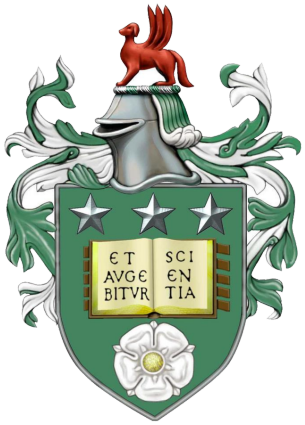
Novel genetic discoveries in rare primary immunodeficiencies

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy



Dylan Lawless

Leeds Institute of Medical Research

School of Medicine

University of Leeds

Under the supervision of

Sinisa Savic MD, PhD and Rashida Anwar, PhD

September 2019

Publication Statement

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Publication Statement

The following chapters are based on work from jointly authored publications.

Chapter 1. Prevalence and clinical challenges among adult PID patients with recombination-activating gene deficiency.

Journal of Allergy and Clinical Immunology. Feb 2018; doi 10.1016/ j.jaci.2018.02.007.
Dylan Lawless*, Christoph B Geier*, Jocelyn R Farmer, Hana Allen Lango, Daniel Thwaites, Faranaz Atschekzei, Matthew Brown, David Buchbinder, Siobhan O Burns, Manish J Butte, et al. (shared *first authors).

Dr Christoph Geier and Dr Jolan Walter completed the clinical and genetic work involved with non-UK cases presented in this study. Dr Hana Allen Lango provided NIHR genetics data. D Lawless performed the genomic analysis of the NIHR study on RAG deficiency, carried out functional work, and compiled phenotypic data used in this study. D Lawless and Dr Christoph Geier wrote this manuscript as shared authors. Dr Jolan Walter, Dr Rashda Anwar, Dr Sinisa Savic all wrote this paper as project leaders. A long list of co-authors, omitted here, also provided individual contributions.

Chapter 2. Predicting the occurrence of variants in *RAG1* and *RAG2*.

Journal of Clinical Immunology. Aug 2019; doi 10.1007/s10875-019-00670-z.
Dylan Lawless, Hana Lango Allen, James Thaventhiran, NIHR BioResource–Rare Diseases Consortium, Flavia Hodel, Rashida Anwar, Jacques Fellay, Jolan E. Walter, and Sinisa Savic.

Dr Hana Allen Lango provided NIHR population genetics data. Dr Jolan Walter provided the summary data on known cases of disease. All remaining work is attributable to D Lawless. D Lawless wrote this paper with Dr Sinisa Savic and Dr Rashda Anwar as project leaders. All co-authors contributed to the manuscript.

Chapter 4. Germline *TET2* loss-of-function causes childhood immunodeficiency and lymphoma.

Manuscript under submission at time of writing.

Jarmila Stremenova Spegarova*, **Dylan Lawless***, Siti Mardhiana Binti Mohamad, Karin R. Engelhardt, Gina Doody, Jennifer Shrimpton, Anne Rensing-Ehl, Stephan Ehl, Frederic Rieux-Laucat, Catherine Cargo, Aneta Mikulasova, Meghan Acres, Helen Griffin, Neil V. Morgan, James A. Poulter, Eamonn G. Sheridan, Philip Chetcuti, Sean O'Riordan, Rashida Anwar, Clive Carter, Stefan Przyborski, Kevin Windebank, Andrew J. Cant, Majlinda Lako, Chris M. Bacon, Sinisa Savic**, Sophie Hambleton** (shared *first and **last authors).

Prof Sophie Hambleton and Dr Sinisa Savic lead this study and collected the majority of clinical data. Dr Jarmila Spegarova-Stremenova, Dr Karin Engelhardt, and several others will smaller roles , omitted here, completed the majority research on one of the two families in this study including the production of iPSCs, flow cytometry, and somatic variant analysis. Prof Gina Doody and her laboratory performed the B cell differentiation assays in this study. Dr Clive Carter provided clinical laboratory data. Dr Christopher Bacon performed histology. D Lawless performed Western blotting for one family, developed and completed the methylation assays used in this study, performed genomic analysis on both families, carried out population genetics analysis, DNA sequencing and RNA quantification. D Lawless and Dr Spegarova-Stremenova wrote this manuscript as shared authors with support from all co-authors.

Acknowledgements

This work was funded by The University of Leeds 110 Anniversary Scholarship. This study makes use of data generated by the NIHR BioResource - Rare Disease Consortium; A full list of the consortium members who contributed to the generation of the data is provided at the end. Partial funding was provided by the National Institute for Health Research (grant number RG65966). Parts of this work was also funded by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases, NIH.

I thank Prof Christian J. Müller for providing pathology specimens and Dr. Karl Waibel for providing pulmonary function testing. I have collaborated with and received clinical data, research data, discussions, and guidance from Dr Hana Lango Allen, Dr James Thaventhiran, Dr Jolan Walter, Dr Christoph Geier, Prof Sophie Hambleton, Dr Jarmila Spegarova-Stremenova, Dr Karin Engelhardt, Prof Gina Doody, Dr Clive Carter, Dr Christopher Bacon, and many others unnamed from additional projects whom I thank sincerely. All specific contributions are acknowledged within.

The University of Leeds has provided an extremely supportive research community with funding and administration. Specifically, the staff in the School of Medicine and Leeds Institute of Medical Research at St James's University Hospital have built a friendly and productive environment where I have happily developed. I thank the examiners of this thesis who have generously provided their time and expertise.

Lastly, I thank my supervisors. Any success and learning is directly due to their guidance. They have each provided a high level of formal scientific rigour while simultaneously allowing the freedom to experiment with ideas, make mistakes, and learn the responsibilities of designing projects. I am grateful for their continued support and thank them for leading the way.

Acknowledgements

To my parents, brothers, and sister who have all provided their support during my studies, for which I am grateful. I would not have completed many of my goals had they not sacrificed and encouraged. Similarly, my grandparents, aunts and uncles, who cannot be named individually have all given their encouragement and time to help out in countless ways. I too am grateful to my partner Denisa for her patience and support that I will never be able to fully repay.

Abstract

Introduction

Rare genetic diseases provide an insight into otherwise obscure mechanisms of human health. Single-case and cohort studies of rare disease can reveal precise and fundamental features of biology that are not as readily apparent in the study of common disease genomics. Furthermore, cases of rare disease also provide a jump start to the incremental scientific method. Statistically robust associations between genetic variation and disease are the most reliable sources of this information. However, since the number of cases in rare disease cohorts is generally low alternative methods must be used to functionally validate genomic findings. Herein, we use best practices in genomic analysis followed by functional validation studies and where possible demonstrate methods for statistically driven analysis of cohorts.

Methods

A combination of genomic sequencing methods were used to uncover the genetic determinants of primary immunodeficiencies (PID). Tailored analysis in single case studies and statistical methods in cohort analysis were used to find candidate causes of disease. Best practices were used for routine analysis of genomic data, complemented by novel bioinformatic approaches. We performed functional investigations using in vitro and in vivo assays to model disease and protein mechanisms and thereby confirm the mode of disease for some patients.

Results

Our results are separated on the basis of patient disorder. First, patients with RAG deficiency may survive into adulthood and the presented findings suggest that prevalence of such cases varies between 1% to 1.9% in adult PID cohorts. Second, we predict a list of amino acid residues for RAG1 and RAG2 that have not been reported to date

but are most likely to present clinically as RAG deficiency. Third, our findings in TET2 deficiency expand the understanding of its critical role within the human hematopoietic system and define a new inborn error of immunity. Fourth, we provide validated methods for the investigation of rare genetic disease.

Conclusion

Genetic investigation in rare PIDs not only provides critical information for clinical care but can provide answers to fundamental questions in basic science.

Contents

List of Figures

List of Tables

5-caC (5-carboxylcytosine)
5-fC (5-formylcytosine)
5ghmC (glucosylated 5hmC)
5hmC (5-hydroxymethylcytosine)
5hmU (5-hydroxymethyluracil)
5mC (5-methylcytosine)
α-KG (α-ketoglutarate)
AI (autoimmunity)
AIC (autoimmune cytopenia)
AIHA (autoimmune haemolytic anaemia)
ALPS (autoimmune lymphoproliferative syndrome)
AML (acute myeloid leukaemia)
AN (autoimmune neutropenia)
ANA (anti-nuclear antibody)
AUC (area under the curve)
BCR (B cell receptor)
BIIa (Basic IIa domain)
BER (base excision repair)
BiB (Born in Bradford)
bGH (Bovine growth hormone)
BWA (Burrows Wheeler aligner)
CADD (combined annotation dependent depletion)
CTD (Carboxy-terminal domain)
CID–G/AI (CID associated with granulomas and/or autoimmunity)
CMML (chronic myelomonocytic leukaemia)
CMV (Cytomegalovirus)
CLP (common lymphoid progenitor)
COSMIC (Catalogue of Somatic Mutations in Cancer)
CpG (cytosine-guanine dinucleotides)
cryo-EM (cryo-electron microscopy)
CVID (common variable immunodeficiency)
D-2-HG (D-2-hydroxyglutarate)
DDSBs (DNA double-strand breaks)
DDBD (Dimerisation and DNA-binding domain)
DME (Demeter)
DNMT (DNA methyltransferase)
DNMT1 (DNA methyltransferase 1)
DNT (double negative T cells)
ESCs (embryonic stem cells)
FBS (fetal bovine serum)
FDR (False discovery rate)
gnomAD (Genome Aggregation Database)
GO (Gene Ontology)
GrCh38 (Genome Reference Consortium Human Build 38)
GVCF (Genomic Variant Call Format)
GWAS (genome-wide association studies)
HGG (hypogammaglobulinaemia)
HMDS (Haematological Malignancy Diagnostic Service)
HOMedU (hydroxymethyldeoxyuridine)
HSC (haematopoietic stem cell)
HSCT (haematopoietic stem cell transplant)
ICL (Idiopathic CD4+ lymphopenia)
IDH1 (isocitrate dehydrogenase 1)
IGV (Integrative genomics viewer)
Ig (immunoglobulin)

Abbreviations

iPSC (induced pluripotent stem cells)
ITP (immune thrombocytopenic purpura)
JBP1 (J-binding protein 1)
KEGG (Kyoto Encyclopedia of Genes and Genomes)
KLD (kinase, kigase and DpnI)
KO (knockout)
KPNA1 (Karyopherin subunit α1)
L (Leader)
LoF (loss-of-function)
MDS (myelodysplastic syndrome)
MFR (mutation rate residue frequency)
MPN (myeloproliferative neoplasms)
M_r (mutation rate)
MLL (Myeloid/lymphoid or mixed-lineage leukaemia 1)
NK (Natural killer)
NCBI (National Center for Biotechnology)
NIHR BR-RD (NIHR-BioResource - Rare Disease)
NHEJ (Non-homologous end joining)
Nonamer-binding domain (NBD)
PAD (primary antibody disorders)
PBMCs (peripheral blood mononuclear cells)
PBS (Phosphate-buffered Saline)
PCA (principal component analysis)
PCGP (Pediatric Cancer Genome Project)
Pfam (Protein families database)
PHD (Plant homeodomain)
pLI (probability of being loss-of-function intolerant)
PPI (Protein-protein interaction)
preR (Pre-RNase H)
qPCR (quantitative RT-PCR)
RSS (Recombination signal sequence)
RAG (Recombination-activating gene)
RF (rheumatoid factor)
R_f (residue frequency)
RNH (RNase H)
ROS1 (Repressor of silencing 1)
RSV (Respiratory Syncytial Virus)
SCID (Severe combined immunodeficiency)
SDM (Site directed mutagenesis)
SDS (Sodium Dodecyl sulfate)
SPAD (Selective PAD)
TCR (T cell receptor)
T4-BGT (T4 β-glucosyltransferase)
TARGET (Therapeutically applicable research to generate effective treatments)
TDG (thymine DNA glycosylase)
TdT (Terminal deoxynucleotidyl transferase)
TET (Ten eleven translocation)
UDP (uridine diphosphate)
VCF (variant call format)
VDJ (variable, diversity, and joining)
ZnBD (Zinc-binding domain)

Introduction

The current classification of primary immunodeficiencies (PIDs) was compiled by the Expert Committee of the International Union of Immunological Societies [1] in *The Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity.* Over approximately 50 years the list of inborn errors of immunity has grown to over 350 disorders. In a growing field of impressive complexity, a variety of conditions present themselves unique ways. The molecular dissection of inborn errors has the potential to reveal key insights into the non-redundant functions of individual genes and pathways.

During this study, over 200 patients with immune disorders (100 with primary immunodeficiencies) from St James's University Hospital had their DNA sequenced locally and were investigated for molecular and genomic determinants of disease. Over 1,000 patients with immune disorders throughout Europe have been assessed in the studies carried out during this work. Herein, we focus on two of the life-threatening disorders found in our cohort of patients.

Like all humans, patients with rare immune diseases carry, on average, twenty thousand rare and common coding variants that can be detected by genomic analysis. It is thus a major challenge to uncover candidate genomic determinants for experimental testing. The first disorder exemplifying the challenges and opportunities for discovery occurs through damaging mutations in Recombination-activating gene 1 (*RAG1*) and *RAG2*. The resulting disease presents at an early age with a distinct phenotype of life-threatening immunodeficiency or autoimmunity. The genetic diagnosis of patients was carried out, findings were functionally validated and new methods of disease prediction were established with hopes of improving preparedness for clinical diagnosis.

The second disorder in this study is due to germline pathogenic variants in the epigenetic regulator Ten-Eleven Translocation methylcytosine dioxygenase 2 (*TET2*). The encoded protein is a known target of somatic loss-of-function mutations associated with clonal haematopoiesis, myeloid, and lymphoid malignancies. Genetic findings were confirmed by extensive functional analysis. To our knowledge, these are the first reported cases of autosomal recessive germline TET2 deficiency in humans, which is compatible with life, but causes a clinically significant immunodeficiency and marked predisposition to lymphoma.

The routine and novel bioinformatic and functional approaches used in these studies are discussed in detail, with examples and potential modifications or improvements also outlined. The major findings of this study are immediately applicable to human health. We hope that others can benefit from the results and protocols which have allowed us to uncover novel genetic discoveries in rare primary immunodeficiencies.

Bibliography

[1] Capucine Picard, H. Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Yanick J. Crow, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, Steven M. Holland, Christoph Klein, Tomohiro Morio, Hans D. Ochs, Eric Oksenhendler, Jennifer Puck, Mimi L. K. Tang, Stuart G. Tangye, Troy R. Torgerson, and Kathleen E. Sullivan. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. *Journal of Clinical Immunology*, 38(1):96–128, Jan 2018. ISSN 1573-2592. doi: 10.1007/s10875-017-0464-9. URL https://doi.org/10.1007/s10875-017-0464-9.

1 RAG deficiency in adult PID patients

1.1 Introduction

The content in this chapter has been peer reviewed in Lawless et al. [1]. Recombination-activating gene 1 (*RAG1*) and *RAG2* encode lymphoid-specific proteins that are essential for diversification of the T and B cell repertoire in the thymus and bone marrow, respectively [2, 3]. The antigen-binding regions of T cell receptors (TCR) and B cell receptors (BCR) depend on the recombination process for joining variable (V), diversity (D) and joining (J) gene segments from which they are encoded. Junctional diversity is introduced during V(D)J recombination, further individualising each receptor. Expression of RAG genes occurs during the early stages of T cell and B cell development [4]. The RAG protein complex then induces DNA double-strand breaks (DSBs) at the junction between V(D)J gene segments by binding recombination signal sequences (RSSs) [5]. RSS sites flank each gene segment and contain consensus nonamer and heptamer elements that are separated by a spacer of either 12 or 23 nucleotides. RAG complex targets and binds to the DNA RSS sites which are digested to form sealed hairpin coding ends and blunt signal ends [6]. The genomic DNA site is eventually joined by the non-homologous end joining pathway (NHEJ pathway) [7, 8].

1.1.1 Foundations in recombination

The foundations of understanding the plasticity of the adaptive immune system can be recounted in several key publications. For the first half of the 20th century, antibody was thought to be produced in the presence of antigens; without antigen the responsible cells would simply produce non-specific globulin. It was not until 1955 that N. K. Jerne formed the theory of naturally selected antibody production [9]. He proposed that antibody of all possible specificities are developed; antigen may encounter a specific antibody and select for its amplification in lymphocytes. This work was later awarded the Nobel prizes in 1984. By 1959 the, now widely accepted, theory of clonal selection was formed by Lederberg, Burnet, Nossel, and Talmage [10]. It was proposed that individual B cells produce an antibody of only one specificity. "Natural antibody" on the lymphocyte surface was thought to bind specific antigen and trigger an unknown mechanism for clonal proliferation and antibody production.

At the time, there was difficulty in explaining the abundance of antibodies which collectively produce a specific affinity for any invading pathogen. The number of genes required to code for each protein would far outweigh the possible capacity of a cell's nucleus (requiring 500 times more than the total DNA volume per cell). W. Dreyer and J. Bennet, in 1965, offered the solution; that the immunoglobulin heavy and light chains, or constant and variable regions, are the products of two separate genes. Furthermore, multiple variable regions could account for the observed diversity [11]. However, this idea faced scepticism as it contradicted the *one gene-one polypeptide principle* (V. Ingram's 1962 offshoot of the *one gene-one enzyme theory* from Beadle and Tatum). Hozumi and Tonegawa [12] identified that B lymphocytes assemble immunoglobulin genes via somatic DNA recombination. Restriction mapping on DNA from embryonic cells and myeloma (representing differentiated B cells), in syngenic mice by S. Tonegawa et al. in 1978 happened upon differentially digested heavy and light chains [13]. This publication showed that alternatively recombined restriction fragments were present in matured but not embryonic cells. Experiments using the newly developed Southern blot and probing confirmed Dreyer and Bennet's two gene theory of *two genes-one polypeptide* for both light and heavy chains, and ultimately won Tonegawa the 1987 Nobel prize for discovery

of the mechanisms of generation of antibody diversity. Tonegawa's work illustrated that constant and variable regions lay separated on germline DNA but after B cell maturation rearrangement brings the gene regions closer.

A decade after Tonegawa's Nobel prize winning research, essentially all antigen receptor loci had been mapped. The genetic architecture of substrates for the V(D)J reactions, as shown in **Figure 1.1** and 1.2, has not required much revision from the contemporaneous definition of 1988. However, the enzymatic process responsible for recombination of these sites had no explanation at the time. The only protein with a known involvement had been theorized, four years before Tonegwa's functional work, by David Baltimore in a letter to Nature; "Is terminal deoxynucleotidyl transferase a somatic mutagen in lymphocytes?" [14]. Terminal deoxynucleotidyl transferase (TdT) was known to add deoxyribonucleotides to the ends of DNA primers and was originally found in lymphocytes of the thymus. The essentiality of the recombination process was clear but none of the involved proteins had been found. The key proteins involved in the complex are illustrated in **Figure 1.3**.

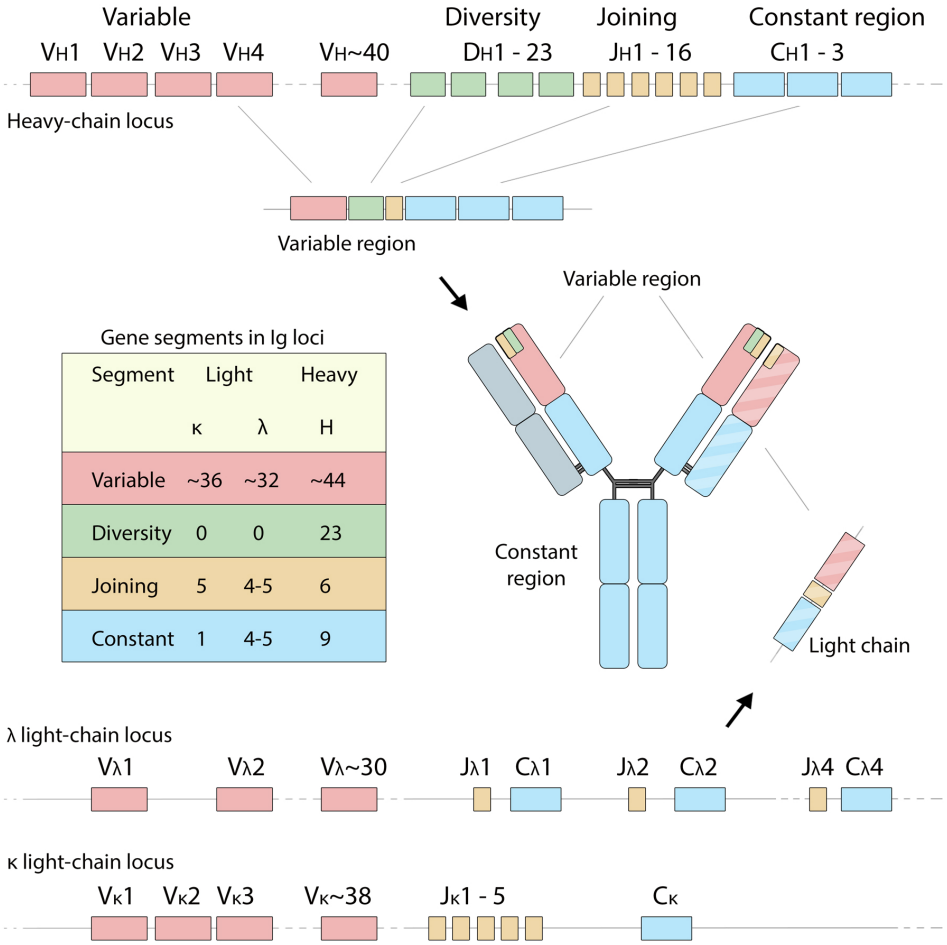


Figure 1.1: **Germline antibody gene locus.** Re-illustration based on several figures from *Janeway's Immunobiology*, Murphy and Weaver [15]. The immunoglobulin (Ig) heavy chain is formed when a diversity (D) gene segment joins to a joining (J) segment; D-J. A variable (V) gene segment rearranges with the DJ to create the variable region exon; V-DJ. Unlike the heavy chain, the light chain has no D segment and proceeds directly to V-J. For both chains, RNA splicing joins the assembled V-region to a neighbouring constant (C) region. The light chain contains a single C region while in the heavy chain it is encoded by several exons. Leader sequences and hinge region are not illustrated. Igs of any class can be produced either as membrane bound receptor or as secreted antibodies. B cells first produce transmembrane IgM. Upon stimulation cells differentiate into either IgM antibody-producing plasma cells, or class switched transmembrane Ig-producing cells followed by antibody secretion of the new class. The last two exons of constant heavy genes encode the peptide sequences for secretion (a hydrophilic secretory sequence) and transmembrane localisation (hydrophobic sequence), respectively. Transcriptional cleavage and poly(A) tagging downstream of both exons results in a membrane bound Ig. However, cleavage upstream of the transmembrane domain will result in protein secretion. Stimulation of surface Ig promotes activation and differentiation into antibody-secreting plasma cells of the same heavy chain isotype. Subsequent transcripts are more likely to splice into the secreted isoform than full-length surface Ig.

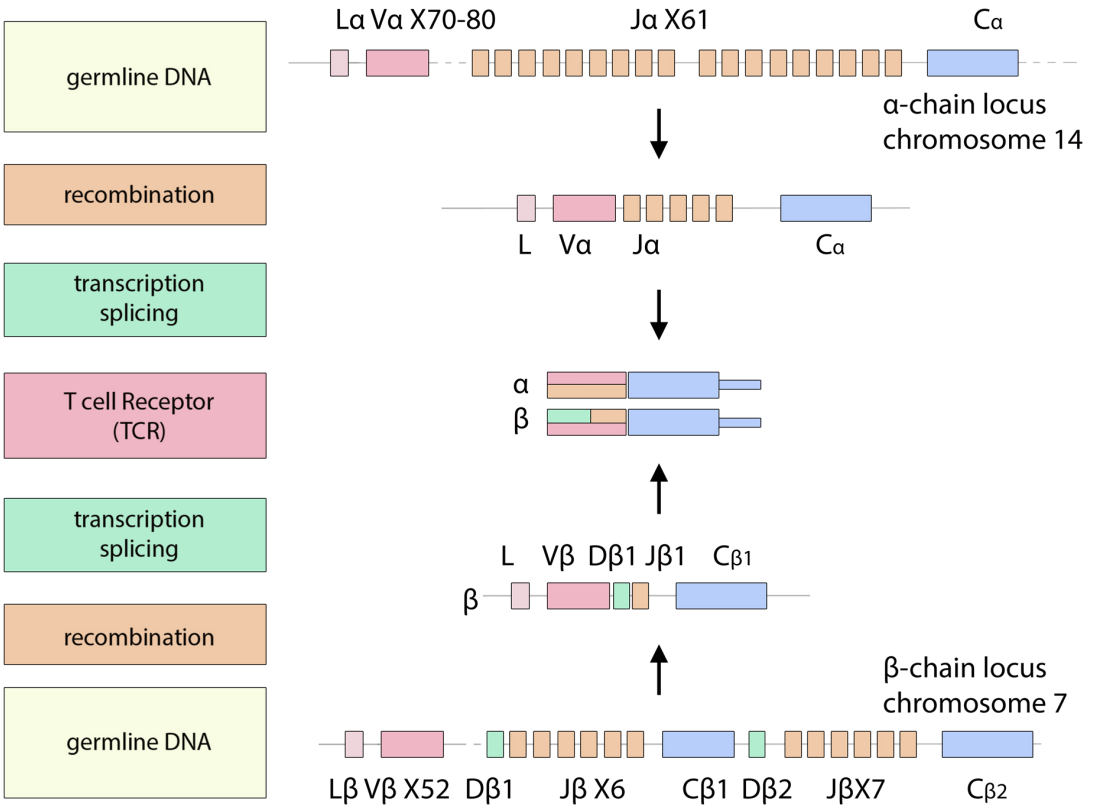


Figure 1.2: **Germline T cell receptor gene locus.** Re-illustration based on figures from Murphy and Weaver [15]. Somatic recombination of the T cell receptor alpha and beta chain genes occurs during T cell development. TCR genes incur recombination similar to the Ig gene locus rearrangement. A variable (V) alpha gene segment recombines to connect a joining (J) alpha segment (V-J), to form the V-region exon. As with Ig rearrangement, transcription-splicing joins the VJ exon to a constant (C) gene region to generate the the VJC transcript. Translation results in the TCR alpha chain. The beta chain contains an additional diversity (D) region, reminiscent of the D-containing Ig heavy chain variable region. From the three TCR beta chain locus D gene segments, recombination with V and D segments produces the VDJ beta V region. Again, transcription and splicing with one of the C gene regions results in a transcript that is translated to form the TCR beta chain. TCR alpha and beta chain proteins pair and translocate to form the alpha beta TCR heterodimer. In this illustration leader (L) sequences are shown, although not all L or J regions are included.

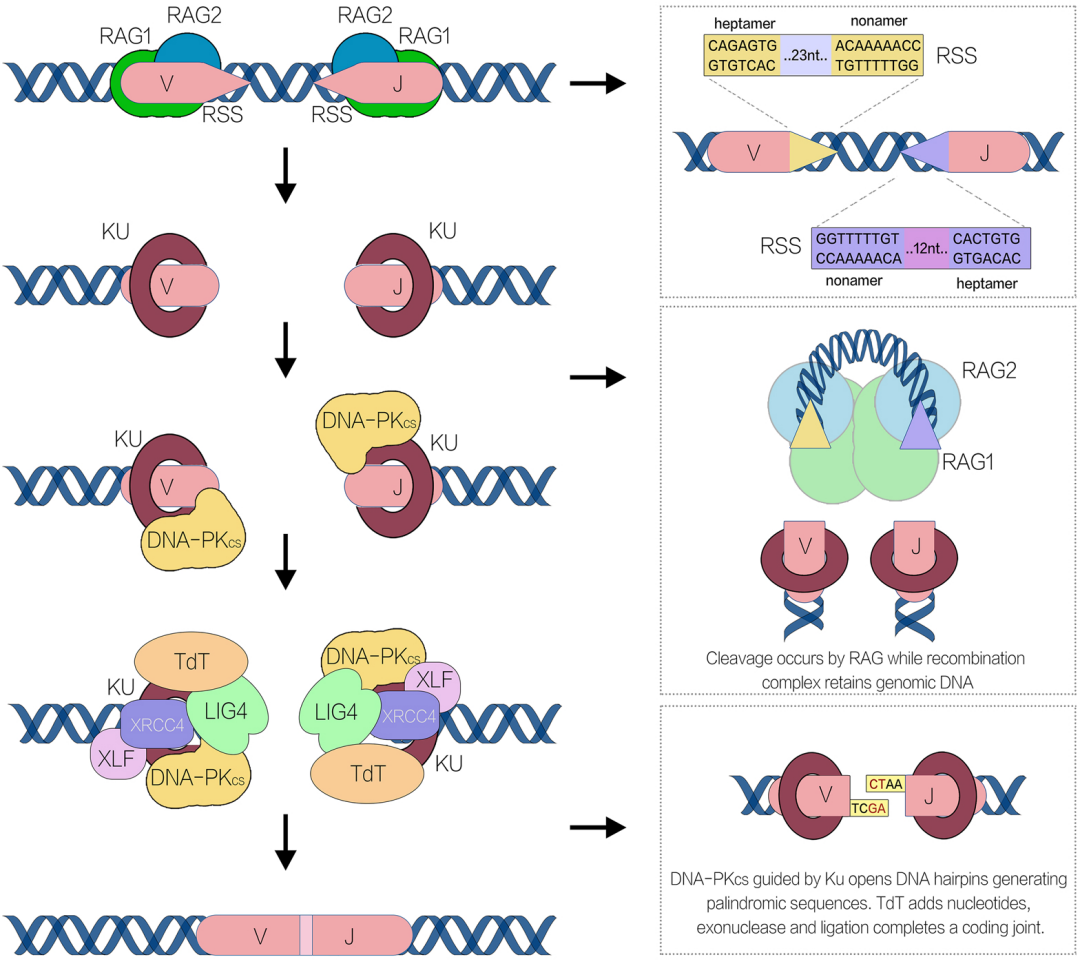


Figure 1.3: **Recombination complex formation.** Illustration based on elements from Khan [16], and Murphy and Weaver [15]. Gene segments involved in V(D)J recombination contain RSSs that are targeted by RAG. An RSS is bound by the RAG1 RAG2 complex and high-mobility group proteins (not shown). The **top right** panel shows the RSS consensus heptamer and nonamer sequences, separated by a 23 nucleotide spacer, which is first bound by RAG. A second RSS is also bound, this time containing a 12 nucleotide spacer. The **middle right** panel shows RAG endonuclease activity producing single-stranded breaks between coding sequences and its RSS. During recombination the digested DNA ends must be retained proximally. The Ku heterodimer binds to DNA double-strand break ends. The formation of a 3'-OH group from the cut DNA causes a reaction with a phosphodiester bond on the opposite DNA strand to generate a hairpin on the coding joints. Blunt double-stranded breaks remain at the end of the RSSs. The coding joint hairpin-bound Ku is targeted by the endonuclease activity of DNA-PKcs. The randomly opened hairpin results in either flush ends or a single extended strand. These ends are modified by TdT and exonuclease, which randomly add and remove nucleotides, respectively; **bottom right**. Final ligation forms a single coding joint.

In the Baltimore lab, Schatz and Baltimore [17] induced stable V(D)J recombinase activity in fibroblasts by transfecting genomic DNA. These transfections happend to carry the unknown genes encoding the illusive recombination machinery. It seemed unlikely that a single protein would carry out all steps for the complex process of recombination alone. However, this set of experiments showed successful recombination from a single locus of transfected DNA. Narrowing down the responsible gene by serial transfection of genomic fractions, Schatz et al. [2] ultimately identified recombination activating gene 1 (*RAG1*). Although successful in identifying a key recombination gene, the expression of a *RAG1* cDNA vector in fibroblasts yielded only meagre recombination activity. The purified cDNA performed no better than the transfections of crude genomic DNA fractions described in their previous paper (which happened to contain *RAG1*). Schatz et al. correctly assumed that a collaborating gene existed close to *RAG1* and quickly reported it as *RAG2* [3]; these three papers cracked the decade long wait for proteins involved in V(D)J recombination.

RAG1 and RAG2 were later confirmed as enzymatically responsible for recombination in both B and T lymphocytes. Soon, van Gent et al. [18] recognised the site-specific DNA cleavage by RAG and illustrated its parallelism to hydrolysis and transesterification reactions carried out by the editing mechanism of some transposases and retroviral integrases. Their idea led to the hypothesis that adaptive immunity evolved through jawed vertebrates after integrating the RAG transposon into an ancestral antigen receptor gene [19, 20]. In 2005, Kapitonov and Jurka [21] found the *Transib* transposon, a 600 amino acid core region of RAG1, and RSS-like (especially heptamer) sequences in many invertebrates. Fugmann et al. [22] identified a linked *RAG1/RAG2* in the lower dueterosome (sea urchin), indicating an earlier common ancestor than the invertebrate [*as described by the same author in a follow-up review* [23]]. Most recently, Huang et al. [24] found a recombinatorially active RAG transposon (ProtoRAG) in the lower chordate amphioxus (or lancelet); the most basal extant chordate and a "living fossil of RAG." This topic is applied practically to human health in *chapter 2* of this thesis. The landmark publications investigating the role of *RAG1* and *RAG2* were, in most cases, only possible because of insightful hypotheses which in some cases could not be tested functionally for

decades.

1.1.2 Recombination accessibility

Four years before the transfection experiments of Schatz et al. [2] using crude genomic fractions to show recombination, Yancopoulos and Alt [25] theorised about how in vivo V(D)J recombination could be developmentally regulated. They wondered why immunoglobulin genes and TCR genes in B and T cells, respectively, rearrange in a particular order; it was known that DH to JH joining preceded VH to DJH joining at the Igh locus, and the rearrangement of Igk occurred after that of Igh. It is not surprising to conclude that the controlled events of these cellular processes would be important to both lymphocyte development and allelic exclusion. Maturation stage and lineage-specific recombination was predicted to occur at controlled developmental stages of lymphocyte maturation, being executed by a single recombinase. They had identified transcripts from unrearranged (germline VH) gene segments which were tissue-specific and developmental stage-specific; in turn these would become substrates for recombination in the maturing B cell. Accessibility to the immunoglobulin and TCR locus chromatin was tightly controlled [25]. The transcription of unrearranged segments that they identified inferred some relation to accessibility [26].

The *accessibility hypothesis* from Yancopoulos and Alt was well-grounded. The functional studies unmasking a single recombinase, RAG, which controlled recombination [2, 3, 27–29] (and subsection 1.1.1) were also reinforced by works coupling *germline transcription* with recombination events, and identifying transcriptional enhancers important for its regulation [30]. Despite the progress there was no clear evidence of accessibility until 1996 when Stanhope-Baker et al. [4] showed that chromatin acted as the critical regulator of in vivo recombination.

In the meantime, Schlissel et al. [31] and Roth et al. [32] had characterised the molecular intermediates of RAG-dependent recombination. Their experiments utilised ligation-mediated PCR [33] to assay purified genomic DNA from developing lymphocytes for recombination–associated double stranded breaks (DSBs). The previously used method

involved Southern blotting with thymus DNA for signs of recombination [6]. Although arduous, it had identified digestion at recombination sites in Dδ2 and Jδ1. The use of ligation-mediated PCRs could allow detection of RAG-dependent 5'-phosphorylated RSS heptamer cut sites. Different developmental stage-specific recombination fragments were confirmed in samples from bone marrow and thymus.

Cleavage of an RSS substrate plasmid occurred with nuclear extract from a cell line which could express recombinant RAG1 (although, not RAG2) [34]. In this system, single RSS sites could be cut rather than the paired-RSS cutting as seen in vivo . The same method meant that the experiments from Stanhope-Baker et al. [4] could measure accessibility at single RSS sites rather than trying to decipher naturally occurring recombination events. Subsequently, McBlane et al. [5] identified that naked DNA was cut by purified recombinant RAG1 and RAG2. These two papers showed that instead of activating recombination indirectly, RAG1 and RAG2 acted as a nuclease.

The initial publications on accessibility led to the most applied investigation so far. Stanhope-Baker et al. [4] prepared nuclear extracts from two cell types; pre-B cell line during its recombination stage and primary bovine thymus expressing recombinant RAG1. The nuclear extracts would have contained RAG1 which was then tested against the nuclei from RAG deficient cells. The cells used were all RAG1 or RAG2-deficient and either pro–B cells, thymocytes, or pre–B cells from transgenic mice.

Lineage-specific DSBs were again shown using ligation-mediated PCR [4]. Cleavage occurred at JH2 in pro–B cells but not at Dδ2. In thymocytes it occurred at Dδ2 but not Jκ1. Jκ1 cleavage didn't occur in thymocytes but it did in LPS-primed pro–B cells and in pre–B cells. Genomic DNA from any source could be cut by these nuclear extracts but extracts from other cell types could not produce DSBs. While conceptualising this sequence of events on-the-fly may be challenging, Stanhope-Baker et al. [4] determined that RSS are only accessible for RAG-dependent recombination in specific lineages, at specific developmental stages. Similarly, VH and DH 5' RSSs of RAG-deficient pro-B cells were cut by the RAG-containing nuclear extract. This lymphocyte development stage matches the time at which VH to DJH rearrangement occurs. The same method applied

to mature B cells resulted in only DH RSS cutting. These experiments addressed Igh allelic exclusion. During late B cell maturation, unrearranged VH gene segments become inaccessible due to chromatin condensation. Therefore, the possibility of VH to DJH recombination is inhibited. The *accessibility hypothesis* has remained valid since [35].

In the experiments from Stanhope-Baker et al. [4] nuclear extract containing endogenous RAG1 could not be mimicked by recombinant RAG1. The main reason is because their recombinant protein produced only the core domain which has since been shown to have less recombination potential than full-length protein [36, 37]. Recombinant core RAG2 also lacks its localisation mechanisms of the plant homeodomain (PHD) which targets open chromatin via trimethylated histone H3 lysine 4 [38–41].

Stanhope-Baker et al. [4] were correct in suggesting that while genomic transcription is correlated, it may not directly affect recombination, which is controlled by chromatin-dependent accessibility. The epigenetics of the antigen receptor locus has remained an important challenge since [42, 43]. Epigenetic mechanisms of histone modifications have important roles in accessibility, including covalent histone modifications [44] histone acetyltransferase [45], and methylation of H3 lysine 4 [46, 47].

