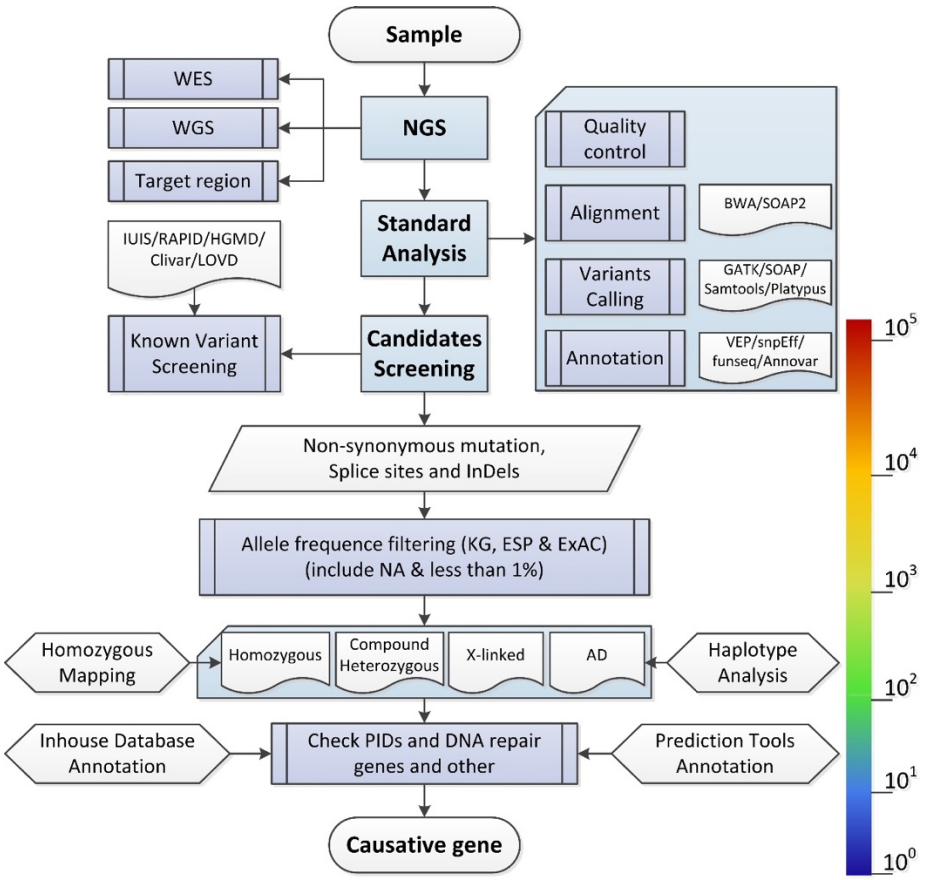

Figure 3. The diagram shows the pipeline of NGS-based disease gene identification in PIDs. The typical number of remaining variants after each prioritization step is indicated in the bar (*44*).

3.2.5 Data mining

In cases where the initial WES/TRS approach did not yield likely disease gene candidates, we considered the following options before proceeding to WGS. 1) relaxing the threshold criteria for statistical significance on the initial screening, followed by replication in additional cases, which has previously been employed successfully; 2) expanding the number of samples, if available, in order to increase the power for discovery and/or; 3) exploring the digenic or polygenic inheritance model for these undiagnosed patients (**Figure 4**).

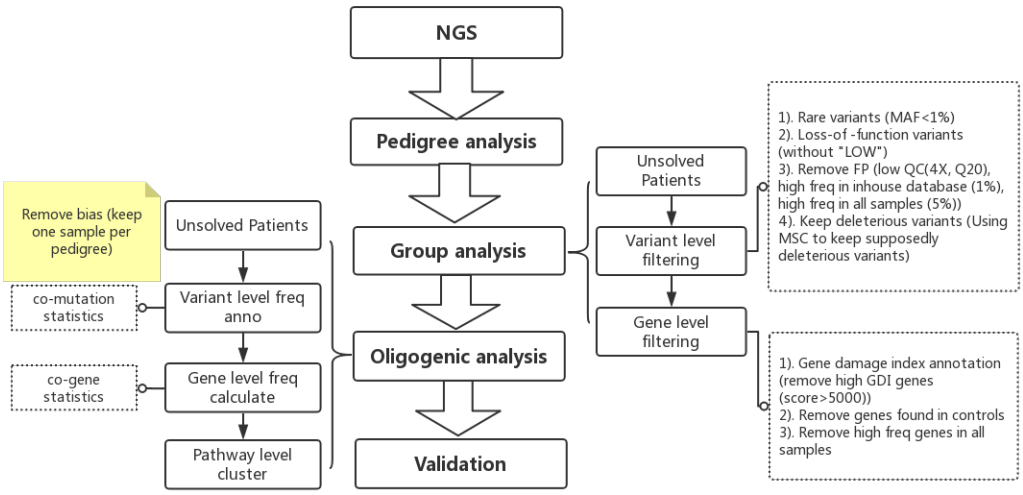

Figure 4. A framework shows the general consideration in PID cohort analysis.

3.2.6 Validation and replication

The ACMG Standards and Guidelines (*56*) were used for evaluating all prioritized candidate variants. Only pathogenic variants, likely pathogenic variants or VUS were considered to have potential clinical relevance.

If any of the above methods yielded disease gene candidates, we would attempt to validate them further to exclude sequencing errors. In addition, replicate relevant associations was also conducted by using additional samples. PCR amplification and Sanger sequencing were performed on the candidate variant. The absence of candidate variations in healthy members of the same family and normal individuals was required to exclude the possibility that the mutations are population specific non-disease-causing rare variation. Sanger sequencing of candidate genes in sporadic patients would also be performed to investigate the pathogenicity of candidate variations.

3.3 PREDICTING PATHOGENICITY OF SINGLE NUCLEOTIDE VARIATIONS FOR PRIMARY IMMUNODEFICIENCY DISEASES

3.3.1 Selection of training dataset for the prediction model

Data on mutations with known pathogenicity leading to PID were collected from the Resource of Asian Primary Immunodeficiency Diseases (RAPID) database (*114*), which is one of the most comprehensive PID genomic variation databases. Mutations from the RAPID database were used as positive samples in the training dataset. To find negative samples of the training dataset, SNVs from the exome samples of gnomAD (*115*) were downloaded and annotated by VEP (*40*). The SNVs were used as negative samples in the training dataset only when the SNVs are not present in RAPID database and were reported as 'benign' or 'uncertain significance' in the ClinVar database (*116*). Exome data from gnomAD was selected because it is the collection of mutation from 125,748 individuals, which contains a large number of rare mutations. This can reduce the risk of overestimating the performance of the model due to only using high-frequency mutations as negative samples in the training dataset.

3.3.2 Annotation of the mutations

To obtain the information of different features of SNVs, SNVBox (*117*) was used for annotation. The tool provided information about the properties of amino acid change, regional sequence composition, evolutionary conservation measurement, properties of SNV in 3D structural context and the functional impact on the translated proteins. All SNVs in the training dataset were also annotated by VarCards (*118*), an integrated annotation engine. Prediction scores from various genomic variation functional impact prediction tools were extracted and were used as features of SNVs in our VIPPID model.

3.3.3 Machine learning and calculation of feature importance

Data were processed by Perl version 5.22 and R version 3.4.4. R package 'caret' was used for Random Forest machine learning. To select features that contribute most to the accuracy of the model and to remove redundant features, R package 'Boruta' was used for feature selection. A score to indicate the importance of the feature was also calculated for each

feature. Function ‘tuneR’ in R package ‘randomForest’ was used for model tuning. The ‘mtry’ parameter was selected by choosing the one with the least Out-of-Bag error. An upper limit of 20 was set for ‘mtry’ to reduce the risk of overfitting. To evaluate the performance of the model, the model was trained and tested using non-overlapping datasets (cross-validation), in order to reduce the potential bias in the evaluation results, ten times cross-validation was performed. R package ‘pROC’ was used for ROC curve generation and AUC value calculation.

In the gene-specific model, genes with 50 or more pathogenic SNVs were trained separately and each of these genes had an independent model. Genes with less than 50 SNVs were combined and trained in a common model. In the non-gene-specific model, all genes were trained in a common model, and mutations in all PID genes shared the same parameters in the model.

Function ‘prcomp’ in R was used for principal component analysis. Package ggplot2 was used for data visualization. The median of feature importance values of each gene was calculated and used for principal component analysis (PCA), and feature importance value of -Inf was assigned as zero in the analysis.

3.4 IMMUNE REPERTOIRE

3.4.1 Sequencing of TCRβ repertoires

To amplify the CDR3 sequences of TCRβ repertoires for equal amounts of DNA samples (1.2 ug) from all TCR subjects, multiplex primers, and two complete sequencing adapter primers were adopted, as described in the previous work (*95*). We used AMPure XP (Beckman, A63882) to clean the first 15 cycles amplicons and purified PCR products were then amplified for a second round (25 cycles) with a pair of common primers. Prepared libraries were then loaded onto a BGIseq500 (BGI-Shenzhen, Shenzhen, China) and underwent 200 bp Single-end for sequencing. We obtained a mean of 5,971,100 (range 1,345,371 - 20,526,738) total reads and 5,159,138 (range 824,569 - 17,542,667) clean reads for the sequenced samples.

3.4.2 Bioinformatics analyses of TCRβ repertoires

IMonitor was used to analyze TCR sequencing data (69). After two rounds of the Basic Local Alignment Search Tool (BLAST), we obtained the final alignment result according to the optimal score (69). The International ImMunoGenneTics database (IMGT, www.imgt.org) was used as a reference. We performed in-frame and out-of-frame determination, V, D and J segments usage calculation, deletion/insertion nucleotide and amino acid sequence determination at the rearrangement with the default parameter, which has been illustrated in our previous work (69). We used “out-of-frame” to tag the rearrangement frame if the sequences with stop-codon or sequences with a length of non-multiple of three (frameshift); and “in-frame” were used as a tag under the alternative circumstance. Evenness is best expressed as a partial order, and this post structure can adequately be illustrated by Shannon-Wiener's diversity (or entropy) index (H’) and the Gini evenness coefficient (G’) measured by considering three basic requirements: permutation invariance, scale invariance and the transfer principle.

3.4.3 Analysis of pathogen-associated sequences

Totally, 20,814 pathology-associated CDR3 records of humans were collected from a manually curated database of T- and B-cell receptors targeting known antigens (TBAdb) (https://db.cngb.org/pird/tbadb/) and online databases McPAS-TCR (http://friedmanlab.weizmann.ac.il/McPAS-TCR/), where all records were derived by searching literature manually. All CDR3β sequences were used as a reference to perform alignment for all studied patients, identical sequences that were perfectly aligned to the reference were then calculated and annotated as specific for various pathogens.

3.4.4 Statistical Analysis of immune repertoire

R (R version 3.3.2) was used for statistical analysis and data visualization. We used normalized metrics from 1 million reads of data to describe the CDR3 sequences of each sample, including Shannon's H index for measurement of repertoire diversity, Gini coefficiency of clonality, frequency of top 100 CDR3 clones, Pielou's evenness index of V-J pairing, frequency of in-frame and out-of-frame rearrangements, nucleotide composition of the CDR3 sequences and V gene usage. We investigated the differences of these metrics between groups and tested their statistical significance using a two-sided Wilcoxon Rank Sum Test.