1.1.3 Structure of the RAG1 and RAG2 complex

In humans, *RAG1* and *RAG2* are found adjacent on GrCh37 Chr11 at positions 36589563-36601310 and 36613493-36619829, respectively. During lymphoctye development cis-enhancer elements facilitate transcription of both single-coding exon genes [48]. The functional domains of RAG1 and RAG2 are described in **Figure 1.5**. It is generally considered that two copies of RAG1 and RAG2 form a heterotetrameric complex. Both crystallographic [49] and cryo-electron microscopy (cryo-EM) structures [50] have been reported for the complex core. The crystal structure has provided important information for both the mechanistic and translational understanding of the relationship between compound heterozygous variants in human disease. The cryo-EM structure shows how RSSs are targeted for cleavage according to the 12–23 rule.

The structural and mechanistic features of the RAG complex are discussed in reference

to **Figure 1.4**, which illustrates DNA bound by RAG complex during recombination. The domains referred to are shown in **Figure 1.5**.

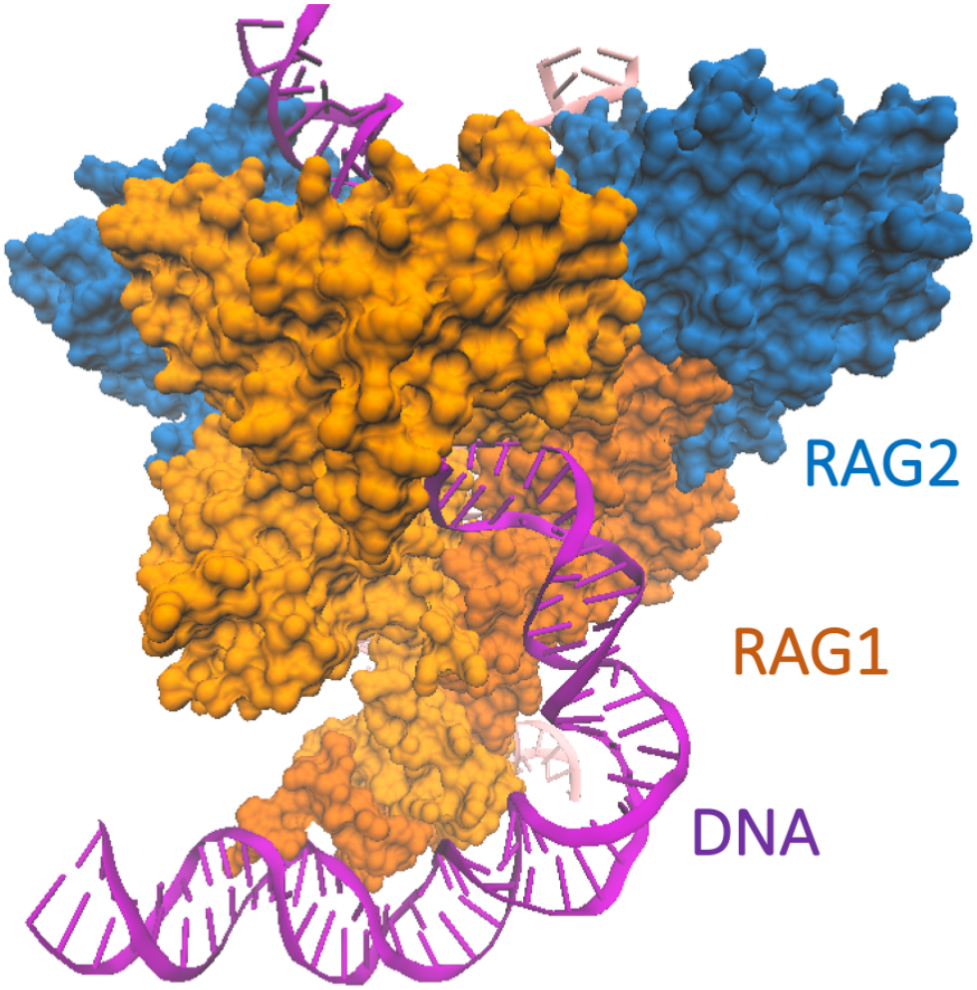


Figure 1.4: **Crystal structure of DNA bound RAG complex. (A)** Crystal structure of DNA bound RAG complex. RAG complex protein structure data from RCSB Protein Data Bank (3jbw.pdb) http://www.rcsb.org [50]. Structure visualised using the software VMD from the Theoretical and Computational Biophysics Group. https://www-s.ks.uiuc.edu/Research/vmd/ Imaged with Tachyon rendering software by Stone [51].

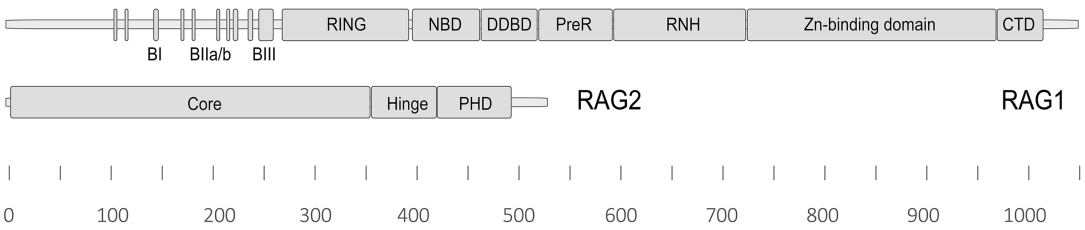


Figure 1.5: **RAG1 and RAG2 protein primary structures.** RAG1 protein consists of 1043 amino acids. Catalytic core contains the nonamer-binding domain (NBD; amino acids 394–460), dimerisation and DNA-binding domain (DDBD; amino acids 461–517), pre-RNase H (preR; amino acids 518–590), catalytic RNase H (RNH; amino acids 591–721), zinc-binding domain (ZnBD; amino acids 722–965), and carboxy-terminal domain (CTD; amino acids 966–1,008). RAG2 protein is composed of 527 amino acids. Core domain (amino acids 1–383) and the non-core region (amino acids 384–527) which includes acidic hinge region (Hinge; amino acids 350 – 410), plant homeodomain (PHD; amino acids 414–487).

RAG1 dimers form a stem, joining at the nonamer-binding domains (NBD) and on top which two RAG2 proteins bind creating a Y shape. DNA-binding domain (DDBD) of RAG1 act as the stem branch point. At the top, the RAG1 molecules again separate to project the the zinc-binding region (ZnBD) while several conserved residues form a catalytic region (D603, D711 and E965) [52, 53] along with the carboxy-terminal domain (CTD) into a Y shaped branch point. Each of the RAG2 cores shown in blue contain *six-bladed β-propellers*. RAG2 binds to its reciprocal RAG1 molecule towards its C-terminal domains, DDBD and CTD [49]. RSS binding causes the protein complex to encase DNA while both RAG1-RAG2 dimers condense [50]. RAG2 stabilised the complex but does not directly interact as RAG1 binds the RSS nonamer and heptamer [50]. RAG1 subnuclear localisation is mediated by basic IIa domain (BIIa; amino acids 219–225). BIIa interacts with karyopherin subunit α1 (KPNA1; also known as importin subunit α5), which is responsible for transport of molecules between the nucleus and cytoplasm [54]. It acts as a putative substrate for the N-terminal RAG1 ubiquitin ligase. The process is mediated by the nuclear pore complex (which allows passive diffusion for molecules up to 70 kD and an active process for larger molecules). This and other non-core domains in RAG1 and RAG2 are not illustrated in the illustrated crystal structure but have been resolved crystallographically and by nuclear magnetic resonance [39, 55].

The (C3HC4) RING finger and zinc finger motif form a domain that regulates zinc ion

interaction and performs as a histone H3 ubiquitin ligase, indicating chromatin-mediated regulation [55]. As the RAG complex targets its recombination site, H3 ubiquitylation must occur for RAG1 catalytic activty to proceed. Unubiquitylated H3 otherwise restrains the process before cleavage can occur [56]. These non-core domains are also required for efficienct recombination activity [56]. Chromatin accessibility has also been attributed to RAG2 non-core domains. RAG2 PHD creates a channel for binding of H3 carrying a trimethylated lysine 4 (H3K4me3) and promotes recombination [38, 39, 57].

Control of RAG gene expression and degradation is likely an important factor for protection against genotoxicity. RAG activity is restricted to lymphocyte G0/G1 phases. While little is published about RAG *over-activity* in human disease, a loss of inhibitory functions might have a pejorative effect as a DNA repair mechanism which, resulting in complex somatic variations, would not be evolutionarily subject to direct selective pressure. Regard, that the phosphorylation of RAG2 residue T490 is known to occur by way of the cyclin-dependent kinase 2 complex before G1 to S phase transition in the cell cycle. This modification promotes poly-ubiquitylation by an S-phase kinase protein complex, tagging RAG2 for proteasomal degradation [58]. Stage-dependent activity also favours the NHEJ pathway which predominantly active during G0/G1 and critical to the consummation of recombination [7].

1.1.4 Human RAG deficiency

Following the work by Oettinger et al. [3] it was found that deficiency of RAG1 [29] and RAG2 [28] in mice causes inhibition of B and T cell development. Schwarz et al. [59] formed the first publication reporting that RAG mutations in humans causes severe combined immunodeficiency (SCID), also deficient in peripheral B and T cells (**Figure 1.6**). Patient studies identified a form of immune dysregulation known as Omenn syndrome [60, 61]. The patient phenotype includes multi-organ infiltration with oligoclonal, activated T cells. The first reported cases of Omenn syndrome identified infants with hypomophic RAG variants which retained partial recombination activity [62].

Human RAG deficiency has traditionally been identified at very young ages due to

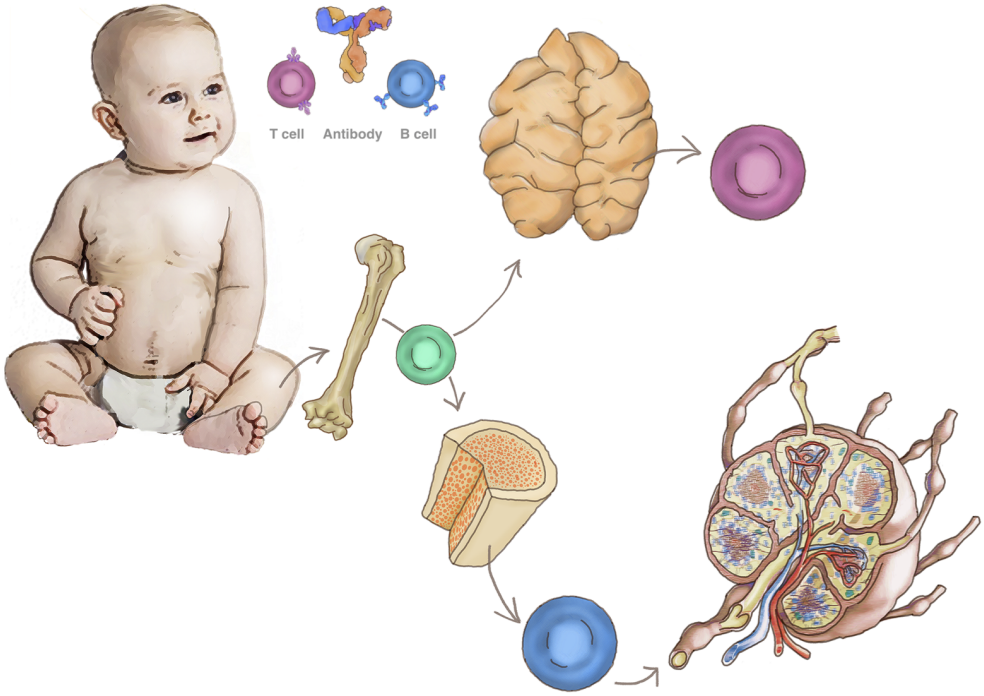


Figure 1.6: **Lymphocyte development.** Progenitor cells must undergo several maturation processes during lymphocyte development. A simplified illustration of diversification of the T and B cell repertoire is shown occurring via the thymus and bone marrow, respectively.

the rapid drop of a maternally-acquired antibody in the first six months of life. A loss of adequate lymphocyte development quickly results in compromised immune responses (**Figure 1.7**). Haematopoietic stem cell transplantation (HSCT) is required in many cases to protect against fatal infections. An increasing understanding of RAG deficiency and modern genomics means that less acute incidents of disease can be identified at later ages. Older children or even adolescents are now found with delayed-onset disease characterized by granulomas and autoimmunity.

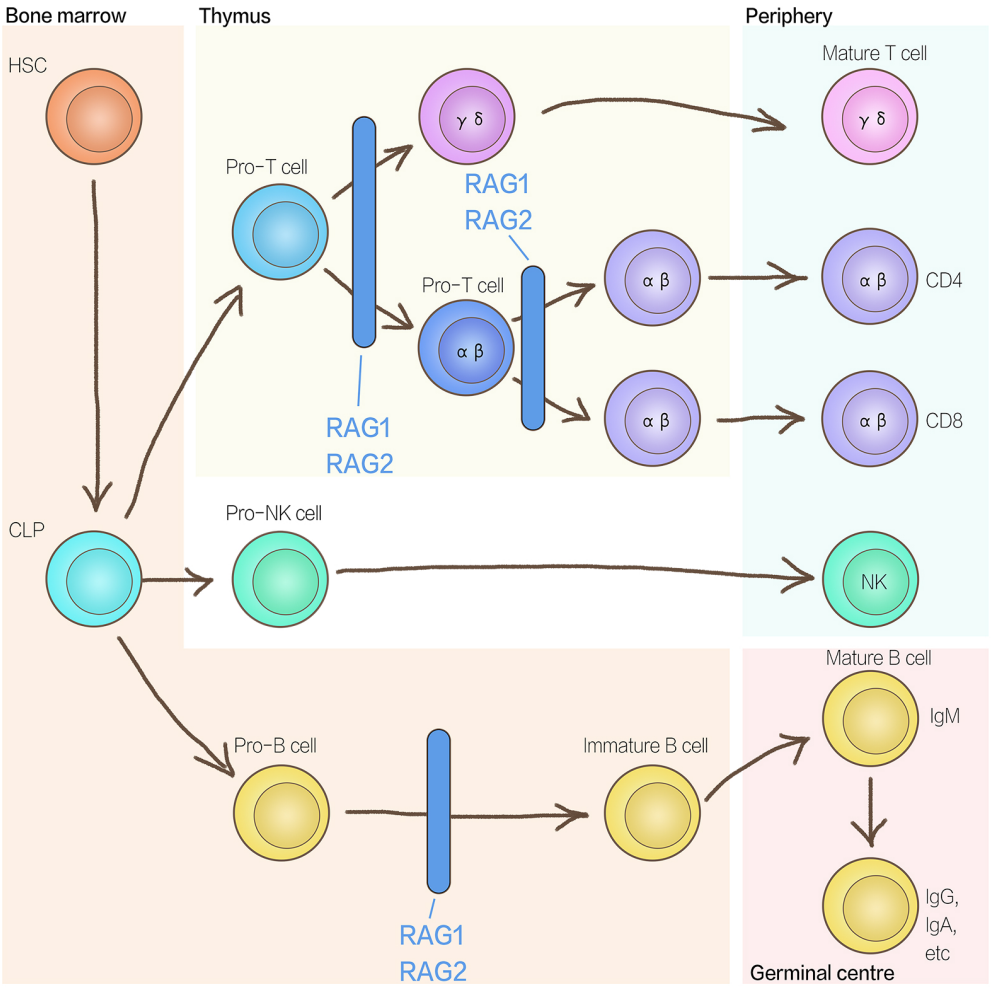


Figure 1.7: ***RAG1* and *RAG2* in lymphocyte development.** Illustration based on De Villartay et al. [63]. Bone marrow-derived haematopoietic stem cell (HSC) give rise to the common lymphoid progenitor (CLP). From this, T cells mature through the thymus, and B cells develop in the bone marrow. Deficiency of proteins involved in the V(D)J recombination system, such as RAG1 and RAG2, can result in severe combined immunodeficiency with an arrest of B and T cell maturation. Lymphocyte maturation restriction may be seen at the stage of pro-B and pre-T or pro-T cells, illustrated as RAG-dependent bars. Genomic variant analysis for immunodeficiency may be guided by a patient phenotype. For example, natural killer (NK) cell development is unrestricted in RAG deficiency. Antigen response normally drives maturation in the germinal centres of peripheral lymphoid organs. Ig isotype switching and variable region somatic mutation are dependent on other components of the adaptive immune system. B-cell maturation arrest may present with a failed transition from cytoplasmic IgM– to IgM+ expression. A patient presenting with hyper-IgM may be more likely to harbour damaging CD40 variants. Interpretation of unknown variants is aided by a detailed clinical description.

RAG protein complex retaining residual activity may produce a repertoire of antibody sufficient to protect against fatal infection, however the disease phenotype takes on autoinflammatory features with the oligoclonal expansion of autoreactive lymphocyte. By identifying the CID associated with granulomas and/or autoimmunity (CID–G/AI), Schuetz et al. [64] expanded the spectrum of investigation. The work in this thesis chapter contributes to the understanding of human RAG deficiency in adults [1].

Hypomorphic *RAG1* and *RAG2* mutations with residual V(D)J recombination activity (on average 5-30%) results in a distinct phenotype of CID-G/AI [64–69]. Besides the work in this chapter, there is no published systematic evaluation for the presence of an underlying RAG deficiency in patients with primary antibody deficiencies. Inflammatory complications are increasingly reported for RAG deficient patients with CID-G/AI phenotype and late diagnosis [70–72]. Allograft rejection and fatal post-transplant complications are more common among SCID patients with RAG variants than in other forms of SCID, especially if harbouring infections [73–75].

There is limited experience with HSCT for partial RAG deficiency with a CID-G/AI phenotype. In a recent multi-centre study, HSCT was offered in 61% of cases, less frequently than in variants of SCID. [68]. CID-G/AI patients with an ongoing or history of infections and/or underlying inflammatory lung disease may also fail to engraft stem cells or die from other post-transplant complications [68, 70, 71, 76]. Overall, early recognition of patients with RAG deficiency and CID-G/AI phenotype and initiation of proper treatment of underlying lung disease may allow for prompt definitive treatment and improved outcomes.

RAG deficiency has an estimated disease incidence of 1:181,000 including SCID at a rate of 1:330,000 [66, 69]. Complete or hypomorphic variants of SCID secondary to low recombinase activity (<5%) present early with severe infections and/or clinical signs of systemic inflammation, such as severe dermatitis and/or colitis [59, 62, 73, 77, 78]. Hypomorphic *RAG1* and *RAG2* mutations result in proteins with residual recombination activity and a phenotype of CID-G/A [64–69]. Recently, hypomorphic *RAG1* mutations were shown to alter the pre-immune repertoire at early stages of lymphoid development

[79]. This is a highly vulnerable patient population with treatment refractory cytopenias, severe vasculitis and increased mortality despite treatment, including hematopoietic stem cell transplantation. In a previous report of a multi-centre study of patients with CID-G/AI phenotype, thirteen patients were described including young adults with a broad spectrum of autoimmunity (cytopenias, vitiligo, psoriasis, vasculitis, neurological complications such as myasthenia gravis, and Guillain–Barré syndrome) (77%), granuloma (54%) and an overall poor survival rate (61%) [68].

Beyond the spectrum of combined immunodeficiency, RAG deficiency has been found in patients with predominantly primary antibody deficiencies such as common variable immunodeficiency [71, 80], agammaglobulinemia [81], selective IgA deficiency [82], and polysaccharide antibody deficiency [72]; however T cell studies eventually confirmed naïve CD4+ T cell lymphopenia in most cases. There are also individual case reports of idiopathic CD4+ T cell lymphopenia [83], hyper-IgM syndrome [84], and sterile chronic multifocal osteomyelitis linked to RAG deficiency [85].

Acute or persistent bacterial and viral infections (especially *Herpesviridae*) of the lung have been reported among patients with variant forms of SCID, which posed a risk for poor transplant outcome but resolved after successful HSCT [73]. RAG deficient patients with a CID-G/AI phenotype and late diagnosis are increasingly reported to have inflammatory complications such as granulomatous-lymphocytic interstitial lung disease and alveolar fibrosis leading to respiratory failure [70–72]. There is great variability among diagnostic modalities for evaluation and treatment to control progression of inflammatory lung disease in case reports of RAG deficient patients with no standardized guidelines in place. Clinical features and lung disease for patients with late presentation of RAG deficiency have not been studied extensively. In addition, no studies have examined the prevalence of RAG deficiency in cohorts of adult primary immunodeficiency patients. This chapter describes a cohort of 15 patients with late presentation of RAG deficiency. The prevalence of RAG deficiency is estimated for adult PID patents following genetic analysis in two separate large cohorts.

1.1.5 Population genetics

The NIHR-BioResource - Rare Disease (NIHR BR-RD) is a study run in the UK whose aim is to assist clinical management of patients with rare diseases and gather insight from large-scale genomics based on disease cohort phenotypes. The NIHR BR-RD study includes whole genome sequence data from about 8,000 individuals. This chapter focuses on 558 unrelated individuals from the PID cohort that were targeted for analysis of genetic determinants of RAG deficiency (NIHR BR-RD PID). Most cases were singletons however family members were included in analysis when possible. Patients who were included in the analysis here were recruited by physicians trained in paediatric or internal medicine specialising in the field of clinical immunology. Participation occurred through 26 hospitals in the UK. Enrolment primarily included those with a clinical diagnosis of CVID according to the current European Society for Immunodeficiencies registry criteria of definitions for clinical diagnosis of PID. Some cases included extreme autoimmunity; or recurrent (and/or unusual) infections suggestive of severely defective innate or cell-mediated immunity. The majority of patients were genetically identified as of European decent (>80%). Two peaks at the age of presentation are found for the NIHR BR-RD PID cohort, early (ages 1-10 years) and middle aged (at ages 30-40) (**Figure 1.8**). Patients and enrolled family members have provided written informed consent with study approval from the East of England Cambridge South national institutional review board. An important source of reference population genetics data was also sourced from whole genome and exome data of approximately 146,000 individuals. This data came from various disease-specific (but unrelated to features of immunodeficiency) and population genetic studies collated as part of GnomAD (version r2.0.2) [86].

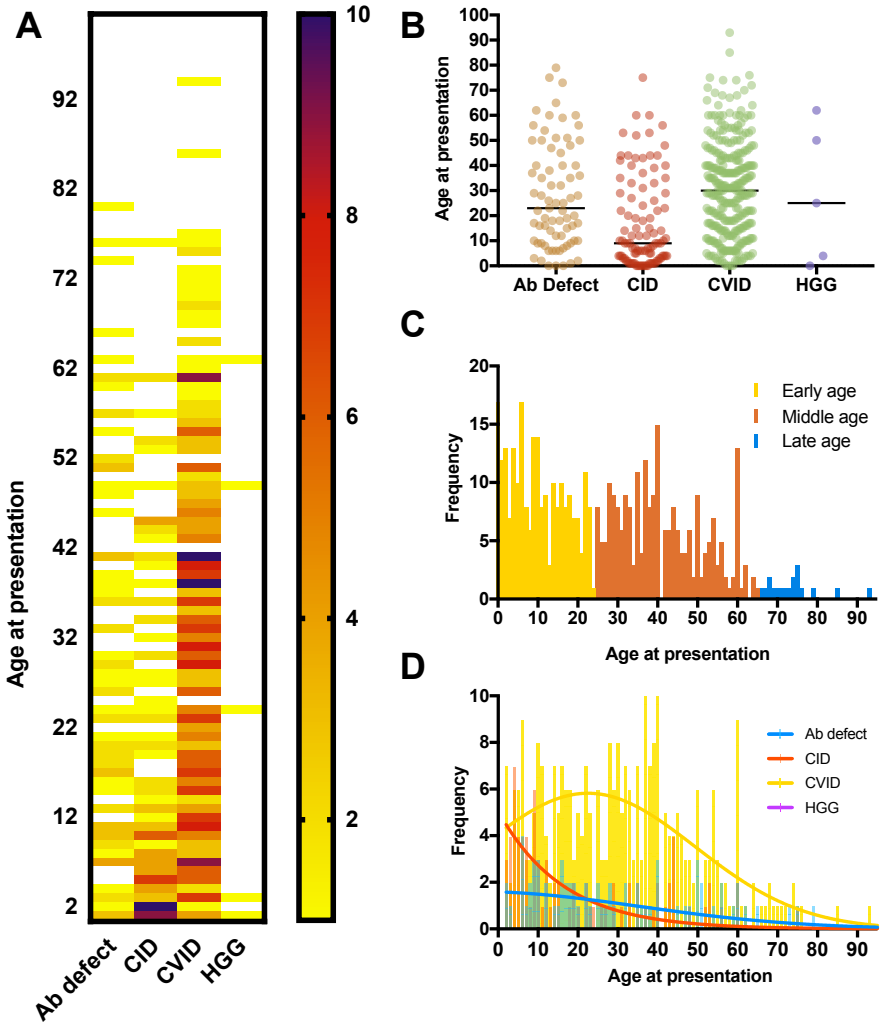


Figure 1.8: **NIHR BR-RD PID cohort age of presentation.** Based on 558 unrelated patients, the age at presentation and general phenotypes are presented. **A.** Heatmap of the number of patients (1-10) per disease presentation age. **B.** Median age of presentation per disease phenotype. **C.** Histogram of total age of presentation with PID separated into three life stages. **D.** Regression curve of total age of presentation with PID. The largest group, CVID, shows two peaks of presentation (early and middle aged). Ab defect (antibody defect, deficiency of specific-Ig), hypogammaglobulinemia (HGG).

1.2 Aims and objectives

To develop a bioinformatic pipeline capable of identifying cases of RAG deficiency in a large cohort of European patients with PID, determine the prevalence of such cases in adult PID cohorts, functionally validate potential candidate variants of unknown significance, and apply confirmed genetic findings to interpret the significance of patients' clinical data.

1.3 Methods

1.3.1 Whole genome sequencing

As part of the NIHR BioResource Rare Disease study, 558 unrelated PID patients had their genomes sequenced. Paired-end whole genome sequencing was performed by Illumina on their HiSeq X Ten system (www.illumina.com). Read pairs were of lengths 100, 125 or 150 base pairs. 95% of bases were covered by at least 15 reads. Sequences were aligned to the GRCh37 genome build using Isaac aligner. Substitutions and InDels up to 50 bp were called and then merged with AGG3 tool, while structural variants were called by Manta and Canvas (all software by Illumina, San Diego, CA, USA). Read depths averaged 35. Only variants with an overall quality score of >80% (based on read depth and variant call confidence) were considered for further analysis. Structural variants were analysed for gene deletions, but none were identified.

1.3.2 Targeted sequencing

The coding regions of *RAG1* and *RAG2* canonical transcripts in 134 patients who were not a part of the NIHR BR-PID cohort were also analysed. Targeted sequencing was coordinated by the Immunology Outpatient Clinic, Vienna, Austria. Genomic DNA was prepared from peripheral blood by spin column purification (QIAamp DNA Blood Mini Kit; QIAGEN, Germany). Targeted resequencing of the canonical region of *RAG1* and *RAG2* was performed using Nextera Custom Enrichment kit according to standard protocols (Illumina, USA) by collaborating colleagues. DNA library was quantified and

validated using Illumina Eco Realtime (Illumina; USA) and Agilent Bioanalyzer (Agilent Technologies; USA). The library was sequenced in a multiplex pool on a single (150 bp paired-end reads) flowcell on the Miseq System (Illumina, USA). Variant analysis was performed on both this and the NIHR BR-RD PID group.

1.3.3 Variant filtration

Detailed descriptions of bioinformatic analysis is presented in *chapter 5*. A brief description of post-VCF analysis consisted of the following steps. PID cohorts were assessed for the region of *RAG1* and *RAG2*; GRCh37 11:36,587,900-36,621,100. Filtering and prediction of functional consequences was performed using using the following tools and databases:
Variant Effect Predictor
(http://www.ensembl.org/info/docs/tools/vep/index.html),
Exome Variant Server
(http://evs.gs.washington.edu/EVS/),
The Single Nucleotide Polymorphism database
(https://www.ncbi.nlm.nih.gov/projects/SNP/)
and ClinVar
(https://www.ncbi.nlm.nih.gov/clinvar/),
The Exome Aggregation Consortium and The Genome Aggregation Database
(http://gnomad.broadinstitute.org).
Filtering of common variations and annotation was performed using vcfhacks
(https://github.com/gantzgraf/vcfhacks)
and in-house scripts (*chapter 5*). Candidate variants were required to pass the following filtering conditions: frequency (count/coverage) between 20-100%, according to VEP-annotation at least one canonical transcript is affected with one of the following consequence: variants of the coding sequence, frameshift, missense, protein altering, splice acceptor, splice donor, or splice region; an inframe insertion or deletion; a start lost, stop gained, or stop retained, or according to VEP an ExAC frequency unknown, <=0.01, or with clinical significance 'path'.

1.3.4 Cell culture and transfection

CV-1 in Origin with SV40 genes (COS-7) cells (fibroblast-like), from African Green Monkey kidney transformed with a mutant SV40 coding for the wild-type T-antigen, were used for recombination assays. Cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Gibco, USA) supplemented with 10% fetal calf serum (FCS, Gibco, USA) and Penicillin-Streptomycin (Gibco, USA). Cells were seeded at 1.5×10^5 per well (6 well plate) in 1.5 mL culture medium 24 hours prior to transfection. Antibiotic free medium was substituted three hours prior to transfection. Transfected cells were cultured for 42 hours after which, supernatant was removed and plasmid DNA was recovered using a Hirt extraction [87].

1.3.5 RAG expression plasmids

Dr Joan Boyes (University of Leeds) kindly provided full length murine *RAG1*, 3192bp, (and *RAG2* 1581bp) which was cloned into the mammalian expression plasmid pCS2+MT, 4352 bp (Clonetech, Takarabio, USA) shown in **Figure 1.9**. RAG2 was expressed from mammalian expression plasmid pEF-XC, 5509bp [88] and is shown in **Figure 1.10**. Both plasmid and coding gene contained restriction sites for XbaI and XhoI; both of which digest palindromic recognition sequences of 6 bp and create 5' overhangs:

XbaI restriction	*XhoI restriction*
5'...T-CTAGA...3'	5'...C-TCGAG...3'
3'...AGATC-T...5	3'...GAGCT-C...5'

pCS2+ is a vector that allows high-level transient expression in vertebrate cells as well as in vitro transcription/translation whose sequence or clone can be obtained from the I.M.A.G.E. Consortium. It includes a strong enhancer/promoter (simian CMV IE94) followed by a polylinker and the SV40 late polyadenlyation site. The PCS2+MT contains 6x Myc tags that can be used protein purification. Myc sequences of 30 bp contain the human c-Myc oncogene epitope tag. Antibodies for recognition of this tag bind to the 10 residue sequence; Glu-Gln-Lys-Leu-Ile-Ser-Glu-Glu-Asp-Leu. Both

mammalian expression plasmids contained sequences necessary for amplification in *E. coli* and subsequent mammalian expression. *AmpR* consists of a 861 bp sequence for β-lactamase which confers resistance to ampicillin, carbenicillin, and other antibiotics which are used to select for plasmid-harbouring *E. coli*; while an *AmpR* promoter sequence of 105 bp precedes. F1 ori is a 456 bp sequence encoding the f1 bacteriophage origin of replication. Ori is a 589 bp sequence for the high-copy-number origin of replication from the plasmids colE1/pMB1/pBR322/pUC. SV40 poly(A) signal is a 135 bp SV40 polyadenylation signal which promotes post-transcriptional addition of multiple adenine nucleotides to the tail of messenger RNA transcript and generally serves to promote transcript longevity after release of synthesized RNA. M13 rev sequence is included as a common sequencing primer site. A 19 bp SP6 promoter is recognised with high specificity by bacteriophage SP6 DNA-dependent RNA polymerase. This 98.5 kDa polymerase catalyzes in vitro RNA synthesis from a cloned DNA template under the SP6 promoter.

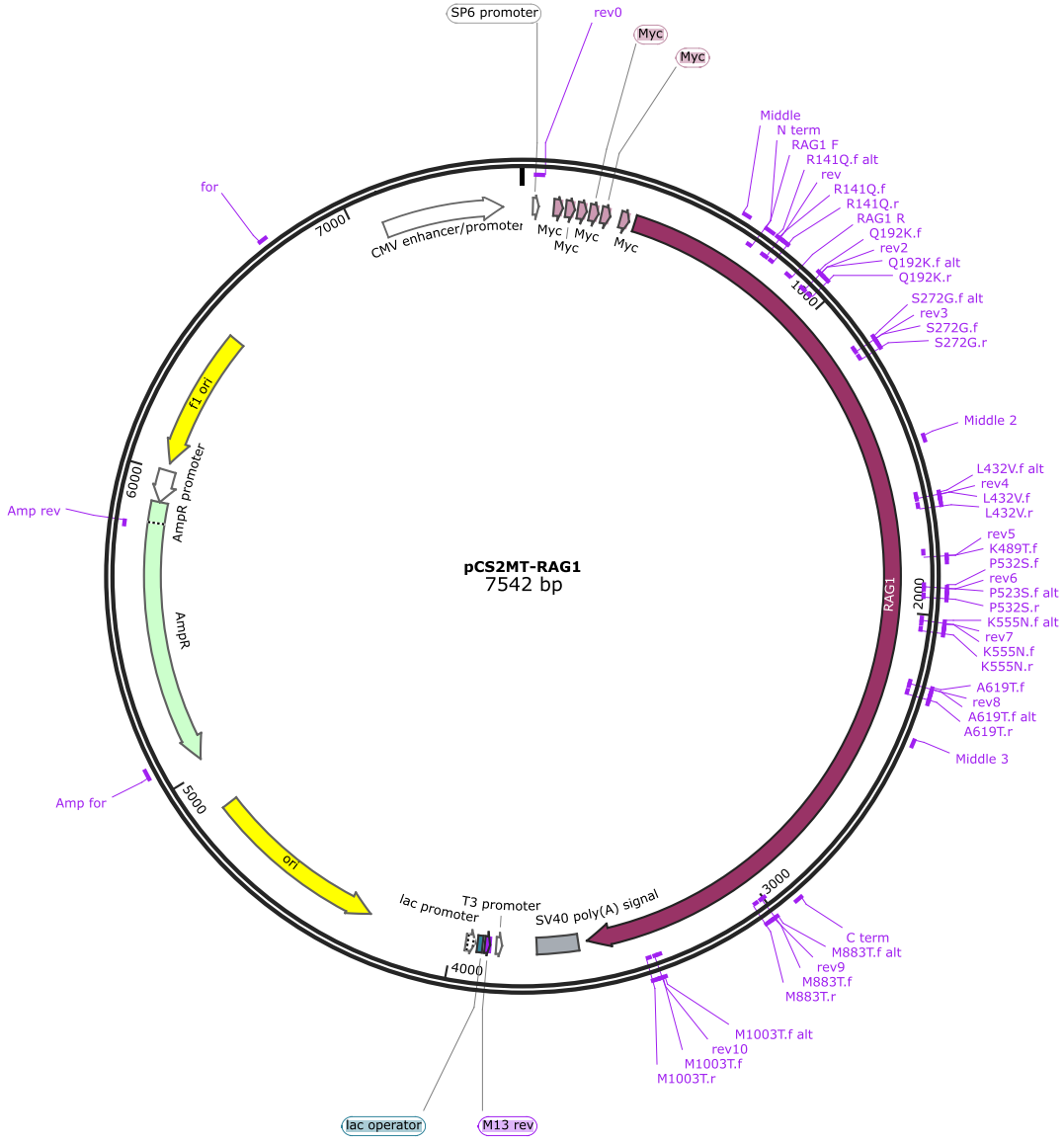


Figure 1.9: **pCS2MT *RAG1* expression plasmid map.** Sites targeted for site directed mutagenesis (SDM) are shown by primer names highlighted in pink. Non-SDM primers on the plasmid backbone were used during PCR and sequencing. pCS2MT-RAG1 contains full-length murine *RAG1* coding sequence. Myc; 30 bp sequence containing human c-Myc oncogene epitope tag, F1 ori; f1 bacteriophage origin of replication, Ori; high- copy-number origin of replication, AmpR; ampicillin antibiotic resistance gene.

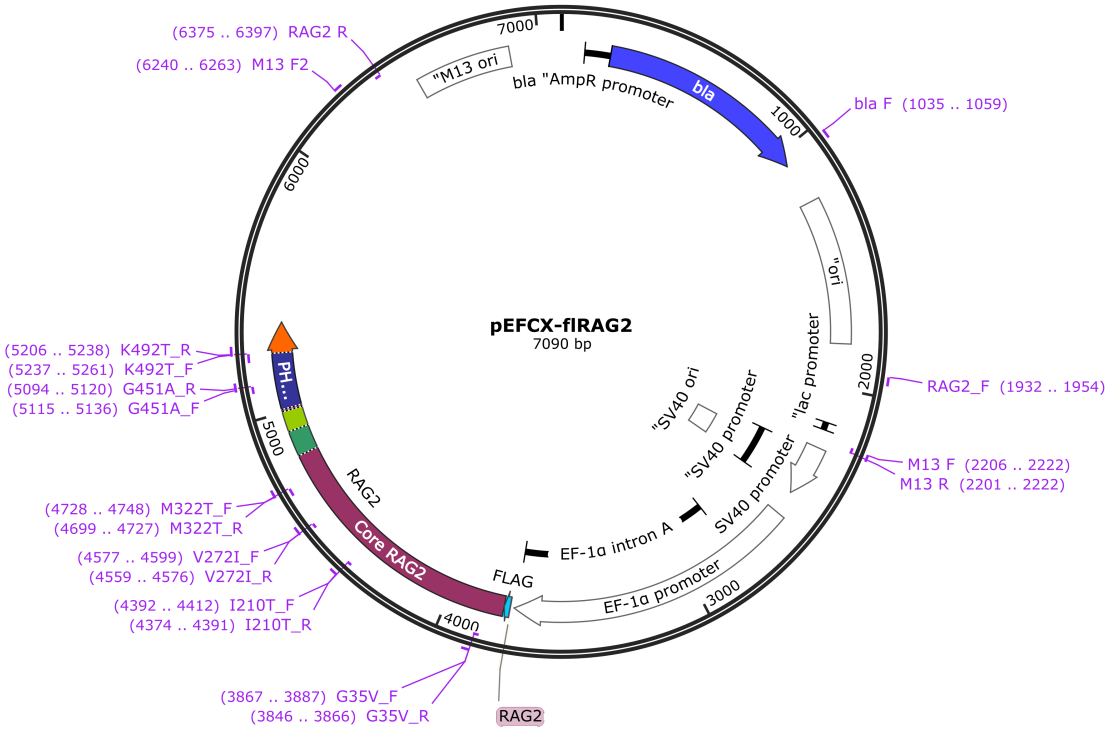


Figure 1.10: **pEFXC-fl*RAG2* expression plasmid map.** Sites targeted for site directed mutagenesis (SDM) are shown by primer names highlighted in pink. Non-SDM primers on the plasmid backbone were used during PCR and sequencing. Full length coding sequence is contained in the core *RAG2* and non-core domains including Hinge and plant homeodomain. Bla; beta-lactamase antibiotic resistance gene, Ori; high-copy-number origin of replication, AmpR; ampicillin antibiotic resistance gene.