We used Bonferroni correction for multiple tests correction in V gene usage comparison. To investigate the patterns in each group during pre-selection and post-selection of TCRβ rearrangement, the length distribution of the CDR3 sequences and insertion/deletion sequences were further examined. For testing the distribution differences, two-sided bootstrap Kolmogorov-Smirnov test was used, and the significance of length difference between each group was tested by one-sided Wilcoxon Rank Sum Test. We performed PCA analysis and visualization by using the built-in R function and function from additional packages ggplot2 downloaded from CRAN (https://cran.r-project.org/). Adjustments of all statistical tests were made to multiple testing corrections, p values ≤ 0.05 or the false discovery rate (FDR) ≤ 0.1 were considered as significant.

3.5 ETHICAL CONSIDERATIONS

3.5.1 Risk Assessment

Potential risks to the participants include medical and psychological concerns. The medical risks stem from blood drawing. Blood is preferable to saliva since the yield of DNA is much higher with blood than with alternatives. Besides, the presence of non-human DNA in saliva can complicate the analysis. Venipuncture may cause discomfort, bruising, or very rarely, infection. Only skilled medical assistants performed venipuncture and drawing the amount of blood which was only sufficient for this study.

The psychological risks included: (1) distress when discussing a history of the disease and (2) loss of confidentiality. For the first concern, subjects were approached by doctors who were trained in counseling and discussing a family history of the disease. This set-up ensured that the participant was fully informed about the study before he/she made a decision about participation. Written consent was obtained from a proband, or in the case of children, both parents/guardians if the family was interested in participating. In terms of the concerns over

confidentiality, each participant was assigned an encrypted sample ID immediately upon recruitment to protect their privacy. Phenotypic information associated with their sample ID was kept in a separate secure locked file only accessible on a need-to-know basis.

3.5.2 Potential Benefits

Our findings may directly benefit the participants as it provides a correct diagnosis. Furthermore, data and conclusions derived from this study may push forward our understanding of the causes of PID and will ultimately lead to improved treatment options for patients, resulting in an enhanced quality of life.

4 RESULTS AND DISCUSSION

4.1 COHORT SPECTRUM

A total of 602 samples (including 88 healthy relatives) from 443 sporadic PID patients and 57 multiple-case families with both affected and normal individuals spanning a broad range of PIDs have been collected. The symptom of patients represents pediatric and adult-onset disorders with a wide variety of phenotypes including CVID, Hypogammaglobulinemia, Agammaglobulinemia, Hyper IgM, combined immune deficiency (CID), SCID, IgA deficiency and other primary immunodeficiency categories according to 2017 IUIS phenotypic Classification for PID (*4*) (see **Figure 5**).

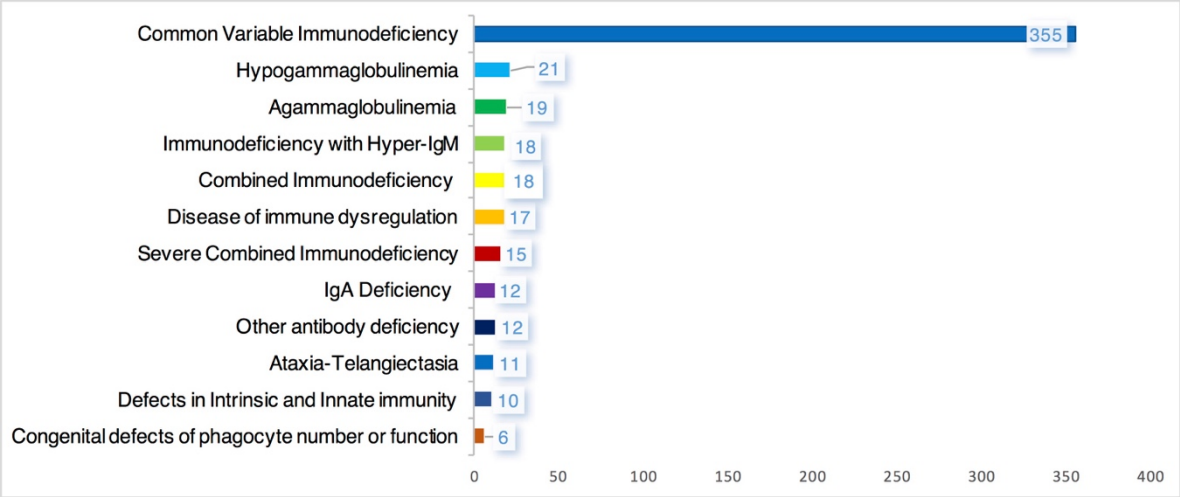

Figure 5. The phenotype distribution of the studied patients (n=514), most of which were Common Variable Immunodeficiency.

Among the 514 affected individuals, 98 of them (19%) were from consanguineous families. Most of the samples were from Iran (44%) and Sweden (31%), 69% of the patients were diagnosed as CVID (**Figure 5**). Four hundred and fifty samples (362 patients and 88 controls) were subjected to WES, whereas 217 samples (215 patients and two controls) were subjected to TRS, 65 samples (63 patients and two controls) went through both TRS and WES (**Table 1**).

Table 1. Number of pedigrees and samples with NGS performed.

No.	Pedigree	Case	Control
WES	348	362	88
TRS	216	215	2
WES&TRS	64	63	2
Total	500	514	88

4.2 NONPATHOGENIC GENETIC VARIANTS

There were in total 1,723,888 variants detected in the 450 samples that underwent WES, and 14,322 variants called in the 217 samples that underwent TRS (**Figure 6**).

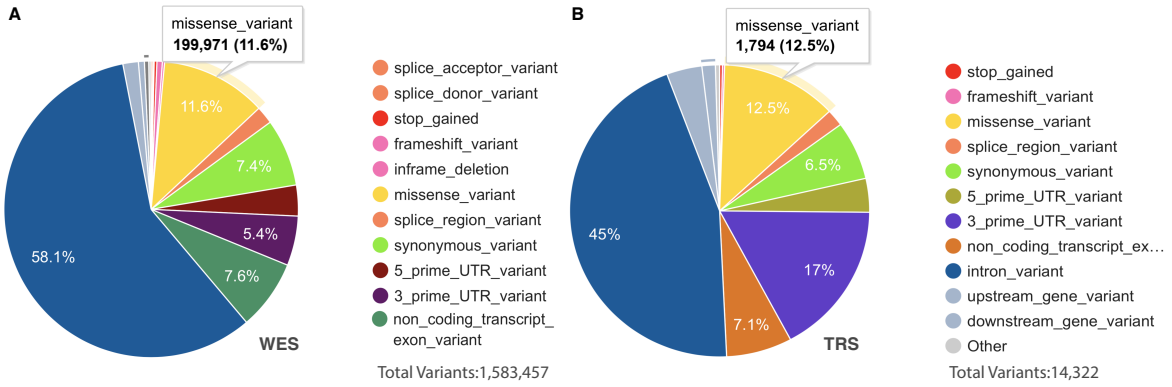

Figure 6. The distribution of consequences of the detected variants.

Among the detected variants, we found that a notable proportion of them was frequently present in the patient cohorts but rare in public databases or in-house databases (**Table 2**). These variants are unlikely to be the pathogenic variants as most of the samples are unrelated and the clinical manifestation of patients is highly heterogeneous. Thus, the variants list was used to remove the artificial variants or nonpathogenic variants.

Table 2. Statistics of genomic variants in mutation filtering.

Filtering criteria	Number of variants	Percentage
All variants	1,723,888	100.00%
Rare in public databases (AF < 1%)*	1,165,663	67.62%
Rare in in-house database (AF < 1%)	1,141,123	66.19%
Common in PID WES cohort (AF > 1%)	130,199	7.55%
Homozygous state in > 1% of samples	51,238	2.97%
Common in PID WES cohort (AF > 5%)	48,655	2.82%
Homozygous state in > 5% of samples	19,495	1.13%

*AF: allele frequency

4.3 MOLECULAR DIAGNOSIS

4.3.1 Pedigree analysis

Thus far, 514 patient samples from 500 pedigrees had undergone WES or TRS, 60% had received a likely molecular diagnosis (**Figure 7**). With the standard approach, we identified 365 causal variants in 260 PID patients (50.6 %) who were submitted to TRS/WES. Copy number variation defect was identified in 16 patients (5.2%) who carries homozygous CNV (4 genes are involved, *LRBA*, *ATM*, *DOCK8* and *PMS2*). A couple of known PID genes have been detected more than twice in different probands in our cohort. Analysis of NGS data revealed that *ATM*, *LRBA*, *BTK*, *TNFRSF13B*, *DNMT3B, PRKDC, DOCK8, PIK3CD, CTLA4*, *IL12RB1*, *IL2RG*, *RAG1* and *WAS* are the most frequent causative genes, which in total explain the disease in approximately 33.6% patients (monogenic model) (**Figure 7C** and **D**). PID known genes (*RAG1* (*18*), *CD27* (*17*) and *RAC2* (*13*)) with a new phenotype have been reported in several of our patients. In addition, two novel genes *CD70* (*15*) and *RAD50* were, for the first time, identified to be associated with primary immunodeficiency.