A third plasmid was used in each transfection assay as an inversion recombination substrate (6009bp), pJH299 which is derived from pJH298 [89] and is shown in **Figure 1.11**. The plasmid was created from a fusion of pUC13, and an SV40 ori promoter and poly (a) signal. It also contains a CMV enhancer and T7 promoter. The plasmid replicates autonomously in *E. coli* and mammalian cells. The multipurpose cloning site of pUC13 allowed insertion of the sequences required for the recombination assay. The poly adenylation sequence downstream of the 23RSS is bovine growth hormone (bGH). The plasmid also contains antibiotic resistance genes *Amp*R and *neo* for NeoR/KanR. Unique SalI and BamHI sites provided a place for insertion of V(D)J recombination signals, RSS. These signals match what is found in Ig and TCR genes. Insertion primers contained either a 12 nucleotide sequence (39 bp RSS) for insertion at the SalI site, or a 23 nucleotide sequence (49 bp RSS) which could bind the overhang at the BamHI site for insertion. Both RSS contained the consensus versions of heptamers and nonamers. The region targeted by the RAG1/RAG2 complex is a 557 nucleotide inversion sequence

flanked with the 12 and 23 nucleotide RSS (12 and 23 bp spacers with heptamer and nonamer flanking sequences). The two sequences used in this recombination plasmid can be annotated as the following:

Position	*heptamer*	*12 or 23 bp spacer*	*nonamer*
5' 12 RSS	CACAGTG	CTACAGACTGGA	ACAAAAACC
3' 23 RSS	CACAGTG	GTAGTACTCCACTGTCTGGCTGT	ACAAAAACC

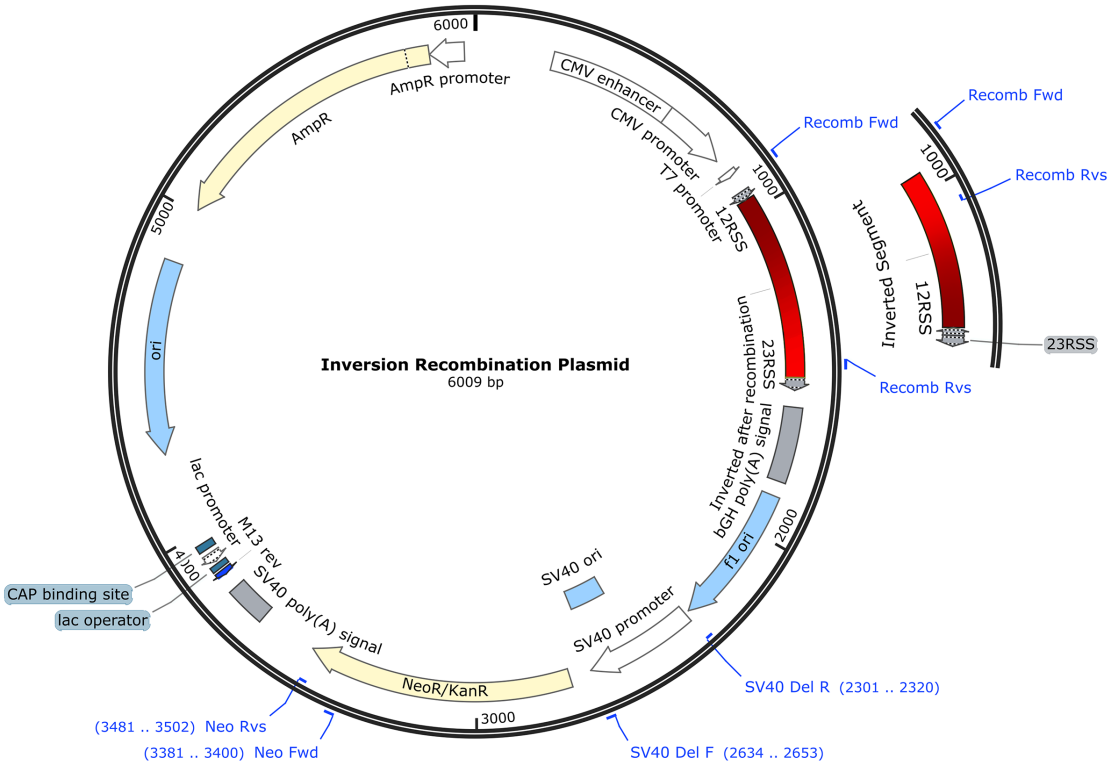


Figure 1.11: **Recombination substrate plasmid map.** The region targeted by the RAG1/RAG2 complex is a 557 nucleotide sequence highlighted in red. It is flanked by RSS (12 and 23 bp spacers with heptamer and nonamer flanking sequences); CACAGTGCTACAGACTGGAACAAAAACC and CACAGTGGTAGTACTCCACTGTCTGGCTGTACAAAAACC, respectively. Wild type RAG protein expression causes in vitro recombination at this site. The cutout illustration shows the resulting configuration (reversed colour red gradient on recombination sequence). The two RSS are left back-to-back. Recombination is measured using the *Recomb Forward* and newly orientated *Recomb Reverse* oligo nucleotide binding site. (Any sites may act as unidirectional primer targets so long as only one is contained within the recombination sequence, which will become inverted through successful recombination.) F1 ori; f1 bacteriophage origin of replication, Ori; high-copy-number origin of replication, AmpR; ampicillin antibiotic resistance gene, NeoR/KanR; *neo* antibiotic resistance gene to neomycin or kanamycin. The primers used for qPCR were (5'-3'): Neo F; TGCTCCTGCCGAGAAAGTATC, Neo R; TTTCGCTTGGTGGTCGAATG, Rec F; ACCCACTGCTTACTGGCTT, Rec R; CAACAGTACTGCGATGAGTGG.

1.3.6 Site directed mutagenesis

SDM was used to produce mutant plasmids using Q5® Site-Directed Mutagenesis Kits (New England Biolabs, UK). Mutagenesis primers used for expression plasmids are listed in Table 1.1. Alternate oligoneucleotide primers are given for variants where SDM proved difficult.

PCR reactions were performed in a total volume of 25 µL as shown in Table 1.2 and using the New England BioLabs Q5 site directed mutagenesis kit (E0554). Table 1.3 lists the thermocycler conditions. The mean recommended manufacturer annealing temperatures were used for each forward and reverse oligonucleotide primer pair. After the SDM PCR was complete, an annealing step was used in the of presence kinase, ligase, and DpnI (KLD) treatment to return plasmids to a closed loop. Table 1.4 lists the KLD reagent concentrations used per reaction. Samples were mixed well and incubated at room temperature for 5 minutes. A transformation procedure was carried out to amplify plasmids in *E. coli* as follows; NEB 5-alpha Competent *E. coli* cells were thawed on ice. 5 µL of the KLD treated products were added to thawed cells. Each tube was mixed by gently flicking 4-5 times (but not vortexed). The mixture was placed on ice for 30 minutes. Cells were heat shocked at 42°C for 30 seconds then placed on ice for 5 minutes. 950 µL of S.O.C medium at room temperature was added into the mixture. Cells were incubated at 37°C for 60 minutes with shaking (250 rpm). Next, the cells were gently mixed by inversion and flicking, and 50-100 µL spread onto an ampicillin (100 µg/mL) selection plate and incubate overnight at 37°C. Plasmid DNA was purified using the QIAprep Spin Miniprep Kit according to manufacturer instructions (Qiagen, cat. 27104) and checked for successful SDM using PCR and Sanger sequencing.

Table 1.1: Site directed mutagenesis primers.

Oligo Name	5'-3' Sequence
RAG1.C term	GGATGAATGGCAACTTTGC
RAG1.middle1	GCGACAAAGCAGTTCACC
RAG1.middle2	ACCACCATGTGTCAAGC
RAG1.middle3	CCATGAGACCCTTACTGC
RAG1.N term	GTCACTCTTGAAACGATTCC
RAG1.rev.f	ACAGCCGGAGATACCCAGTCCACG
RAG1.rev.r	GGTGGATGGAGTCAACATCTGCCT
RAG1.R141Q.f	AGTCTTTTCCAAAAGAAGGAAAAAAGAG
RAG1.R141Q.r	TTGGGTTTTAGCGTCCAC

Table 1.1: Site directed mutagenesis primers.

RAG1.R141Q.f	AAAACCCAAAGTCTTTTCCAAAAGAAGGAAAAAAGAGTC
RAG1.R141Q.r	GACTCTTTTTTCCTTCTTTTGGAAAAGACTTTGGGTTTT
RAG1.Q192K.f	TTCAGCAGTTCCCACAGTAAGGTCTACTTCCCAAGGAAA
RAG1.Q192K.r	TTTCCTTGGGAAGTAGACCTTACTGTGGGAACTGCTGAA
RAG1.Q192K.f	TTCCCACAGTAAGGTCTACTTC
RAG1.Q192K.r	CTGCTGAACTTTCTGTGC
RAG1.S272G.f	GATTCATCTCGGTACCAAGCTTC
RAG1.S272G.r	TTACTGCAGTTGGAGATC
RAG1.S272G.f	TGCAGTAAGATTCATCTCGGTACCAAGCTTCTTGCCGTG
RAG1.S272G.r	CACGGCAAGAAGCTTGGTACCGAGATGAATCTTACTGCA
RAG1.R404W.f	ACTGACGTGGAGGGCGCAGAAA
RAG1.R404W.r	GACAGGAGATGCTGGCGAGG
RAG1.L432V.f	GGCTGTCTGCGTGACATTGTT
RAG1.L432V.r	TTCACATCTCCACCTTCTTC
RAG1.L432V.f	GATGTGAAGGCTGTCTGCGTGACATTGTTTCTCCTGGCA
RAG1.L432V.r	TGCCAGGAGAAACAATGTCACGCAGACAGCCTTCACATC
RAG1.K489T.f	AGGACTGTGAcAGCTATCACTG
RAG1.K489T.r	GTACATCTTATGGTATTGGC
RAG1.R507Q.f	CATGCTCTTCAGAATGCCGAGAAAG
RAG1.R507Q.r	CAAAGGTTGAAAAATCTGCCTCCCAGT
RAG1.P532S.f	CCCTTTGAGTGGCAGCCCTCACTGAAGAATGTGTCCTCC
RAG1.P532S.r	GGAGGACACATTCTTCAGTGAGGGCTGCCACTCAAAGGG
RAG1.P523S.f	GTGGCAGCCCTCACTGAAGAA
RAG1.P523S.r	TCAAAGGGATGGTAGCCTG
RAG1.K555N.f	CCATTGCGAACAGGTTCCGCT
RAG1.K555N.r	TATCTACTGGGTACTCATCC
RAG1.K555N.f	CCAGTAGATACCATTGCGAACAGGTTCCGCTACGACTCT
RAG1.K555N.r	AGAGTCGTAGCGGAACCTGTTCGCAATGGTATCTACTGG
RAG1.A619T.f	CCCGCAGTTCCAGAAAAGACCGTTCGTTTCTCTTTCACA
RAG1.A619T.r	TGTGAAAGAGAAACGAACGGTCTTTTCTGGAACTGCGGG
RAG1.A619T.f	TCCAGAAAAGACCGTTCGTTTCTCTTTCACAGTCATGAG
RAG1.A619T.r	ACTGCGGGCCCACTCCCG
RAG1.M883T.f	AGGGAGCTCACGGACCTTTAC
RAG1.M883T.r	GAGAGCTTCATGCCTCTC
RAG1.M883T.f	GAAGCTCTCAGGGAGCTCACGGACCTTTACCTGAAGATG
RAG1.M883T.r	CATCTTCAGGTAAAGGTCCGTGAGCTCCCTGAGAGCTTC
RAG1.M1003T.f	CAGAAGTTTACGAATGCTCATAAC
RAG1.M1003T.r	GAGGTATTTTGAAGTATACAG
RAG1.M1003T.f	AAATACCTCCAGAAGTTTACGAATGCTCATAACGCGTTA
RAG1.M1003T.r	TAACGCGTTATGAGCATTCGTAAACTTCTGGAGGTATTT
RAG2.G35V.f	GGCCAGAAAGtCTGGCCTAAG
RAG2.G35V.r	AAAGAAGAAAACTTGGCCATC
RAG2.V272I.f	TGATGAATTTaTTATTGTGGGTG
RAG2.V272I.r	TTGTTTGTTTGAGTGAGG
RAG2.M322T.f	GGAAGCAACAcGGGAAACGGG
RAG2.M322T.r	AAACCATATTTTGCTATGCTTAATATCTG
RAG2.I210T.f	CATGTTTCTAcTGCCAGAAAC
RAG2.I210T.r	AAAAGACAGCCCATCCTG
RAG2.K492T.f	GCAAACTCCCAcAAGAAACCCCCCC
RAG2.K492T.r	GCAATGCTCTTGCTATCTGTACATGTTCATTGC
RAG2.G451A.f	CATGGGGATGcGCACTGGGTAC
RAG2.G451A.r	CCCATGAGAACAATAGATCATGGCGGG
RAG2.rev.f	GAGTGAGCTGATACCGCTCGCCG
RAG2.rev.r	GAGAAAATACCGCATCAGGCGCC
RAG2.M13.f	CAGGAAACAGCTATGAC

Table 1.1: Site directed mutagenesis primers.

RAG2.M13.r	GTCATAGCTGTTTCCTGTGTGA
RAG2.M13.f2	CGCCAGGGTTTTCCCAGTCACGAC
RAG2.bla.f	GGTGCCTCACTGATTAAGCATTGGT

Table 1.2: Site directed mutagenesis reagents.

Reagent	25 µL *Rxn*	*Final Conc*
Q5 Hot Start High-Fidelity 2X Master Mix	12.5 µL	1X
10 µM Forward Primer	1.25 µL	0.5 µM
10 µM Reverse Primer	1.25 µL	0.5 µM
Template DNA 1-(25 ng)	1 µL	1-25 ng
Nuclease-free water	9.0 µL	

1.3.7 Transfection

Transfection assays used a combination of wild type or mutant *RAG1* and *RAG2* plasmids to reflect those of patient genotypes. Cells were transfected with wild type or mutant form of murine *RAG1*, *RAG2*, and recombination plasmids using 5uL Lipofectamine 2000 transfection reagent (Invitrogen) and 300 uL Opti-MEM I reduced serum medium (Gibco) per 1.5 mL culture. Serial dilution of wild-type expression plasmids and subsequent transfection experiments identified a suitable range of concentrations to evoke efficient recombination events. Experiments used a total concentration (per 1.5 mL well with 1.5×10^5 cells) 400 ng/µL *RAG1* construct, 200 ng/µL *RAG2* construct, and 1000 ng/µL of inversion recombination substrate. These concentrations provided

Table 1.3: Site directed mutagenesis cycle conditions.

Step	*Temp*	*Time*
1. Initial Denaturation	98°C	30 seconds
2. Denature	98°C	10 seconds
3. Anneal	50–72°C	10 seconds
4. Elongate	72°C	20-30 seconds/kb
Step 2-4 x25 cycles	-	-
Final extend	72°C	2 minutes
Hold	4°C - RT	

Table 1.4: Kinase, ligase and DpnI (KLD) reaction.

	Volume	*Final Conc.*
PCR Product	1 µL	
2X KLD Reaction Buffer	5 µL	1X
10X KLD Enzyme Mix	1 µL	1X
Nuclease-free Water	3 µL	

efficient recombination which was measureable by PCR and quantitative real-time PCR (qPCR) and also typically allowed single mini-preps to produce enough construct for each individual experiment. Experiments testing compound heterozygous mutations used half concentrations of both plasmids for *RAG1* or *RAG2*; i.e.,
Mutation A: 400 ng/µL $RAG1_{WT}$, 200 ng/µL $RAG2_{mutA}$, 1000 ng/µL RSS substrate.
Mutation B: 400 ng/µL $RAG1_{WT}$, 200 ng/µL $RAG2_{mutB}$, 1000 ng/µL RSS substrate.
Compound: 400 ng/µL $RAG1_{WT}$, 100 ng/µL $RAG2_{mutA}$, 100 ng/µL $RAG2_{mutB}$, 1000 ng/µL RSS substrate.

1.3.8 Recombination assay

Triple-transfection of WT/mutant *RAG1*/*RAG2* and a recombination substrate was used to assess the functional activity matching that of patients with homozygous or compound heterozygous genotypes. Transfection assays used a combination of *RAG1* (400 ng/uL) co-transfected with *RAG2* (200 ng/uL). Homozygous expression of a mutant gene was mimicked by co-transfecting a *RAG1* mutant with wild type *RAG2* and vice versa. To mimic compound heterozygous genotypes two equal half concentrations of mutant *RAG1* plasmids were co-transfected with wild type *RAG2* and vice versa (i.e.200ng/uL WT *RAG1*, 200ng/uL mutant *RAG1*, and 200ng/uL WT *RAG2*).

To measure the activity of these wild type or mutant proteins a third (or fourth in compound heterozygous instances) construct was used as an inversion recombination substrate (1000 ng/µL). The DNA sites targeted on this recombination substrate by RAG1/RAG2 complex are 12 and 23 nucleotide RSS flanking a 557 nucleotide inversion

sequence (**Figure 1.12** and also depicted in **Figure 1.11**).

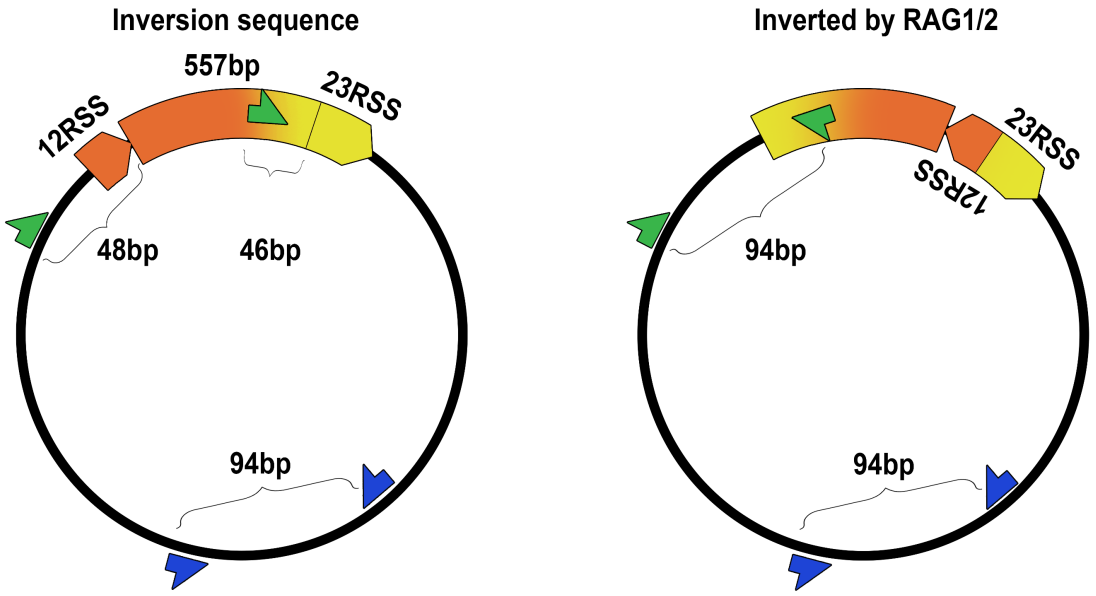


Figure 1.12: **The recombination substrate containing RSS sequences.** (**Left**) The sequence enclosed between the beginning of 12 recombination signal sequences (RSS) up to the beginning of 23 RSS is targeted by RAG1/RAG2 for recombination (inclusive of the 12RSS and 577 bp inversion sequence). (**Right**) Successful inversions in cell culture are measured by production of a 94 bp fragment during qPCR. Primer sites used to measure recombination activity are indicated by green arrows. A second 94 bp sequence on the substrate, indicated with blue arrows, is not affected by RAG activity and is used during qPCR for normalisation. Comparative CT ($\Delta\Delta$CT) measures the recombination activity of mutant protein as a percentage of wild type.

Successful recombination events, represented by a reversal of inversion sequence, were assessed by quantitative real-time PCR (qPCR) using comparative CT (delta delta CT). qPCR primer sites were selected on the recombination substrate plasmid at 48 bp upstream of the 12RSS, prior to the inversion sequence, and a primer site 46 upstream of the of the 23RSS, laying internally on the inversion sequence, with both in the forward direction. A successful recombination event resulted in the reverse of inversion sequence and allowing amplification of a 94 bp product. A second sequence of 94 bp on the recombination plasmid backbone, which is not affected by RAG activity, was used as a reference to calculate $\Delta\Delta$CT values and assess relative recombination activity. The relative recombination activity is measured against wild type RAG1/RAG2 to calculate the activity % of WT with mean ± SEM.

Gene expression of mutant *RAG1/RAG2* was assessed by qPCR. Protein expression was not assessed for each SNV mutant here, although in parallel, the same system was used to separately study the in vivo mechanisms of RAG deficiency. In this case, protein expression was assayed by Daniel Thwaites (University of Leeds) and is reported in detail in Thwaites et al. [90].

1.3.9 Quantitative real-time PCR

Successful recombination events were assessed by quantitative real-time PCR (qPCR) using comparative CT (delta delta CT). To recover recombination plasmid a modified Hirt's cell lysis extraction for low molecular weight DNA was performed before a phenol chloroform extraction (Thermo Fischer Scientific, 17909) and ethanol precipitation. DNA was diluted 50 times and Fast SYBR® Green Master Mix was used for qPCR. Experiments were performed on a QuantStudio 5 Real-Time PCR System (Thermo Fischer Scientific, 4385610 and A28573).

1.3.10 Laboratory evaluation of immune phenotypes

Lymphocyte panel and immunoglobulin levels are briefly referred to in this chapter for discussion of patients' clinical features. These findings were kindly provided by Dr. Jolan Walter through clinical laboratory testing. Anti-cytokine antibodies were detected by Enzyme-Linked Immunosorbent Assay (ELISA) as previously described [68].

1.4 Results

1.4.1 Whole genome sequencing and the prevalence of RAG deficiency

This chapter investigated the prevalence of RAG deficiency in patients who have been diagnosed with a primary immunodeficiency due to an unknown genetic determinant. Patients in this study were recruited from two European projects:

1. **NIHR BR-RD PID (UK)** and
2. **the Vienna PID cohort (Austria).**

1. The majority of patients were from the UK cohort as part of the NIHR BR-RD PID study with 558 cases of antibody deficiency. The specific age groups are listed in table 1.5.
2. To increase the scope of this project a collaboration was made with Dr Jolan Walter (US) and Dr. Christoph Geier (Austria). Genetic data from patients in Vienna, Austria was collected by Dr. Christoph Geier and analysed along with clinical data in collaboration. The Vienna-sourced cohort consisted of 134 patients whose sequencing data was assessed together with that of NIHR BR-RR PID.

These two cohorts were combined to make up the joint study of 692 European PID patients. The complete cohort count is shown in **table 1.5**. **Figure 1.13** illustrates the NIHR BR-RD PID cohort based on age of presentation and phenotype. Less phenotypic data was available from the Austrian cohort of patients and is therefore not included in this figure.

The canonical regions of *RAG1* and *RAG2* were analysed for functional variants in 692 PID patients. The bioinformatic approach is discussed in detail in *chapter 5*.

Table 1.5: Number of PID patients in the cohort. Two PID patient cohorts from the UK and Austria were combined to identify cases of RAG deficiency. In the NIHR BR-RD PID cohort, 124 related, healthy family members were also included for segregation analysis to potentially identify cis-inherited alleles.

Cohort source		Number of individuals
NIHR BR-RD	adult cases	299
	child cases	188
	unconfirmed age	71
	total cases	**558**
Vienna cohort	adult cases	106
	child cases	28
	total cases	**134**
Combined PID cohort		**692**

The combined cohort contains many variations in phenotypic description. While a detailed record facilitates the study of disease mechanism, it impedes a blind analysis approach. Therefore, patients were briefly categorised based on their lowest denominator phenotype. Following recent classification proposals [92, 93] the cohort included patients with the phenotypes listed in table 1.6.

Table 1.6: PID patient phenotypes. Some patients could only by classified as having some antibody deficiency and are labelled as "other". CID (combined immunodeficiency), CVID (common variable immunodeficiency), PAD (primary antibody disorders (including hypogammaglobinemia, selective PAD (SPAD) and agammaglobulinemia)), SPAD (Specific polysaccharide antibody deficiency).

Phenotype	UK	Austria	Total
CVID	305	57	362
CID	101	36	137
PAD	78	41	119
Other	74		74
Total	558	134	692

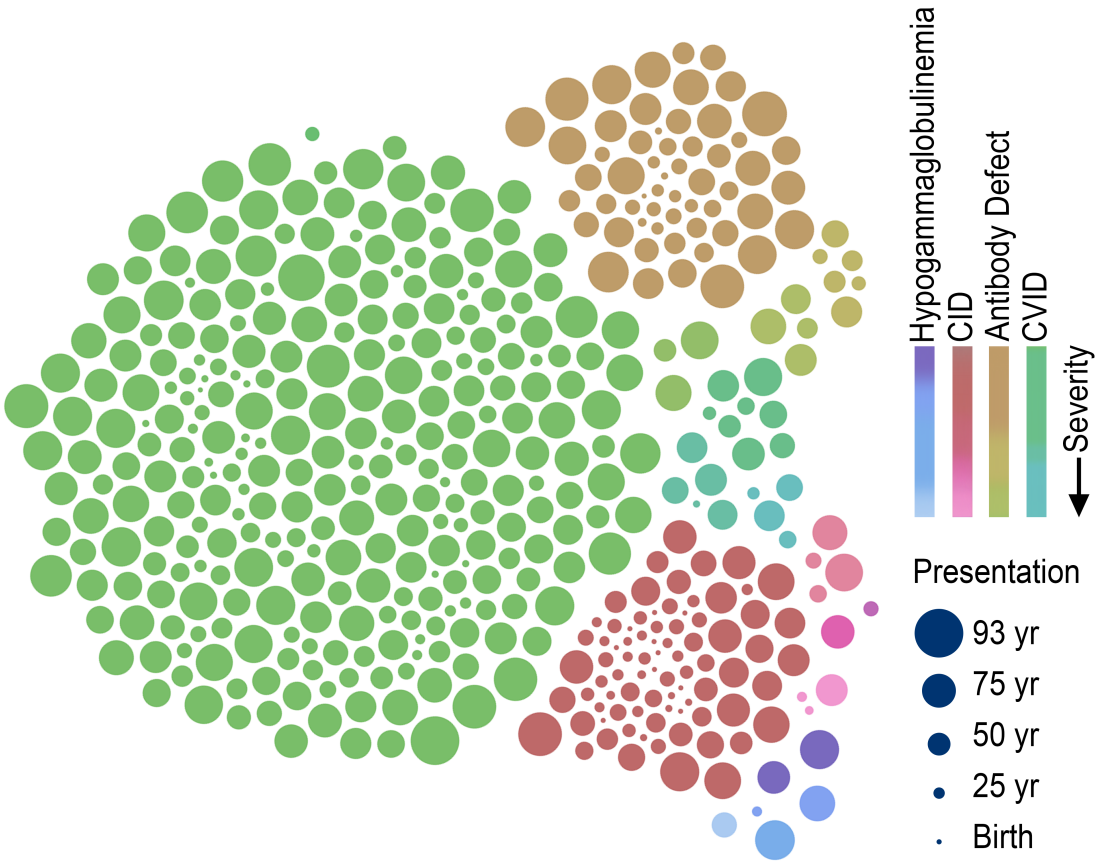


Figure 1.13: **The NIHR BR-RD PID cohort.** 558 unrelated patients were investigated by whole genome sequencing and targeted analysis was performed on *RAG1* and *RAG2*. Each dot represents one genome. The size of dots indicates age (ranging from 1 to 93 years old). Specific phenotypes are illustrated by colour and severity, based on descriptive enrolment data, is represented by colour gradient. Forced cluster using the software "Raw" [91].

After routine analysis steps, *RAG1* and *RAG2* for these patients contained 335 known SNPs and 199 unknown variants. 237 variants were filtered and 297 variants were retained for tailored investigation. To illustrate the number of variants, a visualisation of the unfiltered variant call format from *RAG1-RAG2* of roughly half of the NIRH-BR PID cohort is shown in **Figure 1.14**. Each blue mark represents one called variant; each row down the y-axis representing individuals. Only coding variants in *RAG1* and *RAG2* were targeted; variants identified in non-coding regions (+/− 5-nucleotides from intronic spice sites) were either not expected to affect protein function, or could not be functionally assessed for their effect on protein expression in this study. There were no instances

Table 1.7: *RAG1* and *RAG2* coding variants of interest identified in patients with primary immunodeficiency.

RAG1		
Amino acid	**cDNA**	**Count**
p.R142Q	c.425 G>A	1
p.E193K	c.577 G>A	2
p.Q242R	c.725 A>G	4
p.H249R	c.746 A>G	38
p.S275N	c.824 G>A	1
p.D302E	c.906 C>A	6
p.M435V	c.1303 A>G	2
p.R449K	c.1346 G>A	26
p.K492T	c.1475 A>C	1
p.P525S	c.1573 C>T	2
p.K558N	c.1674 G>C	2
p.A622T	c.1864 G>A	1
p.K820R	c.2459 A>G	8
p.E880K	c.2638 G>A	7
p.M886T	c.2657 T>C	1
p.D887N	c.2659 G>A	2
p.M1006V	c.3016 A>G	5

RAG2		
Amino acid	**cDNA**	**Count**
p.M5K	c.14 T>A	1
p.V8I	c.22 G>A	11
p.R123C	c.367 C>T	1
p.T215I	c.644 C>T	2
p.I210T	c.629 T>C	1
p.V238I	c.712 G>A	1
p.V272I	c.814 G>A	1
p.E293G	c.878 A>G	6
p.D310N	c.928 G>A	1
p.F386L	c.1158 C>A	27
p.D400H	c.1198 G>C	1
p.G451A	c.1352 G>C	2
p.K498*	c.1492 A>T	1
p.M502V	c.1504 A>G	5
p.R506C	c.1516 C>T	1

where a patient had a damaging mutation in both *RAG1* and *RAG2*. The initial set of coding variants of interest are listed in **Table 1.7**.

Samples in this study were assessed by either exome or whole genome sequencing. An analysis plan focusing on interpretable or actionable results was required. Confirmatory functional analysis was only possible for variants that produced a protein coding change (substitutions, insertions, deletions, splicing variants). Non-coding variants were only assessed when occurring in splice regions (+5 / -5 bp of splice junction). Non-coding variants outside of splice regions such as promoter sites, introns, etc. could not be assessed in this study. No splice variants were identified. Variants were of interest when found as homozygous, compound heterozygous, or had the potential to produce a damaging effect by other means. Although candidate heterozygous variants were found, the potential for monoallelic causes of RAG deficiency would require separate study. No patients with a candidate heterozygous variant had any potential pathogenic secondary variants that could be interpreted or functionally validated in this study. The probability of

loss-of-function intolerance in both genes is zero indicating that haploinsufficiency does not cause a selective disadvantage. However, functional variants are still quite infrequent in these genes. See *chapter 5* for detailed discussion. Therefore, segregating disease and control groups based on potentially accumulative, compound heterozygous variants was not a major difficulty. For genes where variant frequency (and population-based allele frequencies of those variants) is high, identifying association can become challenging.

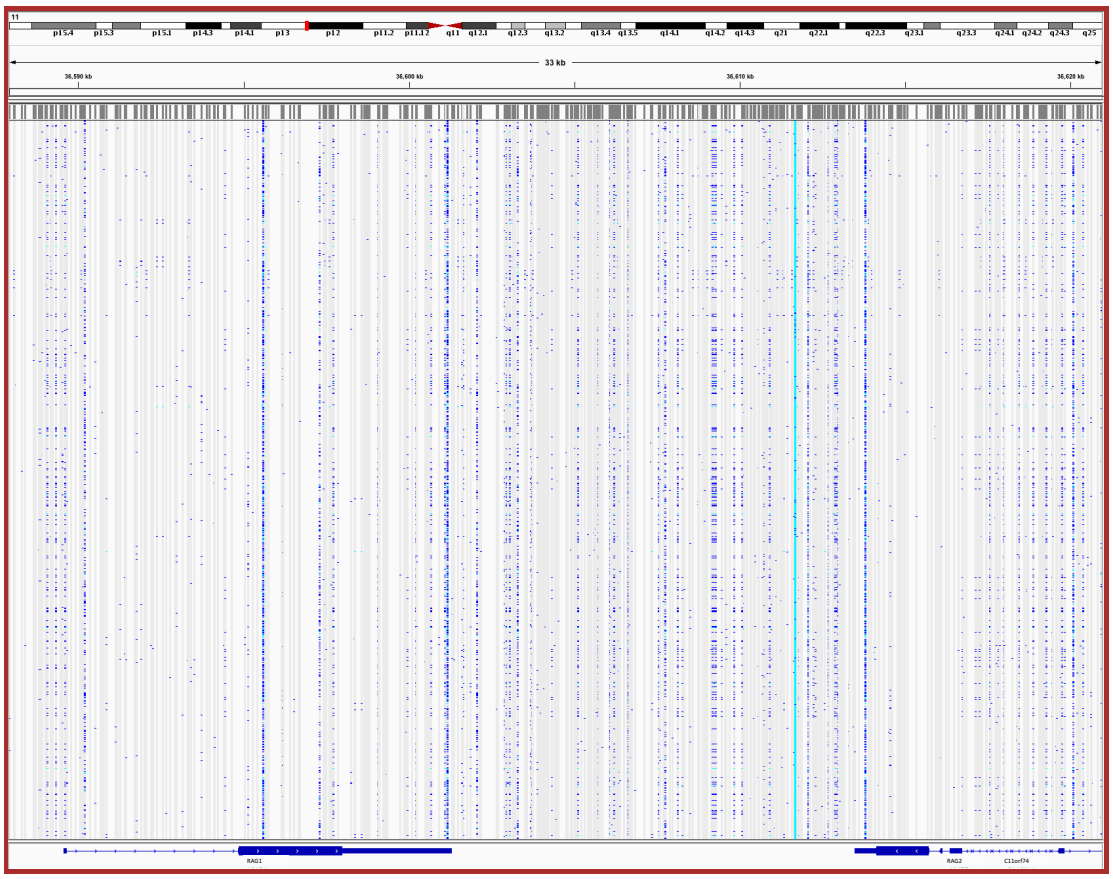


Figure 1.14: **View of NIHR BR-RD PID cohort VCF** View of NIHR BR-RD PID cohort VCF. A variant call format (VCF) file containing variant calls including SNPs, indels, and genomic rearrangements. At the top, chromosomal coordinates are provided relative to GrCh37. On the bottom, the genomic structures of *RAG1* and *RAG2* are shown. Vertically, each horizontal row present variants for individual patients. Blue dots indicate a called variant; minor allele frequency/fraction is known from annotation and genotype data. Common SNPs can be seen as blue vertical lines, created when the majority of patients harbour a polymorphism at that site. The VCF is generally an endpoint for routine bioinformatic pipelines and the start-point for tailored investigations.

Figure 1.15 illustrates the case of two patients with compound heterozygous variants

in *RAG1* (M435V and M1006V), amongst a background of benign or potentially damaging variants in other patients, some of whom also harbour heterozygous M1006V. The challenge of interpreting large datasets to uncover complex associations is discussed in Chapter 5.

Four patients were identified from the UK arm of this study and functional analysis was pursued to confirm RAG deficiency. One of these four (patient 16) was identified with homozygous RAG2 p.L492T, however this variant had been identified simultaneously by NIHR BR-PID primary investigators and ultimately excluded from the resulting publication and no additional clinical data was collected for use in Section 1.4.3. Other known causes of PID were also excluded in the remaining three patients. Two patients from the Austrian arm of the study were identified as potentially RAG deficient cases, which were also followed up functionally. A total five newly identified cases of RAG deficiency are reported in this study in detail. Based on these findings (3 cases from 299 adults and 2 cases from 106 adults), the prevalence of RAG deficiency in adult PID can be estimated as ranging from 1%-1.9%. For all adult PID patients that are currently registered with the UK Primary immunodeficiency network database (3,294 patients over age of 18 years), this estimate means that an additional 32-63 (+/- 1) cases of RAG deficiency are expected.

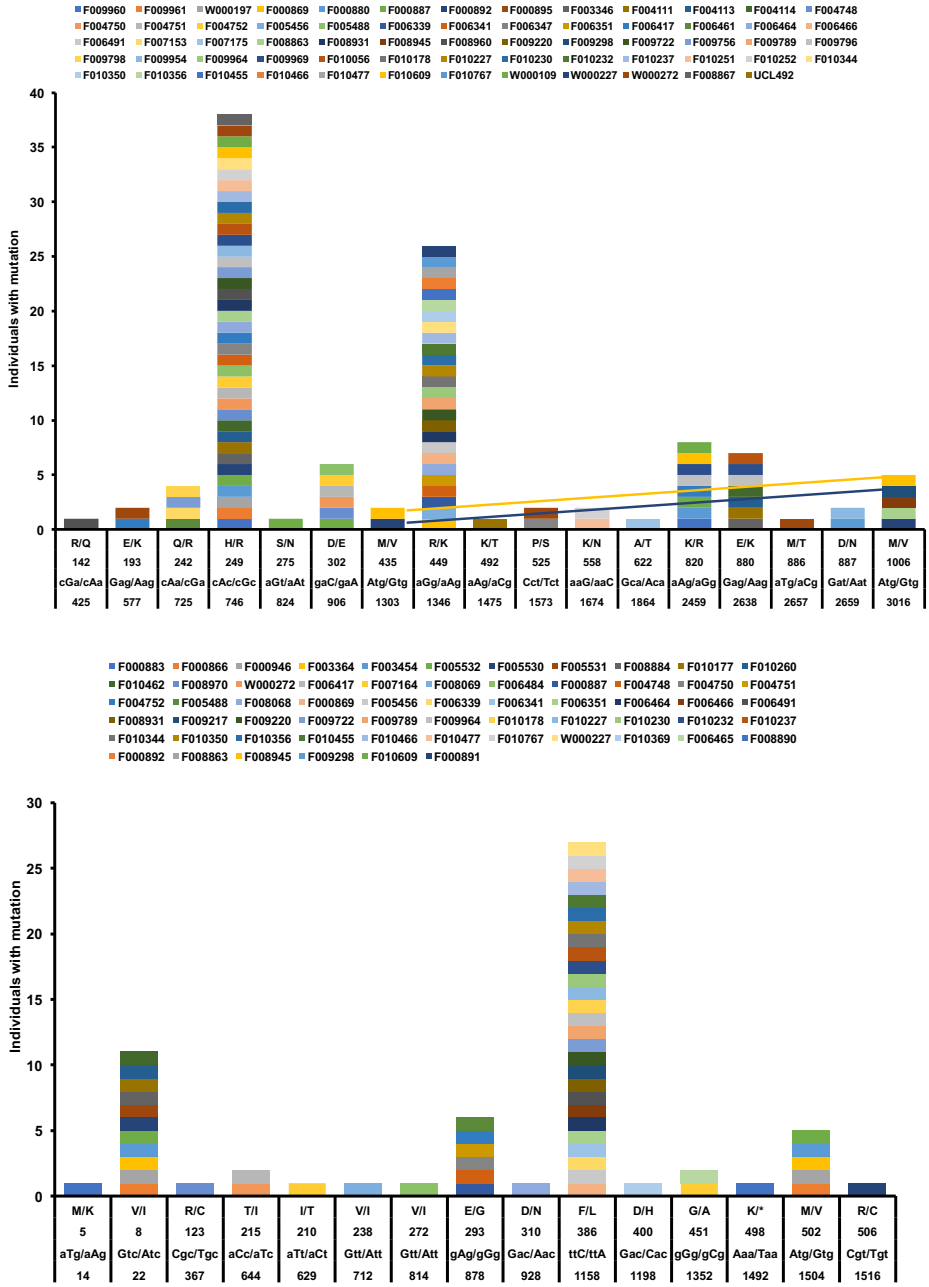


Figure 1.15: **Identifying NIHR BR-PID compound heterzygous cases.** Coding variants that were identified and of potential interest are shown for *RAG1* and *RAG2*. Patients harbouring each variant have colour-coded IDs and are stacked to indicate variant frequency. Two patients had compound heterozygous *RAG1* p.M435V and p.M1006V. Identifying more complex variant phasing or larger datasets (>500 patients) requires alternative analysis methods discussed in *chapter 5*.

During analysis and functional investigation, ten additional cases were added to this study with reference to the referring physician and acknowledgement where clinical data was sourced collaboratively. These ten patients were not identified as part of the bioinformatic investigation but had been identified by genetic sequencing, case by case. All 17 cases had phenotypic data, and some had clinically-sourced data available from their referring physicians, which contributed to the project. **Figure 1.16** illustrates the gene structures of *RAG1* and *RAG2* with all validated, pathogenic variants investigated in this study. For reference, previously reported variants are also indicated as small blue dots on the structure.

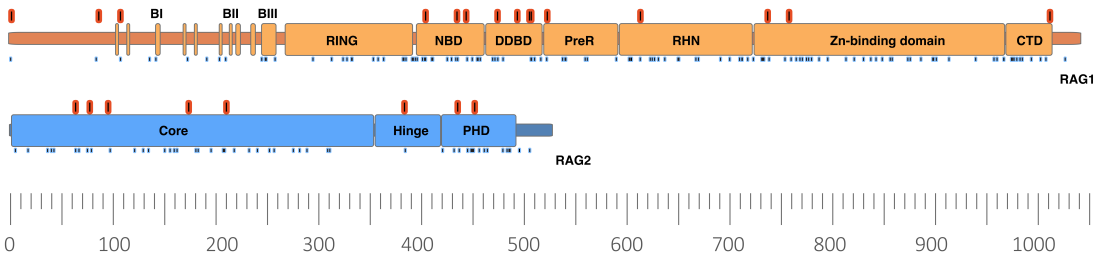


Figure 1.16: **Distribution of variants that cause RAG deficiency.** Schematic representation of RAG1 and RAG2 adapted from Notarangelo et al. [94]. Amino acid positions are shown on the scale bar. Variants in this cohort (17 in RAG1 and 8 in RAG2) are illustrated in red. Known pathogenic variants previously reviewed are shown as blue dots [94]. RAG1 protein consists of 1043 amino acids. Catalytic core contains the nonamer-binding domain (NBD; amino acids 394–460), dimerisation and DNA-binding domain (DDBD; amino acids 461–517), pre-RNase H (preR; amino acids 518–590), catalytic RNase H (RNH; amino acids 591–721), zinc-binding domain (ZnBD; amino acids 722–965), carboxy-terminal domain (CTD; amino acids 966–1,008). RAG2 protein is composed of 527 amino acids. Core domain (amino acids 1–383) and the non-core region (amino acids 384–527) which includes acidic hinge region (Hinge; amino acids 350 – 410), plant homeodomain (PHD; amino acids 414–487).

1.4.2 Functional characterisation of novel RAG variants

The recombination activity of mutant and wild-type RAG1 and RAG2 proteins normally required for catalyzing V(D)J recombination events are shown in **Figures 1.17** and **1.18**. In addition to the method previously described, [95] a system was employed in this chapter to measure recombination activity in compound heterozygous cases by in vitro expression of murine RAG1 and RAG2. Over half of the mutant proteins tested show almost complete loss of activity. All patients' variants tested had an overall low combined RAG activity (6.4-28%). Eighteen variants were assessed with in vitro expression of murine RAG1 and RAG2 in COS7 cells. The relative recombination activity was measured against wild type RAG1/RAG2 to calculate the activity % of WT with mean ± SEM shown in **Figures 1.17** and **1.18**. Several variants were assessed as previously described [95] as part of another study on other predicted pathogenic variants. Both systems simulate the efficiency of protein expressed in patients in their ability to produce a diverse repertoire of TCR and BCR coding for immunoglobulins. The mutations found in patients 8-10 have also been tested by Lee et al. [95] and show very similar levels of recombination activity for individual mutations [95]. Patients 2, 4, 8, 9, and 15 all carry one allele with mutations that individually do not indicate any major loss of function. However, the assay used here also had the ability to measure recombination activity in compound heterozygous cases and found a striking decrease in a compound heterozygous state (Fig. 1.17 and 1.18).

The cases reported in this chapter have been discussed in Lawless et al. [1] with the exception of patients 16 and 17. These were excluded from our publication to prevent duplicate reports since our collaborators were also preparing manuscripts that included these cases. Patient 16 was identified through analysis of the NIHR BR-PID cohort with homozygous RAG1 p.Lys492Thr. Functional analysis found 23.3% recombination activity for this variant. This patient received a HPSC transplant at Great Ormond Street Hospital. Because this variant was also identified by the NIHR BR-PID group the patient was not included in the resulting publication, nor was any further clinical data sought. Similarly, patient 17 had been identified with three potentially pathogenic *RAG2* variants by a collaborating group and was investigated functionally as part of this study. However,

this patient was not included in the publication, nor was further clinical data collected. The patient was a 19 year old with atypical CID who had a bone marrow transplant after receiving methotrexate with Ig replacement. Panhypogammagloubulinaemia and T cell lymphopenia with cutaneous granulomas were reported by our collaborator for this person. The variants tested functionally had recombination activities of: p.Gly35Val - 25.4%, p.Val272Ile - 81.7% (considered benign), p.Met322Thr - 52.4%, compound - 33.7% (Gly35Val and Met322Thr).

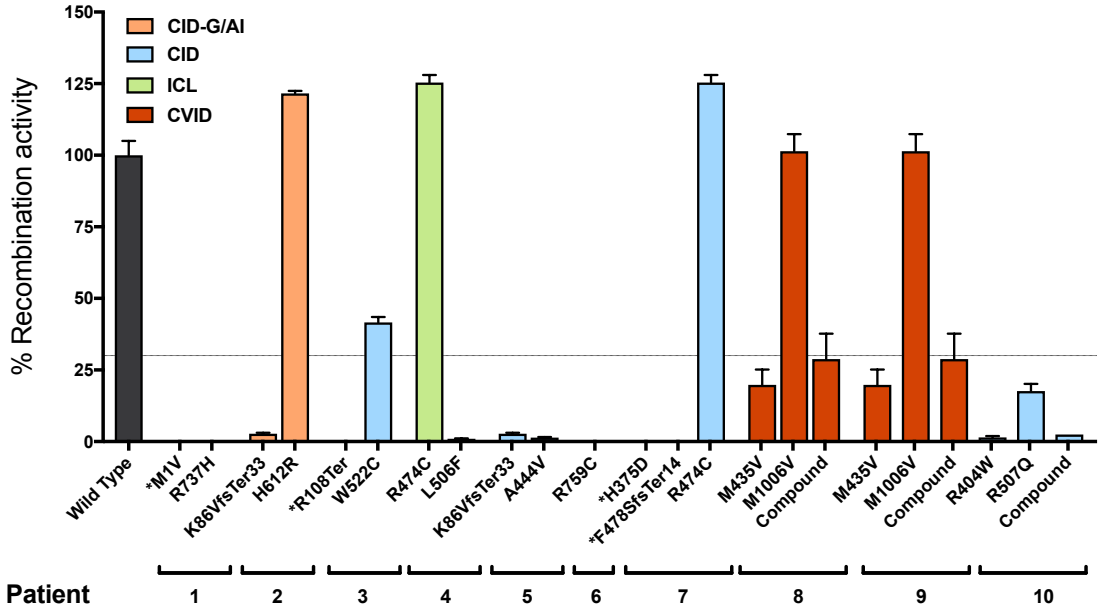


Figure 1.17: **Recombination activity of RAG1 mutants in vitro.** Expression of WT or mutant forms of RAG1 in complex with wild type RAG2 in vitro. A recombination substrate containing an RSS mimics TCR, or Ig locus genes. Expression plasmids were transfected to represent a wild type, homozygous, or compound heterozygous genotype as found in patients. Procedure detailed in Section 1.3.7. Successful recombination events assessed by qPCR (delta delta CT). * Residues annotated with an asterisk were assessed using an in-vivo mouse model by Lee et al. [95] and therefore we not tested in compound heterozygous states in the recombination assay.

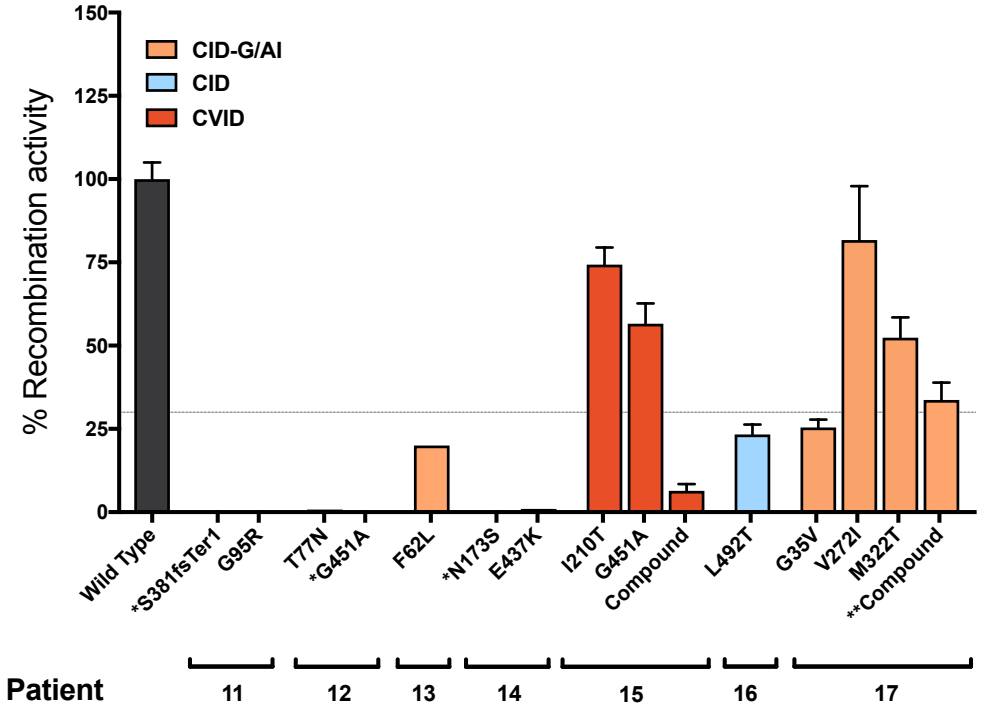


Figure 1.18: **Recombination activity of RAG2 mutants in vitro.** Expression of WT or mutant forms of RAG2 in complex with wild type RAG1 in vitro. A recombination substrate containing an RSS mimics TCR, or Ig locus genes. Expression plasmids were transfected to represent a wild type, homozygous, or compound heterozygous genotype as found in patients. Procedure detailed in Section 1.3.7. Successful recombination events assessed by qPCR (delta delta CT). *Assessed using an in-vivo mouse model by Lee et al. [95] and not tested in compound heterozygous states. ** For case seventeen RAG2 V272I was considered benign and not included in the compound heterozygous state assay.

1.4.3 Clinical collaboration

1.4.3.1 Immune phenotypes

In addition to the five newly identified cases of RAG deficiency, clinical data was collected for ten additional cases (>15 years of age) by Dr Christoph Geier and Dr Jolan Walter. This data was assessed collaboratively and is included here to provide the plenitude to accurately describe RAG deficiency in adult PID. The immune phenotypes and clinical diagnoses are shown in **Figure 1.19**. Of the seventeen patients described here the median age is 37 years (15-73 years), with five patients already deceased at ages 15, 22, 25, 37, and 43. There are nine female (60%) and six male (40%) patients. Clinical phenotype was predominantly CID (n=9, 60%), of whom three had clinical history of autoimmunity and/or the presence of granulomas, followed by SPAD (3 patients, 20%), CVID (2 patients, 13%), and a single case of idiopathic CD4+ lymphopenia (ICL) (7%) (**Figure 1.19**). Most had late presentation. Although recurrent infections and lung disease were commonly seen in adolescence, severe disease and PID diagnosis generally occurred in adulthood. Persistently low IgG and/or low IgA and IgM levels are seen in approximately 50% of cases (**Table 1.8**). The dominant laboratory features were naïve CD4+ T cell lymphopenia with low absolute number and fraction of naïve CD4 cells (CD4+CD45RA+), and B cell counts were variably low (**Table 1.8**). Enzyme-linked immunoassay was tested by Dr. Jolan Walter for anti-cytokine antibodies (targeting IFNα, ω and IL-12) on plasma from seven patients (not shown). Four patients were positive which is comparable to a previous report (56%) [68].

Although patients identified by whole genome sequencing were primarily diagnosed with antibody deficiency, closer examination of the T cell compartment revealed low absolute number of naïve CD4 fraction (CD4+CD45RA+) suggestive of a CID phenotype (**Table 1.8**).

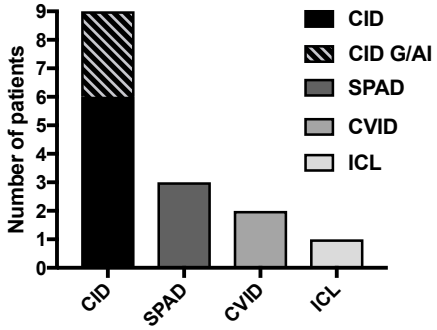


Figure 1.19: **Clinical phenotypes.** The clinical phenotypes of patients 1-15 consisted of CID (combined immunodeficiency), CID-G/AI (combined immunodeficiency with granuloma autoimmunity), SPAD (selective polysaccharide antibody), CVID (common variable immunodeficiency), and ICL (idiopathic CD4 T cell lymphopenia).

Table 1.8: Relative immunoglobulin count. Most of this cohort (73%) demonstrates low serum immunoglobulins. 14-27% have normal or borderline low Ig measurements respectively. Low IgG subtypes (not shown) occur at roughly the same rate in this cohort where measured.

ID	IgG	IgA	IgM	IgE
Patient 1	Δ Normal	Normal	Normal	High
Patient 2	Δ Normal	Low	Low	High
Patient 3	Δ Normal	Low	Low	Normal
Patient 4	Δ Normal	Normal	Normal	Normal
Patient 5	Δ Normal	Normal	Normal	Normal
Patient 6	Low	Low	Low	Normal
Patient 7	Δ Low*	Low	Low	Normal
Patient 8	Low	Low	Low	Normal
Patient 9	Low	Low	Low	Normal
Patient 10	n/a***	Low	Low	-
Patient 11	Low	Low	Low	-
Patient 12	Low	Normal	Δ Low	-
Patient 13	Low	Δ Low	Δ Low	-
Patient 14	Δ Normal	Low	Low	Normal
Patient 15	Low	Variable*	Normal**	-

1.4.3.2 Autoimmune complications

Inflammatory autoimmune complications developed in 87% of patients (**Figure 1.20**). Organ-specific manifestations were the most common autoimmune complications affecting 73% of this adult cohort (**Figure 1.21**) similar to previously described reports with 48-77% [68, 93]. Gastrointestinal complications were the most prevalent organ specific manifestation followed by dermatological manifestations (**Figure 1.22**). Granulomatous disease was seen in 40% of patients, with five out of six patients showing granuloma localisation within interstitial lung tissue. Other complications included myopathies (14%), endocrine abnormalities, sarcoidosis, and polyarthritis was seen in one patient (**Figure 1.21**). Cytopenias occurred in 40% of adult RAG deficient patients, similar to recently reported cohorts (21-77%) [68, 93]. Autoimmune hemolytic anaemia was the most frequent (27%) followed by immune thrombocytopenic purpura (20%), and autoimmune neutropenia in one patient (**Figure 1.23**). Further studies may determine the underlining pathophysiology that drives autoimmunity in RAG deficient patients. 57% of the patients developed antibodies to cytokines, which may serve as a potential biomarker for adults with PID due to RAG1 and/or RAG2 mutations [68] (**Table 1.8**). It was recently demonstrated, that RAG deficient patients show a restriction of Treg repertoire diversity and a molecular signature of self-reactive conventional CD4+ T cells [96].

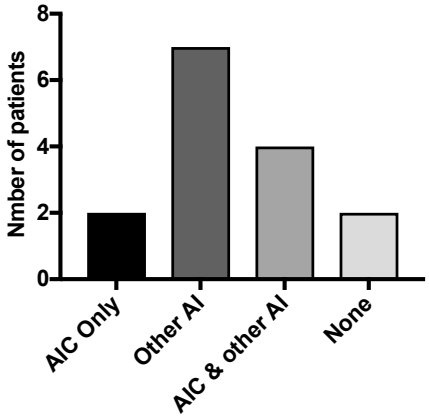


Figure 1.20: **Autoimmunity.** Overall frequency of autoimmune complications in adult patients with RAG deficiency. Autoimmune occurrence in patients 1-15 consisted of AI (autoimmune features), AIC (autoimmune cytopenia), or none.

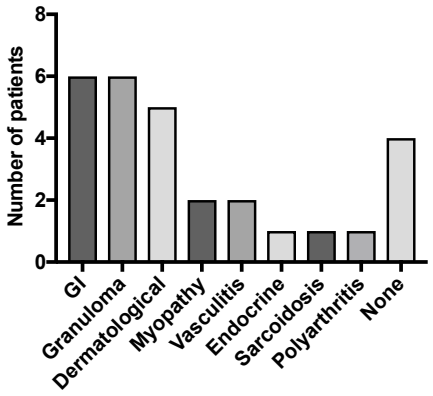


Figure 1.21: **Organ-specific manifestation.** The frequency of organ-specific autoinflammatory features are shown for 11 patients. Gastrointestinal includes duodenitits, collagenous colitis, coeliac disease, enteropathy, gastritis; Dermatological includes alopecia, psoriasis, dyskeratosis, vitiligo; Myopathy includes myopathy, myoatrophy, myasthenia gravis, myofasciitis; Endocrine include hypogonadism, hypothyroidism.

1.4.3.3 Pulmonary disease in adult RAG deficiency

Progressive pulmonary disease was prominent in the cohort of adults with RAG deficiency and was the leading cause of mortality (93%). Of seventeen unrelated adult patients recruited, the median age at onset of lung disease was 11 years. The five patients (33%) that were deceased at the time of the study had a median age of 23.5 years. A diverse spectrum of pulmonary manifestations were observed; pneumonia being the most common, followed by bronchiectasis, chronic bronchitis, granuloma, fibrosis, chronic obstructive pulmonary disease, and bronchiolitis. Clinical symptoms persisted for an average of 13 years. Of the five deceased patients two cases were due to progressive lung disease with pulmonary fibrosis. Two more patients died due to infections post-transplant. One patient died due to PML caused by John Cunningham virus infection. There was no significant difference in overall survival between patients presenting with pneumonia, bronchiectasis or granuloma, fibrosis, or chronic obstructive pulmonary disease (**Figure 1.24**). High-resolution computed tomography imaging of lung revealed bronchiectasis and granuloma. Histology of lung biopsies (patient 1 and 2) revealed atypical lymphoid hyperplasia with granulomatous features and giant cells formation (**Figure 1.24**) (Figure credit of Dr C Geier). Germinal center formations in patient 2 were comprised of CD3+ T-cells and CD20+ B-cells. Patient 1 had peribronchial fibrosis. Retrospective analysis of

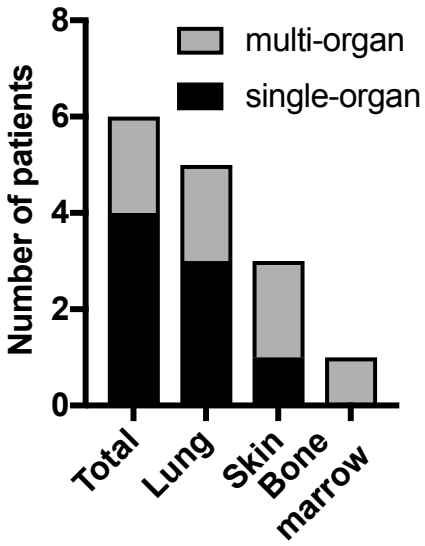


Figure 1.22: **Multiorgan inflammatory conditions.** Distribution of single or multiple organs effected with granulomatous disease for 6 patients.

pulmonary lung function tests over two or more years to assess the decline of respiratory function. Two of four patients had a significant decline indicating a variable degree of lung function in adult patients with RAG deficiency. Based on literature searches there are only four additional cases of similar adults [64, 67–69]. Of these four, two have died and two had a history of severe lung disease. High resolution chest CT every 1-3 years for signs of progressive pulmonary disease is recommended for similar patients. Treatment of choice should be tailored to both infectious and inflammatory components.

1.4.3.4 Treatment

Thirteen out of fourteen patients (93%) received first line immunoglobulin replacement therapy (**Figure 1.25**). 57% received antibiotic prophylaxis, 21% antiviral drugs, and 14% disease-modifying anti-rheumatic drug. Five patients (36%) were considered for HSCT. Comparisons of therapeutics approaches revealed no statistically significant difference in survival. Three out of eight patients who only received Ig replacement therapy were deceased. Among transplanted patients the major mortality cause was infections post HSCT.

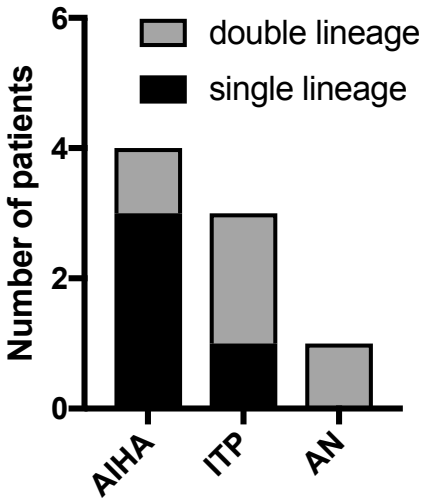


Figure 1.23: **Autoimmune lineage distribution.** Distribution and lineages effected by autoimmune cytopenias in 6 patients. AI (other autoimmunity), AIC (autoimmune cytopenia), AIHA (autoimmune haemolytic anaemia), AN (autoimmune neutropenia), ITP (immune thrombocytopenic purpura).

1.4.4 Addressing phenotype-genotype correlations

Phenotype-genotype correlations are reported for regions of RAG1 and RAG2 [94]. Pathogenic missense variants in RAG1 most frequently occur in the catalytic core (amino acids 387–1,011) (Fig 1B), predominantly in the zinc-binding domain. When normalized for domain length, a higher pathogenic variant rate is observed in the NBD and CTD [94]. Forty percent of RAG1 patients reported here have disease-causing variants in NBD or CTD. These two domains constitute the highest reported pathogenic mutation rates [94]. A few RAG1 missense mutations are associated with CID–G/AI. These variants are predominantly located in the domains DDBD, PreR and CTD [94]. Deviation from the phenotype-genotype correlation is illustrated by three patients found to have CID-G/AI; patient 2 was found to have compound heterozygous RAG1 mutations in non-core (frameshift stop variant) and the catalytic RNase H (RNH) domain while presenting with CID-G/AI, and patients 12 and 13 reported as CID-G/AI due to compound heterozygous core/plant homeodomain (PHD) and homozygous core RAG2 mutations, respectively (**table 1.9**). **Figure 1.16** illustrates the distribution of mutations reported in this study amongst RAG1 and RAG2 functional domains.

Table 1.9: Phenotype-genotype distribution. CID (combined immunodeficiency), CID-G/AI (combined immunodeficiency with granuloma autoimmunity), CTD (carboxy-terminal domain), CVID (common variable immunodeficiency), DDBD (dimerisation and DNA-binding domain), Hypo (Hypogammaglobinemia), ICL (idiopathic CD4+ lymphopenia), NBD (nonamer-binding domain) PHD (plant homeodomain), PreR (pre-RNase H), RNH (RNase H), SPAD (Specific polysaccharide antibody deficiency), ZnBD (zinc-binding domain).

RAG1

Domain	NBD	DDBD	PreR	RNH	ZnBD
NBD					
DDBD		ICL			
PreR					
RNH					
ZnBD					CID
CTD	CVID Hypo/SPAD				
Non-core	CID	CID	CID	CID-G/AI	SPAD

RAG2

Domain	Core	PHD
Core	CID-G/AI	CID, CVID, CID-G/AI
Hinge	SPAD	

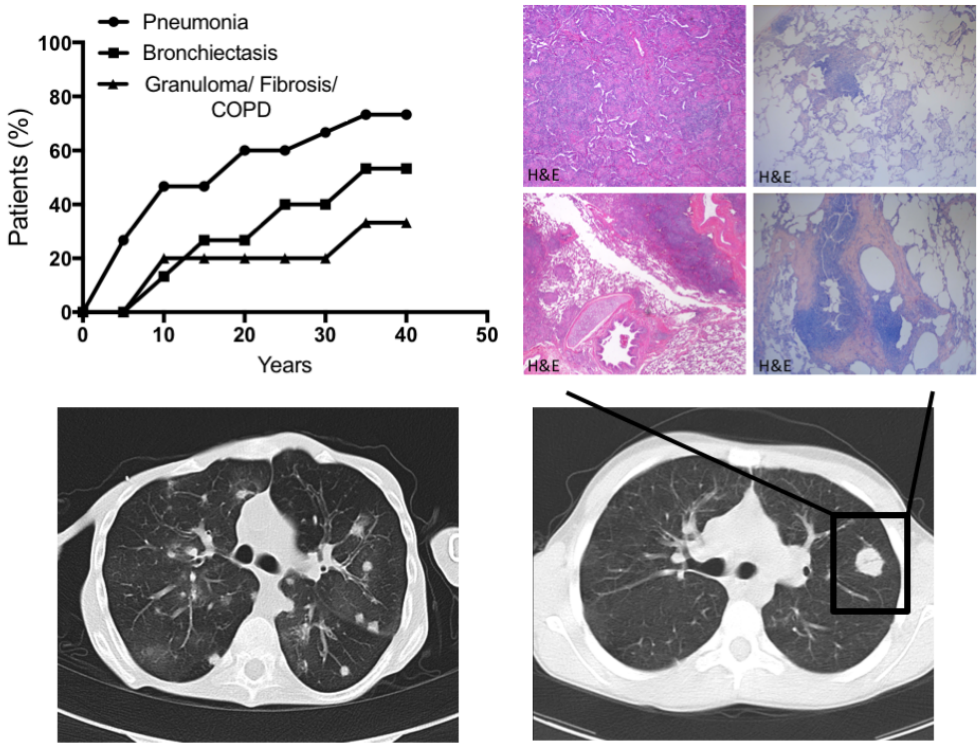


Figure 1.24: **Features of lung disease. (Top left)** Onset of pneumonia, bronchiectasis and granuloma/fibrosis/COPD (n=15). **(Top right)** Histologic examination of lung biopsies from patient 1 with atypical lymphoid hyperplasia with granuloma and fibrosis. **(Bottom)** High resolution computed tomography of 2 patients with lung granuloma. Figure credit of Dr. C Geier.

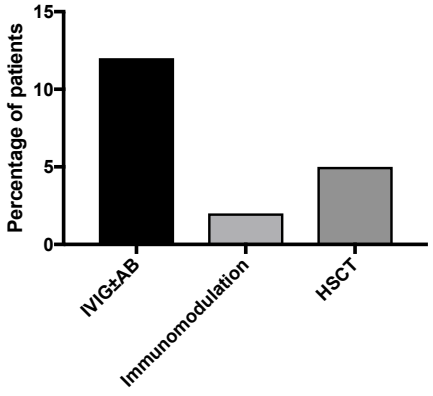


Figure 1.25: **Treatment strategies** provided by primary physicians for the 15 adult RAG deficient patients. IVIG+AB (immunoglobulin replacement therapy and/or antibiotic prophylaxis), HSTC (haematopoietic stem cell transplantation).

1.5 Discussion

Newborn screening for SCID and related conditions has identified *RAG1* and *RAG2* as the most common defective genes associated with atypical SCID [93, 97]. Systemic analysis of PID related genes such as RAG1 provides a reliable reference to incentivize genetic screening for newborns [95]. Based on ExAC data analysis, the incidence of RAG deficiency has been predicted at 1:181,000 individuals who are homozygous or compound heterozygous for pathogenic *RAG1* and *RAG2* mutations [69].

Based on two large cohort studies in Europe, the estimate for the prevalence of RAG deficiency in adult PID was found to be between 1%-1.9%. With this estimate an additional 32-63 (+/- 1) cases of RAG deficiency are expected in the UK based on number of adult PID patients registered with the UKPIN database. A robust systemic analysis of 79 individual mutations in RAG1 was conducted previously [95]. This study utilized a flow cytometry–based assay that allows analysis of RAG recombination activity based on green fluorescent protein expression. The assay utilized in the present study is a qPCR-based assay for the measurement of substrate inversion. A combination of murine RAG1 and RAG2 was used due to high homology between mouse and human RAG genes. Murine sequence expression plasmids were used firstly because they were readily availability from our collaborator and secondly because a high-quality crystallographic and cryo-EM structures are available for mapping and potentially predicting the effects of novel mutations [49, 50]. The results of recombination assays presented in this study closely matched the recombination of measurements in another published assay system of human RAG variants when the same variants were tested in both studies [95]. Furthermore, the activity levels in RAG1 and RAG2 compound heterozygous cases were assessed and found, in some cases, lower activity levels than the average of each allele measured separately. Further investigations of compound heterozygous effects in RAG1 and RAG2 may find evidence of other cooperative loss of function mutations as seen in patient 8, 9 and 15. It is still unclear if these adults could have been identified at birth with newborn screening for SCID or naïve CD4+ T cell lymphopenia that is a hallmark of adult RAG deficiency. The relative absence of RAG deficiency in the paediatric cohort of 216 patients suggest that

milder forms of RAG deficiency may not be diagnosed as readily as a PID in childhood. The low number of naïve CD4 cells may appear as idiopathic T cell lymphopenia subset when screened at birth [83]. The late presentation of RAG deficiencies presented here indicate that screening for RAG1 and RAG2 defects in early childhood among milder PID phenotypes may provide an opportunity for intervention before clinical manifestations including the onset of the pulmonary disease.

Progressive T cell lymphopenia (total and naïve T cell) might serve as a possible biomarker to identify patients with a variety of CID, including RAG deficiency [78, 93]. RAG dependent immunodeficiency known to produce immune dysregulation; granulomas and autoimmunity occurs with mutations retaining residual recombination activity [68, 71, 95]. The majority of patients investigated have developed inflammatory and/or autoimmune complications, at similar frequencies to previously described studies [68, 94]. Total and naïve CD4+ T cell lymphopenia, [78, 93] autoimmunity, and progressive inflammatory lung disease should all prompt further investigations for RAG deficiency in adult PID patients.

Further studies may determine the underlining pathophysiology that drives autoimmunity in RAG deficient patients. A number of the patients (57%) developed antibodies to cytokines, which may serve as a potential biomarker for adults with PID due to RAG1 and/or RAG2 mutations (Table 3) [68]. Further, a recent report has shown profound abnormalities of central and peripheral T cell tolerance, with low expression of AIRE in the thymus of patients with CID-G/A [67]. Additionally, it was recently demonstrated, that RAG deficient patients show a restriction of Treg repertoire diversity and a molecular signature of self-reactive conventional CD4+ T cells [96].

Lung disease was the most prominent feature in the seventeen cases of adolescent or adult patients with RAG deficiency. Based on literature searches there are only four additional cases of similar adults [64, 67–69]. Of these four, two have died and two had a history of severe lung disease.

The clinical spectrum of lung disease may range from infections to immune dysregulation and is a key cause of morbidity and mortality among patients with hypomorphic

RAG1 or *RAG2* variants. Severe non-infectious complications including fibrosis are likely underestimated in the absence of close monitoring of lung disease. Although large cohort-based whole exome studies among patients with pulmonary fibrosis have not revealed known damaging variants in *RAG1/RAG2*, the data presented here suggests that RAG deficiency should be considered, especially if the immune phenotype is suggestive. To promote early intervention, high resolution chest CT every 1-3 years for signs of progressive pulmonary disease is recommend. Serial lung evaluation with pulmonary function testing (PFT) including DLCO and lung volumes, especially TLC as an accurate indicator for restrictive processes, should be considered in these patients. Treatment of choice should be tailored to both infectious and inflammatory components.

Careful analysis of HSCT decision is needed and should be considered before onset of rapid or progressive decline in lung function [76, 98]. Interim analysis has been provided to guide difficult HSCT decisions for patients with SCID and profound CID [93, 97].

Phenotype-genotype correlations are reported for regions of RAG1 and RAG2 [94]. Pathogenic missense variants in *RAG1* most frequently occur in the catalytic core (amino acids 387–1,011), predominantly in the zinc-binding domain (amino acids 722–965). When normalized for domain length, a higher pathogenic variant rate is observed in the NBD (amino acids 394–460) and CTD (amino acids 966–1,008) [94]. Forty percent of RAG1 patients reported here have disease-causing variants in NBD or CTD. A few *RAG1* missense mutations are associated with CID–G/AI. These variants are predominantly located in the domains DDBD (amino acids 461–517), PreR (amino acids 518–590) and CTD [94]. Deviation from the phenotype-genotype correlation is illustrated by three patients found to have CID-G/AI due to variants in non-core and RNH domain (amino acids 591–721) in RAG1 and patients with core (amino acids 1–383) and PHD (amino acids 414–487) RAG2 variants.

Patients with more severe phenotypes generally have a progressively pronounced restriction of their BCR and TCR repertoire diversity. Analysis of the TCR and BCR repertoire identifies skewed usage of V(D)J segment genes and abnormalities of CDR3 length distribution [99] Collaboration with the primary physicians allowed analysis of

immunoglobulin measures. It was found that, as recently published [100], that low IgA and IgM is associated with bronchiectasis in PID. Mutant RAG1 and RAG2 proteins with residual recombination activity in these patients likely provides antibody repertoire that may be sufficient during early childhood but immunodeficiency and progressive autoimmunity becomes apparent towards early adolescence. Many countries do not yet have newborn screening for SCID which might prompt HSCT [101].

Phenotypic heterogeneity impedes prediction of the clinical phenotype, although at least 150 and 57 disease-causing variants which are likely to result in clinical intervention are reported for RAG1 and RAG2, respectively [94].

In the era of whole exome sequencing, the spectrum of RAG deficiency further broadens to include adults with autoimmune and inflammatory manifestations that may result in progressive decline. Systemic analysis of PID related genes [95] and functional in vitro assays that confirm decreased recombination activity are essential. Laboratory features of naïve CD4+ T cell lymphopenia and presence of anti-cytokine antibodies can further support the diagnosis of partial RAG deficiency. Where RAG deficiency is confirmed, therapy may be adjusted based on the mechanistic understanding and may ultimately provide targeted strategy for early intervention. The characterisation of recombination activity affected by RAG1 and RAG2 variants (especially with systems that can test compound heterozygous forms) in combination with newborn or prenatal screening, may ultimately provide a strategy for early intervention in RAG deficiency.

1.6 Conclusion

Patients with RAG deficiency may survive in adulthood and the presented findings suggest that prevalence of such cases varies between 1% to 1.9% in adult PID cohorts.

Bibliography

[1] Dylan Lawless, Christoph B Geier, Jocelyn R Farmer, Hana Allen Lango, Daniel Thwaites, Faranaz Atschekzei, Matthew Brown, David Buchbinder, Siobhan O

Burns, Manish J Butte, et al. Prevalence and clinical challenges among adults with primary immunodeficiency and recombination-activating gene deficiency. *Journal of Allergy and Clinical Immunology*, 2018.