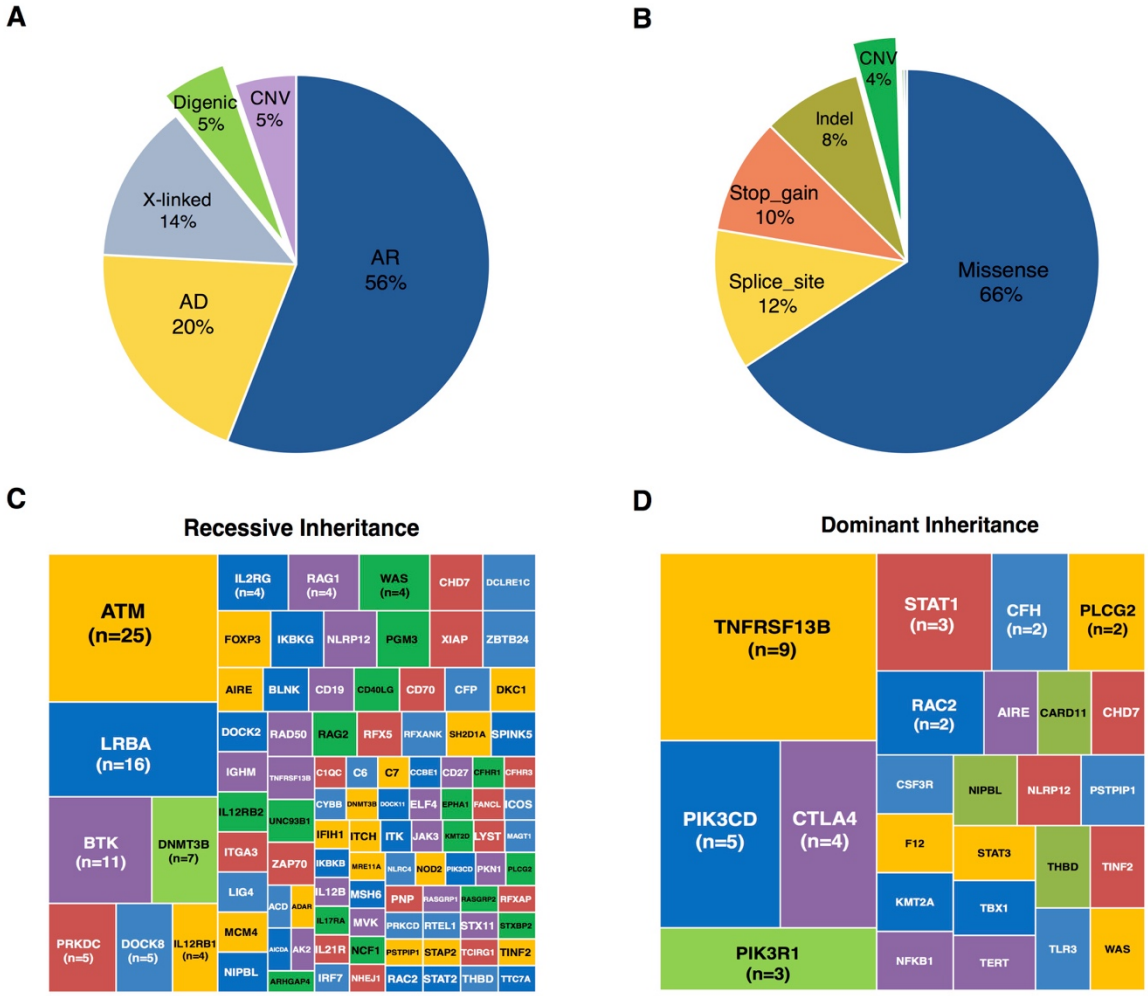

Figure 7. Inheritance pattern (**A**) and the spectrum of potential disease-associated variants (**B**) observed among 514 patients, and the treemaps show the frequency of the likely causative genes identified in the patients, either in a recessive (**C**) or dominant inheritance (**D**) model.

4.3.2 Mutation frequency analysis

We did not find any promising variants of in known PID genes for some samples, suggesting that there might be novel genes that contribute to immunodeficiency. As a large number of our patients were diagnosed as CVID, and we only found the disease-associated gene in a minority of them, we suspected that there might be some new, yet unidentified disease-associated genes among a subset of CVID patients.

Conventional analysis strategy failed to find disease genes in those unsolved CVID samples, and we hypothesized that this might be due to variation in penetrance or expressivity, or heterogeneity of onset ages, which led to imperfect gene-phenotype co-segregations. For instance, disease-causative variant carriers could have no disease manifestation, which is mistakenly treated as healthy individuals, and those variants could be interpreted as benign since they presented in control samples. Thus, we employed a new method, which was to look for genes that had significantly higher mutation rate in unsolved samples than in others, to overcome the limitation. Since samples with similar phenotype have a higher possibility of sharing the same disease genes, we also further divided the 450 WES samples into four categories: CVID-unsolved, non-CVID-unsolved, non-CVID-solved and controls. Then we

looked for genes with likely pathogenic mutations in a significantly higher proportion of samples in the CVID-unsolved group as compared to the others (Fisher exact test) and selected them as candidate disease genes for further analysis.

After applying the mutation frequency comparison strategy on the unsolved CVID and other groups, we identified 17 potential causal variants in 17 patients (5.5%). We found that the CVID-unsolved group has a higher proportion of individuals who carried rare deleterious mutations in multiple genes (*TNFRSF18, PIK3CG*, *TXNIP, CD5*, *NLRP5* and *EPHB2* for autosomal dominant mode; *LILRB1* for autosomal recessive model) than the other groups (unpublished data).

4.3.3 Digenic analysis

The heterogeneity of the phenotype of patients from the same pedigree suggests that multiple genes might contribute to the disease. We used Fisher exact test to detect the statistical evidence of gene-gene interaction in patients, which was driven by the assumption that the proportion of cases carrying mutations in both of interacting genes is higher than expected in pairs of genes without interaction. This case-only approach combined with the phenotype-genotype correlations revealed the potential digenic inheritance model in our cohort. Remaining unsolved PID cases were submitted to the digenic inheritance model for further analysis, and this cohort-based approach suggested 32 possible causal mutations in 16 patients (5.2% in potential solved patients, 3.1% in all patients) who may follow a digenic inheritance model (unpublished data).

4.3.4 The importance of early molecular diagnosis

Among the patients who underwent TRS, a Swedish female patient who was diagnosed CVID at the age of 18. However, at that time, her condition was not a "typical" case of the disease, and thus no molecular diagnosis was made after the clinical diagnosis. It was not until 2017, when the patient revisited us for medical suggestions due to the progressive symptoms, including Lymphoma and Hemophagocytic Lymphohistiocytosis (HLH). A genetic investigation was launched soon thereafter. Unfortunately, at the time when the genetic result became available, which suggested ITK deficiency, the patient had already passed away due to her aggressive B cell lymphoma (*119*).

We presented this case report in **Paper III**, which adds to the growing amount of evidence supporting the importance of genetic investigations initiated at an early stage of the patient´s disease. If the mutation in our patient had been known, she would have been referred for stem cell transplantation, which may have saved her life, instead of merely being given gammaglobulin substitution. In the conventional clinical setting, the clinical misclassification and inappropriate immunological phenotyping occur, which has hampered targeted therapy. Therefore, it is paramount to employ NGS as an initiatory procedure of the molecular diagnostic approach in dysgammaglobulinemia patients. Additionally, this tool could be used in aiding the therapy of patients and in genetic consulting for patients' family members.

In all, 117 genes and 430 variants have been identified by using a pedigree-based strategy and seven genes and 17 variants by a cohort-based approach. The total genetic diagnostic yield is 60%, which is much higher than previously reported. A number of 204 patients remained

undiagnosed at the molecular level. These findings indicated that NGS genomic approach, together with large sample size and novel strategies, is powerful in decoding genetic characteristics of PID and provide insight into molecular mechanisms that cause the disease, and the utility of exome sequencing in the identification of novel genes underlying primary immunodeficiency.

4.4 RANDOM FOREST-BASED VARIANT IMPACT PREDICTOR

We also developed a Random Forest-based discriminator, VIPPID, to predict the impact of the gene mutation for the PID phenotype. We utilized 85 features scores from SNVBOX, combined with 24 impact scores from existing tools to train the model in either non-gene-specific model or gene-specific manner. A total of 1,664 pathogenic variants in 132 known PID genes obtained from RAPID database and 3,373 non-pathogenic mutations from gnomAD database were used as positive and negative samples in the training dataset, respectively.

The non-gene specific model achieved an AUC value of 0.94, and the gene-specific model achieved a better accuracy with AUC of 0.95, showing considerable superior accuracy over all existing methods (**Figure 8**).

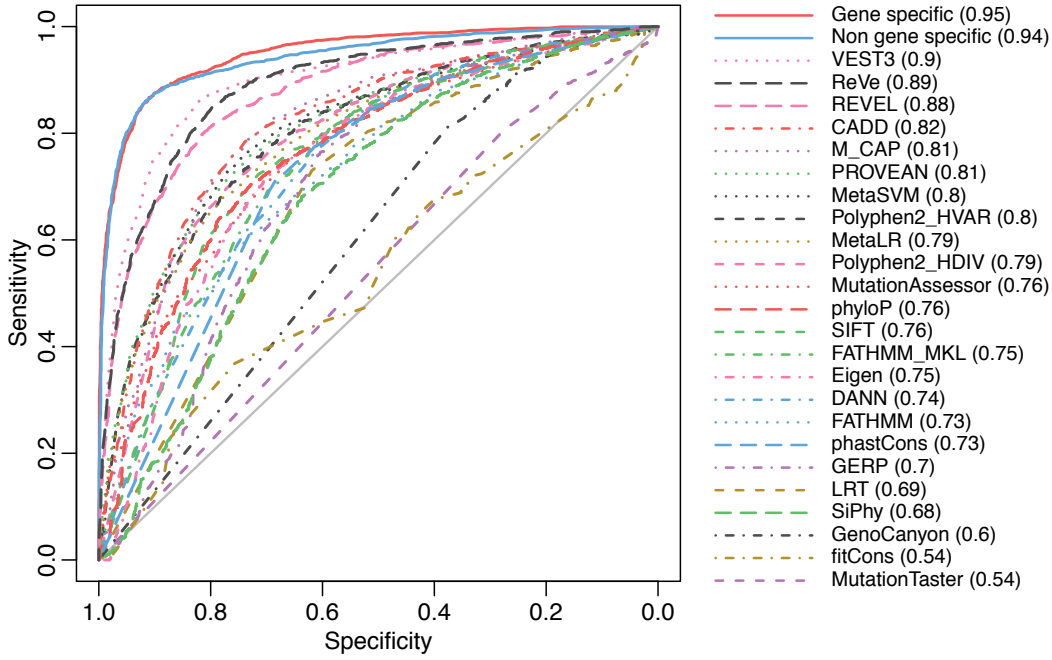

Figure 8. Receiver operating characteristic curve (ROC) represent the predictions from VIPPID model and existing SNV effect prediction tools. Gene-specific VIPPID model (solid red line) and non-gene specific VIPPID model (solid blue line) have superior accuracy compared with existing prediction tools (non-solid lines). Numbers in parentheses indicate AUC value.

Moreover, the machine learning algorithm can calculate an importance score for each feature of different genes, features with high importance scores were those that could best separate pathogenic mutations from non-pathogenic mutations, and they were also more likely to have a high impact on the function of the genes (**Paper IV**).

4.5 IMMUNE REPERTOIRE

Repair of DNA double-strand breaks is important during antigen receptor recombination and creates a diversified repertoire for the developing lymphocytes. Mutations in multiple DNA repairing genes have a variable effect on adaptive immunity, resulting in a broad spectrum of immunological and clinical phenotypes. We investigated 19 patients with monogenic DNA repair defects and performed deep immune repertoire sequencing and bioinformatics analysis. Subsequently, patients were further categorized into 4 groups (atypical T^+ SCID (n=3), AT (n=6), ICF1 (n=6) and ICF2 (n=4)) based on their molecular diagnosis and a comparison was made with 14 age-matched healthy controls to evaluate the roles of identified genetic defects during T cell receptor (TCR) recombination in **Paper V**.

4.5.1 Restriction of TCR correlates with the pathogenic cause and the clinical phenotype

Patients with different molecular diagnoses exhibited distinct repertoire diversity, clonality and V-J pairing patterns. Hypomorphic variants in a component of non-homologous end joining lead to atypical severe combined immunodeficiency affecting variable-joining (V-J) sections paring, and length and amino acids composition of complementarity-determining region 3 (CDR3). Patients with mutant *ATM* and *ZBTB24* genes exhibited restriction of repertoire diversity, decreased clonality, skewed V-J pairing accompanied by aberrant CDR3 length, whereas the patients with *DNMT3B* mutations presented longer CDR3 lengths and a lower percentage of out-of-frame rearrangements with a stop codon. Altered CDR3 length distribution was observed in unproductive and productive TCR in all patients, suggesting that it predominantly arises before thymic selection. Shorter CDR3 lengths in AT patients resulted from a decreased number of insertions, leading to an increase in the number of shared clonotypes, whereas patients with *DNMT3B* and *ZBTB24* mutations presented longer CDR3 lengths and reduced specificity for pathogen-associated CDR3 sequences.