[2] David G Schatz, Marjorie A Oettinger, and David Baltimore. The v (d) j recombination activating gene, rag-1. *Cell*, 59(6):1035–1048, 1989.

[3] Marjorie A Oettinger, David G Schatz, Carolyn Gorka, and David Baltimore. Rag-1 and rag-2, adjacent genes that synergistically activate v (d) j recombination. *Science*, 248(4962):1517–1523, 1990.

[4] Patricia Stanhope-Baker, Karen M Hudson, Arthur L Shaffer, Andrei Constantinescu, and Mark S Schlissel. Cell type–specific chromatin structure determines the targeting of v (d) j recombinase activity in vitro. *Cell*, 85(6):887–897, 1996.

[5] J Fraser McBlane, Dik C van Gent, Dale A Ramsden, Charles Romeo, Christina A Cuomo, Martin Gellert, and Marjorie A Oettinger. Cleavage at a v (d) j recombination signal requires only rag1 and rag2 proteins and occurs in two steps. *Cell*, 83 (3):387–395, 1995.

[6] David B Roth, Joseph P Menetski, Pamela B Nakajima, Melvin J Bosma, and Martin Gellert. V (d) j recombination: broken dna molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. *Cell*, 70(6):983–991, 1992.

[7] Dana Branzei and Marco Foiani. Regulation of dna repair throughout the cell cycle. *Nature reviews Molecular cell biology*, 9(4):297, 2008.

[8] Shruti Malu, Vidyasagar Malshetty, Dailia Francis, and Patricia Cortes. Role of non-homologous end joining in v (d) j recombination. *Immunologic research*, 54 (1-3):233–246, 2012.

[9] Niels K Jerne. The natural-selection theory of antibody formation. *Proceedings of the National Academy of Sciences*, 41(11):849–857, 1955.

[10] Frank Macfarlane Burnet et al. A modification of jerne's theory of antibody production using the concept of clonal selection. *Australian J. Sci.*, 20(3):67–9, 1957.

[11] William J Dreyer and J Claude Bennett. The molecular basis of antibody formation: a paradox. *Proceedings of the National Academy of Sciences*, 54(3):864–869, 1965.

[12] Nobumichi Hozumi and Susumu Tonegawa. Evidence for somatic rearrangement of

immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences*, 73(10):3628–3632, 1976.

[13] Christine Brack, Minoru Hirama, Rita Lenhard-Schuller, and Susumu Tonegawa. A complete immunoglobulin gene is created by somatic recombination. *Cell*, 15(1): 1–14, 1978.

[14] David Baltimore. Is terminal deoxynucleotidyl transferase a somatic mutagen in lymphocytes? *Nature*, 248(5447):409, 1974.

[15] Ken Murphy and Casey Weaver. *Janeway's immunobiology*. Garland Science, 2016.

[16] Fahim Halim Khan. *The elements of immunology*. Pearson Education India, 2009.

[17] David G Schatz and David Baltimore. Stable expression of immunoglobulin gene v (d) j recombinase activity by gene transfer into 3t3 fibroblasts. *Cell*, 53(1):107–115, 1988.

[18] Dik C van Gent, Kiyoshi Mizuuchi, and Martin Gellert. Similarities between initiation of v (d) j recombination and retroviral integration. *Science*, 271(5255): 1592–1594, 1996.

[19] Alka Agrawal, Quinn M Eastman, and David G Schatz. Transposition mediated by rag1 and rag2 and its implications for the evolution of the immune system. *Nature*, 394(6695):744, 1998.

[20] Kevin Hiom, Meni Melek, and Martin Gellert. Dna transposition by the rag1 and rag2 proteins: a possible source of oncogenic translocations. *Cell*, 94(4):463–470, 1998.

[21] Vladimir V Kapitonov and Jerzy Jurka. Rag1 core and v (d) j recombination signal sequences were derived from transib transposons. *PLoS biology*, 3(6):e181, 2005.

[22] Sebastian D Fugmann, Cynthia Messier, Laura A Novack, R Andrew Cameron, and Jonathan P Rast. An ancient evolutionary origin of the rag1/2 gene locus. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3728–3733, 2006.

[23] Sebastian D Fugmann. The origins of the rag genes from transposition to v (d) j recombination. *Seminars in immunology*, 22(1):10–16, 2010.

[24] Shengfeng Huang, Xin Tao, Shaochun Yuan, Yuhang Zhang, Peiyi Li, Helen A Beilinson, Ya Zhang, Wenjuan Yu, Pierre Pontarotti, Hector Escriva, et al. Discovery

of an active rag transposon illuminates the origins of v (d) j recombination. *Cell*, 166(1):102–114, 2016.

[25] George D Yancopoulos and Frederick W Alt. Developmentally controlled and tissue-specific expression of unrearranged vh gene segments. *Cell*, 40(2):271–281, 1985.

[26] T Keith Blackwell, Mark W Moore, George D Yancopoulos, Heikyung Suh, Stuart Lutzker, Erik Selsing, and Frederick W Alt. Recombination between immunoglobulin variable region gene segments is enhanced by transcription. *Nature*, 324(6097): 585–589, 1986.

[27] George D Yancopoulos, T Keith Blackwell, Heikyung Suh, Leroy Hood, and Frederick W Alt. Introduced t cell receptor variable region gene segments recombine in pre-b cells: evidence that b and t cells use a common recombinase. *Cell*, 44(2): 251–259, 1986.

[28] Yoichi Shinkai, Kong-Peng Lam, Eugene M Oltz, Valerie Stewart, Monica Mendelsohn, Jean Charron, Milton Datta, Faith Young, Alan M Stall, Frederick W Alt, et al. Rag-2-deficient mice lack mature lymphocytes owing to inability to initiate v (d) j rearrangement. *Cell*, 68(5):855–867, 1992.

[29] Peter Mombaerts, John Iacomini, Randall S Johnson, Karl Herrup, Susumu Tonegawa, and Virginia E Papaioannou. Rag-1-deficient mice have no mature b and t lymphocytes. *Cell*, 68(5):869–877, 1992.

[30] Barry P Sleckman, James R Gorman, and Frederick W Alt. Accessibility control of antigen-receptor variable-region gene assembly: role of cis-acting elements. *Annual review of immunology*, 14(1):459–481, 1996.

[31] Mark Schlissel, Andrei Constantinescu, Terri Morrow, Mike Baxter, and Albert Peng. Double-strand signal sequence breaks in v (d) j recombination are blunt, 5'-phosphorylated, rag-dependent, and cell cycle regulated. *Genes & development*, 7(12b):2520–2532, 1993.

[32] David B Roth, Chengming Zhu, and Martin Gellert. Characterization of broken dna molecules associated with v (d) j recombination. *Proceedings of the National Academy of Sciences*, 90(22):10788–10792, 1993.

[33] Paul R Mueller and Barbara Wold. In vivo footprinting of a muscle specific enhancer

by ligation mediated pcr. *Science*, 246(4931):780–786, 1989.

[34] Dik C van Gent, J Fraser McBlane, Dale A Ramsden, Moshe J Sadofsky, Joanne E Hesse, and Martin Gellert. Initiation of v (d) j recombination in a cell-free system. *Cell*, 81(6):925–934, 1995.

[35] Robin Milley Cobb, Kenneth J Oestreich, Oleg A Osipovich, and Eugene M Oltz. Accessibility control of v (d) j recombination. *Advances in immunology*, 91:45–109, 2006.

[36] Hong-Erh Liang, Lih-Yun Hsu, Dragana Cado, Lindsay G Cowell, Garnett Kelsoe, and Mark S Schlissel. The "dispensable" portion of rag2 is necessary for efficient v-to-dj rearrangement during b and t cell development. *Immunity*, 17(5):639–651, 2002.

[37] Darryll D Dudley, JoAnn Sekiguchi, Chengming Zhu, Moshe J Sadofsky, Scott Whitlow, Jeffrey DeVido, Robert J Monroe, Craig H Bassing, and Frederick W Alt. Impaired v (d) j recombination and lymphocyte development in core rag1-expressing mice. *Journal of Experimental Medicine*, 198(9):1439–1450, 2003.

[38] Yun Liu, Ramesh Subrahmanyam, Tirtha Chakraborty, Ranjan Sen, and Stephen Desiderio. A plant homeodomain in rag-2 that binds hypermethylated lysine 4 of histone h3 is necessary for efficient antigen-receptor-gene rearrangement. *Immunity*, 27(4):561–571, 2007.

[39] Adam GW Matthews, Alex J Kuo, Santiago Ramón-Maiques, Sunmi Han, Karen S Champagne, Dmitri Ivanov, Mercedes Gallardo, Dylan Carney, Peggie Cheung, David N Ciccone, et al. Rag2 phd finger couples histone h3 lysine 4 trimethylation with v (d) j recombination. *Nature*, 450(7172):1106, 2007.

[40] Yanhong Ji, Wolfgang Resch, Elizabeth Corbett, Arito Yamane, Rafael Casellas, and David G Schatz. The in vivo pattern of binding of rag1 and rag2 to antigen receptor loci. *Cell*, 141(3):419–431, 2010.

[41] Grace Teng, Yaakov Maman, Wolfgang Resch, Min Kim, Arito Yamane, Jason Qian, Kyong-Rim Kieffer-Kwon, Malay Mandal, Yanhong Ji, Eric Meffre, et al. Rag represents a widespread threat to the lymphocyte genome. *Cell*, 162(4):751–765, 2015.

[42] David G Schatz and Yanhong Ji. Recombination centres and the orchestration of v

(d) j recombination. *Nature Reviews Immunology*, 11(4):251, 2011.

[43] Claudia Bossen, Robert Mansson, and Cornelis Murre. Chromatin topology and the regulation of antigen receptor assembly. *Annual review of immunology*, 30:337–356, 2012.

[44] Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41, 2000.

[45] Birgitte Ø Wittschieben, Gabriel Otero, Therese de Bizemont, Jane Fellows, Hediye Erdjument-Bromage, Reiko Ohba, Yang Li, C David Allis, Paul Tempst, and Jesper Q Svejstrup. A novel histone acetyltransferase is an integral subunit of elongating rna polymerase ii holoenzyme. *Molecular cell*, 4(1):123–128, 1999.

[46] Nevan J Krogan, Jim Dover, Adam Wood, Jessica Schneider, Jonathan Heidt, Marry Ann Boateng, Kimberly Dean, Owen W Ryan, Ashkan Golshani, Mark Johnston, et al. The paf1 complex is required for histone h3 methylation by compass and dot1p: linking transcriptional elongation to histone methylation. *Molecular cell*, 11(3):721–729, 2003.

[47] Huck Hui Ng, François Robert, Richard A Young, and Kevin Struhl. Targeted recruitment of set1 histone methylase by elongating pol ii provides a localized mark and memory of recent transcriptional activity. *Molecular cell*, 11(3):709–719, 2003.

[48] Tracy C Kuo and Mark S Schlissel. Mechanisms controlling expression of the rag locus during lymphocyte development. *Current opinion in immunology*, 21(2): 173–178, 2009.

[49] Min-Sung Kim, Mikalai Lapkouski, Wei Yang, and Martin Gellert. Crystal structure of the v (d) j recombinase rag1–rag2. *Nature*, 518(7540):507, 2015.

[50] Heng Ru, Melissa G Chambers, Tian-Min Fu, Alexander B Tong, Maofu Liao, and Hao Wu. Molecular mechanism of v (d) j recombination from synaptic rag1-rag2 complex structures. *Cell*, 163(5):1138–1152, 2015.

[51] John Stone. An efficient library for parallel ray tracing and animation. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[52] Mark A Landree, Jamie A Wibbenmeyer, and David B Roth. Mutational analysis of rag1 and rag2 identifies three catalytic amino acids in rag1 critical for both cleavage steps of v (d) j recombination. *Genes & development*, 13(23):3059–3069, 1999.

[53] Deok Ryong Kim, Yan Dai, Cynthia L Mundy, Wei Yang, and Marjorie A Oettinger. Mutations of acidic residues in rag1 define the active site of the v (d) j recombinase. *Genes & development*, 13(23):3070–3080, 1999.

[54] Patricia Cortes, Zheng-Sheng Ye, and David Baltimore. Rag-1 interacts with the repeated amino acid motif of the human homologue of the yeast protein srp1. *Proceedings of the National Academy of Sciences*, 91(16):7633–7637, 1994.

[55] Steven F Bellon, Karla K Rodgers, David G Schatz, Joseph E Coleman, and Thomas A Steitz. Crystal structure of the rag1 dimerization domain reveals multiple zinc-binding motifs including a novel zinc binuclear cluster. *Nature Structural and Molecular Biology*, 4(7):586, 1997.

[56] Zimu Deng, Haifeng Liu, and Xiaolong Liu. Rag1-mediated ubiquitylation of histone h3 is required for chromosomal v (d) j recombination. *Cell research*, 25(2):181, 2015.

[57] Noriko Shimazaki, Albert G Tsai, and Michael R Lieber. H3k4me3 stimulates the v (d) j rag complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. *Molecular cell*, 34(5):535–544, 2009.

[58] Hao Jiang, Fu-Chung Chang, Ashley E Ross, Jihyun Lee, Keiichi Nakayama, Keiko Nakayama, and Stephen Desiderio. Ubiquitylation of rag-2 by skp2-scf links destruction of the v (d) j recombinase to the cell cycle. *Molecular cell*, 18(6):699–709, 2005.

[59] Klaus Schwarz, George H Gauss, Leopold Ludwig, Ulrich Pannicke, Zhong Li, Doris Lindner, Wilhelm Friedrich, Reinhard A Seger, Thomas E Hansen-Hagge, Stephen Desiderio, et al. Rag mutations in human b cell-negative scid. *Science*, 274(5284): 97–99, 1996.

[60] G de Saint-Basile, F Le Deist, JP De Villartay, N Cerf-Bensussan, O Journet, N Brousse, C Griscelli, and A Fischer. Restricted heterogeneity of t lymphocytes in combined immunodeficiency with hypereosinophilia (omenn's syndrome). *The journal of clinical investigation*, 87(4):1352–1359, 1991.

[61] Frédéric Rieux-Laucat, Philippe Bahadoran, Nicole Brousse, Françoise Selz, Alain Fischer, Françoise Le Deist, and Jean Pierre De Villartay. Highly restricted human t cell repertoire in peripheral blood and tissue-infiltrating lymphocytes in omenn's

syndrome. *The journal of clinical investigation*, 102(2):312–321, 1998.

[62] Anna Villa, Sandro Santagata, Fabio Bozzi, Silvia Giliani, Annalisa Frattini, Luisa Imberti, Luisa Benerini Gatta, Hans D Ochs, Klaus Schwarz, Luigi D Notarangelo, et al. Partial v (d) j recombination activity leads to omenn syndrome. *Cell*, 93(5): 885–896, 1998.

[63] Jean-Pierre De Villartay, Alain Fischer, and Anne Durandy. The mechanisms of immune diversification and their disorders. *Nature Reviews Immunology*, 3(12):962, 2003.

[64] Catharina Schuetz, Kirsten Huck, Sonja Gudowius, Mosaad Megahed, Oliver Feyen, Bernd Hubner, Dominik T Schneider, Burkhard Manfras, Ulrich Pannicke, Rein Willemze, et al. An immunodeficiency disease with rag mutations and granulomas. *New England journal of Medicine*, 358(19):2030–2038, 2008.

[65] Lauren A Henderson, Francesco Frugoni, Gregory Hopkins, Helen de Boer, Sung-Yun Pai, Yu Nee Lee, Jolan E Walter, Melissa M Hazen, and Luigi D Notarangelo. Expanding the spectrum of recombination-activating gene 1 deficiency: a family with early-onset autoimmunity. *Journal of Allergy and Clinical Immunology*, 132 (4):969–971, 2013.

[66] Antonia Kwan, Roshini S Abraham, Robert Currier, Amy Brower, Karen Andruszewski, Jordan K Abbott, Mei Baker, Mark Ballow, Louis E Bartoshesky, Vincent R Bonagura, et al. Newborn screening for severe combined immunodeficiency in 11 screening programs in the united states. *Jama*, 312(7):729–738, 2014.

[67] Suk See De Ravin, Edward W Cowen, Kol A Zarember, Narda L Whiting-Theobald, Douglas B Kuhns, Netanya G Sandler, Daniel C Douek, Stefania Pittaluga, Pietro L Poliani, Yu Nee Lee, et al. Hypomorphic rag mutations can cause destructive midline granulomatous disease. *Blood*, 116(8):1263–1271, 2010.

[68] Jolan E Walter, Lindsey B Rosen, Krisztian Csomos, Jacob M Rosenberg, Divij Mathew, Marton Keszei, Boglarka Ujhazi, Karin Chen, Yu Nee Lee, Irit Tirosh, et al. Broad-spectrum antibodies against self-antigens and cytokines in rag deficiency. *The journal of clinical investigation*, 125(11):4135–4148, 2015.

[69] Attila Kumánovics, Yu Nee Lee, Devin W Close, Emily M Coonrod, Boglarka

Ujhazi, Karin Chen, Daniel G MacArthur, Gergely Krivan, Luigi D Notarangelo, and Jolan E Walter. Estimated disease incidence of rag1/2 mutations: A case report and querying the exome aggregation consortium. *Journal of Allergy and Clinical Immunology*, 139(2):690–692, 2017.

[70] Svetlana O Sharapova, Alexandr Migas, Irina Guryanova, Svetlana Aleshkevich, Semen Kletski, Anne Durandy, and Michael Belevtsev. Late-onset combined immune deficiency associated to skin granuloma due to heterozygous compound mutations in rag1 gene in a 14 years old male. *Human immunology*, 74(1):18–22, 2013.

[71] David Buchbinder, Rebecca Baker, Yu Nee Lee, Juan Ravell, Yu Zhang, Joshua McElwee, Diane Nugent, Emily M Coonrod, Jacob D Durtschi, Nancy H Augustine, et al. Identification of patients with rag mutations previously diagnosed with common variable immunodeficiency disorders. *Journal of clinical immunology*, 35 (2):119–124, 2015.

[72] Christoph B Geier, Alexander Piller, Angela Linder, Kai MT Sauerwein, Martha M Eibl, and Hermann M Wolf. Leaky rag deficiency in adult patients with impaired antibody production against bacterial polysaccharide antigens. *PloS one*, 10(7): e0133220, 2015.

[73] Jean-Pierre De Villartay, Annick Lim, Hamoud Al-Mousa, Sophie Dupont, Julie Déchanet-Merville, Edith Coumau-Gatbois, Marie-Lise Gougeon, Arnaud Lemainque, Céline Eidenschenk, Emmanuelle Jouanguy, et al. A novel immunodeficiency associated with hypomorphic rag1 mutations and cmv infection. *The journal of clinical investigation*, 115(11):3291–3299, 2005.

[74] Sung-Yun Pai and Morton J Cowan. Stem cell transplantation for primary immunodeficiency diseases: the north american experience. *Current opinion in allergy and clinical immunology*, 14(6):521, 2014.

[75] Catharina Schuetz, Benedicte Neven, Christopher C Dvorak, Sandrine Leroy, Markus J Ege, Ulrich Pannicke, Klaus Schwarz, Ansgar S Schulz, Manfred Hoenig, Monika Sparber-Sauer, et al. Scid patients with artemis vs rag deficiencies following hct: increased risk of late toxicity in artemis-deficient scid. *Blood*, 123(2):281–289, 2014.

[76] Tami John, Jolan E Walter, Catherina Schuetz, Karin Chen, Roshini S Abra-

ham, Carmem Bonfim, Thomas G Boyce, Avni Y Joshi, Elizabeth Kang, Beatriz Tavares Costa Carvalho, et al. Unrelated hematopoietic cell transplantation in a patient with combined immunodeficiency with granulomatous disease and autoimmunity secondary to rag deficiency. *Journal of clinical immunology*, 36(7):725–732, 2016.

[77] Stephan Ehl, Klaus Schwarz, Anselm Enders, Ulrich Duffner, Ulrich Pannicke, Joachim Kühr, Françoise Mascart, Annette Schmitt-Graeff, Charlotte Niemeyer, and Paul Fisch. A variant of scid with specific immune responses and predominance of $\gamma\delta$ t cells. *The journal of clinical investigation*, 115(11):3140–3148, 2005.

[78] Kerstin Felgentreff, Ruy Perez-Becker, Carsten Speckmann, Klaus Schwarz, Krzysztof Kalwak, Gasper Markelj, Tadej Avcin, Waseem Qasim, EG Davies, Tim Niehues, et al. Clinical and immunological manifestations of patients with atypical severe combined immunodeficiency. *Clinical immunology*, 141(1):73–82, 2011.

[79] Lisa M Ott de Bruin, Marita Bosticardo, Alessandro Barbieri, Sherry G Lin, Jared H Rowe, Pietro L Poliani, Kimberly Ching, Daniel Eriksson, Nils Landegren, Olle Kämpe, et al. Hypomorphic rag1 mutations alter the pre-immune repertoire at early stages of lymphoid development. *Blood*, pages blood–2017, 2018.

[80] Hassan Abolhassani, Ning Wang, Asghar Aghamohammadi, Nima Rezaei, Yu Nee Lee, Francesco Frugoni, Luigi D Notarangelo, Qiang Pan-Hammarström, and Lennart Hammarström. A hypomorphic recombination-activating gene 1 (rag1) mutation resulting in a phenotype resembling common variable immunodeficiency. *Journal of Allergy and Clinical Immunology*, 134(6):1375–1380, 2014.

[81] Mona Hedayat, Michel J Massaad, Yu Nee Lee, Mary Ellen Conley, Jordan S Orange, Toshiro K Ohsumi, Waleed Al-Herz, Luigi D Notarangelo, Raif S Geha, and Janet Chou. Lessons in gene hunting: A rag1 mutation presenting with agammaglobulinemia and absence of b cells. *Journal of Allergy and Clinical Immunology*, 134(4):983–985, 2014.

[82] Tamaki Kato, Elena Crestani, Chikako Kamae, Kenichi Honma, Tomoko Yokosuka, Takeshi Ikegawa, Naonori Nishida, Hirokazu Kanegane, Taizo Wada, Akihiro Yachie, et al. Rag1 deficiency may present clinically as selective iga deficiency. *Journal of*

clinical immunology, 35(3):280–288, 2015.

[83] Taco W Kuijpers, Hanna IJspeert, Ester MM van Leeuwen, Machiel H Jansen, Mette D Hazenberg, Kees C Weijer, Rene AW Van Lier, and Mirjam van der Burg. Idiopathic cd4+ t lymphopenia without autoimmunity or granulomatous disease in the slipstream of rag mutations. *Blood*, pages blood–2011, 2011.

[84] Janet Chou, Rima Hanna-Wakim, Irit Tirosh, Jennifer Kane, David Fraulino, Yu Nee Lee, Soha Ghanem, Iman Mahfouz, André Mégarbané, Gérard Lefranc, et al. A novel homozygous mutation in recombination activating gene 2 in 2 relatives with different clinical phenotypes: Omenn syndrome and hyper-igm syndrome. *Journal of Allergy and Clinical Immunology*, 130(6):1414–1416, 2012.

[85] Andreas Reiff, Alexander G Bassuk, Joseph A Church, Elizabeth Campbell, Xinyu Bing, and Polly J Ferguson. Exome sequencing reveals rag1 mutations in a child with autoimmunity and sterile chronic multifocal osteomyelitis evolving into disseminated granulomatous disease. *Journal of clinical immunology*, 33(8):1289–1292, 2013.

[86] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[87] Bernhard Hirt. Selective extraction of polyoma dna from infected mouse cell cultures. *Journal of molecular biology*, 26(2):365–369, 1967.

[88] Seiichi Mizushima and Shigekazu Nagata. pef-bos, a powerful mammalian expression vector. *Nucleic acids research*, 18(17):5322, 1990.

[89] Joanne E Hesse, Michael R Lieber, Kiyoshi Mizuuchi, and Martin Gellert. V (d) j recombination: a functional definition of the joining signals. *Genes & development*, 3(7):1053–1061, 1989.

[90] Daniel T Thwaites, Clive Carter, Dylan Lawless, Sinisa Savic, and Joan M Boyes. A novel rag1 mutation reveals a critical in vivo role for hmgb1/2 during v (d) j recombination. *Blood*, 133(8):820–829, 2019.

[91] Michele Mauri, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. Rawgraphs: A visualisation platform to create open outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, CHItaly '17, pages 28:1–28:5,

New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5237-6. doi: 10.1145/3125571.3125585. URL http://doi.acm.org/10.1145/3125571.3125585.

[92] Rémi Bertinchamp, Laurence Gérard, David Boutboul, Marion Malphettes, Claire Fieschi, Eric Oksenhendler, E Oksenhendler, C Fieschi, M Malphettes, L Galicier, et al. Exclusion of patients with a severe t-cell defect improves the definition of common variable immunodeficiency. *The journal of Allergy and Clinical Immunology: In Practice*, 4(6):1147–1157, 2016.

[93] Carsten Speckmann, Sam Doerken, Alessandro Aiuti, Michael H Albert, Waleed Al-Herz, Luis M Allende, Alessia Scarselli, Tadej Avcin, Ruy Perez-Becker, Caterina Cancrini, et al. A prospective study on the natural history of patients with profound combined immunodeficiency: an interim analysis. *Journal of Allergy and Clinical Immunology*, 139(4):1302–1310, 2017.

[94] Luigi D Notarangelo, Min-Sung Kim, Jolan E Walter, and Yu Nee Lee. Human rag mutations: biochemistry and clinical implications. *Nature Reviews Immunology*, 16 (4):234, 2016.

[95] Yu Nee Lee, Francesco Frugoni, Kerry Dobbs, Jolan E Walter, Silvia Giliani, Andrew R Gennery, Waleed Al-Herz, Elie Haddad, Francoise LeDeist, Jack H Bleesing, et al. A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. *Journal of Allergy and Clinical Immunology*, 133(4):1099–1108, 2014.

[96] Jared H Rowe, Brian D Stadinski, Lauren A Henderson, Lisa Ott de Bruin, Ottavia Delmonte, Yu Nee Lee, M Teresa de la Morena, Rakesh K Goyal, Anthony Hayward, Chiung-Hui Huang, et al. Abnormalities of t-cell receptor repertoire in cd4+ regulatory and conventional t cells in patients with rag mutations: Implications for autoimmunity. *Journal of Allergy and Clinical Immunology*, 140(6):1739–1743, 2017.

[97] Christopher C Dvorak, Morton J Cowan, Brent R Logan, Luigi D Notarangelo, Linda M Griffith, Jennifer M Puck, Donald B Kohn, William T Shearer, Richard J O'Reilly, Thomas A Fleisher, et al. The natural history of children with severe combined immunodeficiency: baseline features of the first fifty patients of the primary immune deficiency treatment consortium prospective study 6901. *Journal*

of clinical immunology, 33(7):1156–1164, 2013.

[98] Kerry Dobbs, Giovanna Tabellini, Enrica Calzoni, Ornella Patrizi, Paula Martinez, Silvia Clara Giliani, Daniele Moratto, Waleed Al-Herz, Caterina Cancrini, Morton Cowan, et al. Natural killer cells from patients with recombinase-activating gene and non-homologous end joining gene defects comprise a higher frequency of cd56bright nkg2a+++ cells, and yet display increased degranulation and higher perforin content. *Frontiers in immunology*, 8:798, 2017.

[99] Yu Nee Lee, Francesco Frugoni, Kerry Dobbs, Irit Tirosh, Likun Du, Francesca A Ververs, Heng Ru, Lisa Ott de Bruin, Mehdi Adeli, Jacob H Bleesing, et al. Characterization of t and b cell repertoire diversity in patients with rag deficiency. *Science immunology*, 1(6), 2016.

[100] John P Hodkinson, Catherine Bangs, Andrea Wartenberg-Demand, Artur Bauhofer, Patrick Langohr, Matthew S Buckland, David Guzman, Patrick FK Yong, and Sorena Kiani-Alikhan. Low iga and igm is associated with a higher prevalence of bronchiectasis in primary antibody deficiency. *Journal of clinical immunology*, 37 (4):329–331, 2017.

[101] Anna Szaflarska, Magdalena Rutkowska-Zapała, Monika Kotula, Anna Gruca, Agnieszka Grabowska, Marzena Lenart, Marta Surman, Elżbieta Trzyna, Anna Mordel, Anna Pituch-Noworolska, et al. Mutation c. 256_257delaa in rag1 gene in polish children with severe combined immunodeficiency: Diversity of clinical manifestations. *Archivum immunologiae et therapiae experimentalis*, 64(1):177–183, 2016.

2 Predicting the occurrence of variants in *RAG1* and *RAG2*

2.1 Introduction

The content in this chapter has been peer reviewed in Lawless et al. [1]. Costs associated with genomic investigations continue to reduce [2] while the richness of data generated increases. Globally, the adoption of wide scale genome sequencing implies that all newborn infants may receive screening for pathogenic genetic mutation in an asymptomatic stage, pre-emptively [3]. The one dimensionality of individual genomes is now being expanded by the possibility of massive parallel sequencing for somatic variant analysis and by single-cell or lineage-specific genotyping; culminating in a genotype spectrum. In whole blood, virtually every nucleotide position may be mutated across 10^5 cells [4]. Mapping one's genotype across multiple cell types and at several periods during a person's life may soon be feasible [5]. Such genotype snapshots might allow for prediction and tracking of somatic, epigenetic, and transcriptomic profiling.

The predictive value of genomic screening highly depends on the computation tools used for data analysis and its correlation with functional assays or prior clinical experience. Interpretation of that data is especially challenging for variants of unknown significance. There is a need for predictive genomic modelling with aims to provide a reliable guidance for therapeutic intervention for patients harbouring genetic defects for life-threatening disease before the illness becomes clinically significant. Although, most genomic investigations

currently are not predictive for clinical outcome. The study of predictive genomics is exemplified by consideration of gene essentiality, accomplished by observing intolerance to loss-of-function variants. Several gene essentiality scoring methods are available for both the coding and non-coding genome [6].

Approximately 3,000 human genes cannot tolerate the loss of one allele [6]. The greatest hurdle in monogenic disease is the interpretation of variants of unknown significance while functional validation is a major time and cost investment for laboratories investigating rare disease. Severe, life-threatening immune diseases are caused by genetic variations in almost 300 genes [7, 8] however, only a small percentage of disease-causing variants have been characterised with functional studies. Several robust tools are in common usage for predicting variant pathogenicity. A void remains for predicting mutations of interest, essential for pre-emptive validation. Our investigation aims to apply predictive genomics as a tool to identify genetic variants that are most likely to be seen in patient cohorts.

This is the first application of the novel approach of predictive genomics using Recombination activating gene 1 (RAG1) and RAG2 deficiency as a model for a rare primary immunodeficiency (PID) caused by autosomal recessive variants. *RAG1* and *RAG2* encode lymphoid-specific proteins that are essential for V(D)J recombination. This genetic recombination mechanism is essential for a robust immune response by diversification the T and B cell repertoire in the thymus and bone marrow, respectively [9, 10]. Deficiency of RAG1 [11] and RAG2 [12] in mice causes inhibition of B and T cell development. Schwarz et al. [13] formed the first publication reporting that RAG mutations in humans causes severe combined immunodeficiency (SCID), where patients were deficient in peripheral B and T cells. Patient studies identified a form of immune dysregulation known as Omenn syndrome [14, 15]. The patient phenotype includes multi-organ infiltration with oligoclonal, activated T cells. The first reported cases of Omenn syndrome identified infants with hypomorphic RAG variants which retained partial recombination activity [16]. RAG deficiency can be measured by in vitro quantification of recombination activity [17–19]. Hypomorphic *RAG1* and *RAG2* mutations, responsible for residual V(D)J recombination activity (in average 5-30%), result in a distinct phenotype of combined immunodeficiency with granuloma and/or

autoimmunity (CID-G/A) [3, 20, 21].

Human RAG deficiency has traditionally been identified at very early ages due to the rapid drop of maternally-acquired antibody in the first six months of life. A loss of adequate lymphocyte development quickly results in compromised immune responses. More recently, we found that RAG deficiency is also found for some adults living with PID [17].

RAG1 and *RAG2* are highly conserved genes but disease is only reported with autosomal recessive inheritance. Only 44% of amino acids in RAG1 and RAG2 are reported as mutated on GnomAD and functional validation of clinically relevant variants is difficult. Pre-emptive selection of residues for functional validation is a major challenge; a selection based on low allele frequency alone is infeasible. A shortened time between genetic analysis and diagnosis means that treatments may be delivered earlier. RAG deficiency may present with very variable phenotypes and treatment strategies vary. With such tools, early intervention may be prompted. Some patients could benefit from haematopoietic stem cell transplant [22] when necessary while others may be provided mechanism-based treatment [23]. Predictive scoring was validated against groups of functional assay values, human disease cases, and population genetics data. Presented is the list of variants most likely seen as future determinants of RAG deficiency, meriting functional investigation.

Work on predictive genomics methods was seeded after initially applying basic population genetics data to variants identified from the NIHRBR RD-PID cohort from chapter 1. Visualising coding variants in *RAG1* against the allele frequencies for mutations against the The Exome Aggregation Consortium (ExAC) database of 60,706 unrelated individuals allowed us to perceive the rate of rare variants (**Figure 2.1**). One of the most basic bioinformatic approaches for identifying potentially damaging variants (or invariant nucleotides) is filtering variants by frequency. Quite a few cases of RAG deficiency have been reported to date [24]. Although much of the RAG genes are conserved evolutionarily, many of the reported variants occur outside of the most conserved regions. Despite being a recessive disease, and both genes being tolerant to loss-of-function, selective pressure

on invariant coding regions means that the most damaging mutations are not often likely to be seen within the normal healthy population. For some recessive diseases, the healthy population can carry a potentially damaging variant (even at low frequency) and remain healthy until a compounding mutation is introduced. For RAG deficiency, very few damaging variants are carried in the general population.

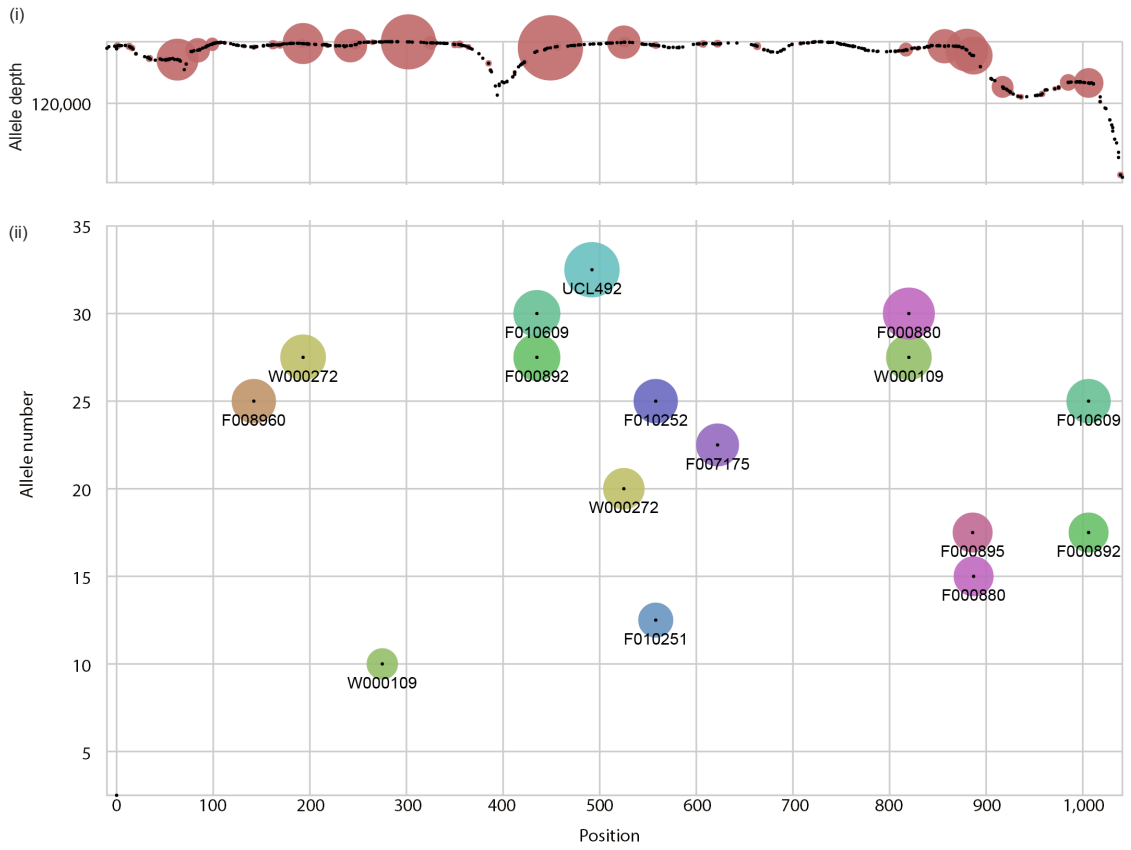


Figure 2.1: **Basic population genetics data applied to *RAG1* variants of interest.** (i) The allele frequency of the general population (primarily European) reported on GnomAD shows very few rare variants in RAG1. (ii) Candidate pathogenic rare variants are shown by their amino acid position and with unique colour and case IDs. The sequencing quality for each candidate variant adequately passed initial filtering thresholds; an allele depth of at least ten is shown for each variant.

To visualise known damaging variants against population genetics data, the conservation level of each residue was annotated with data from case reports (**Figure 2.2**). This initial population genetics and gene mapping showed that conservation rates did not necessarily predict for disease-causing mutations in *RAG1* and *RAG2*. The score of

conservation is separated into three categories. Highly conserved, rare variants reported (frequency <0.001), and multiple or variants that are not rare (frequency >0.001). From this "rough work", it was apparent that mutability has an important role that is not typically accounted for in popular genomics.