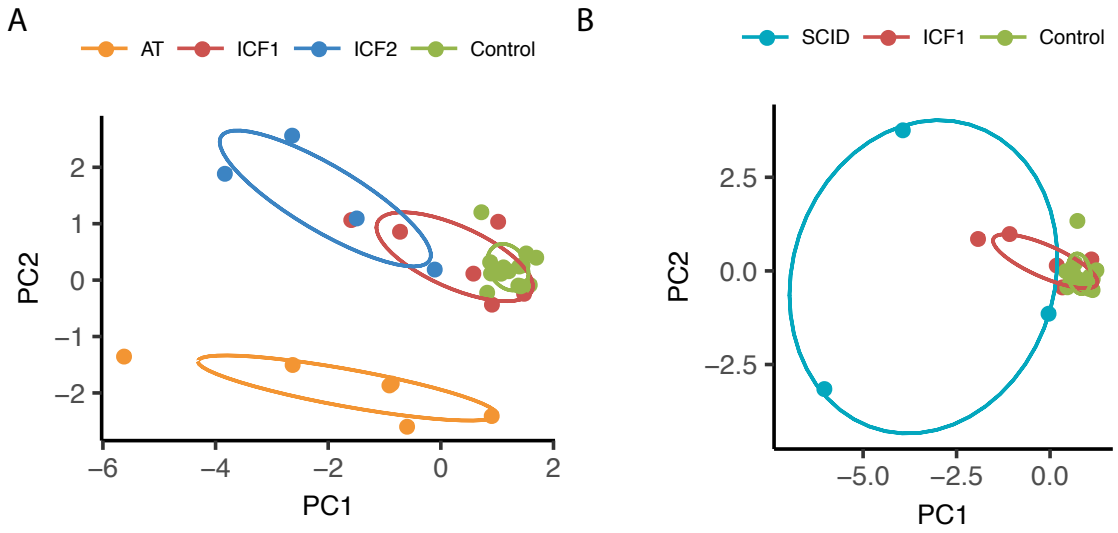

Figure 9. PCA plots illustrated the segregation of the AT and ICF2 patient groups from healthy controls on principal component 1 (PC1) and PC2 based on six variables (Pielou's evenness (TCR), Gini skewing index (TCR), Gini skewing index (V-J paring), the percentage of in-frame rearrangement, mean CDR3 length of out-of-frame rearrangement and the percentage of tyrosine in unique clones).

Principal components analysis (PCA) demonstrated the separation of AT patients, ICF2 patients and healthy donors (green circles). It was based on six variables (Pielou's evenness (TCR), Gini skewing index (TCR), Gini skewing index (V-J paring), the percentage of in-frame rearrangement, mean CDR3 length of out-of-frame rearrangement and the percentage of tyrosines in unique clones) (**Figure 9A**). On the other hand, the ICF1 patients clustered together with the ICF2 group and the healthy individuals, the atypical SCID patients did not gather together in these characteristics, which may be due to the differences in causative genes. Neither the ICF1 group nor the atypical SCID patients could be differentiated from other groups (**Figure 9**).

4.5.2 Discrepant enrichment of pathology associated T cells contributes to the phenotypic heterogeneity of immunodeficiencies

Shorter CDR3 length in ATM-deficient patients resulted from significantly decreased nucleotide insertions, led to an increase in the pathology associated T cells clonotypes. The pathology associated TCRs showed a dramatically higher proportion of short CDR3 length compared to other unknown clonotypes, suggesting that pathology associated TCRs were preferentially shorter.

The proportion of shared pathogen, autoimmunity and cancer-associated clonotypes was dramatically increased among the unique clones in the AT group. In contrast to AT patients, an extremely low frequency of shared pathogens specific clonotypes was recorded in the ICF1 patients, and a similar downward trend was also observed for the proportion of shared common clonotypes in both the ICF1 and ICF2 groups.

Notably, a high frequency of pathogen-specific clonotypes was observed in the AT patients compared with healthy controls which might explain why opportunistic infections are rarely reported in AT patients. The rate of the total shared cancer specific clonotypes in our pediatric AT patients was reduced, perhaps reflected that none of the AT patients in this study had as yet developed cancer.

In summary, the characteristics of the TCR repertoire observed in our study provided novel insights into the role of different components of the DNA repair/methylation machinery in TCR repertoire dysfunction and diverse phenotypes of immunodeficiency. Furthermore, for the first time, it shed light on to the mechanical etiology of T cell dysfunction in clinically similar diseases associated with DNMT3B and ZBTB24 deficiencies.

5 CONCLUSION AND FUTURE PERSPECTIVES

WES/TRS and the bioinformatics used for this study had led to the discovery of the disease-associated genes in over half of the families and cohorts, which was much higher than previously reported. In all, 430 variants in 117 genes were identified using a pedigree-based strategy, and 17 potential disease-causing variants in 7 genes were identified in a cohort-based approach in the patients we studied. Around 200 patients remain undiagnosed at the molecular level.

For the majority of the remaining "unsolved" patients, there were several factors which could have contributed to such an undiagnosed situation. 1) TRS and WES are not able to capture variants located in non-coding regions which could have resulted in the phenotypes (*120*); 2) Structure variations that contribute to PID (*121*) failed to be identified by adopting the existing library construction and data analysis method which was based on gene panels; 3) Pre-existing and complex PID genotype-phenotype relations may confuse our analysis of the heterozygous mutations in a number of known genes that previously reported as under a recessive model or related to haploinsufficiency. 4) Variation in penetrance or existence of modifying genes in PID patients led to the failure of co-segregation and thus impeded the identification of the genetic cause (*122-124*). 5) Furthermore, the limitation of existing mutation pathogenicity prediction tools might also lead to some false negatives in disease mutation identification. The new Random Forest based model we developed was a proof of concept of using gene-specific and disease-specific model can improve the accuracy of pathogenicity prediction for variants in the PID genes.

We should learn from **Paper II** and **Paper III** that early and appropriate molecular diagnosis is of utmost need, and genetic testing is strongly advocated for all CVID patients associated with HLH for taking prompt action and adjust treatment to improve survival. Thus, we will employ a variety of methods on our undiagnosed patients and try to make a molecular diagnosis.

WGS will be used to detect large insertions or deletions, copy number variations or structural variations, which are not easily traceable using WES. Meanwhile, increasing the sample size and integrating rare nonpathogenic genetic variants found in public or private databases, especially for the ethnically diverse populations (*125, 126*), this is capable of improving the analytical power for the novel genes or gene-gene interaction identification.

Other factors may also contribute to the development of the disease, for instance, alterations in the epigenome, transcriptome, proteome or microbiome can influence the disease phenotype (*127*), and we will, therefore, consider using alternative strategies to look for other explanations for the disorders in our patients.

Inspired by the recent findings suggesting that cardiovascular-disease-related protein alteration can be driven by genetics and the gut flora (*128*), we would like to explore the potential interrelationship and the joint effects between the genetics and the microbiome in

IgA deficiency, in the hope of providing new therapeutic paradigms for future applications in personalized medicine. Moreover, given that PID is especially susceptible to be infected by the pathogen, integrating other multi-omics studies could help us to establish a comprehensive understanding about PID, which can be useful in guiding individualized treatment in the clinical setting.

Some new technologies could also be helpful for disease screening or diagnosis, for instance, recently developed single tube long fragment read (stLFR) (*129*), provide possibilities of non-invasive prenatal testing of PIDs by sequencing the maternal plasma DNA. T-cell receptor excision circles (TRECs) and κ-deleting recombination excision circles (KRECs) assay combined with conventional genome sequencing may also enable a direct identification of the genetic cause of PIDs, and this will allow a more rapid and accurate diagnosis and provide clues for therapeutic interventions.

In summary, the application of NGS and other new technologies, along with innovation in research strategies, will facilitate the identification of more disease genes. Integrated knowledge of PID genes or mutations and phenotypes will be helpful to elucidate the genetic causes of the diseases and to simplify the procedure of the genetic diagnostic setting while improving the diagnostic yield.

6 ACKNOWLEDGMENTS

First and foremost, I would like to express my utmost gratitude to my principal supervisor **Prof. Lennart Hammarström** for giving me the opportunity to undertake this Ph.D. journey. Your invaluable guidance, coaching and encouragements throughout my doctoral research are greatly appreciated. I sincerely thank you for your inspirations, motivations and critical reviews on the multiple research projects we conducted. I would like to thank you for all your suggestions on my presentations, funding applications and manuscripts preparations, which helps me tremendously to improve my academic writing skills.

Special thanks to my co-supervisor **Prof. Xiuqing Zhang**. Thank you very much for your support, advice, and the freedom you gave for my research projects.

I am especially grateful for two of my supervisors in BGI-Shenzhen: **Dr. Xiao Liu** and **Dr. Jianguo Zhang**. Thanks for introducing me to the fascinating research area of human genomics, genetics and immune repertoire. Thanks for teaching me how to conduct excellent academic research. Thanks to **Prof. Qiang Pan-Hammarström** for your valuable scientific suggestions.

Thanks to all my co-authors for your time, efforts and valuable comments, without your contributions, my thesis would never have been accomplished. And I would like to express my sincere gratitude to all patients and their family. Without your support, this work would not have been possible.

I want to thank all the group members from Lennart's group and Pan's group (past and present). In particular, I would like to thank **Hassan**, for his assistance in the interpretation of results and writing. To **Kerstin**, **Nina** and **Harold**, I thank you for all of your help and kind suggestions. To **CK**, **Xiaofei**, **Meggie**, **Rosa**, **Yin**, **Qian**, **Du**, **Bo**, **Weichen**, **Mandy** and **Mohammad,** thanks for all your help and encouragement.

I am grateful for all the support and help from to my former and current colleagues in the Immunogenetics Group, Immune and Health Group, Rare Disease Group, and colleagues based in Sweden (**Vivien**, **Yulin** and **Tharshany**). I also would like to especially thank for **Chongyi**, **Bochen**, **Yulan**, **Xiao Dang**, **Lanlan**, **Xue**, **Prof. Tao Li**, **Wei**, **Tao**, **Fengping**, **Wenyan**, **Jinghua**, **Liya**, **Xie**, **Ziyun**, **Ruifang, Lei, Yong Hou** and **Hui Jiang**, for your kind help, support, and encouragement.

Thanks to the division administrators: **Lili**, **Lisa,** and **Moa**, for their kindly administrative supports. To **Prof. Wright**, **Arja**, **Marlene** and **Helen**, thanks for your valuable advice and academic guidance.

I want to take this opportunity to thank my friends. I would like to thank **Hui Wang** for her encouragements and suggestions on the grant application. I would also like to acknowledge

Shanshan for her sharing in creative insights and taking me to try out different cuisines. I also must express my gratitude to **Hui Liu**, who helped me a lot during my time in Sweden and helped me hand in my thesis application. Many thanks to **Wenting** for your supporting, encouragements and sharing your experience of Ph.D. studies. I also appreciate the support from **Qian**, who kindly proofread the thesis.

Last but not least, I would like to express my special gratitude to my beloved family: my husband **Zheng**, thank you for all your love and support. To **our parents**, without your support, my Ph.D. career would never have been completed. Mostly, to my little angel **Kiki** (**Yuluo**), who is the fountain of my power and happiness, your existence is the happiest thing in my life.