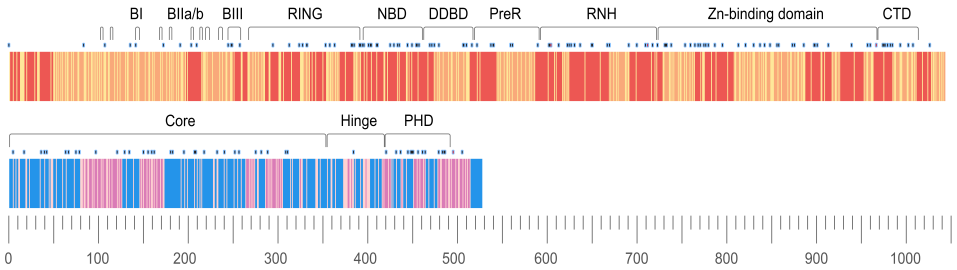


Figure 2.2: ***RAG1* and *RAG2* raw conservation rates.** Amino acid residues are mapped based on conservation rate. RAG1 (top) and RAG2 (bottom) are illustrated in colour with respect to conservation; highly conserved (red/blue), rare variants reported (frequency <0.001) (orange/purple), and multiple or variants that are not rare (frequency >0.001) (yellow/pink). Blue dots over each map indicate reported damaging variants in humans with PID [24]. Scale bar represents amino acid number in the 5'-3' coding sequence. CTD (carboxy-terminal domain), DDBD (dimerisation and DNA-binding domain), NBD (nonamer-binding domain), PHD (plant homeodomain), PreR (pre-RNase H), RNH (RNase H), ZnBD (zinc-binding domain).

2.2 Aims and objectives

To determine the mutation frequency of RAG genes and rank the likelihood of de novo mutation at each residue. Compare public and private databases of pathogenic and benign variants with these probabilities. Test predictions on known recorded cases of disease. Combine predicted mutation likelihoods with pathogenicity predictions.

2.3 Methods

2.3.1 Population genetics and data sources

GnomAD (version r2.0.2) [25] was queried for the canonical transcripts of *RAG1* and *RAG2* from population genetics data of approximately 146,000 individuals; ENST00000299440 (*RAG1*) 1586 variants, GRCh37 11:36532259-36614706 and ENST00000311485 (*RAG2*) 831 variants, GRCh37 11:36597124 - 36619829. Data was filtered to contain the variant effect identifiers: frameshift, inframe deletion, inframe insertion, missense, stop lost, or stop gained. Reference transcripts were sourced from Ensembl in the FASTA format amino acid sequence for transcript RAG1-201 ENST00000299440.5 [HGNC:9831] and transcript RAG2-201 ENST00000311485.7 [HGNC:9832]. These sequences were converted to their three-letter code format using *One to Three* from the Sequence Manipulation Suite (SMS2) [26]. Combined Annotation Dependent Depletion (CADD) scores were sourced from https://cadd.gs.washington.edu/download (Nov 2018) and are reported by Kircher et al. [27]. The dataset used was "All possible SNVs" from whole genome data, from which I extracted the data for coding regions of *RAG1* and *RAG2*. I used the Human Gene Mutation Database (HGMD) from the Institute of Medical Genetics in Cardiff as a pre-defined source of known RAG deficiency cases http://www.hgmd.cf.ac.uk/ac/index.php (Feb 2019, free access version to NM_000448.2.) [28]. Data was formatted into CSV and imported into R for combined analysis with PHRED-scaled CADD scores and the main dataframe. The crystal structure render of DNA bound RAG complex was produced with data from RCSB Protein Data Bank (3jbw.pdb) [29]. Structures were visualised using the software VMD from the Theoretical and Computational Biophysics Group [30],

imaged with Tachyon rendering [31], and colour mapped using our scoring method.

2.3.2 Data processing

The population genetics input dataset used GnomAD variant allele frequencies and reference sequences processed as CSV files, cleaned and sorted to contain only amino acid codes, residue numbers, alternate residues, alternate allele frequencies, and a score of 0 or 1 to indicate presence or absence of variants where 1 represented none reported. An annotation column was also provided to label where multiple alternate variants existed. Statistics and calculation steps are listed in order in the appendix chapter 6 table 6.3.

The percentage of conserved residues was calculated (55.99% of amino acids contained no reported variants in RAG1, 55.98% in RAG2 (table 6.1)). Basic protein statistics were generated using canonical reference transcript sequences of *RAG1* and *RAG2* with the SMS2 tool *Protein Stats* [26]. The resulting pattern percentage value was converted to a frequency based on the number of residues per protein to generate the residue frequency (R_f). The R_f values were found for both proteins as shown in the appendix chapter 6 table 6.2.

The count of variants per residue were found for both proteins and the mutation rates (M_r) per residue were calculated as shown in the appendix chapter 6. M_r was found by counting the number of mutations per residue in a window, sized to contain each protein individually. For genome-wide application the window size may be increased or decreased. In this case the window consisted of only the coding regions. The M_r values were then converted to the frequencies based on the number of residues per protein. Separate, and overlapping windows could also be used based on genome phase data and regions of linkage disequilibrium to account for non-random association of alleles at different loci; this might be particularly important for disorders with multiple genetic determinants.

The M_r and R_f multiply to give the raw mutation rate residue frequency (MRF) value. This value is also shown in the appendix chapter 6 table 6.3. Our investigation used a Boolean score C to account for the presence or absence of a mutation in the general population; 0 for any variant existing in the population and 1 for conserved residues.

$C \times M_r \times R_f$, in our case, produced the MRF score for conserved residues. **Figure 2.4 (ii)** illustrates the raw MRF as a histogram and the MRF, after applying C, as a heatmap.

An important consideration for future application is whether to use this Boolean score or instead use a discrete variable which accounts for the true allele frequency in the general population. In the clinical setting, the likelihood of *de novo* mutations and inherited mutations have different impacts when considering recessive and dominant diseases. A patient is more likely to inherit a variant that exists even at a very low frequency than to acquire a random *de novo* mutation. Therefore, a value representing an allele frequency may be used to replace C in many investigations, particularity when considering variants that exist at low rates. PRHED-scaled CADD score data consisted of nucleotide level values. For comparison with MRF, the median CADD scores were averaged per codon as demonstrated in 2.3.3 Median CADD score per residue. A summary of data processing and analysis is illustrated in **Figure 2.3**.

2.3.3 Median CADD score per residue

The sourced PHRED-scaled CADD score data consisted of nucleotide level values. We were interested in CADD scores averaged per codon. For every nucleotide position there were three alternative variants to consider, e.g.

Chrom	Pos	Ref	Alt1	Alt2	Alt3	PHRED1	PHRED2	PHRED3
11	36594855	A	C	G	T	22.3	18.81	22.4

The PHRED-scaled scores are listed here; raw CADD scores are also included in the original database. To produce a working input we used the median score per codon, that is three scores per nucleotide and three nucleotides per codon. This produced median PHRED-scaled score per codon / residue, e.g.:

Chrom	Pos	PHRED1	PHRED2	PHRED3
11	36594855	22.3	18.81	22.4
11	36594856	25.3	23.6	24.6
11	36594857	24.8	24.3	24.5

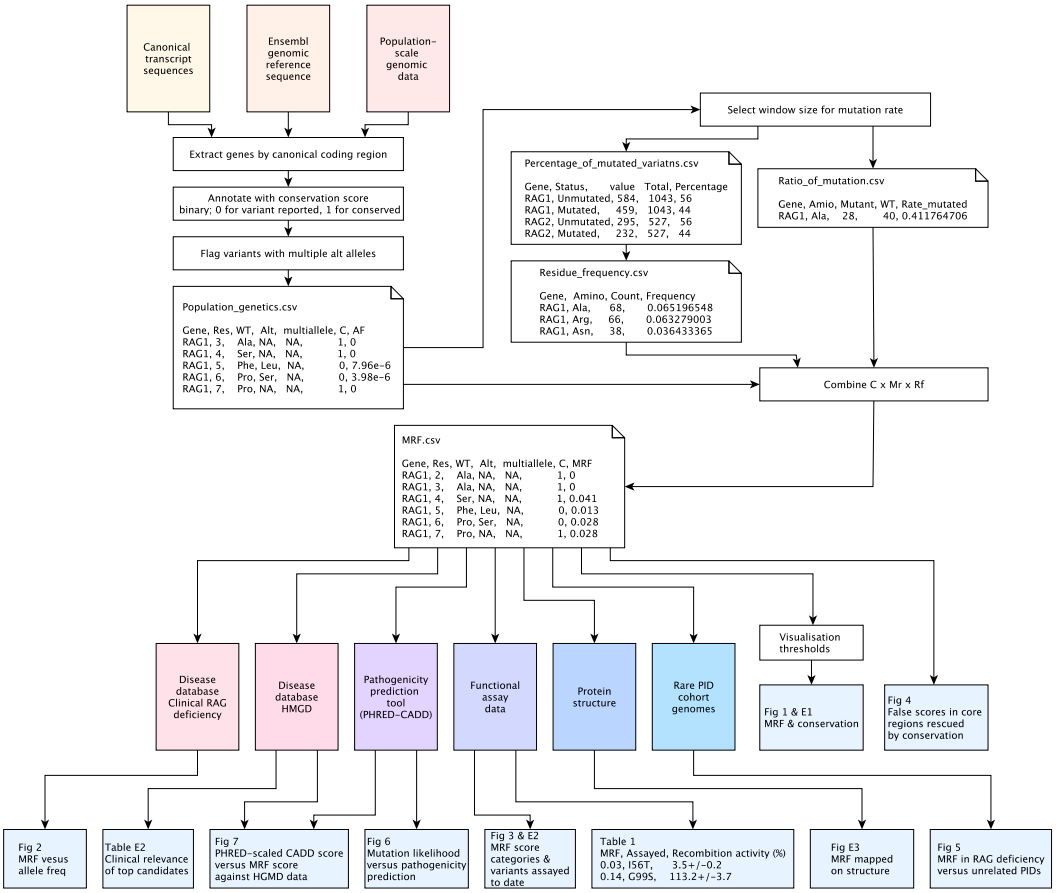


Figure 2.3: **Data analysis summary map.** Raw data and analysis scripts are provided in the methods. Analysis steps and data sources for each procedure described in *methods*. MRF; mutation rate residue frequency, PID; primary immunodeficiency.

Median PHRED = 24.3

Repository file *"RAG1.cadd.amino.csv"* within the analysis data *"Raw data R analysis for figures"* contains the median values over a three-nucleotide window, starting at nucleotide 1 to produce input data with the correct reading frame. The "PHRED-scaled" values are used as a normalised and externally comparable unit of analysis, rather than raw CADD scores. The area under the curve was calculated for density plots to quantify the difference between pathogenic and unreported variants with high scores, above the intersects >0.0409 and >22.84 for MRF and CADD, respectively, using score value (x) versus density (y) (Fig. 2.12 (i-ii)) with $\int_a^b f(x)\,\mathrm{d}x \approx\ (b-a)\left[\frac{f(a)\ +f(b)}{2}\right]$.

2.3.4 Raw data availability and analysis script

The files *"Raw data R analysis for figures"* contains all raw data and analysis methods used to produce figures (except illustrations in Figures 2.4 and 2.11). These are available from the data repository from https://www.biorxiv.org/content/10.1101/272609v3. *"data analysis.R"* is an R script that contains the methods used to produce figures. Each of the input data CSV files are explained on first usage within the analysis script. Running *"data analysis.R"* from within the same directory as the associated input data CSV files will replicate analysis.

2.3.5 Data visualisation

For our visualisation of MRF scores, small clusters of high MRF values were of more appealing than individual highly conserved residues. Therefore, I applied a 1% average filter where values were averaged over a sliding window of N number of residues (10 in the case of RAG1, 6 in the case of RAG2). For a clear distinction of MRF clusters, a cut-off threshold was applied at the 75^{th} percentile (e.g. 0.0168 in RAG1) as shown in heatmaps in **Figure 2.4 (iii)** and **2.11**. The gene heatmaps for coding regions in *RAG1* and *RAG2* (**Fig. 2.4**) were populated with (i) Boolean C score from population genetics data, (ii) raw MRF scores, and (iii) MRF clusters with 1% average and cut-off threshold. GraphPad Prism was used for heatmaps. The data used for heatmaps is available in the appendix chapter 6 table 6.3 and in the R source to allow for alternative visualisations. An example of alternative output for non-R users is shown in **Figure 2.5**. Adobe Illustrator and Photoshop were used for protein domain illustrations in **Figure 2.4 (iv)**.

2.3.6 Validation of MRF against functional data

The recombination activity of RAG1 and RAG2 was previously measured on known or candidate pathogenic variants [17–19]. Briefly, the pathogenicity of variants in *RAG1* and *RAG2* was measured functionally *in vitro* by either expression of RAG1 and RAG2 in combination with a recombination substrate plasmid containing recombination signal

sequence (RSS) sites which are targeted by RAG complex during normal V(D)J recombination, or Abelson virus-transformed Rag2-/- pro-B cells with an RSS-flanked inverted GFP cassette. Recombination events were assessed by quantitative real-time PCR using comparative CT or expression of GFP evaluated by flow cytometry, respectively. The inverse score of recombination activity (0-100%) was used to quantify pathogenicity of variants in our study. Comparison between known pathogenicity scores and MRF was done by scaling MRF scores from 0-100% (100% being highest probability of occurring as damaging). A data and analysis is summarised in **Figure 2.3**.

2.3.7 Supplemental data tables

Data tables that are used in the published version of this chapter [1] can also be found in the appendix chapter 6 tables 6.1-6.3. These tables are not necessary for interpretation, however they summarise the raw data used in this study which can therefore be replicated.

2.4 Results

2.4.1 RAG1 and RAG2 conservation and mutation rate residue frequency

Variant probability prediction is dependent on population genetics data, among other factors. Our study queried GnomAD [25] to identify conserved residues using a Boolean score C of 0 (present in population) or 1 (conserved). The gene-specific mutation rate M_r of each residue was calculated from allele frequencies. The gene-specific residue frequency R_f represented the frequency of a residue occurring per gene, acquired by converting gene residue percentage (from the SMS2 tool *Protein stats*) to a frequency [26]. Together the values were used to calculate the most probable disease-causing variants which have not yet been identified in patients. I termed the resulting score a mutation rate residue frequency, where $MRF = C \times M_r \times R_f$. This score represents the likelihood that a clinically relevant mutation will occur.

Figure 2.4 presents the most probable unidentified disease-causing variants in

RAG1/2. Variants with a low MRF may still be damaging but resources for functional validation are best spent on gene regions with high MRF. Clusters of conserved residues are shown in **Figure 2.4 (i)** and are generally considered important for protein structure or function. However, these clusters do not predict the likelihood of mutation. Raw MRF scores are presented in **Figure 2.4 (ii)**. Histograms illustrates the MRF without Boolean scoring applied and **Figure 2.4 (iii)** provides a clearer illustration of top MRF score clusters. For visualisation, a noise reduction method was applied; a sliding window was used to find the average MRF per 1% interval of each gene. The resulting scores displayed in **Figure 2.4 (iii)** contain a cut-off threshold to highlight the top scoring residues (using the 75^{th} percentile). Variant sites most likely to present in disease cases are identified by high MRF scoring. This model may be expanded by the addition of phenotypic or epigenetic data (**Supplemental data tables**).

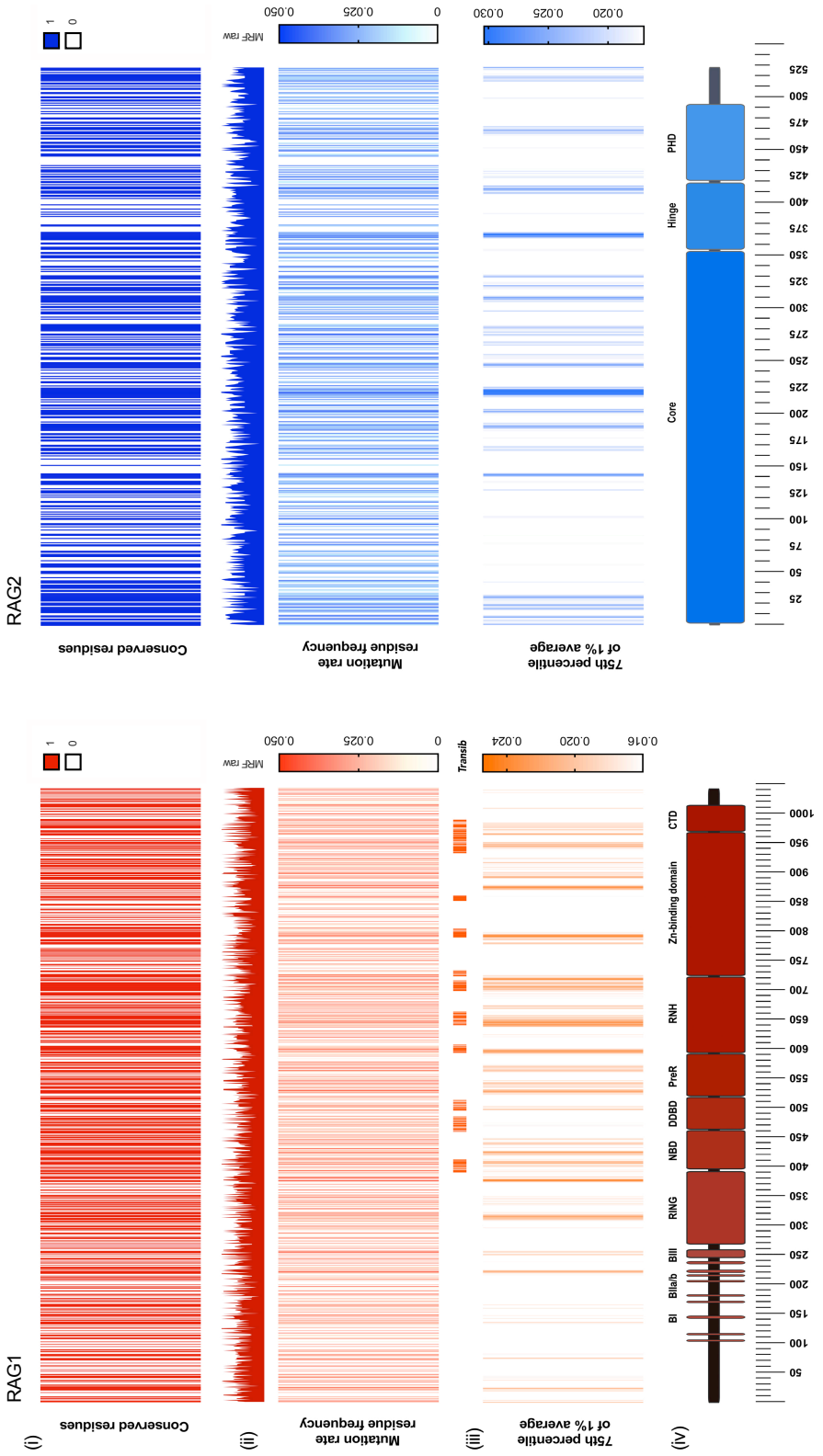


Figure 2.4: **RAG1 (red, left) and RAG2 (blue, right) conservation and mutation rate residue frequency (continued).**

Figure 2.4: **RAG1 (red, left) and RAG2 (blue, right) conservation and mutation rate residue frequency.** (i) Gene conservation score; non-conserved 0 and conserved 1. Colour indicates no known mutations in humans. (ii) Histogram; raw MRF score. Heatmap; MRF prediction for conserved residues, graded 0 to 0.05 (scale of increasing mutation likelihood with human disease). (iii) Coloured bars indicate most likely clinically relevant variant clusters. MRF score averaged with 1% intervals for each gene and cut-off below 75th percentile, graded 0 to 0.03 (noise reduction method). (iv) Gene structure with functional domains. Full list of residues and scores available in the appendix chapter 6.

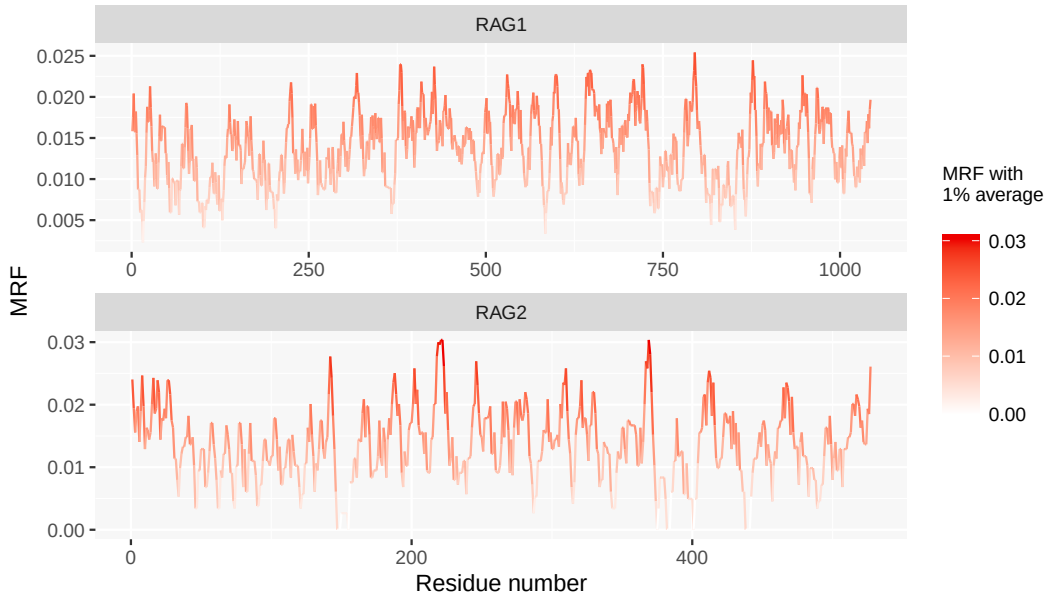


Figure 2.5: **An alternative visualisation of MRF scores for RAG1 and RAG2 proteins.** The data from the appendix chapter 6 table 6.3, "Average over 1%" is displayed on both the y-axis and colour scale. Raw data is also provided the appendix chapter 6.

The appendix chapter 6 provides all MRF scores for both proteins. Raw data used for calculations and the list of validated residues of RAG1 and RAG2 are available in order in the appendix chapter 6. **Table 2.3** shows the MRF mutation likelihood score for mutations that have also been reported as tested for recombination activity in functional assays. Analysis-ready files are also available in methods to the on-line data along and the associated R source file to allow for alternative visualisations as shown in **Figure 2.5**.

Table 2.3: **MRF likelihood scores for variants functionally assayed to date** [17–19]. Increased MRF score indicate higher likelihood of occurrence. Recombination activity measured by functional assay is shown as a percentage of wildtype (% SEM). Residues with multiple mutations are shown with both alternative variants and values. $MRF_{max} = 0.043$ and $MRF_{min} = 0.004$. The full table of all protein positions can be found in the appendix chapter 6.

RAG1

MRF	Residue	Assayed	Recombition activity (%)	MRF	Residue	Assayed	Recombition activity (%)
0.030	56	I56T	3.5 ± 0.2	0.022	539	D539V	3.2 ± 0.2
0.030	86	K86VfsX33	2.7 ± 0.3	0.025	541	L541CfsX30	1.2 ± 0.9
0.014	99	G99S	113.2 ± 3.7	0.043	559	R559S	1.0 ± 0.4
0.012	106	N106K	80.4 ± 16.4	0.043	561	R561H	2.0 ± 0.6
0.043	108	R108X	1.8 ± 0.3	0.041	601	S601P	0.0 ± 0.0
0.043	142	R142X	9.0 ± 4.0	0.026	612	H612R	121.6 ± 0.9
0.032	174	E174SfsX27	0.5 ± 0.2	0.043	624	R624H	0.0 ± 0.4
0.027	246	A246TfsX17	0.8 ± 0.1	0.041	626	S626X	0.0 ± 0.0
0.012	248	Q248X	1.2 ± 0.2	0.041	651	S651P	0.5 ± 0.5
0.026	249	H249R	112.2 ± 3.5	0.043	699	R699Q,W	45.9 ± 1.5, 19.3 ± 1.8
0.043	314	R314W	24.3 ± 5.2	0.032	722	E722K	0.0 ± 0.2
0.012	328	C328Y	16.0 ± 2.9	0.012	730	C730F	0.0 ± 0.0
0.030	383	K383RfsX7	0.1 ± 0.0	0.025	732	L732P	0.0 ± 0.0
0.013	386	F386CfsX4	0.2 ± 0.1	0.043	737	R737H	0.2 ± 0.0
0.030	391	K391E	6.5 ± 1.6	0.043	759	R759C	17.2 ± 3.3
0.043	394	R394Q	0.1 ± 0-0.1	0.043	764	R764P	0.0 ± 0.0
0.043	396	R396C	0.4-0.6 ± 0-0.1	0.008	768	Y768X	0.0 ± 0.0
0.041	401	S401P	0.0 ± 0.0	0.032	770	E770K	21.0 ± 0.4
0.020	403	T403P	0.0 ± 0.0	0.043	778	R778Q,W	8.6 ± 1.0, 4.6 ± 0.6
0.043	404	R404Q	1.2 ± 0.1	0.028	786	P786L	0.0 ± 0.1
0.043	410	R410Q	0.0 ± 0.0	0.030	820	K820R	117.9 ± 6.3
0.025	411	L411P	0.0 ± 0.0	0.025	836	L836V	75.0 ± 1.3
0.022	429	D429G	0.1 ± 0.0	0.043	841	R841Q,W	0.0 ± 0.0, 10.0 ± 0.5
0.028	433	V433M	0.2 ± 0.0	0.027	868	A868V	100.0 ± 5.0
0.019	435	M435V	23.6 ± 4.8	0.005	896	W896R	0.9 ± 0.1
0.027	444	A444V	1.4 ± 0.2	0.008	912	Y912C	6.9 ± 0.4
0.043	449	R449K	92.1 ± 3.6	0.005	959	W959X	0.0 ± 0.0
0.025	454	L454Q	5.4 ± 0.7	0.032	965	E965X	0.0 ± 0.0
0.019	458	M458SfsX34	0.0 ± 0.0	0.043	973	R973C	0.0 ± 0.2
0.027	472	A472V	0.4 ± 0.0	0.013	974	F974L	56.5 ± 0.8
0.043	474	R474C	125.4 ± 2.6	0.043	975	R975W,Q	57.9 ± 1.6, 53.5 ± 3.6

0.028	475	V475AfsX17	0.1 ± 0.0	0.012	981	Q981P	7.2 ± 0.1
0.025	506	L506F	1.0 ± 0.1	0.030	983	K983NfsX9	0.1 ± 0.0
0.043	507	R507W	15.9 ± 0.8	0.030	992	K992E	9.1 ± 1.2
0.014	516	G516A	40.2 ± 1.3	0.030	1006	M1006V	105.6 ± 6.8
0.005	522	W522C	41.6 ± 1.9				

RAG2

MRF	Residue	Assayed	Recombition activity (%)	MRF	Residue	Assayed	Recombition activity (%)
0.013	1	M1T	65.3 ± 2.2	0.013	285	M285R	24.7 ± 0.8
0.006	16	Q16X	1.7 ± 0.4	0.004	307	W307X	0.2 ± 0.2
0.038	35	G35A,V	22.1 ± 3.1, 0.4 ± 0.3	0.017	386	F386L	109.1 ± 5
0.023	39	R39G	0.2 ± 0.1	0.025	407	E407X	2.9 ± 0.4
0.011	41	C41W	0.2 ± 0.4	0.004	416	W416L	1.4 ± 0.2
0.017	62	F62L	19.6 ± 3	0.025	437	E437K	0.9 ± 0.2
0.028	65	D65Y	6.8 ± 1.2	0.017	440	K440N	26.7 ± 2.4
0.023	73	R73H	12.4 ± 1.4	0.013	443	M443I	0.4 ± 0.2
0.034	77	T77N	42.6 ± 2.7	0.027	444	I444M	2.7 ± 0.3
0.038	95	G95R	0.3 ± 0.2	0.011	446	C446W	2.9 ± 0.1
0.013	110	M110L	74.6 ± 1.8	0.038	451	G451A	66.3 ± 4.8
0.017	127	K127X	0.1 ± 0	0.004	453	W453R	0.6 ± 0.1
0.038	157	G157V	0.4 ± 0.2	0.017	456	A456T	16 ± 2.9
0.030	160	S160L	5.8 ± 0.6	0.013	459	M459L	30.8 ± 0.6
0.023	180	P180H	31.1 ± 0.5	0.034	474	N474S	97.5 ± 5.9
0.019	195	Y195D	2 ± 0.3	0.011	478	C478Y	0.2 ± 0.1
0.034	215	T215I	67.2 ± 1	0.025	480	E480X	2.8 ± 0.6
0.023	229	R229Q,W	8.9 ± 1, 10.5 ± 0.5	0.017	481	H481P	23.8 ± 3.9
0.023	253	P253R	95.4 ± 2.3	0.013	502	M502V	99.6 ± 3.4
0.006	278	Q278X	0.1 ± 0.1				

2.4.2 MRF scores select for confirmed variants in human disease

I have applied MRF scores to known damaging mutations from other extensive reports in cases of human disease [13, 16, 18, 20, 21, 32–55] [*originally compiled by* Notarangelo et al. [24]]. This dataset compares a total of 44 variants. I expected that functionally damaging variants (resulting in low recombination activity in vitro) that have the highest probability of occurrence would be identified with high MRF scores. MRF prediction correctly identified clinically relevant mutations in *RAG1* and *RAG2* (**Fig. 2.6 (i)**). Variants reported on GnomAD which are clinically found to cause disease had significantly higher MRF scores than variants which have not been reported to cause disease. I observed that rare and likely mutations provided high scores while rare but unlikely or common variants had low scores (**Fig. 2.6 (i)**).

Allele frequency is generally the single most important filtering method for rare disease

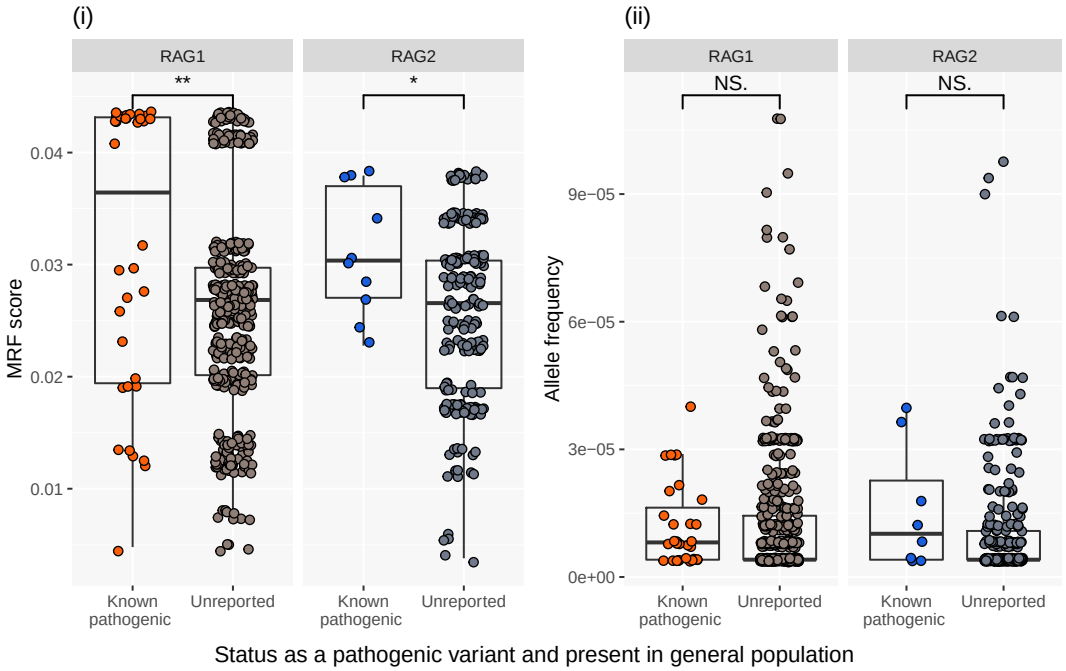


Figure 2.6: **RAG1 and RAG2 MRF score predict the likelihood of mutations that are clinically relevant.** (i) Known damaging variants (clinically diagnosed with genetic confirmation) reported on GnomAD have significantly higher MRF scores than unreported variants. (ii) GnomAD rare variant allele frequency <0.0001. No significant difference in allele frequency is found between known damaging and non-clinically reported variants. Unpaired t-test. RAG1 P-value 0.002** RAG2 P-value 0.0339*. MRF; mutation rate residue frequency, ns; non-significant.

in whole genome (and exome) sequencing experiments. Variants under pressure from purifying selection are more likely to cause disease than common variants. However, most RAG mutations are rare. Therefore, allele frequencies of rare variants reported on GnomAD cannot differentially predict the likelihood of causing disease (**Fig. 2.6 (ii)**). As such, we found no significant difference between known damaging variants and those that have not yet been reported as disease-causing. The comparison between **Figure 2.6 (i) and (ii)** illustrates the reasoning for the design of our method.

Many non-clinically-reported rare variants may cause disease; the MRF score identifies the top clinically relevant candidates. Based on the frequency of protein-truncating variants in the general population, *RAG1* and *RAG2* are considered to be tolerant to the loss of one allele, as indicated by their low probability of being loss-of-function intolerant (pLI) scores of 0.00 and 0.01, respectively [25]. This is particularly important for recessive

diseases such as RAG deficiency where most new missense variants will be of unknown significance until functionally validated.

2.4.3 Top candidate variants require validation

Functionally characterising protein activity is both costly and time consuming. RAG1 and RAG2 have now been investigated by multiple functional assays for at least 110 coding variants [17–19]. In each case, researchers selected variants in *RAG1* and *RAG2* that were potentially damaging or were identified from PID patients as the most probable genetic determinant of disease. Functional assays for RAG deficiency in those cases, and generally, measured a loss of recombination activity as a percentage of wild type function (0-100%).

Pre-emptively performing functional variant studies benefits those who will be identified with the same variants in the future, before the onset of disease complications. While more than 100 variants have been assayed in vitro, we calculated that only one-quarter of them are most probable candidates for clinical presentation. **Figure 2.7** illustrates that while functional work targeted "hand-picked" variants that were ultimately confirmed as damaging, many of them may be unlikely to arise based on population genetics data. **Figure 2.7** presents, in increasing order, the number of potential variants based on likelihood of presentation and stacked by the number of variants per score category. Variants that have been measured for their loss of protein activity are coloured by severity. Potential variants that remain untested are coloured in grey. Only 21 of the top 66 most probable clinically relevant variants have been assayed in *RAG1*.

Figure **2.8** further illustrates the individual variants which have been tested functionally (the coloured *recombination activity* subset of Fig 2.7). We compared predicted MRF scores to assay measurements for 71 *RAG1* and 39 *RAG2* mutants. Most mutations tested showed severe loss of protein function (bottom panel of Figure **2.8**), while the likelihood each mutation occurring in humans varied significantly (top panels).

If MRF scoring was used in the same cases pre-emptively, the loss of investment would be minimal; only 8 variants out of 71 mutants tested had an above-average MRF

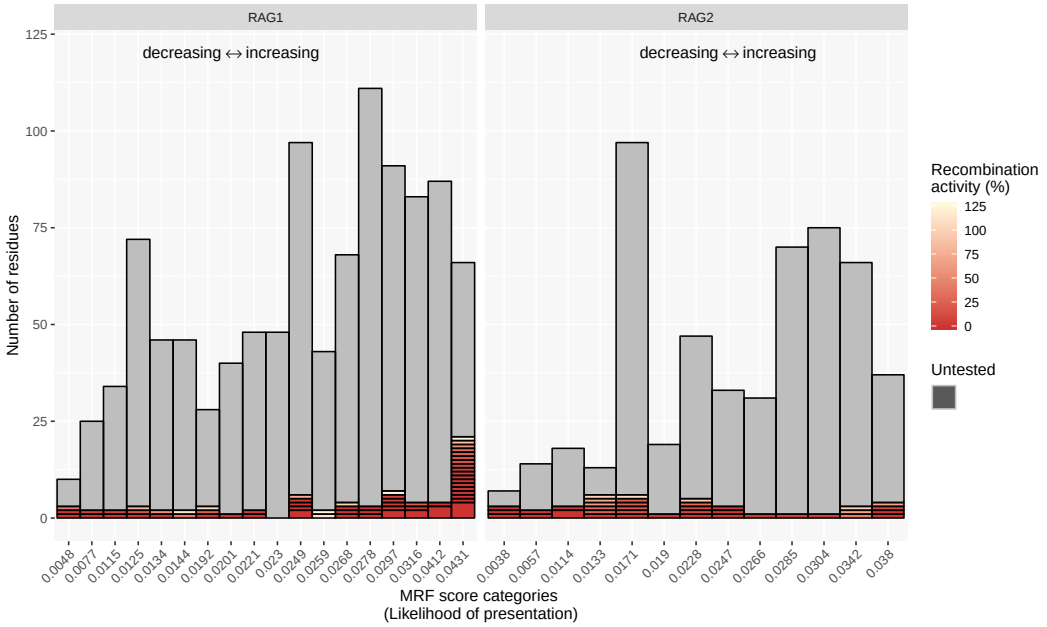


Figure 2.7: **RAG1 and RAG2 MRF score categories and variants assayed to date.** Protein residues are ranked and stacked into categories based on their MRF score. High scores (0.043 and 0.038 in RAG1 and RAG2, respectively) represent a greater mutation likelihood. Functional assays have measured recombination activity (as its inverse; % loss of activity) in a total of 110 mutants. The severity of protein loss of function is represented by a red gradient. Residues that have not been functionally tested are shown in grey. While many protein residues are critical to protein function, their mutation is less probable than many of the top MRF candidates. Data further expanded in Figure 2.8. MRF; mutation rate residue frequency.

score while being measured as functionally benign (a rate of 11.27%). RAG2 had only 3 out of 39 variants (7.69%) with an above-average MRF score while functionally benign. For the expended resources, approximately 30% more top candidates would have been tested in place of unlikely and functionally non-damaging mutations. However, the true measurement of accuracy is limited in that very few of the most likely clinically relevant variants predicted by MRF scoring have been tested to date.

2.4.4 False positives in *Transib* domains do not negatively impact prediction

Adaptive immunity is considered to have evolved through jawed vertebrates after integration of the RAG transposon into an ancestral antigen receptor gene [56, 57]. The *Transib*

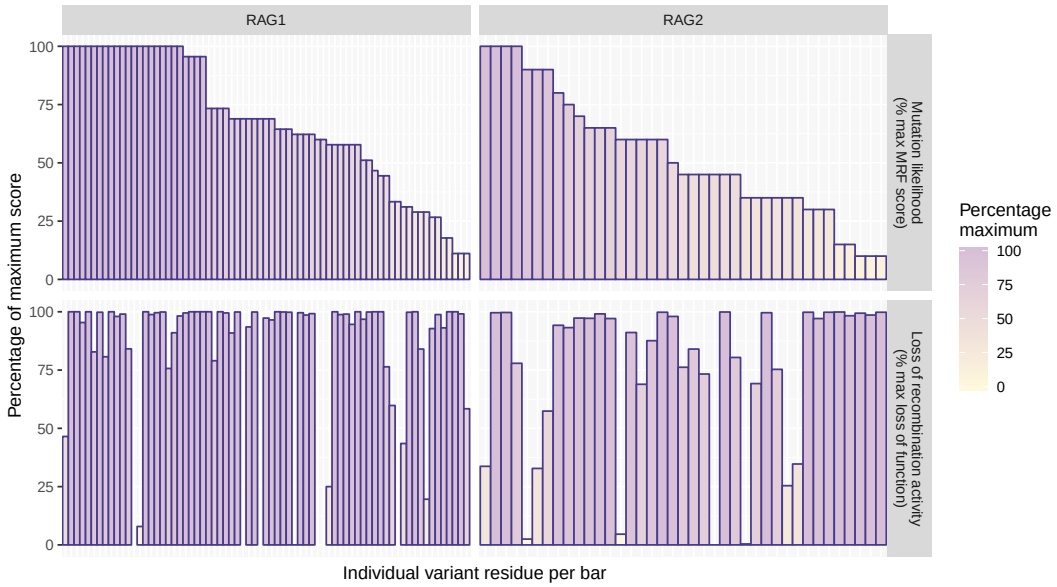


Figure 2.8: **MRF likelihood score versus known functional activity.** We compiled all variants that we know to have been assayed for protein function to date. The inverse of functional assay measurements were used, where 0% activity represents 100% loss of activity. MRF scores are presented as a percentage of the maximum score per gene (i.e., for RAG1 $MRF_{max} = 0.043$ (100%) and $MRF_{min} = 0.0048$ (0%)). **Top panels** show how likely each mutation is predicted to occur in humans. **Bottom panels** show the loss of protein activity as a percentage compared to wild-type (% SEM); most mutations tested produced severe loss of protein function, regardless of their mutation likelihood. Subset of *Recombination activity* data from Figure 2.7.

transposon is a 600 amino acid core region of RAG1 that targets RSS-like sequences in many invertebrates. A linked *RAG1/RAG2* was shown in the lower dueterosome (sea urchin), indicating an earlier common ancestor than the invertebrate [58], and more recently, a recombinatorially active RAG transposon (ProtoRAG) was found in the lower chordate amphioxus (or lancelet); the most basal extant chordate and a "living fossil of RAG" [59].

A set of conserved motifs in core *RAG1* are shared with the *Transib* transposase, including the critical DDE residue catalytic triad (residues 603, 711, and 965) [60]. Ten *RAG1* core motifs are conserved amongst a set of diverse species including human [60]. This evolutionarily conserved region is considered as most important to protein function. Therefore, we chose this region to determine if MRF scoring would have a negative impact if mutations were falsely predicted as clinically important. To assess the influence

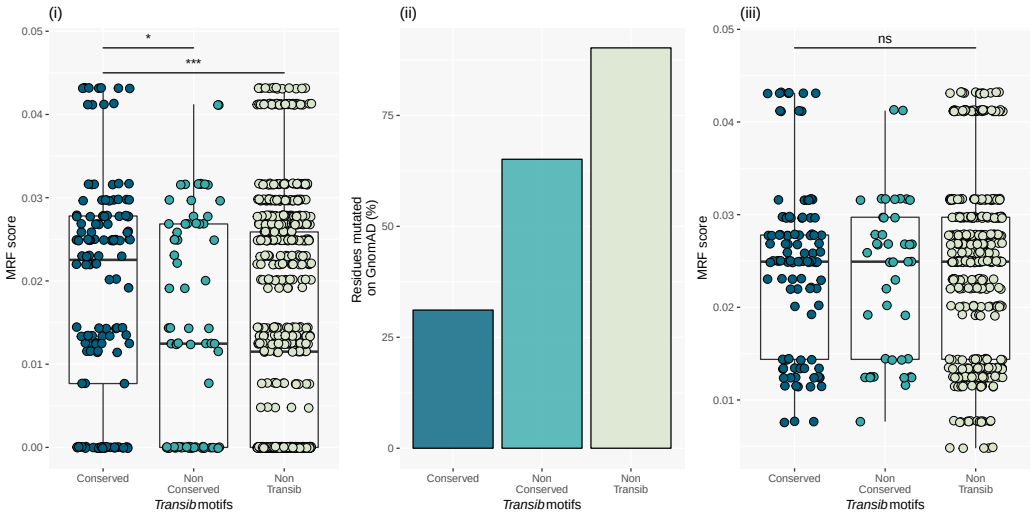


Figure 2.9: **False positives in *Transib* domains do not worsen probability prediction.** The *Transib* domains contain critical conserved protein residues. (i) False positives were simulated by scoring *Transib* domain MRF without omitting Boolean conservation weight $C = 0$. (ii) Allele frequencies on GnomAD had conservation levels inversely proportional to simulated false-positive MRF scoring. (iii) When testing for all Boolean component $C > 0$ after MRF calculation the effect of false positives remained non-significant, illustrating the non-negative impact of MRF for predicting the mutation. Unpaired t-test, * P = 0.0195, *** P < 0.0001. MRF; mutation rate residue frequency, ns; non-significant.

of a false positive effect on prediction, the MRF scores for conserved residues in this group were compared to GnomAD allele frequencies. **Figure 2.9 (i)** plots the MRF (without omitting the Boolean component $C = 0$) for conserved *Transib* motif residues, non-conserved *Transib* motif residues, and non-*Transib* residues. **Figure 2.9 (ii)** shows the percentage of these which were reported as mutated on GnomAD. By accounting for unreported variants by applying $C > 0$, the resulting effect on incorrectly scoring MRF in the conserved *Transib* motifs remained neutral.

2.4.5 MRF predicts RAG deficiency amongst PID patients harbouring rare variants

We have previously measured the recombination activity of RAG1 and RAG2 disease-causing variants in several patients [17]. We have compiled our own and other functional assay data from Lee et al. [18] and Tirosh et al. [19] to produce a panel of recombination

activity measurements for coding variants in both *RAG1* and *RAG2*. RAG deficiency was measured as the level of recombination potential produced by the protein complex. Each method of investigation simulated the efficiency of wild-type or mutant proteins expressed by patients for their ability to produce a diverse repertoire of T-cell receptor (TCR) and B-cell receptor (BCR) and coding for immunoglobulins. In functional experiments, mutant proteins were assayed for their ability to perform recombination on a substrate which mimics the RSS of TCR and BCR in comparison to wild-type protein complex (as % SEM).

By gathering confirmed RAG deficiency cases, we compiled the MRF scores for 43 damaging *RAG1* variants in 77 PID cases and 14 damaging *RAG2* variants in 21 PID cases (MRF scores spanning over 22 categories). To test our method against a strong control group, we identified coding variants in patients with PID where RAG deficiency due to coding variants has been ruled out as the cause of disease. We obtained *RAG1/2* variants in 558 PID patients who had their genomes sequenced as part of the NIHR BioResource - Rare Diseases study [17]. Filtering initially identified 32 variants in 166 people. This set was trimmed to contain only rare variants; 29 variants over 26 MRF scoring categories from 72 cases of non-RAG-deficient PID. Linear regression on this control group produced negative or near-zero slopes for *RAG1* and *RAG2*, respectively. The same analysis for known-damaging mutations in disease cases had a significant prediction accuracy for *RAG1*. Analysis for *RAG2* was not significant. However, the sample size to date may be too small to significantly measure *RAG2* MRF scoring although a positive correlation was inferred in **Figure 2.10** [61]. A link to the R source and raw data can be found in methods.

2.4.6 MRF supplements pathogenicity prediction tools for translational research

CADD scoring [27] is an important bioinformatics tool that exemplifies pathogenicity prediction. While CADD is a valuable scoring method, its purpose is not to predict likelihood of variation. Similarly, MRF scoring is not a measure of pathogenicity. MRF scoring may be complemented by tools for scoring variant deleteriousness. We compare

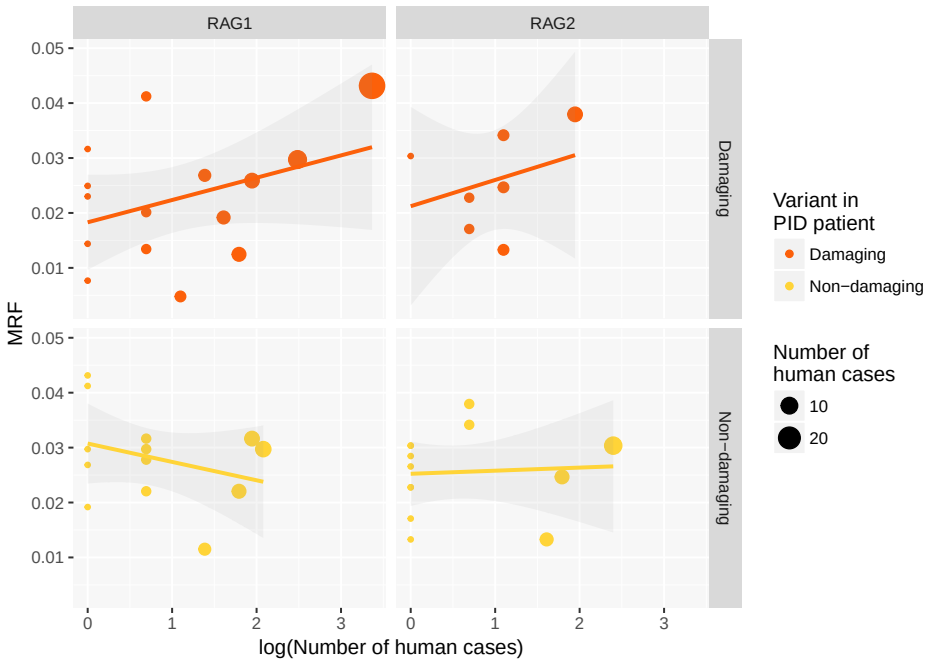


Figure 2.10: **A linear regression model of RAG1/2 MRF scoring in cases of primary immune deficiency.** MRF prediction correlates with clinical presentation. Damaging variants identified in confirmed RAG deficiency cases. Non-damaging variants sourced from cases of PID with rare variants but not responsible for disease. (Slopes of RAG1: Damaging: 0.0008* (± 0.0004) P<0.05, intercept 5.82e-05 ***, Non-damaging: -0.0007 (± 0.001). Slopes of RAG2; Damaging: 0.0023 (± 0.0018), intercept 0.0312 *, Non-damaging 0.0001 (± 0.0008). Source data and script in methods).

MRF to the PHRED-scaled CADD scores for all possible SNV positions in *RAG1* (**Fig. 2.11**) illustrating that pathogenicity prediction cannot account for mutation probability. Combining both methods allows researchers to identify highly probable mutations before querying predicted pathogenicity.

To further develop this concept, we first annotated variants with MRF likelihood scores and pathogenic prediction PHRED-scaled CADD scores (**Figure 2.12**), and secondly, performed a manual investigation of the clinical relevance of top candidates (**Table 2.4**). We used HGMD as an unbiased source of known RAG deficiency cases in both instances. CADD score was very successful at predicting the pathogenicity of a variant, (a high-density cluster of variants with CADD scores >25) as shown in **red** in **Figure 2.12 (i)**. At about the same rate, CADD score also predicted variants as pathogenic that are, to date, unreported (as **pink** in **Fig. 2.12 (i)**). Indeed, those unreported variants

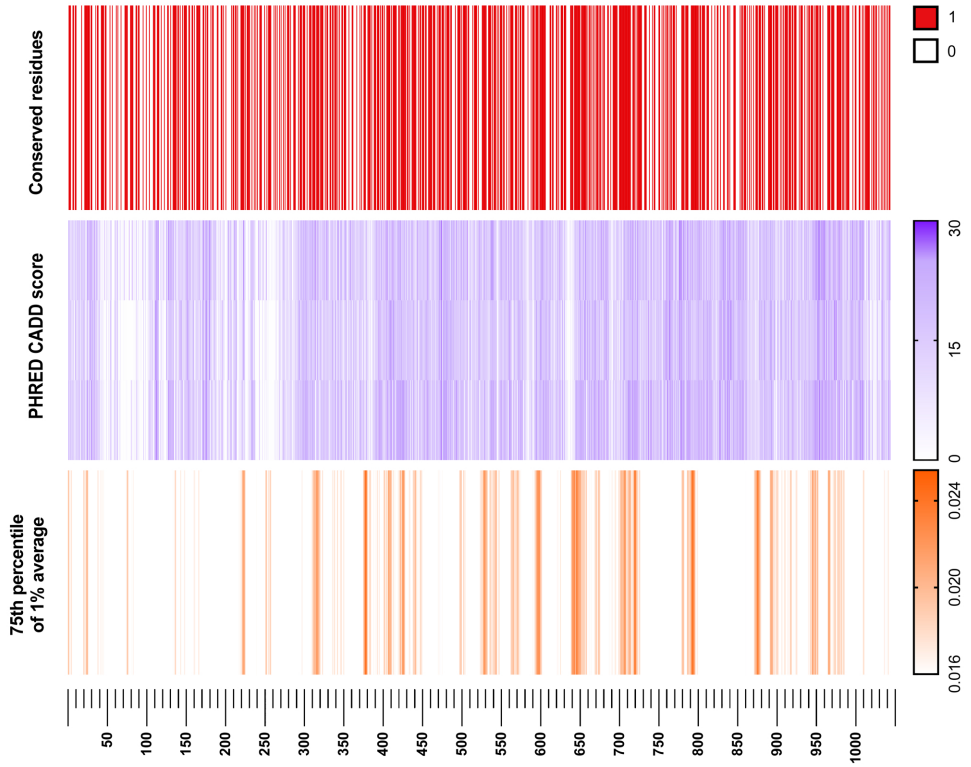


Figure 2.11: ***RAG1* PHRED-scaled CADD score versus GnomAD conservation rate and MRF score.** Allele frequency conservation rate (**top**) is vastly important for identifying critical structural and functional protein regions. The impact of mutation in one of these conserved regions is often estimated using CADD scoring (**middle**). CADD score heatmap is aligned by codon and separated into three layers for individual nucleotide positions. The MRF score (**bottom**)(visualised using the 75^{th} percentile with 1% averaging) highlights protein regions which are most likely to present clinically and may require pre-emptive functional investigation.

may very well be pathogenic. However, the likelihood of each mutation varies. As such, we developed the MRF score to account for that likelihood. As expected, the likelihood of mutations occurring that were unreported was low according to MRF (**Fig. 2.12 (ii)**, **pink**), while the mutations which did occur were highly enriched in at high MRF scores (**Fig. 2.12 (ii)**, **red** high-density cluster >0.043). Combining mutation prediction (MRF) with pathogenicity prediction (tools like CADD) increases the accuracy of pre-emptively targeting clinically relevant variants. **Figure 2.12 (iii)** shows that while the number of variants presented to date is relatively small, they already account for 36% of the top MRF score candidates.

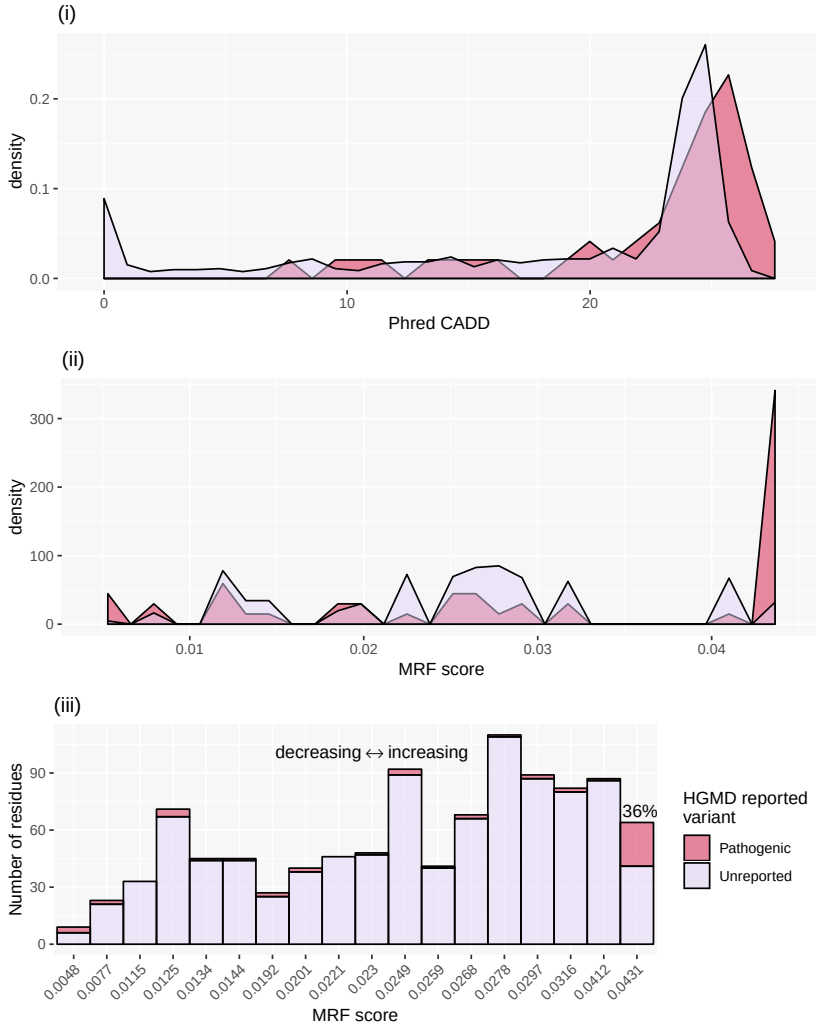


Figure 2.12: ***RAG1* PHRED-scaled CADD score versus MRF score against HGMD data.** (i) A high CADD score is a predictor of deleteriousness. Both reported (red) and non-reported residues (pink) have a high density of high CADD score. (ii) MRF scores only show a high-density cluster for high-likelihood variants, reflected by the high MRF score observed for known RAG deficiency variants. The number of pathogenic variants is outweighed by conserved residues; (i-ii) shows density of scores to normalise between groups. AUC overlap difference in CADD score of 21.43% and MRF score of 74.28% (above intersects >22.84 and >0.0409, in *(i-ii)* respectively). (iii) The number of residues per MRF category shows that disease reported on HGMD accounts for 36% of top MRF candidates. AUC; area under curve, CADD; Combined Annotation Dependent Depletion, HGMD; Human Gene Mutation Database.

2.4.7 Clinical relevance of top candidates

The top scoring candidates in RAG1 were assessed for potential clinical relevance (**Table 2.4**). HGMD was chosen as a reliable, curated source of identifying pathogenic variants. 45% of RAG1 variants reported on HGMD (23 of 51) were predicted by our model as the most likely candidates seen clinically (the top scoring MRF group of had 66 residues total). The remaining variants in the top MRF group, which were not reported by HGMD (43 of 66), were assessed manually for their likelihood as potentially disease causing. 21 (49%) were highly conserved, not reported on GnomAD, and would be considered probable RAG deficiency on presentation as homozygous or compound heterozygous with a second damaging variant. The remainder had allele frequencies <0.0006, were only found as low frequency heterozygous in the general population and justify functional validation. We expect that none of the top candidate mutations are benign.

Table 2.4: **Clinical relevance of top candidates.** 23 top MRF score variants were reported as pathogenic on HGMD to date. The remaining variants (the 43 not reported) were assessed by their frequency in population based on GnomAD (allele frequencies vary between individual variants but equate to approximately <6* and 9-77** heterozygous per 125,000 individuals). Therefore, no top candidates should be considered benign without functional validation. HGMD; Human Gene Mutation Database, MRF; mutation rate residue frequency.

Most likely mutation candidates	
Variant type	**Number variants**
Top MRF score candidates total	66
(i) Of which are reported on HGMD	23
(ii) Not reported on HGMD to date	43

Top MRF score candidates not reported on HGMD			
Frequency in Population	**Number variants**	**Unreported top candidate (%)**	**GnomAD allele frequency**
Not found	21 of 43	49%	0
Very rare	15 of 43	35%	<0.00002*
Very rare	7 of 43	16%	<0.00006**

2.4.8 Protein structure application

With the availability of a structured protein complex, modelling can be carried out prior to functional assays. Residues with the highest MRF for both RAG1 and RAG2 were mapped in **Figure 2.13**.

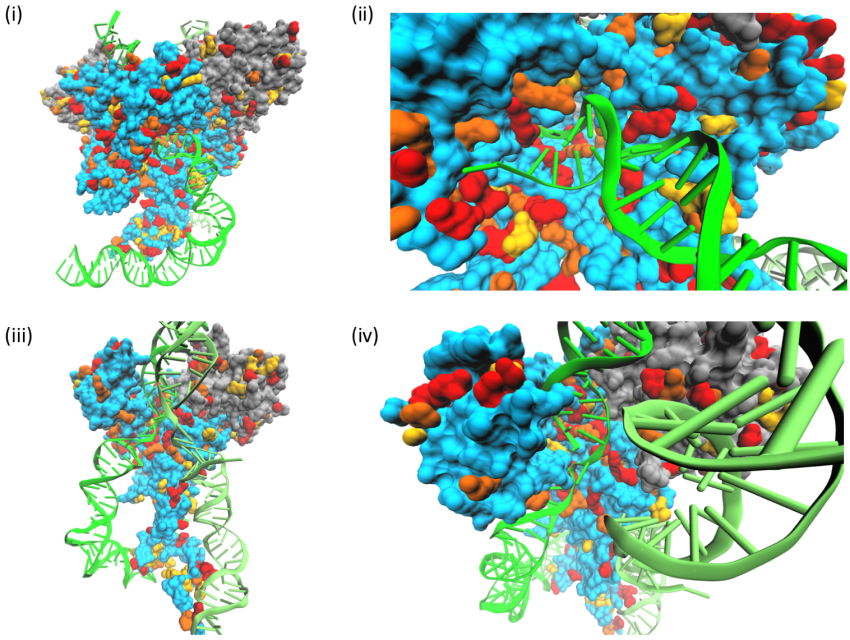


Figure 2.13: **The RAG1 (blue) and RAG2 (grey) protein structure with top candidate MRF scores.** (i) Protein dimers and (ii=iv) monomers illustrating the three highest category MRF scores for predicted clinically relevant variants. Increasing in score the top three MRF categories (illustrated in **Figure 2.7**) for each protein are highlighted; yellow, orange, red. DNA (green) is bound by the RAG protein complex at recombination signal sequences (PDB:3jbw).

2.4.9 Genome-wide and disease-specific application

Weighting data can also be applied to the MRF score model to amplify the selectivity. The mutation rate can be applied genome wide with a process common in the study of information retrieval; term frequency, inverse document frequency ($tf - idf$). In this case the "term" and "document" are replaced by amino acid residue r and gene g , respectively such that,

$$rf - igf_{r,g} = rf_{r,g} \times igf_r \tag{2.1}$$

We may view each gene as a vector with one component corresponding to each residue mutation in the gene, together with a weight for each component that is given by (1). Therefore, we can find the overlap score measure with the $rf - igf$ weight of each term in g, for a query q;

$$\text{Score}(q, g) = \sum_{r \in q} \text{rf-igf}_{r,g}.$$

In respect to MRF scoring, this information retrieval method might be applied as follows; the $rf - igf$ weight of a term is the product of its rf weight and its igf weight ($W_{r,g} = rf_{r,g} \times \log \frac{N}{gf_r}$) or ($W_{r,g} = (1 + \log rf_{r,g}) \times \log \frac{N}{gf_r}$). That is, firstly, the number of times a residue mutates in a gene ($rf = rf_{r,g}$) and secondly, the rarity of the mutation genome-wide in N number of genes ($igf = N/gf_r$). Finally, ranking the score of genes for a mutation query q by;

$$\text{Score}(q, g) = \sum_{r \in q \cap g} \text{rf-igf}_{r,g}$$

The score of the query (Score(q, g)) equals the mutations (terms) that appear in both the query and the gene ($r \in q \cap g$). Working out the $rf - igf$ weight for each of those variants ($rf.igf_{r,g}$) and then summing them ($\sum$) to give the score for the specific gene with respect to the query.

2.4.10 Bayesian probability

MRF score may provide a limiting component required for applying Bayesian probability to disease prediction. A clinician may ask for the likelihood of RAG deficiency (or any Mendelian disease of interest) for a patient given a set of gene variants $P(H|E)$ using Bayes' theorem,

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where $P(H)$ is the probability of a patient having RAG deficiency, $P(E|H)$ is the probability of RAG deficiency due to a set of variants that have been pre-emptively assayed, and $P(E)$ is the probability of having a set of gene variants.

$P(H)$ is known since the rate of RAG deficiency is estimated at an incidence of 1:181,000 [62], SCID at a rate of 1:330,000 [3], and we also recently show the rate of RAG deficiency in adults with PID [17]. Being a recessive disease, $P(E)$ must account for biallelic variants and is the most difficult value to determine. This may be found from population genetics data for (i) the rate of two separate, compound heterozygous variants, (ii) the rate of a homozygous variant or potential consanguinity, or (iii) the rate of de novo variation [25]. $P(E|H)$ would be identified where all variants are functionally validated. This requires a major investment, however the MRF score provides a good approximation.

2.5 Discussion

Determining disease-causing variants for functional analysis typically aims to target conserved gene regions. On GnomAD 56% of *RAG1* (approx. 246,000 alleles) is conserved with no reported variants. Functional validation of unknown variants in genes with this level purifying selection is generally infeasible. Furthermore, we saw that a vast number of candidates are "predicted pathogenic" by commonly used pathogenicity tools, which may indeed be damaging but unlikely to occur. To overcome the challenge of manual selection we quantified the likelihood of mutation for each candidate variant.

Targeting clearly defined regions with high MRF scores allows for functional validation studies tailored to the most clinically relevant protein regions. An example of high MRF score clustering occurred in the RAG1 catalytic RNase H (RNH) domain at p.Ser638-Leu658 which is also considered a conserved *Transib* motif.

While many hypothetical variants with low MRF scores may be uncovered as functionally damaging, our findings suggest that human genomic studies will benefit by first targeting variants with the highest probability of occurrence (gene regions with high

MRF). The appendix chapter 6 lists the values for calculated MRFs for RAG1 and RAG2.

We have presented a basic application of MRF scoring for RAG deficiency. The method can be applied genome-wide. This can include phenotypically derived weights to target candidate genes or tissue-specific epigenetic features. In the state presented here, MRF scores are used for pre-clinical studies. A more advanced development may allow for use in single cases. During clinical investigations using personalised analysis of patient data, further scoring methods may be applied based on disease features. A patient phenotype can contribute a weight based on known genotype correlations separating primary immunodeficiencies or autoinflammatory diseases [7]. For example, a patient with autoinflammatory features may require a selection that favours genes associated with proinflammatory disease such as *MEFV* and *TNFAIP3*, whereas a patient with mainly immunodeficiency may have preferential scoring for genes such as *BTK* and *DOCK8*. In this way, a check-list of most likely candidates can be confirmed or excluded by whole genome or panel sequencing. However, validation of these expanded implementations requires a deeper consolidation of functional studies than is currently available.

Havrilla et al. [63] have recently developed a method with similar possible applications for human health mapping constrained coding regions. Their study employed a method that included weighting by sequencing depth. Similarly, genome-wide scoring may benefit from mutation significance cut-off, which is applied for tools such as CADD, PolyPhen-2, and SIFT [64]. We have not included an adjustment method as our analysis was gene-specific but implementation is advised when calculating genome-wide MRF scores.

The MRF score was developed to identify the top most probable variants that have the potential to cause disease. It is not a predictor of pathogenicity. However, MRF may contribute to disease prediction; a clinician may ask for the likelihood of RAG deficiency (or any other Mendelian disease of interest) prior to examination (**2.4.10 Bayesian probability**).

Predicting the likelihood of discovering novel mutations has implications in genome-wide association studies (GWAS). Variants with low minor allele frequencies have a low discovery rate and low probability of disease association [65], an important consideration

for rare diseases such as RAG deficiency. An analysis of the NHGRI-EBI catalogue data highlighted diseases whose average risk allele frequency was low [65]. Autoimmune diseases had risk allele frequencies considered low at approximately 0.4. Without a method to rank most probable novel disease-causing variants, it is unlikely that GWAS will identify very rare disease alleles (with frequencies <0.001). It is conceivable that a number of rare immune diseases are attributable to polygenic rare variants. However, evidence for low-frequency polygenic compounding mutations will not be available until large, accessible genetics databases are available, exemplified by the NIHR BioResource Rare Diseases study [17]. An interesting consideration when predicting probabilities of variant frequency, is that of protective mutations. Disease risk variants are quelled at low frequency by negative selection, while protective variants may drift at higher allele frequencies [66].

The cost-effectiveness of genomic diagnostic tests is already outperforming traditional, targeted sequencing [2]. Even with substantial increases in data sharing capabilities and adoption of clinical genomics, rare diseases due to variants of unknown significance and low allele frequencies will remain non-actionable until reliable predictive genomics practices are developed. Bioinformatics as a whole has made staggering advances in the field of genetics [67]. Challenges that remain unsolved, hindering the benefit of national or global genomics databases, include DNA data storage and random-access retrieval [68], data privacy management [69], and predictive genomics analysis methods. Variant filtration in rare disease is based on reference allele frequency, yet the result is not clinically actionable in many cases. Development of predictive genomics tools may provide a critical role for single patient studies and timely diagnosis [23].

2.6 Conclusion

We provide a list of amino acid residues for RAG1 and RAG2 that have not been reported to date but are most likely to present clinically as RAG deficiency. This method may be applied to other diseases with hopes of improving preparedness for clinical diagnosis.

Bibliography

[1] Dylan Lawless, Hana Lango Allen, James Thaventhiran, NIHR BioResource–Rare Diseases Consortium, Flavia Hodel, Rashida Anwar, Jacques Fellay, Jolan E. Walter, and Sinisa Savic. Predicting the occurrence of variants in rag1 and rag2. *Journal of Clinical Immunology*, 2019.

[2] Katherine Payne, Sean P Gavan, Stuart J Wright, and Alexander J Thompson. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nature Reviews Genetics*, 2018.

[3] Antonia Kwan, Roshini S Abraham, Robert Currier, Amy Brower, Karen Andruszewski, Jordan K Abbott, Mei Baker, Mark Ballow, Louis E Bartoshesky, Vincent R Bonagura, et al. Newborn screening for severe combined immunodeficiency in 11 screening programs in the united states. *Jama*, 312(7):729–738, 2014.

[4] L. Alexander Liggett, Anchal Sharma, Subhajyoti De, and James DeGregori. Conserved patterns of somatic mutations in human peripheral blood cells. *bioRxiv*, 2017. doi: 10.1101/208066.

[5] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):781, 2018.

[6] István Bartha, Julia di Iulio, J Craig Venter, and Amalio Telenti. Human gene essentiality. *Nature Reviews Genetics*, pages nrg–2017, 2017.

[7] Capucine Picard, H Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Yanick J Crow, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, et al. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. *Journal of clinical immunology*, 38(1):96–128, 2018.

[8] Mary Ellen Conley and Jean-Laurent Casanova. Discovery of single-gene inborn errors of immunity by next generation sequencing. *Current opinion in immunology*, 30:17–23, 2014.

[9] David G Schatz, Marjorie A Oettinger, and David Baltimore. The v (d) j recombination activating gene, rag-1. *Cell*, 59(6):1035–1048, 1989.

[10] Marjorie A Oettinger, David G Schatz, Carolyn Gorka, and David Baltimore. Rag-1 and rag-2, adjacent genes that synergistically activate v (d) j recombination. *Science*, 248(4962):1517–1523, 1990.

[11] Peter Mombaerts, John Iacomini, Randall S Johnson, Karl Herrup, Susumu Tonegawa, and Virginia E Papaioannou. Rag-1-deficient mice have no mature b and t lymphocytes. *Cell*, 68(5):869–877, 1992.

[12] Yoichi Shinkai, Kong-Peng Lam, Eugene M Oltz, Valerie Stewart, Monica Mendelsohn, Jean Charron, Milton Datta, Faith Young, Alan M Stall, Frederick W Alt, et al. Rag-2-deficient mice lack mature lymphocytes owing to inability to initiate v (d) j rearrangement. *Cell*, 68(5):855–867, 1992.

[13] Klaus Schwarz, George H Gauss, Leopold Ludwig, Ulrich Pannicke, Zhong Li, Doris Lindner, Wilhelm Friedrich, Reinhard A Seger, Thomas E Hansen-Hagge, Stephen Desiderio, et al. Rag mutations in human b cell-negative scid. *Science*, 274(5284): 97–99, 1996.

[14] G de Saint-Basile, F Le Deist, JP De Villartay, N Cerf-Bensussan, O Journet, N Brousse, C Griscelli, and A Fischer. Restricted heterogeneity of t lymphocytes in combined immunodeficiency with hypereosinophilia (omenn's syndrome). *The journal of clinical investigation*, 87(4):1352–1359, 1991.

[15] Frédéric Rieux-Laucat, Philippe Bahadoran, Nicole Brousse, Françoise Selz, Alain Fischer, Françoise Le Deist, and Jean Pierre De Villartay. Highly restricted human t cell repertoire in peripheral blood and tissue-infiltrating lymphocytes in omenn's syndrome. *The journal of clinical investigation*, 102(2):312–321, 1998.

[16] Anna Villa, Sandro Santagata, Fabio Bozzi, Silvia Giliani, Annalisa Frattini, Luisa Imberti, Luisa Benerini Gatta, Hans D Ochs, Klaus Schwarz, Luigi D Notarangelo, et al. Partial v (d) j recombination activity leads to omenn syndrome. *Cell*, 93(5): 885–896, 1998.

[17] Dylan Lawless, Christoph B Geier, Jocelyn R Farmer, Hana Allen Lango, Daniel Thwaites, Faranaz Atschekzei, Matthew Brown, David Buchbinder, Siobhan O Burns, Manish J Butte, et al. Prevalence and clinical challenges among adults with primary

immunodeficiency and recombination-activating gene deficiency. *Journal of Allergy and Clinical Immunology*, 2018.

[18] Yu Nee Lee, Francesco Frugoni, Kerry Dobbs, Jolan E Walter, Silvia Giliani, Andrew R Gennery, Waleed Al-Herz, Elie Haddad, Francoise LeDeist, Jack H Bleesing, et al. A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. *Journal of Allergy and Clinical Immunology*, 133(4):1099–1108, 2014.

[19] Irit Tirosh, Yasuhiro Yamazaki, Francesco Frugoni, Francesca A Ververs, Eric J Allenspach, Yu Zhang, Siobhan Burns, Waleed Al-Herz, Lenora Noroski, Jolan E Walter, et al. Recombination activity of human rag2 mutations and correlation with the clinical phenotype. *Journal of Allergy and Clinical Immunology*, 2018.

[20] Jolan E Walter, Lindsey B Rosen, Krisztian Csomos, Jacob M Rosenberg, Divij Mathew, Marton Keszei, Boglarka Ujhazi, Karin Chen, Yu Nee Lee, Irit Tirosh, et al. Broad-spectrum antibodies against self-antigens and cytokines in rag deficiency. *The journal of clinical investigation*, 125(11):4135–4148, 2015.

[21] Catharina Schuetz, Kirsten Huck, Sonja Gudowius, Mosaad Megahed, Oliver Feyen, Bernd Hubner, Dominik T Schneider, Burkhard Manfras, Ulrich Pannicke, Rein Willemze, et al. An immunodeficiency disease with rag mutations and granulomas. *New England journal of Medicine*, 358(19):2030–2038, 2008.

[22] Tami John, Jolan E Walter, Catherina Schuetz, Karin Chen, Roshini S Abraham, Carmem Bonfim, Thomas G Boyce, Avni Y Joshi, Elizabeth Kang, Beatriz Tavares Costa Carvalho, et al. Unrelated hematopoietic cell transplantation in a patient with combined immunodeficiency with granulomatous disease and autoimmunity secondary to rag deficiency. *Journal of clinical immunology*, 36(7):725–732, 2016.

[23] Jean-Laurent Casanova, Mary Ellen Conley, Stephen J Seligman, Laurent Abel, and Luigi D Notarangelo. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *Journal of Experimental Medicine*, pages jem–20140520, 2014.

[24] Luigi D Notarangelo, Min-Sung Kim, Jolan E Walter, and Yu Nee Lee. Human rag mutations: biochemistry and clinical implications. *Nature Reviews Immunology*, 16

(4):234, 2016.

[25] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[26] Paul Stothard. The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *University of Alberta, Education and Research Archive*, 2000.

[27] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[28] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Shaw, Andrew D Phillips, and David N Cooper. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics*, 133(1):1–9, 2014.

[29] Heng Ru, Melissa G Chambers, Tian-Min Fu, Alexander B Tong, Maofu Liao, and Hao Wu. Molecular mechanism of v (d) j recombination from synaptic rag1-rag2 complex structures. *Cell*, 163(5):1138–1152, 2015.

[30] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[31] John Stone. An efficient library for parallel ray tracing and animation. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[32] Anna Villa, Cristina Sobacchi, Luigi D Notarangelo, Fabio Bozzi, Mario Abinun, Tore G Abrahamsen, Peter D Arkwright, Michal Baniyash, Edward G Brooks, Mary Ellen Conley, et al. V (d) j recombination defects in lymphocytes due torag mutations: severe immunodeficiency with a spectrum of clinical presentations. *Blood*, 97(1):81–88, 2001.

[33] Hassan Abolhassani, Ning Wang, Asghar Aghamohammadi, Nima Rezaei, Yu Nee Lee, Francesco Frugoni, Luigi D Notarangelo, Qiang Pan-Hammarström, and Lennart Hammarström. A hypomorphic recombination-activating gene 1 (rag1) mutation resulting in a phenotype resembling common variable immunodeficiency. *Journal of*

Allergy and Clinical Immunology, 134(6):1375–1380, 2014.