7 REFERENCES

1. A. Fischer, Human primary immunodeficiency diseases: a perspective. *Nat Immunol* **5**, 23-30 (2004).
2. A. A. Bousfiha, L. Jeddane, F. Ailal, I. Benhsaien, N. Mahlaoui, J. L. Casanova, L. Abel, Primary immunodeficiency diseases worldwide: more common than generally thought. *Journal of clinical immunology* **33**, 1-7 (2013).
3. C. Picard, H. Bobby Gaspar, W. Al-Herz, A. Bousfiha, J. L. Casanova, T. Chatila, Y. J. Crow, C. Cunningham-Rundles, A. Etzioni, J. L. Franco, S. M. Holland, C. Klein, T. Morio, H. D. Ochs, E. Oksenhendler, J. Puck, M. L. K. Tang, S. G. Tangye, T. R. Torgerson, K. E. Sullivan, International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. *Journal of clinical immunology* **38**, 96-128 (2018).
4. A. Bousfiha, L. Jeddane, C. Picard, F. Ailal, H. Bobby Gaspar, W. Al-Herz, T. Chatila, Y. J. Crow, C. Cunningham-Rundles, A. Etzioni, J. L. Franco, S. M. Holland, C. Klein, T. Morio, H. D. Ochs, E. Oksenhendler, J. Puck, M. L. K. Tang, S. G. Tangye, T. R. Torgerson, J. L. Casanova, K. E. Sullivan, The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. *Journal of clinical immunology* **38**, 129-143 (2018).
5. Y. Itan, J. L. Casanova, Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Frontiers in immunology* **6**, 142 (2015).
6. E. Mamcarz, S. Zhou, T. Lockey, H. Abdelsamed, S. J. Cross, G. Kang, Z. Ma, J. Condori, J. Dowdy, B. Triplett, C. Li, G. Maron, J. C. Aldave Becerra, J. A. Church, E. Dokmeci, J. T. Love, A. C. da Matta Ain, H. van der Watt, X. Tang, W. Janssen, B. Y. Ryu, S. S. De Ravin, M. J. Weiss, B. Youngblood, J. R. Long-Boyle, S. Gottschalk, M. M. Meagher, H. L. Malech, J. M. Puck, M. J. Cowan, B. P. Sorrentino, Lentiviral Gene Therapy Combined with Low-Dose Busulfan in Infants with SCID-X1. *The New England journal of medicine* **380**, 1525-1534 (2019).
7. N. Principi, S. Esposito, Vaccine use in primary immunodeficiency disorders. *Vaccine* **32**, 3725-3731 (2014).
8. S. Borte, U. von Dobeln, L. Hammarstrom, Guidelines for newborn screening of primary immunodeficiency diseases. *Current opinion in hematology* **20**, 48-54 (2013).
9. S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, M. J. Bamshad, Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-35 (2010).
10. K. A. Wetterstrand, DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) *Available at:* www.genome.gov/sequencingcostsdata., (2019).
11. J. Shendure, G. M. Findlay, M. W. Snyder, Genomic Medicine–Progress, Pitfalls, and Promise. *Cell* **177**, 45-57 (2019).
12. L. D. Notarangelo, Primary immunodeficiencies. *The Journal of allergy and clinical immunology* **125**, S182-194 (2010).
13. O. K. Alkhairy, N. Rezaei, R. R. Graham, H. Abolhassani, S. Borte, K. Hultenby, C. Wu, A. Aghamohammadi, D. A. Williams, T. W. Behrens, L. Hammarstrom, Q. Pan-Hammarstrom, RAC2 loss-of-function mutation in 2 siblings with characteristics of

common variable immunodeficiency. *The Journal of allergy and clinical immunology* **135**, 1380-1384.e1381-1385 (2015).

14. H. Abolhassani, T. Cheraghi, N. Rezaei, A. Aghamohammadi, L. Hammarstrom, Common Variable Immunodeficiency or Late-Onset Combined Immunodeficiency: A New Hypomorphic JAK3 Patient and Review of the Literature. *Journal of investigational allergology & clinical immunology* **25**, 218-220 (2015).

15. H. Abolhassani, E. S. Edwards, A. Ikinciogullari, H. Jing, S. Borte, M. Buggert, L. Du, M. Matsuda-Lennikov, R. Romano, R. Caridha, S. Bade, Y. Zhang, J. Frederiksen, M. Fang, S. K. Bal, S. Haskologlu, F. Dogu, N. Tacyildiz, H. F. Matthews, J. J. McElwee, E. Gostick, D. A. Price, U. Palendira, A. Aghamohammadi, B. Boisson, N. Rezaei, A. C. Karlsson, M. J. Lenardo, J. L. Casanova, L. Hammarstrom, S. G. Tangye, H. C. Su, Q. Pan-Hammarstrom, Combined immunodeficiency and Epstein-Barr virus-induced B cell malignancy in humans with inherited CD70 deficiency. *The Journal of experimental medicine* **214**, 91-106 (2017).

16. O. K. Alkhairy, H. Abolhassani, N. Rezaei, M. Fang, K. K. Andersen, Z. Chavoshzadeh, I. Mohammadzadeh, M. A. El-Rajab, M. Massaad, J. Chou, A. Aghamohammadi, R. S. Geha, L. Hammarstrom, Spectrum of Phenotypes Associated with Mutations in LRBA. *Journal of clinical immunology* **36**, 33-45 (2016).

17. O. K. Alkhairy, R. Perez-Becker, G. J. Driessen, H. Abolhassani, J. van Montfrans, S. Borte, S. Choo, N. Wang, K. Tesselaar, M. Fang, K. Bienemann, K. Boztug, A. Daneva, F. Mechinaud, T. Wiesel, C. Becker, G. Duckers, K. Siepermann, M. C. van Zelm, N. Rezaei, M. van der Burg, A. Aghamohammadi, M. G. Seidel, T. Niehues, L. Hammarstrom, Novel mutations in TNFRSF7/CD27: Clinical, immunologic, and genetic characterization of human CD27 deficiency. *The Journal of allergy and clinical immunology* **136**, 703-712.e710 (2015).

18. H. Abolhassani, N. Wang, A. Aghamohammadi, N. Rezaei, Y. N. Lee, F. Frugoni, L. D. Notarangelo, Q. Pan-Hammarstrom, L. Hammarstrom, A hypomorphic recombination-activating gene 1 (RAG1) mutation resulting in a phenotype resembling common variable immunodeficiency. *The Journal of allergy and clinical immunology* **134**, 1375-1380 (2014).

19. S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, J. Shendure, Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

20. J. L. Wang, X. Yang, K. Xia, Z. M. Hu, L. Weng, X. Jin, H. Jiang, P. Zhang, L. Shen, J. F. Guo, N. Li, Y. R. Li, L. F. Lei, J. Zhou, J. Du, Y. F. Zhou, Q. Pan, J. Wang, J. Wang, R. Q. Li, B. S. Tang, TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* **133**, 3510-3518 (2010).

21. S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, J. Shendure, Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-793 (2010).

22. K. Musunuru, J. P. Pirruccello, R. Do, G. M. Peloso, C. Guiducci, C. Sougnez, K. V. Garimella, S. Fisher, J. Abreu, A. J. Barry, T. Fennell, E. Banks, L. Ambrogio, K. Cibulskis, A. Kernytsky, E. Gonzalez, N. Rudzicz, J. C. Engert, M. A. DePristo, M. J. Daly, J. C. Cohen, H. H. Hobbs, D. Altshuler, G. Schonfeld, S. B. Gabriel, P. Yue, S. Kathiresan, Exome sequencing, ANGPTL3 mutations, and familial combined

hypolipidemia. *The New England journal of medicine* **363**, 2220-2227 (2010).

23. N. Pillar, O. Pleniceanu, M. Fang, L. Ziv, E. Lahav, S. Botchan, L. Cheng, B. Dekel, N. Shomron, A rare variant in the FHL1 gene associated with X-linked recessive hypoparathyroidism. *Human genetics* **136**, 835-845 (2017).

24. R. F. Schindler, C. Scotton, J. Zhang, C. Passarelli, B. Ortiz-Bonnin, S. Simrick, T. Schwerte, K. L. Poon, M. Fang, S. Rinne, A. Froese, V. O. Nikolaev, C. Grunert, T. Muller, G. Tasca, P. Sarathchandra, F. Drago, B. Dallapiccola, C. Rapezzi, E. Arbustini, F. R. Di Raimo, M. Neri, R. Selvatici, F. Gualandi, F. Fattori, A. Pietrangelo, W. Li, H. Jiang, X. Xu, E. Bertini, N. Decher, J. Wang, T. Brand, A. Ferlini, POPDC1(S201F) causes muscular dystrophy and arrhythmia by affecting protein trafficking. *The Journal of clinical investigation* **126**, 239-253 (2016).

25. S. Olgiati, M. Quadri, M. Fang, J. P. Rood, J. A. Saute, H. F. Chien, C. G. Bouwkamp, J. Graafland, M. Minneboo, G. J. Breedveld, J. Zhang, F. W. Verheijen, A. J. Boon, A. J. Kievit, L. B. Jardim, W. Mandemakers, E. R. Barbosa, C. R. Rieder, K. L. Leenders, J. Wang, V. Bonifati, DNAJC6 Mutations Associated With Early-Onset Parkinson's Disease. *Annals of neurology* **79**, 244-256 (2016).

26. E. Gregianin, G. Pallafacchina, S. Zanin, V. Crippa, P. Rusmini, A. Poletti, M. Fang, Z. Li, L. Diano, A. Petrucci, L. Lispi, T. Cavallaro, G. M. Fabrizi, M. Muglia, F. Boaretto, A. Vettori, R. Rizzuto, M. L. Mostacciuolo, G. Vazza, Loss-of-function mutations in the SIGMAR1 gene cause distal hereditary motor neuropathy by impairing ER-mitochondria tethering and Ca2+ signalling. *Human molecular genetics* **25**, 3741-3753 (2016).

27. C. Dallabona, T. E. Abbink, R. Carrozzo, A. Torraco, A. Legati, C. G. van Berkel, M. Niceta, T. Langella, D. Verrigni, T. Rizza, D. Diodato, F. Piemonte, E. Lamantea, M. Fang, J. Zhang, D. Martinelli, E. Bevivino, C. Dionisi-Vici, A. Vanderver, S. G. Philip, M. A. Kurian, I. C. Verma, S. Bijarnia-Mahay, S. Jacinto, F. Furtado, P. Accorsi, A. Ardissone, I. Moroni, I. Ferrero, M. Tartaglia, P. Goffrini, D. Ghezzi, M. S. van der Knaap, E. Bertini, LYRM7 mutations cause a multifocal cavitating leukoencephalopathy with distinct MRI appearance. *Brain* **139**, 782-794 (2016).

28. A. Reyes, L. Melchionda, A. Nasca, F. Carrara, E. Lamantea, A. Zanolini, C. Lamperti, M. Fang, J. Zhang, D. Ronchi, S. Bonato, G. Fagiolari, M. Moggio, D. Ghezzi, M. Zeviani, RNASEH1 Mutations Impair mtDNA Replication and Cause Adult-Onset Mitochondrial Encephalomyopathy. *American journal of human genetics* **97**, 186-193 (2015).

29. A. Masotti, P. Uva, L. Davis-Keppen, L. Basel-Vanagaite, L. Cohen, E. Pisaneschi, A. Celluzzi, P. Bencivenga, M. Fang, M. Tian, X. Xu, M. Cappa, B. Dallapiccola, Keppen-Lubinsky syndrome is caused by mutations in the inwardly rectifying K+ channel encoded by KCNJ6. *American journal of human genetics* **96**, 295-300 (2015).

30. A. Belkadi, A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko, L. Shang, B. Boisson, J. L. Casanova, L. Abel, Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**, 5473-5478 (2015).

31. C. A. Steward, A. P. J. Parker, B. A. Minassian, S. M. Sisodiya, A. Frankish, J. Harrow, Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome medicine* **9**, 49 (2017).

32. X. Li, Y. Kim, E. K. Tsang, J. R. Davis, F. N. Damani, C. Chiang, G. T. Hess, Z. Zappala, B. J. Strober, A. J. Scott, A. Li, A. Ganna, M. C. Bassik, J. D. Merker, I. M. Hall, A. Battle, S. B. Montgomery, The impact of rare variation on gene expression

across tissues. *Nature* **550**, 239-243 (2017).

33. A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

34. L. J. Carithers, H. M. Moore, The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and biobanking* **13**, 307-308 (2015).

35. I. J. Nijman, J. M. van Montfrans, M. Hoogstraat, M. L. Boes, L. van de Corput, E. D. Renner, P. van Zon, S. van Lieshout, M. G. Elferink, M. van der Burg, C. L. Vermont, B. van der Zwaag, E. Janson, E. Cuppen, J. K. Ploos van Amstel, M. E. van Gijn, Targeted next-generation sequencing: a novel diagnostic tool for primary immunodeficiencies. *The Journal of allergy and clinical immunology* **133**, 529-534 (2014).

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

37. R. Li, Y. Li, K. Kristiansen, J. Wang, SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714 (2008).

38. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

39. R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, J. Wang, SNP detection for massively parallel whole-genome resequencing. *Genome research* **19**, 1124-1132 (2009).

40. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122 (2016).

41. P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012).

42. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

43. M. Yandell, C. Huff, H. Hu, M. Singleton, B. Moore, J. Xing, L. B. Jorde, M. G. Reese, A probabilistic disease-gene finder for personal genomes. *Genome research* **21**, 1529-1542 (2011).

44. M. Fang, H. Abolhassani, C. K. Lim, J. Zhang, L. Hammarstrom, Next Generation Sequencing Data Analysis in Primary Immunodeficiency Disorders - Future Directions. *Journal of clinical immunology* **36 Suppl 1**, 68-75 (2016).

45. Y. Itan, L. Shang, B. Boisson, M. J. Ciancanelli, J. G. Markle, R. Martinez-Barricarte, E. Scott, I. Shah, P. D. Stenson, J. Gleeson, D. N. Cooper, L. Quintana-Murci, S. Y. Zhang, L. Abel, J. L. Casanova, The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature methods* **13**, 109-110 (2016).

46. Y. Itan, L. Shang, B. Boisson, E. Patin, A. Bolze, M. Moncada-Velez, E. Scott, M. J. Ciancanelli, F. G. Lafaille, J. G. Markle, R. Martinez-Barricarte, S. J. de Jong, X. F. Kong, P. Nitschke, A. Belkadi, J. Bustamante, A. Puel, S. Boisson-Dupuis, P. D. Stenson, J. G. Gleeson, D. N. Cooper, L. Quintana-Murci, J. M. Claverie, S. Y. Zhang, L. Abel, J. L. Casanova, The human gene damage index as a gene-level approach to

prioritizing exome variants. *Proc Natl Acad Sci U S A* **112**, 13615-13620 (2015).

47. D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, R. K. Wilson, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576 (2012).

48. A. V. Dalca, M. Brudno, Genome variation discovery with high-throughput sequencing data. *Briefings in bioinformatics* **11**, 3-14 (2010).

49. P. Medvedev, M. Stanciu, M. Brudno, Computational methods for discovering structural variation with next-generation sequencing. *Nature methods* **6**, S13-20 (2009).

50. A. Valsesia, A. Mace, S. Jacquemont, J. S. Beckmann, Z. Kutalik, The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Frontiers in genetics* **4**, 92 (2013).

51. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, E. E. Eichler, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646 (2010).

52. R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).

53. G. Klambauer, K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker, U. Bodenhofer, S. Hochreiter, cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research* **40**, e69 (2012).

54. M. Y. Hwang, S. Moon, L. Heo, Y. J. Kim, J. H. Oh, Y. J. Kim, Y. K. Kim, J. Lee, B. G. Han, B. J. Kim, Combinatorial approach to estimate copy number genotype using whole-exome sequencing data. *Genomics* **105**, 145-149 (2015).

55. N. Krumm, P. H. Sudmant, A. Ko, B. J. O'Roak, M. Malig, B. P. Coe, A. R. Quinlan, D. A. Nickerson, E. E. Eichler, Copy number variation detection and genotyping from exome sequence data. *Genome research* **22**, 1525-1532 (2012).

56. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, A. L. Q. A. Committee, 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. *Genetics in medicine* **17**, 405-424 (2015).

57. H. Houlleberghs, M. Dekker, H. Lantermans, R. Kleinendorst, H. J. Dubbink, R. M. Hofstra, S. Verhoef, H. Te Riele, Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. *Proc Natl Acad Sci U S A* **113**, 4128-4133 (2016).

58. G. M. Findlay, R. M. Daza, B. Martin, M. D. Zhang, A. P. Leith, M. Gasperini, J. D. Janizek, X. Huang, L. M. Starita, J. Shendure, Accurate classification of BRCA1

variants with saturation genome editing. *Nature* **562**, 217-222 (2018).

59. V. P. Mauro, S. A. Chappell, A critical analysis of codon optimization in human therapeutics. *Trends Mol Med* **20**, 604-613 (2014).

60. A. Stray-Pedersen, H. S. Sorte, P. Samarakoon, T. Gambin, I. K. Chinn, Z. H. Coban Akdemir, H. C. Erichsen, L. R. Forbes, S. Gu, B. Yuan, S. N. Jhangiani, D. M. Muzny, O. K. Rodningen, Y. Sheng, S. K. Nicholas, L. M. Noroski, F. O. Seeborg, C. M. Davis, D. L. Canter, E. M. Mace, T. J. Vece, C. E. Allen, H. A. Abhyankar, P. M. Boone, C. R. Beck, W. Wiszniewski, B. Fevang, P. Aukrust, G. E. Tjonnfjord, T. Gedde-Dahl, H. Hjorth-Hansen, I. Dybedal, I. Nordoy, S. F. Jorgensen, T. G. Abrahamsen, T. Overland, A. G. Bechensteen, V. Skogen, L. T. Osnes, M. A. Kulseth, T. E. Prescott, C. F. Rustad, K. R. Heimdal, J. W. Belmont, N. L. Rider, J. Chinen, T. N. Cao, E. A. Smith, M. S. Caldirola, L. Bezrodnik, S. O. Lugo Reyes, F. J. Espinosa Rosales, N. D. Guerrero-Cursaru, L. A. Pedroza, C. M. Poli, J. L. Franco, C. M. Trujillo Vargas, J. C. Aldave Becerra, N. Wright, T. B. Issekutz, A. C. Issekutz, J. Abbott, J. W. Caldwell, D. K. Bayer, A. Y. Chan, A. Aiuti, C. Cancrini, E. Holmberg, C. West, M. Burstedt, E. Karaca, G. Yesil, H. Artac, Y. Bayram, M. M. Atik, M. K. Eldomery, M. S. Ehlayel, S. Jolles, B. Flato, A. A. Bertuch, I. C. Hanson, V. W. Zhang, L. J. Wong, J. Hu, M. Walkiewicz, Y. Yang, C. M. Eng, E. Boerwinkle, R. A. Gibbs, W. T. Shearer, R. Lyle, J. S. Orange, J. R. Lupski, Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders. *The Journal of allergy and clinical immunology* **139**, 232-245 (2017).

61. K. Zhang, S. Chandrakasan, H. Chapman, C. A. Valencia, A. Husami, D. Kissell, J. A. Johnson, A. H. Filipovich, Synergistic defects of different molecules in the cytotoxic pathway lead to clinical familial hemophagocytic lymphohistiocytosis. *Blood* **124**, 1331-1334 (2014).

62. M. Germeshausen, C. Zeidler, M. Stuhrmann, M. Lanciotti, M. Ballmaier, K. Welte, Digenic mutations in severe congenital neutropenia. *Haematologica* **95**, 1207-1210 (2010).

63. L. Gao, X. Dang, L. Huang, L. Zhu, M. Fang, J. Zhang, X. Xu, L. Zhu, T. Li, L. Zhao, J. Wei, J. Zhou, Search for the potential "second-hit" mechanism underlying the onset of familial hemophagocytic lymphohistiocytosis type 2 by whole-exome sequencing analysis. *Translational research* **170**, 26-39 (2016).

64. D. Schubert, C. Bode, R. Kenefeck, T. Z. Hou, J. B. Wing, A. Kennedy, A. Bulashevska, B. S. Petersen, A. A. Schaffer, B. A. Gruning, S. Unger, N. Frede, U. Baumann, T. Witte, R. E. Schmidt, G. Dueckers, T. Niehues, S. Seneviratne, M. Kanariou, C. Speckmann, S. Ehl, A. Rensing-Ehl, K. Warnatz, M. Rakhmanov, R. Thimme, P. Hasselblatt, F. Emmerich, T. Cathomen, R. Backofen, P. Fisch, M. Seidl, A. May, A. Schmitt-Graeff, S. Ikemizu, U. Salzer, A. Franke, S. Sakaguchi, L. S. Walker, D. M. Sansom, B. Grimbacher, Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. *Nature medicine* **20**, 1410-1416 (2014).

65. H. S. Kuehn, W. Ouyang, B. Lo, E. K. Deenick, J. E. Niemela, D. T. Avery, J. N. Schickel, D. Q. Tran, J. Stoddard, Y. Zhang, D. M. Frucht, B. Dumitriu, P. Scheinberg, L. R. Folio, C. A. Frein, S. Price, C. Koh, T. Heller, C. M. Seroogy, A. Huttenlocher, V. K. Rao, H. C. Su, D. Kleiner, L. D. Notarangelo, Y. Rampertaap, K. N. Olivier, J. McElwee, J. Hughes, S. Pittaluga, J. B. Oliveira, E. Meffre, T. A. Fleisher, S. M. Holland, M. J. Lenardo, S. G. Tangye, G. Uzel, Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. *Science* **345**, 1623-1627 (2014).

66. B. S. Boland, C. E. Widjaja, A. Banno, B. Zhang, S. H. Kim, S. Stoven, M. R. Peterson,

M. C. Jones, H. I. Su, S. E. Crowe, J. D. Bui, S. B. Ho, Y. Okugawa, A. Goel, E. V. Marietta, M. Khosroheidari, K. Jepsen, J. Aramburu, C. Lopez-Rodriguez, W. J. Sandborn, J. A. Murray, O. Harismendy, J. T. Chang, Immunodeficiency and autoimmune enterocolopathy linked to NFAT5 haploinsufficiency. *Journal of immunology* **194**, 2551-2560 (2015).

67. H. S. Kuehn, B. Boisson, C. Cunningham-Rundles, J. Reichenbach, A. Stray-Pedersen, E. W. Gelfand, P. Maffucci, K. R. Pierce, J. K. Abbott, K. V. Voelkerding, S. T. South, N. H. Augustine, J. S. Bush, W. K. Dolen, B. B. Wray, Y. Itan, A. Cobat, H. S. Sorte, S. Ganesan, S. Prader, T. B. Martins, M. G. Lawrence, J. S. Orange, K. R. Calvo, J. E. Niemela, J. L. Casanova, T. A. Fleisher, H. R. Hill, A. Kumanovics, M. E. Conley, S. D. Rosenzweig, Loss of B Cells in Patients with Heterozygous Mutations in IKAROS. *The New England journal of medicine* **374**, 1032-1043 (2016).

68. A. P. Hsu, L. J. McReynolds, S. M. Holland, GATA2 deficiency. *Current opinion in allergy and clinical immunology* **15**, 104-109 (2015).

69. R. Colobran, R. Pujol-Borrell, M. Hernandez-Gonzalez, M. Guilarte, A novel splice site mutation in the SERPING1 gene leads to haploinsufficiency by complete degradation of the mutant allele mRNA in a case of familial hereditary angioedema. *Journal of clinical immunology* **34**, 521-523 (2014).

70. M. T. Wong, E. H. Scholvinck, A. J. Lambeck, C. M. van Ravenswaaij-Arts, CHARGE syndrome: a review of the immunological aspects. *European journal of human genetics* **23**, 1451-1459 (2015).

71. M. G. de Bielke, L. Perez, J. Yancoski, J. B. Oliveira, S. Danielian, FAS Haploinsufficiency Caused by Extracellular Missense Mutations Underlying Autoimmune Lymphoproliferative Syndrome. *Journal of clinical immunology* **35**, 769-776 (2015).

72. X. F. Kong, G. Vogt, Y. Itan, A. Macura-Biegun, A. Szaflarska, D. Kowalczyk, A. Chapgier, A. Abhyankar, D. Furthner, C. Djambas Khayat, S. Okada, V. L. Bryant, D. Bogunovic, A. Kreins, M. Moncada-Velez, M. Migaud, S. Al-Ajaji, S. Al-Muhsen, S. M. Holland, L. Abel, C. Picard, D. Chaussabel, J. Bustamante, J. L. Casanova, S. Boisson-Dupuis, Haploinsufficiency at the human IFNGR2 locus contributes to mycobacterial disease. *Human molecular genetics* **22**, 769-781 (2013).

73. R. Rachid, E. Castigli, R. S. Geha, F. A. Bonilla, TACI mutation in common variable immunodeficiency and IgA deficiency. *Current allergy and asthma reports* **6**, 357-362 (2006).

74. M. Fliegauf, V. L. Bryant, N. Frede, C. Slade, S. T. Woon, K. Lehnert, S. Winzer, A. Bulashevska, T. Scerri, E. Leung, A. Jordan, B. Keller, E. de Vries, H. Cao, F. Yang, A. A. Schaffer, K. Warnatz, P. Browett, J. Douglass, R. V. Ameratunga, J. W. van der Meer, B. Grimbacher, Haploinsufficiency of the NF-kappaB1 Subunit p50 in Common Variable Immunodeficiency. *American journal of human genetics* **97**, 389-403 (2015).

75. K. Chen, E. M. Coonrod, A. Kumanovics, Z. F. Franks, J. D. Durtschi, R. L. Margraf, W. Wu, N. M. Heikal, N. H. Augustine, P. G. Ridge, H. R. Hill, L. B. Jorde, A. S. Weyrich, G. A. Zimmerman, A. V. Gundlapalli, J. F. Bohnsack, K. V. Voelkerding, Germline mutations in NFKB2 implicate the noncanonical NF-kappaB pathway in the pathogenesis of common variable immunodeficiency. *American journal of human genetics* **93**, 812-824 (2013).

76. V. C. Rodriguez-Cortez, H. Hernando, L. de la Rica, R. Vento, E. Ballestar, Epigenomic deregulation in the immune system. *Epigenomics* **3**, 697-713 (2011).

77. J. C. Knight, Genomic modulators of the immune response. *Trends in genetics* **29**, 74-83 (2013).

78. V. C. Rodriguez-Cortez, L. Del Pino-Molina, J. Rodriguez-Ubreva, L. Ciudad, D. Gomez-Cabrero, C. Company, J. M. Urquiza, J. Tegner, C. Rodriguez-Gallego, E. Lopez-Granados, E. Ballestar, Monozygotic twins discordant for common variable immunodeficiency reveal impaired DNA demethylation during naive-to-memory B-cell transition. *Nature communications* **6**, 7335 (2015).

79. H. Zan, P. Casali, Editorial: Epigenetics of B Cells and Antibody Responses. *Frontiers in immunology* **6**, 656 (2015).

80. S. R. Pulivarthy, M. Lion, G. Kuzu, A. G. Matthews, M. L. Borowsky, J. Morris, R. E. Kingston, J. H. Dennis, M. Y. Tolstorukov, M. A. Oettinger, Regulated large-scale nucleosome density patterns and precise nucleosome positioning correlate with V(D)J recombination. *Proc Natl Acad Sci U S A* **113**, E6427-E6436 (2016).

81. S. Gatto, M. Gagliardi, M. Franzese, S. Leppert, M. Papa, M. Cammisa, G. Grillo, G. Velasco, C. Francastel, S. Toubiana, M. D'Esposito, C. Angelini, M. R. Matarazzo, ICF-specific DNMT3B dysfunction interferes with intragenic regulation of mRNA transcription and alternative splicing. *Nucleic acids research* **45**, 5739-5756 (2017).

82. K. Huang, Z. Wu, Z. Liu, G. Hu, J. Yu, K. H. Chang, K. P. Kim, T. Le, K. F. Faull, N. Rao, A. Gennery, Z. Xue, C. Y. Wang, M. Pellegrini, G. Fan, Selective demethylation and altered gene expression are associated with ICF syndrome in human-induced pluripotent stem cells and mesenchymal stem cells. *Human molecular genetics* **23**, 6448-6457 (2014).

83. D. Jiang, Y. Zhang, R. P. Hart, J. Chen, K. Herrup, J. Li, Alteration in 5-hydroxymethylcytosine-mediated epigenetic regulation leads to Purkinje cell vulnerability in ATM deficiency. *Brain* **138**, 3520-3536 (2015).

84. L. S. Kremer, D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, E. Konarikova, B. Repp, G. Kastenmuller, J. Adamski, P. Lichtner, C. Leonhardt, B. Funalot, A. Donati, V. Tiranti, A. Lombes, C. Jardel, D. Glaser, R. W. Taylor, D. Ghezzi, J. A. Mayr, A. Rotig, P. Freisinger, F. Distelmaier, T. M. Strom, T. Meitinger, J. Gagneur, H. Prokisch, Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature communications* **8**, 15824 (2017).

85. A. Latosinska, M. Frantzi, A. Vlahou, A. S. Merseburger, H. Mischak, Clinical Proteomics for Precision Medicine: The Bladder Cancer Case. *Proteomics. Clinical applications* **12**, (2018).

86. X. Li, W. Wang, J. Chen, Recent progress in mass spectrometry proteomics for biomedical research. *Science China. Life sciences* **60**, 1093-1113 (2017).

87. K. Berer, L. A. Gerdes, E. Cekanaviciute, X. Jia, L. Xiao, Z. Xia, C. Liu, L. Klotz, U. Stauffer, S. E. Baranzini, T. Kumpfel, R. Hohlfeld, G. Krishnamoorthy, H. Wekerle, Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice. *Proc Natl Acad Sci U S A* **114**, 10719-10724 (2017).

88. E. Cekanaviciute, B. B. Yoo, T. F. Runia, J. W. Debelius, S. Singh, C. A. Nelson, R. Kanner, Y. Bencosme, Y. K. Lee, S. L. Hauser, E. Crabtree-Hartman, I. K. Sand, M. Gacias, Y. Zhu, P. Casaccia, B. A. C. Cree, R. Knight, S. K. Mazmanian, S. E. Baranzini, Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. *Proc Natl Acad Sci U S A* **114**, 10713-10718 (2017).

89. D. Zhong, C. Wu, X. Zeng, Q. Wang, The role of gut microbiota in the pathogenesis

of rheumatic diseases. *Clinical rheumatology* **37**, 25-34 (2018).

90. J. Rowin, Y. Xia, B. Jung, J. Sun, Gut inflammation and dysbiosis in human motor neuron disease. *Physiological reports* **5**, (2017).

91. Z. Jie, H. Xia, S. L. Zhong, Q. Feng, S. Li, S. Liang, H. Zhong, Z. Liu, Y. Gao, H. Zhao, D. Zhang, Z. Su, Z. Fang, Z. Lan, J. Li, L. Xiao, J. Li, R. Li, X. Li, F. Li, H. Ren, Y. Huang, Y. Peng, G. Li, B. Wen, B. Dong, J. Y. Chen, Q. S. Geng, Z. W. Zhang, H. Yang, J. Wang, J. Wang, X. Zhang, L. Madsen, S. Brix, G. Ning, X. Xu, X. Liu, Y. Hou, H. Jia, K. He, K. Kristiansen, The gut microbiome in atherosclerotic cardiovascular disease. *Nature communications* **8**, 845 (2017).

92. J. Cabrera-Perez, J. C. Babcock, T. Dileepan, K. A. Murphy, T. A. Kucaba, V. P. Badovinac, T. S. Griffith, Gut Microbial Membership Modulates CD4 T Cell Reconstitution and Function after Sepsis. *Journal of immunology* **197**, 1692-1698 (2016).

93. A. Ray, B. N. Dittel, Interrelatedness between dysbiosis in the gut microbiota due to immunodeficiency and disease penetrance of colitis. *Immunology* **146**, 359-368 (2015).

94. R. Rigoni, E. Fontana, S. Guglielmetti, B. Fosso, A. M. D'Erchia, V. Maina, V. Taverniti, M. C. Castiello, S. Mantero, G. Pacchiana, S. Musio, R. Pedotti, C. Selmi, J. R. Mora, G. Pesole, P. Vezzoni, P. L. Poliani, F. Grassi, A. Villa, B. Cassani, Intestinal microbiota sustains inflammation and autoimmunity induced by hypomorphic RAG defects. *The Journal of experimental medicine* **213**, 355-375 (2016).

95. W. Zhang, Y. Du, Z. Su, C. Wang, X. Zeng, R. Zhang, X. Hong, C. Nie, J. Wu, H. Cao, X. Xu, X. Liu, IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics* **201**, 459-472 (2015).

96. I. J. H, J. Rozmus, K. Schwarz, R. L. Warren, D. van Zessen, R. A. Holt, I. Pico-Knijnenburg, E. Simons, I. Jerchel, A. Wawer, M. Lorenz, T. Patiroglu, H. H. Akar, R. Leite, N. S. Verkaik, A. P. Stubbs, D. C. van Gent, J. J. van Dongen, M. van der Burg, XLF deficiency results in reduced N-nucleotide addition during V(D)J recombination. *Blood* **128**, 650-659 (2016).

97. G. K. Wong, J. M. Heather, S. Barmettler, M. Cobbold, Immune dysregulation in immunodeficiency disorders: The role of T-cell receptor sequencing. *J Autoimmun* **80**, 1-9 (2017).

98. S. P. Jackson, Sensing and repairing DNA double-strand breaks. *Carcinogenesis* **23**, 687-696 (2002).

99. E. R. Phillips, P. J. McKinnon, DNA double-strand break repair and development. *Oncogene* **26**, 7799-7808 (2007).

100. H. D. Kondilis-Mangum, P. A. Wade, Epigenetics and the adaptive immune response. *Mol Aspects Med* **34**, 813-825 (2013).

101. A. H. Moarefi, F. Chedin, ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3B-mediated de novo DNA methylation. *J Mol Biol* **409**, 758-772 (2011).

102. B. Jin, Q. Tao, J. Peng, H. M. Soo, W. Wu, J. Ying, C. R. Fields, A. L. Delmas, X. Liu, J. Qiu, K. D. Robertson, DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Human molecular genetics* **17**, 690-709 (2008).

103. M. M. Hagleitner, A. Lankester, P. Maraschio, M. Hulten, J. P. Fryns, C. Schuetz, G. Gimelli, E. G. Davies, A. Gennery, B. H. Belohradsky, R. de Groot, E. J. Gerritsen, T. Mattina, P. J. Howard, A. Fasth, I. Reisli, D. Furthner, M. A. Slatter, A. J. Cant, G. Cazzola, P. J. van Dijken, M. van Deuren, J. C. de Greef, S. M. van der Maarel, C. M. Weemaes, Clinical spectrum of immunodeficiency, centromeric instability and facial dysmorphism (ICF syndrome). *J Med Genet* **45**, 93-99 (2008).

104. G. L. Xu, T. H. Bestor, D. Bourc'his, C. L. Hsieh, N. Tommerup, M. Bugge, M. Hulten, X. Qu, J. J. Russo, E. Viegas-Pequignot, Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187-191 (1999).

105. M. Okano, D. W. Bell, D. A. Haber, E. Li, DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

106. E. Chouery, J. Abou-Ghoch, S. Corbani, N. El Ali, R. Korban, N. Salem, C. Castro, S. Klayme, M. Azoury-Abou Rjeily, R. Khoury-Matar, G. Debo, M. Germanos-Haddad, V. Delague, G. Lefranc, A. Megarbane, A novel deletion in ZBTB24 in a Lebanese family with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2. *Clinical genetics* **82**, 489-493 (2012).

107. J. C. de Greef, J. Wang, J. Balog, J. T. den Dunnen, R. R. Frants, K. R. Straasheijm, C. Aytekin, M. van der Burg, L. Duprez, A. Ferster, A. R. Gennery, G. Gimelli, I. Reisli, C. Schuetz, A. Schulz, D. Smeets, Y. Sznajer, C. Wijmenga, M. C. van Eggermond, M. M. van Ostaijen-Ten Dam, A. C. Lankester, M. J. D. van Tol, P. J. van den Elsen, C. M. Weemaes, S. M. van der Maarel, Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2. *American journal of human genetics* **88**, 796-804 (2011).

108. Y. N. Lee, F. Frugoni, K. Dobbs, I. Tirosh, L. Du, F. A. Ververs, H. Ru, L. Ott de Bruin, M. Adeli, J. H. Bleesing, D. Buchbinder, M. J. Butte, C. Cancrini, K. Chen, S. Choo, R. A. Elfeky, A. Finocchi, R. L. Fuleihan, A. R. Gennery, D. H. El-Ghoneimy, L. A. Henderson, W. Al-Herz, E. Hossny, R. P. Nelson, S. Y. Pai, N. C. Patel, S. M. Reda, P. Soler-Palacin, R. Somech, P. Palma, H. Wu, S. Giliani, J. E. Walter, L. D. Notarangelo, Characterization of T and B cell repertoire diversity in patients with RAG deficiency. *Science immunology* **1**, (2016).

109. A. Berland, J. Rosain, S. Kaltenbach, V. Allain, N. Mahlaoui, I. Melki, A. Fievet, C. Dubois d'Enghien, M. Ouachee-Chardin, L. Perrin, N. Auger, F. E. Cipe, A. Finocchi, F. Dogu, F. Suarez, D. Moshous, T. Leblanc, A. Belot, C. Fieschi, D. Boutboul, M. Malphettes, L. Galicier, E. Oksenhendler, S. Blanche, A. Fischer, P. Revy, D. Stoppa-Lyonnet, C. Picard, J. P. de Villartay, PROMIDISalpha: A T-cell receptor alpha signature associated with immunodeficiencies caused by V(D)J recombination defects. *The Journal of allergy and clinical immunology* **143**, 325-334 e322 (2019).

110. G. J. Driessen, H. Ijspeert, C. M. Weemaes, A. Haraldsson, M. Trip, A. Warris, M. van der Flier, N. Wulffraat, M. M. Verhagen, M. A. Taylor, M. C. van Zelm, J. J. van Dongen, M. van Deuren, M. van der Burg, Antibody deficiency in patients with ataxia telangiectasia is caused by disturbed B- and T-cell homeostasis and reduced immune repertoire diversity. *The Journal of allergy and clinical immunology* **131**, 1367-1375 e1369 (2013).

111. J. Wu, D. Liu, W. Tu, W. Song, X. Zhao, T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome. *The Journal of allergy and clinical immunology* **135**, 209-216 (2015).

112. G. K. Wong, D. Millar, S. Penny, J. M. Heather, P. Mistry, N. Buettner, J. Bryon, A. P. Huissoon, M. Cobbold, Accelerated Loss of TCR Repertoire Diversity in Common Variable Immunodeficiency. *Journal of immunology* **197**, 1642-1649 (2016).

113. K. M. Roskin, N. Simchoni, Y. Liu, J. Y. Lee, K. Seo, R. A. Hoh, T. Pham, J. H. Park, D. Furman, C. L. Dekker, M. M. Davis, J. A. James, K. C. Nadeau, C. Cunningham-Rundles, S. D. Boyd, IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Science translational medicine* **7**, 302ra135 (2015).

114. S. Keerthikumar, R. Raju, K. Kandasamy, A. Hijikata, S. Ramabadran, L. Balakrishnan, M. Ahmed, S. Rani, L. D. Selvan, D. S. Somanathan, S. Ray, M. Bhattacharjee, S. Gollapudi, Y. L. Ramachandra, S. Bhadra, C. Bhattacharyya, K. Imai, S. Nonoyama, H. Kanegane, T. Miyawaki, A. Pandey, O. Ohara, S. Mohan, RAPID: Resource of Asian Primary Immunodeficiency Diseases. *Nucleic acids research* **37**, D863-867 (2009).

115. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, C. Exome Aggregation, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

116. M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, D. R. Maglott, ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-985 (2014).

117. W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, R. Karchin, CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27**, 2147-2148 (2011).

118. J. Li, L. Shi, K. Zhang, Y. Zhang, S. Hu, T. Zhao, H. Teng, X. Li, Y. Jiang, L. Ji, Z. Sun, VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic acids research* **46**, D1039-D1048 (2018).

119. M. Fang, H. Abolhassani, Q. Pan-Hammarstrom, E. Sandholm, X. Liu, L. Hammarstrom, Compound Heterozygous Mutations of IL2-Inducible T cell Kinase in a Swedish Patient: the Importance of Early Genetic Diagnosis. *Journal of clinical immunology* **39**, 131-134 (2019).

120. J. E. D. Thaventhiran, H. Lango Allen, O. S. Burren, J. H. R. Farmery, E. Staples, Z. Zhang, W. Rae, D. Greene, I. Simeoni, J. Maimaris, C. Penkett, J. Stephens, S. V. V. Deevi, A. Sanchis-Juan, N. S. Gleadall, M. J. Thomas, R. B. Sargur, P. Gordins, H. E. Baxendale, M. Brown, P. Tuijnenburg, A. Worth, S. Hanson, R. Linger, M. S. Buckland, P. J. Rayner-Matthews, K. C. Gilmour, C. Samarghitean, S. L. Seneviratne, P. A. Lyons, D. M. Sansom, A. G. Lynch, K. Megy, E. Ellinghaus, D. Ellinghaus, S. F. Jorgensen, T. H. Karlsen, K. E. Stirrups, A. J. Cutler, D. S. Kumararatne, S. Savic, S. O. Burns, T. W. Kuijpers, E. Turro, W. H. Ouwehand, A. J. Thrasher, K. G. C. Smith,

Whole Genome Sequencing of Primary Immunodeficiency reveals a role for common and rare variants in coding and non-coding sequences. *bioRxiv*, 499988 (2018).

121. K. R. Engelhardt, Y. Xu, A. Grainger, M. G. Germani Batacchi, D. J. Swan, J. D. Willet, I. J. Abd Hamid, P. Agyeman, D. Barge, S. Bibi, L. Jenkins, T. J. Flood, M. Abinun, M. A. Slatter, A. R. Gennery, A. J. Cant, M. Santibanez Koref, K. Gilmour, S. Hambleton, Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by Whole Exome Sequencing. *Journal of clinical immunology* **37**, 42-50 (2017).

122. F. Rieux-Laucat, J. L. Casanova, Immunology. Autoimmunity by haploinsufficiency. *Science* **345**, 1560-1561 (2014).

123. J. L. Casanova, M. E. Conley, S. J. Seligman, L. Abel, L. D. Notarangelo, Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *The Journal of experimental medicine* **211**, 2137-2149 (2014).

124. D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, H. Kehrer-Sawatzki, Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human genetics* **132**, 1077-1130 (2013).

125. G. Sirugo, S. M. Williams, S. A. Tishkoff, The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31 (2019).

126. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* **51**, 584-591 (2019).

127. L. D. Notarangelo, T. A. Fleisher, Targeted strategies directed at the molecular defect: Toward precision medicine for select primary immunodeficiency disorders. *The Journal of allergy and clinical immunology* **139**, 715-723 (2017).

128. D. V. Zhernakova, T. H. Le, A. Kurilshikov, B. Atanasovska, M. J. Bonder, S. Sanna, A. Claringbould, U. Vosa, P. Deelen, L. Franke, R. A. de Boer, F. Kuipers, M. G. Netea, M. H. Hofker, C. Wijmenga, A. Zhernakova, J. Fu, s. LifeLines cohort, B. consortium, Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nature genetics* **50**, 1524-1532 (2018).

129. O. Wang, R. Chin, X. Cheng, M. Wu, Q. Mao, J. Tang, Y. Sun, E. Anderson, H. Lam, D. Chen, Y. Zhou, L. Wang, F. Fan, Y. Zou, Y. Xie, R. Zhang, S. Drmanac, D. Nguyen, C. Xu, C. Villarosa, S. Gablenz, N. Barua, S. Nguyen, W. Tian, J. Liu, J. Wang, X. Liu, X. Qi, A. Chen, H. Wang, Y. Dong, W. Zhang, A. Alexeev, H. Yang, J. Wang, K. Kristiansen, X. Xu, R. Drmanac, B. Peters, Efficient and unique co-barcoding of second-generation sequencing reads from long DNA molecules enabling cost effective and accurate sequencing, haplotyping, and de novo assembly. *Genome research*, (2019).