[34] Necil Kutukculer, Nesrin Gulez, Neslihan Edeer Karaca, Guzide Aksu, and Afig Berdeli. Novel mutatıons and diverse clinical phenotypes in recombınase-activating gene 1 deficiency. *Italian journal of pediatrics*, 38(1):8, 2012.

[35] Cristina Sobacchi, Veronica Marrella, Francesca Rucci, Paolo Vezzoni, and Anna Villa. Rag-dependent primary immunodeficiencies. *Human mutation*, 27(12):1174–1184, 2006.

[36] Jeroen G Noordzij, Sandra de Bruin-Versteeg, Nicole S Verkaik, Jaak MJJ Vossen, Ronald de Groot, Ewa Bernatowska, Anton W Langerak, Dik C van Gent, and Jacques JM van Dongen. The immunophenotypic and immunogenotypic b-cell differentiation arrest in bone marrow of rag-deficient scid patients corresponds to residual recombination activities of mutated rag proteins. *Blood*, 100(6):2145–2152, 2002.

[37] Elena Crestani, Sharon Choo, Francesco Frugoni, Yu Nee Lee, Stephanie Richards, Joanne Smart, and Luigi D Notarangelo. Rag1 reversion mosaicism in a patient with omenn syndrome. *Journal of clinical immunology*, 34(5):551–554, 2014.

[38] Ilan Dalal, Uri Tabori, Bela Bielorai, Hana Golan, Eli Rosenthal, Ninette Amariglio, Gidi Rechavi, and Amos Toren. Evolution of a tb-scid into an omenn syndrome phenotype following parainfluenza 3 virus infection. *Clinical Immunology*, 115(1): 70–73, 2005.

[39] Taco W Kuijpers, Hanna IJspeert, Ester MM van Leeuwen, Machiel H Jansen, Mette D Hazenberg, Kees C Weijer, Rene AW Van Lier, and Mirjam van der Burg. Idiopathic cd4+ t lymphopenia without autoimmunity or granulomatous disease in the slipstream of rag mutations. *Blood*, pages blood–2011, 2011.

[40] Tanja A Gruber, Ami J Shah, Michelle Hernandez, Gay M Crooks, Hisham Abdel-Azim, Sudhir Gupta, Sean McKnight, Drew White, Neena Kapoor, and Donald B Kohn. Clinical and genetic heterogeneity in omenn syndrome and severe combined immune deficiency. *Pediatric transplantation*, 13(2):244–250, 2009.

[41] Suk See De Ravin, Edward W Cowen, Kol A Zarember, Narda L Whiting-Theobald, Douglas B Kuhns, Netanya G Sandler, Daniel C Douek, Stefania Pittaluga, Pietro L Poliani, Yu Nee Lee, et al. Hypomorphic rag mutations can cause destructive midline

granulomatous disease. *Blood*, 116(8):1263–1271, 2010.

[42] David Buchbinder, Rebecca Baker, Yu Nee Lee, Juan Ravell, Yu Zhang, Joshua McElwee, Diane Nugent, Emily M Coonrod, Jacob D Durtschi, Nancy H Augustine, et al. Identification of patients with rag mutations previously diagnosed with common variable immunodeficiency disorders. *Journal of clinical immunology*, 35(2):119–124, 2015.

[43] Kerstin Felgentreff, Ruy Perez-Becker, Carsten Speckmann, Klaus Schwarz, Krzysztof Kalwak, Gasper Markelj, Tadej Avcin, Waseem Qasim, EG Davies, Tim Niehues, et al. Clinical and immunological manifestations of patients with atypical severe combined immunodeficiency. *Clinical immunology*, 141(1):73–82, 2011.

[44] Andreas Reiff, Alexander G Bassuk, Joseph A Church, Elizabeth Campbell, Xinyu Bing, and Polly J Ferguson. Exome sequencing reveals rag1 mutations in a child with autoimmunity and sterile chronic multifocal osteomyelitis evolving into disseminated granulomatous disease. *Journal of clinical immunology*, 33(8):1289–1292, 2013.

[45] Barbara Corneo, Despina Moshous, Tayfun Güngör, Nicolas Wulffraat, Pierre Philippet, Françoise Le Deist, Alain Fischer, and Jean-Pierre de Villartay. Identical mutations in rag1 or rag2 genes leading to defective v (d) j recombinase activity can cause either tb–severe combined immune deficiency or omenn syndrome. *Blood*, 97 (9):2772–2776, 2001.

[46] Erika Asai, Taizo Wada, Yasuhisa Sakakibara, Akiko Toga, Tomoko Toma, Takashi Shimizu, Sheela Nampoothiri, Kohsuke Imai, Shigeaki Nonoyama, Tomohiro Morio, et al. Analysis of mutations and recombination activity in rag-deficient patients. *Clinical Immunology*, 138(2):172–177, 2011.

[47] Tamaki Kato, Elena Crestani, Chikako Kamae, Kenichi Honma, Tomoko Yokosuka, Takeshi Ikegawa, Naonori Nishida, Hirokazu Kanegane, Taizo Wada, Akihiro Yachie, et al. Rag1 deficiency may present clinically as selective iga deficiency. *Journal of clinical immunology*, 35(3):280–288, 2015.

[48] Xiaomin Yu, Jorge R Almeida, Sam Darko, Mirjam van der Burg, Suk See DeRavin, Harry Malech, Andrew Gennery, Ivan Chinn, Mary Louise Markert, Daniel C Douek, et al. Human syndromes of immunodeficiency and dysregulation are characterized by distinct defects in t-cell receptor repertoire development. *Journal of Allergy and*

Clinical Immunology, 133(4):1109–1115, 2014.

[49] Jean-Pierre De Villartay, Annick Lim, Hamoud Al-Mousa, Sophie Dupont, Julie Déchanet-Merville, Edith Coumau-Gatbois, Marie-Lise Gougeon, Arnaud Lemainque, Céline Eidenschenk, Emmanuelle Jouanguy, et al. A novel immunodeficiency associated with hypomorphic rag1 mutations and cmv infection. *The journal of clinical investigation*, 115(11):3291–3299, 2005.

[50] Junyan Zhang, Linda Quintal, Adelle Atkinson, Brent Williams, Eyal Grunebaum, and Chaim M Roifman. Novel rag1 mutation in a case of severe combined immunodeficiency. *Pediatrics*, 116(3):e445–e449, 2005.

[51] Lauren A Henderson, Francesco Frugoni, Gregory Hopkins, Helen de Boer, Sung-Yun Pai, Yu Nee Lee, Jolan E Walter, Melissa M Hazen, and Luigi D Notarangelo. Expanding the spectrum of recombination-activating gene 1 deficiency: a family with early-onset autoimmunity. *Journal of Allergy and Clinical Immunology*, 132(4): 969–971, 2013.

[52] Elizabeth Mannino Avila, Gulbu Uzel, Amy Hsu, Joshua D Milner, Maria L Turner, Stefania Pittaluga, Alexandra F Freeman, and Steven M Holland. Highly variable clinical phenotypes of hypomorphic rag1 mutations. *Pediatrics*, 126(5):e1248–e1252, 2010.

[53] AGL Riccetto, M Buzolin, JF Fernandes, F Traina, MLR Barjas-de Castro, MTN Silva, JB Oliveira, and MM Vilela. Compound heterozygous rag2 mutations mimicking hyper igm syndrome. *Journal of clinical immunology*, 34(1):7–9, 2014.

[54] Carlos A Gomez, Leon M Ptaszek, Anna Villa, Fabio Bozzi, Cristina Sobacchi, Edward G Brooks, Luigi D Notarangelo, Eugenia Spanopoulou, ZQ Pan, Paolo Vezzoni, et al. Mutations in conserved regions of the predicted rag2 kelch repeats block initiation of v (d) j recombination and result in primary immunodeficiencies. *Molecular and cellular biology*, 20(15):5653–5664, 2000.

[55] Janet Chou, Rima Hanna-Wakim, Irit Tirosh, Jennifer Kane, David Fraulino, Yu Nee Lee, Soha Ghanem, Iman Mahfouz, André Mégarbané, Gérard Lefranc, et al. A novel homozygous mutation in recombination activating gene 2 in 2 relatives with different clinical phenotypes: Omenn syndrome and hyper-igm syndrome. *Journal of Allergy and Clinical Immunology*, 130(6):1414–1416, 2012.

[56] Alka Agrawal, Quinn M Eastman, and David G Schatz. Transposition mediated by rag1 and rag2 and its implications for the evolution of the immune system. *Nature*, 394(6695):744, 1998.

[57] Kevin Hiom, Meni Melek, and Martin Gellert. Dna transposition by the rag1 and rag2 proteins: a possible source of oncogenic translocations. *Cell*, 94(4):463–470, 1998.

[58] Sebastian D Fugmann, Cynthia Messier, Laura A Novack, R Andrew Cameron, and Jonathan P Rast. An ancient evolutionary origin of the rag1/2 gene locus. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3728–3733, 2006.

[59] Shengfeng Huang, Xin Tao, Shaochun Yuan, Yuhang Zhang, Peiyi Li, Helen A Beilinson, Ya Zhang, Wenjuan Yu, Pierre Pontarotti, Hector Escriva, et al. Discovery of an active rag transposon illuminates the origins of v (d) j recombination. *Cell*, 166(1):102–114, 2016.

[60] Vladimir V Kapitonov and Jerzy Jurka. Rag1 core and v (d) j recombination signal sequences were derived from transib transposons. *PLoS biology*, 3(6):e181, 2005.

[61] Douglas G Altman and J Martin Bland. Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, 311(7003):485, 1995.

[62] Attila Kumánovics, Yu Nee Lee, Devin W Close, Emily M Coonrod, Boglarka Ujhazi, Karin Chen, Daniel G MacArthur, Gergely Krivan, Luigi D Notarangelo, and Jolan E Walter. Estimated disease incidence of rag1/2 mutations: A case report and querying the exome aggregation consortium. *Journal of Allergy and Clinical Immunology*, 139 (2):690–692, 2017.

[63] James M Havrilla, Brent S Pedersen, Ryan M Layer, and Aaron R Quinlan. A map of constrained coding regions in the human genome. *bioRxiv*, 2017. doi: 10.1101/220814.

[64] Yuval Itan, Lei Shang, Bertrand Boisson, Michael J Ciancanelli, Janet G Markle, Ruben Martinez-Barricarte, Eric Scott, Ishaan Shah, Peter D Stenson, Joseph Gleeson, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature methods*, 13(2):109, 2016.

[65] Takashi Kido, Weronika Sikora-Wohlfeld, Minae Kawashima, Shinichi Kikuchi, Naoyuki Kamatani, Anil Patwardhan, Richard Chen, Marina Sirota, Keiichi Kodama,

Dexter Hadley, et al. Are minor alleles more likely to be risk alleles? *BMC medical genomics*, 11(1):3, 2018.

[66] Yingleong Chan, Elaine T Lim, Niina Sandholm, Sophie R Wang, Amy Jayne McKnight, Stephan Ripke, Mark J Daly, Benjamin M Neale, Rany M Salem, Joel N Hirschhorn, et al. An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *The American journal of Human Genetics*, 94(3):437–452, 2014.

[67] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321, 2015.

[68] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z. Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss. Scaling up dna data storage and random access retrieval. *bioRxiv*, 2017. doi: 10.1101/114553.

[69] Zhicong Huang, Erman Ayday, Huang Lin, Raeka S. Aiyar, Adam Molyneaux, Zhenyu Xu, Jacques Fellay, Lars M. Steinmetz, and Jean-Pierre Hubaux. A privacy-preserving solution for compressed storage and selective retrieval of genomic data. *Genome Research*, 26(12):10. 1687–1696, 2016.

3 Methylation status assay; theory and example

Preface

A sub-chapter to chapter 4, provided here as a stand-alone to assist dissemination of future works. Readers may not recall some of the common usage for notations in this section as it is not often presented in the primary literature of biological fields. This includes calculations such as measuring the area under a curve and weighted multiplication of curves. We can tackle some of the notations with a few simple explanations. [1]

(1) $f(x)$ = "something" is the classic way of writing a function, where the output, "something", indicates that a calculation was performed on the input "x".

(2) Sometimes we need to consider smaller slices of "x". To label these small parts we use the symbol d, to say "a little bit (or element) of x", or $d(x)$.

(3) When we want to put all of those little bits back together we say, "the sum of $d(x)$" and write the symbol as a tall s. The notation will be written as $\int d(x)$. A better description of that notation is "the integral of $d(x)$." With these simple explanations in mind, any formulas used henceforth should be readable.

[1] This simplification is influenced by the preface to "Calculus Made Easy, Silvanus Thompson, 2nd ed 1914". [1]

3.1 Introduction

Epigenetic modification has a significant effect on gene expression and chromatin remodelling. The activity of DNA methyltransferases such as DNMT1, DNMT3a, and DNMT3b can control expression of many genes. Global changes occur during DNA replication and early in development. Tissue and gene specific changes continually occur over the lifespan of a cell. TET enzymes are involved in methyl group oxidation with the production of 5-hmC as an intermediate. Mutations are frequently identified in methyltransferase and TET genes for patients with haematological malignancies. While many options are available for investigation of methylation status, DNA material is often limited for young or deceased individuals and sensitive assays are generally an expensive investment. Described here is a method for high sensitivity relative quantification of global 5-mC and 5-hmC from minimal sample quantity. Genomic DNA conversion of 5-hmC to glucosylated 5-hmC by T4 β-glucosyltransferase is measured by high-sensitivity electrophoresis and densitometry.

The major form of DNA methylation in eukaryotes most often consists of 5-methylcytosine (5mC) occurring in the context of cytosine-guanine dinucleotides (CpG) [2, 3]. Mammalian genomes are reportedly modified by 5-mC at about 60-80% of CpG sites [2]. Epigenetic modification by 5-mC plays an important role in gene transcriptional regulation. Methylation is mitotically heritable and has important functions for regulation of gene expression [4]. Loss of function variants in Ten eleven translocation (TET) proteins are widely reported to contribute to down-regulated expression of tumour-suppressor genes. Somatic mutations in TET proteins, as well as other epigenetic modifier proteins are often reported in malignancies such as lymphoma.

Although stable, 5-mC may be reverted to its unmodified state in several ways. Passive DNA demethylation (or passive dilution of 5-mC) occurs during DNA replication when there is a lack of functional methylation maintenance mechanisms. Active DNA demethylation is controlled by TET proteins which mediate oxidation of 5-mC to 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). Passively, replication-dependent dilution of these oxidized forms of 5-mC result

in demethylation. Actively, thymine DNA glycosylase (TDG)-mediated excision of 5-fC and 5-caC is followed by base excision repair (BER) to result in demethylated DNA. TET-TDG-independent mechanisms of active DNA demethylation have also been proposed [5–7].

5-hmC is most frequently measured as the intermediate substrate during the demethylation of 5-mC. Bisulfite sequencing is the most commonly used technique for identifying 5-mC in genomic DNA, although this is laborious and cannot distinguish between 5-mC from 5-hmC. Anti-5-mC and anti-5-hmC are used in several applications including immunoblotting for detection of DNA with 5-mC/5-hmC on nitrocellulose, while immunocytochemistry and immuno-fluorescence use the same antibodies to detect epigenetic modification within cells.

Many high-accuracy sequencing techniques exist for genome-scale mapping of oxidized 5-mC. A summary of these techniques has been published in a recent review by Wu and Zhang [8]. The authors also provide a complete overview of mechanisms and function of TET-mediated active DNA demethylation. Furthermore, Teschendorff and Relton [9] have recently reviewed the "statistical challenges and algorithms associated with drawing inferences from DNA methylation data".

The *HpaII* gene from *Haemophilus parainfluenzae* produces a restriction enzyme which digests the sequence 5' CCGG 3' but is blocked by CpG methylation. The *MspI* gene from *Moraxella* species encodes an enzyme which recognises the same sequence on DNA but is not sensitive to methylation. This pair of enzymes have been used extensively for methylation-sensitive differentiation of DNA. The enzyme DNA β-glucosyltransferase catalyses a reaction where a β-D-glucosyl residue is transferred from uridine diphosphate glucose (UDP-glucose) to an hydroxymethylcytosine residue on DNA. *Escherichia coli* virus, bacteriophage T4, produces the phage T4 β-glucosyltransferase (T4-BGT) which functions to modify DNA by transfer of UDP to 5-hmC of phage T4 DNA. T4-BGT is exploited for use with human DNA by complete conversion of 5-hmC to glucosylated 5-hmC (5-ghmC). This process is sequence-independent, and unmodified or 5-mC-containing DNA is unaffected. Both HpaII and MspI are sensitive to 5-ghmC which therefore

differentiates 5-mC from 5-hmC. These enzymes were used in combination with high-sensitivity DNA fragment detection for a novel method of genome-wide 5-mC and 5-hmC relative quantification.

While many methods exist for accurate methylation quantification or methylation sequencing, it can be very expensive to assay patient DNA material that is limited for patients who are deceased or perhaps are of a young age and too unwell to provide blood samples often. The following method can be used in such circumstances.

3.2 Materials

QIAamp DNA Blood Mini Kit (cat. 51104) (Qiagen, CA, USA). Qubit (previously Quant-iT) dsDNA BR Assay Kit (cat. Q32850) with use of the Qubit Fluorometer (ThermoFischer, MA, USA). Enzymes sourced from New England Biolabs; MspI (R0106S) and HpaII (R0171S) with or without CutSmart buffer (B704S). T4-BGT sourced from New England Biolabs (M0357S) including NEB buffer 4 and uridine diphosphate glucose. Agilent Genomic DNA ScreenTape assay (cat. 5067-5365/6) with ladder and sample buffer (5067-5366) run on the Agilent 2200 (or 4400 cat. G2991AA) TapeStation system. Other miscellaneous accessories required for this system are listed by the manufacturer. ImageJ open source software [10]. Spreadsheet software capable defining simple equations such as LibreOffice [11] is required.

3.3 Sample preparation

Genomic DNA was purified from 2ml whole blood PBMCs using QIAamp DNA Blood Mini Kit (Qiagen). Purified DNA was quantified with Qubit dsDNA BR Assay Kit with use of the Qubit Fluorometer (ThermoFischer). Quantified DNA was then diluted to 100 ng/µL in water.

Three 2 µL aliquots of DNA were prepared for each sample (healthy control or patient DNA). Aliquot one was treated with T4-BGT to convert 5-hmC to glucosylated 5-hmC. Table 3.1 lists the reagent and enzyme concentrations added to aliquot one of each sample.

Table 3.1: Glycosylation reagents

Genomic DNA	20 ng
UDP-Glucose (50x 2mM)	1.24 µL
NEB buffer 4 (10x)	3.1 µL
T4-BGT	1 µL (10 units)
H20	2 µL
Reagent volume	7.34 µL

Aliquot one for each sample was heated at 37°C for 6 hours to allow glucosylation of 5-hmC to form 5-ghmC. At this point samples may be stored at 4°C or -20°C before the next step. After glucosylation of genomic DNA in aliquot one, enzyme restriction was performed on all three aliquots for each sample. MspI and HpaII recognise the same DNA sequence ('5 CCGG 3') but are differentially sensitive to methylation status. MspI cleaves both 5-mC and 5-hmC. However, MspI cleavage is blocked by 5-ghmC. HpaII cannot cleave modified sites; any modification with 5-mC, 5-hmC, or 5-ghmC at either cytosine will prevent cleavage. Table 3.2 lists the restriction enzyme added to each aliquot of genomic DNA for methylation-specific digestion. All samples were heated at 37°C for 6 hours to allow for complete digestion. Samples may be stored at 4°C or -20°C before the next step.

Table 3.2: Restriction enzymes

Aliquot	MspI	HpaII	H20	Total volume
1	1 µL	-	-	10.34 µL
2	1 µL	-	7.34	10.34 µL
3	-	1 µL	7.34	10.34 µL

3.4 Measuring methylation-specific digestion

After digestion the separation and analysis of DNA samples up to greater than 60,000 base pairs was performed using the Agilent Genomic DNA ScreenTape assay (cat. 5067-5365/6) with the Agilent 2200 (or alternatively 4400 cat. G2991AA) TapeStation system (Agilent) according to manufacturer specifications. This system is most commonly used for library preparation of next generation sequencing samples and is currently used in many sequencing facilities. Quantification is sensitive down to 5 pg/µL.

3.4.1 DNA fragment density plot

A high-resolution assay report was created for each Genomic DNA ScreenTape which were run in batches of 15 samples per run. In the event of a ladder failure, all samples on that run cannot be quantified. Therefore, large sample numbers can be run in batches to minimise potential loss due to ladder failure (**Figure 3.1**). Each sample was analysed with the use of ImageJ software [10]. The feature "Analyze" "Gels" was utilised to "select and plot lanes" for each sample. Any capture window size may be used to cover the desired lane region once all samples are processed in the same fashion per experiment. In this case the captured window size used height 650, width 60, and an equal Y coordinate across all samples. The Y coordinate is the capture window position (yellow box seen in **Figure 3.1**), displayed on-screen in ImageJ.

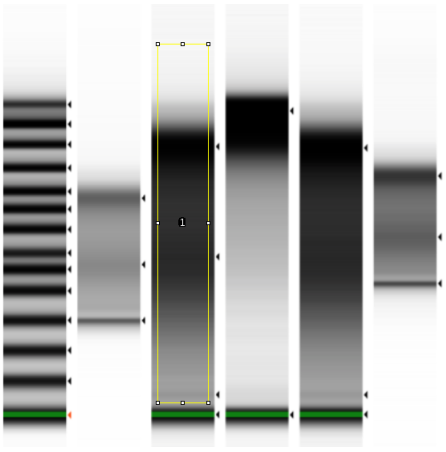


Figure 3.1: **TapeStation report with ImageJ lane analysis.** High sensitivity gel eletrophoresis shows a ladder in lane one, with digested genomic DNA in lanes two-six. Lane three shows an ImageJ capture window which will be used to measure density from high to low molecular weight DNA. Each lane shown contains DNA at different concentrations and digestion methods, used during optimisation.

The molecular weight ladder used to measure DNA is shown in Figure 3.2. Along with the ladder, the plotted density data for each lane was collected. **Figure 3.3** shows the plot of a typical lane. The plot was divided into ten equally spaced intervals. DNA runs from high to low molecular weight on the gel from top to bottom. The density plot presents this data from left to right. The area under the curve (AUC) at each interval was recorded and used for calculation of methylation difference.

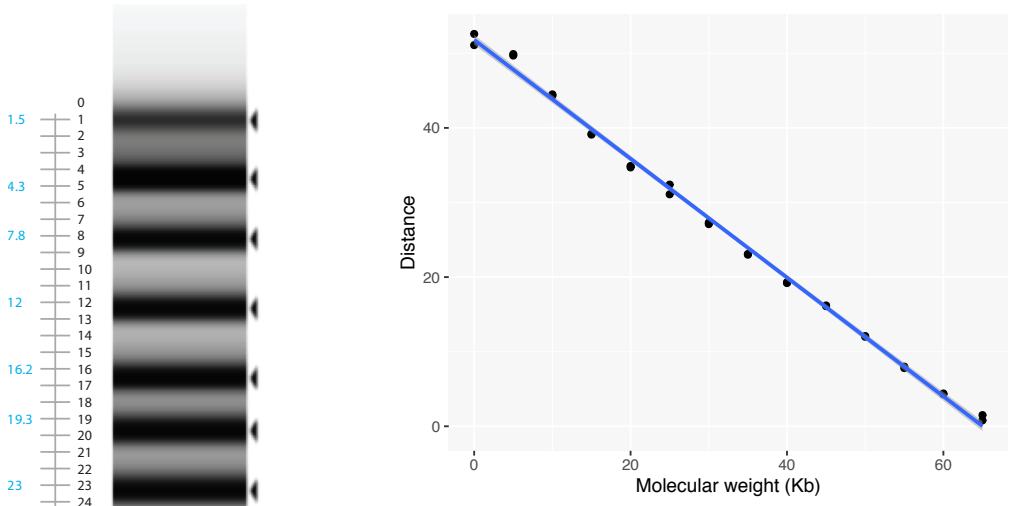


Figure 3.2: **Genomic DNA ladder.** The ladder used for genomic DNA analysis is shown on the left, with markers indicating the molecular weight set against a relative, arbitrary distance scale. On the right, the linear migration pattern of distance versus molecular weight ladder is plotted.

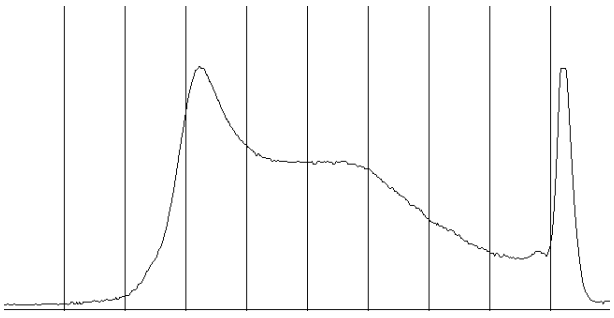


Figure 3.3: **Typical density plot of a lane.** The on-screen output density plot from the capture window shown in lane three of Figure 3.1 is shown. The plot is divided into ten sections to allow measurement of area under the curve per fragment. More divisions can be used for higher sensitivity, although ten measurements are generally sufficient.

3.5 Theory and calculation

3.5.1 Example calculation of methylation difference

An extremely simplified illustration is provided before performing the same analysis on assay data. The cartoon illustration of density curves represent those as seen in **Figure 3.3**. **Figure 3.4** (a-b) presents example results with a curve $f(x)_1$ for MspI-treated DNA from (a) healthy control and (b) DNA from a patient with a loss of 5-hmC production. No difference is seen for methylation-insensitive digestion. **Figure 3.4** (c-d) illustrates the effect of T4-BGT pre-treatment for production of 5-ghmC with curve $f(x)_2$. Firstly, healthy control (c) produces 5-hmC which is converted to 5-ghmC and protected from MspI cleavage. This results in more uncut, high molecular weight DNA and reduced low molecular weight fragments. The patient (d) produces very little 5-hmC and therefore has no change in enzymatic digestion. Secondly, the treatment conditions cause partial degradation of DNA non-specifically across all samples resulting in a rightward shift in fragment size. The difference in MspI digestion due to 5-ghmC may be illustrated when the corresponding values at each interval are subtracted; $f(x)_1 - f(x)_2$. **Figure 3.4** (e-f) illustrates this difference with red +/- symbols and the resulting curve is shown in **Figure 3.5**.

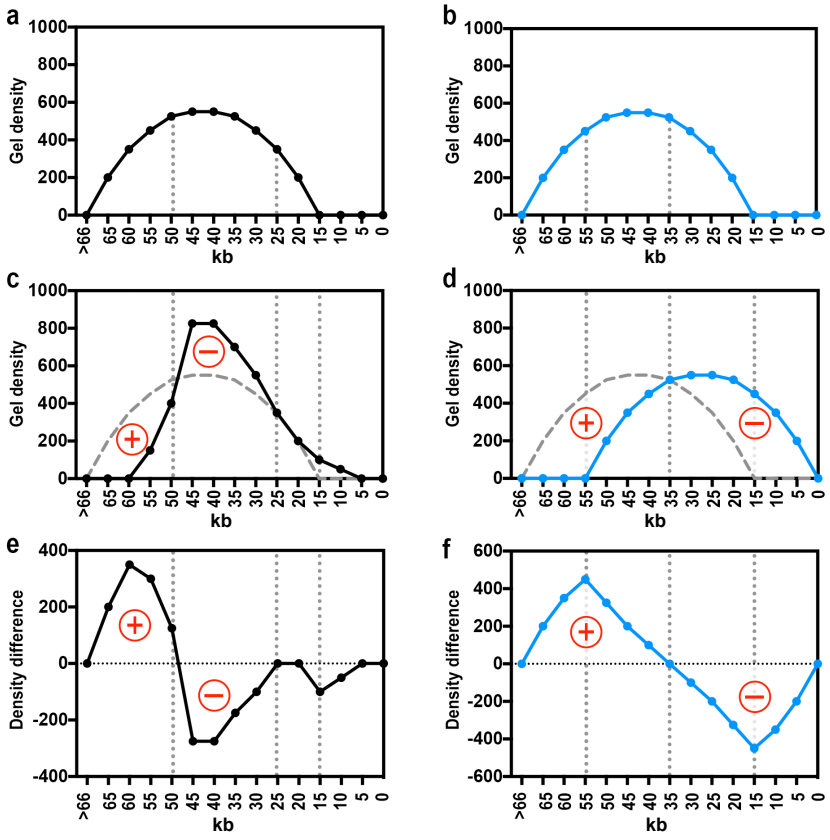


Figure 3.4: **Example data density plots.** A cartoon version of density curves representing those seen in figure 3.3. (a-b) Example data from density plot of MspI-treated DNA; Healthy control (black) and patient (blue). (c-d) Example data from density plot of T4-BGT and MspI-treated DNA. (e-f) Resulting curve of difference between MspI treatment with and without T4-BGT 5-hmC conversion.

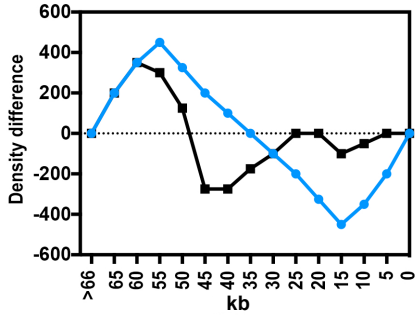


Figure 3.5: **Overlay of normal 5-hmC in healthy control versus reduced 5-hmC in patient.** The curves shown in Fig. 3.4 e-f are overlaid to demonstrate the difference between methylation-dependent digestion between "typical" genomic DNA and a DNA that is deficient for 5-hmC.

Figure 3.5 shows this difference with overlay of normal levels of 5-hmC produced in healthy control (a) and reduced levels of 5-hmC found in a patient (b). **Table 3.3** shows the data used to illustrate digestion with MspI and MspI after T4-BGT glycosylation (**Figure 3.5** with curve $f(x)_1 - f(x)_2$). The same method may be applied to compare HpaII-treated DNA (aliquot 3) to MspI-treated DNA (aliquot 2). However, the resulting curve from this process provides comparison of general 5-mC levels, rather than 5-hmC specifically. Note: *this illustrative version of comparison is provided only to visualise the density difference and not used in the final analysis. The method for producing a summary measurement for comparing groups is shown in subsection 3.5.3.*

Table 3.3: Calculation of values for example density difference $f(x)_1 - f(x)_2$ used in Figure 3.4 and 3.5. This illustrative version of comparison is provided only to visualise the density difference and not used in the final analysis.

Density value calculations							
$f(x)$	Fragement size (Kb)	Control MspI $f(x)_1$	Control MspI+T4 $f(x)_2$	Control $f(x)_1 - f(x)_2$	Patient MspI $f(x)_1$	Patient MspI+T4 $f(x)_2$	Patient $f(x)_1 - f(x)_2$
a	>66	0	0	0	0	0	0
b	65	200	0	200	200	0	200
c	60	350	0	350	350	0	350
d	55	450	150	300	450	0	450
r	50	525	400	125	525	200	325
f	45	550	825	-275	550	350	200
g	40	550	825	-275	550	450	100
h	35	525	700	-175	525	525	0
i	30	450	550	-100	450	550	-100
j	25	350	350	0	350	550	-200
k	20	200	200	0	200	525	-325
l	15	0	100	-100	0	450	-450
m	10	0	50	-50	0	350	-350
n	5	0	0	0	0	200	-200
o	0	0	0	0	0	0	0

3.5.2 Quantification weighting

A critical component of accurate relative quantification relies on weighting image density relative to molecular weight. Fragment shift at higher molecular weights are both subtler and impart greater information than low molecular weight fragments. Figure 3.6 illustrates this idea. This hypothesis has been recently shown by Ito et al. [12], where the % of differentially expressed genes were measured in mouse models of TET2 WT, KO, and enzymatically inactive protein. The authors show that the effect of TET2 on gene methylation was seen as occurring *approximately*; 85% at promoters (+/− 1kb), 5% in gene bodies, and 10% at distal regions (+/− 50kb).

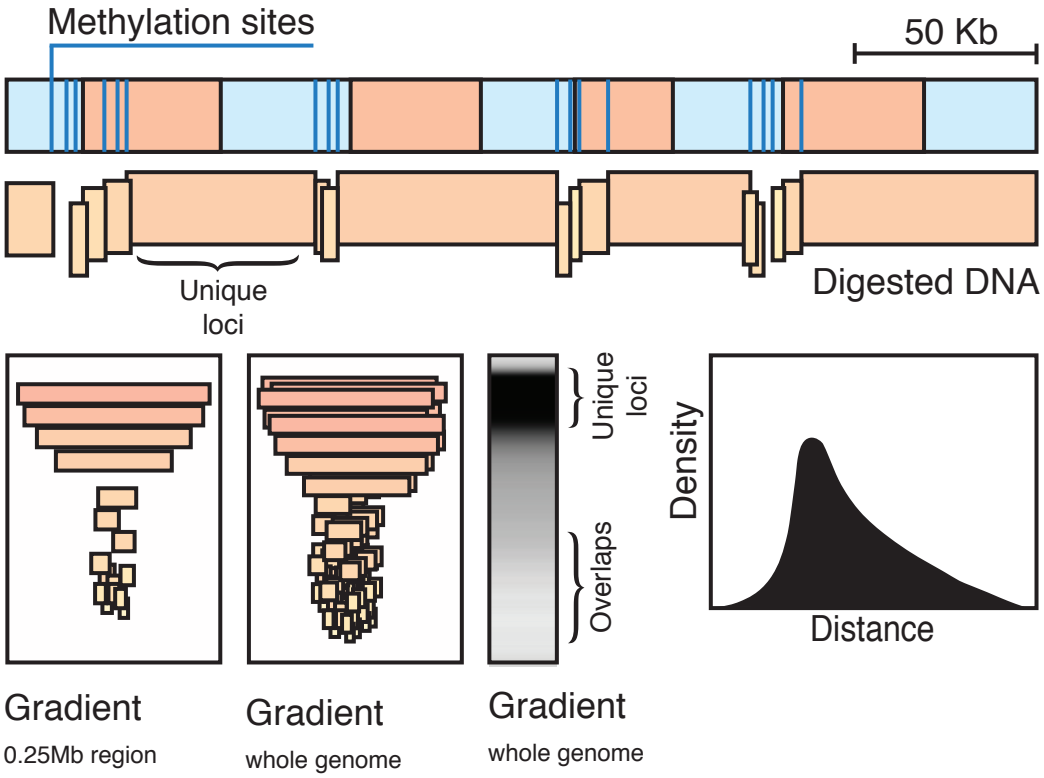


Figure 3.6: **Fragment shift and image density.** The majority of TET2 activity occurs on relatively short confined regions around gene promoters. These regions are illustrated by light blue vertical bands. Long regions of genomic DNA separate these promoter regions and incur few interactions with TET2. Methylation-sensitive digestion therefore produces high molecular weight unique loci and low molecular weight DNA from "overlapping" areas of the same promoter. With only global methylation comparisons between individuals, the unique distal regions offer more insight than smaller overlapping regions and therefore receive proportionally higher weighting during quantification.

The illustration in **Figure 3.6** represents the bulk of protein-function-per-loci at promoter regions as "overlaps". Ideally, methylation sequencing would be used to map genome-wide methylation. In our case this was not an option. Therefore, the high molecular weights from methylation-specific digestion at *unique* distal regions offer more insight than smaller overlapping regions. The molecular weight ladder used to measure DNA is shown in Figure 3.2.

The method for multiplying the molecular weight at each fragment interval by the relevant density score is shown in **equation** 3.1 where w represents the fragment size (Kb) interval that will be listed in **Table 3.4**.

$$\begin{pmatrix} f(x)_a \\ f(x)_b \\ \vdots \\ f(x)_o \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{15} \end{pmatrix} = \begin{pmatrix} w_1 f(x)_a \\ w_2 f(x)_b \\ \vdots \\ w_{15} f(x)_o \end{pmatrix} \tag{3.1}$$

3.5.3 5-mC and 5-hmC potential

Visual representation of results can provide information such as an increase in a particular fragment size amongst a patient group. This occurrence would be pronounced if assaying particular loci (as a single band restriction fragment in a typical agarose gel) rather than genome-wide methylation levels. To simplify the characterisation of methylation status a percentage difference between patient and healthy control is beneficial. The theoretical minimum level of 5-hmC genome wide is found when difference before and after T4-BGT treatment approaches zero; $f(x)_1 - (f(x)_1 - f(x)_2) \approx f(x)_1$. However the simplest general usage will graph $f(x)_1$healthy control and patient versus $f(x)_2$ healthy control and patient. The percentage difference is used to compare samples and can be found with the use of the trapezoidal rule by approximating the definite integral $\int_a^b f(x)\,\mathrm{d}x$. The trapezoidal rule approximates the region under the graph of the weighted function

$f(x)$ as a trapezoid to calculate its area. It follows that

$$\int_a^b f(x)\,\mathrm{d}x \approx \ (b-a)\left[\frac{f\,(a)\ + f\,(b)}{2}\right].$$

The integral is better approximated by partitioning the integration interval, applying the trapezoidal rule to each subinterval, and summing the results (the composite). If unweighted, let k_x be a partition of $[a, b]$ of such that

$$a = x_0 < x_1 < \cdots < x_{N-1} < x_N < k_{n+1} = b.$$

and Δk_x be the length of the k-th subinterval (that is $\Delta k_x = k_x - k_{x-1}$) then,

$$\int_a^b f(x)\,\mathrm{d}x \approx \ \sum_{k-1}^{N} \frac{f\,(x_{k-1})\ + f\,(x_k)}{2}\Delta x_k.$$

The approximation becomes more accurate as the resolution of the partition increases (that is, for larger N). When the partition has a regular spacing, as in this case, the formula can be simplified for calculation efficiency. Ten to fifteen intervals are sufficienct for densitometry of Genomic TapeStation lanes. It is possible to place error bounds on the accuracy of the value of a definite interval estimated using the trapezoidal rule although this process is not necessary for the method outlined here.

Indeed, the output of imageJ software used to quantify the density of DNA per lane already consists of a curve, with the AUC data present. To determine the relative difference between 5-mC and 5-hmC levels the weighted AUC for $f(x)_1$ and $f(x)_2$ are compared. First, for each interval measured the average values $f(x)a - o$ are found for all control or patient samples in both $f(x)_1$ and $f(x)_2$. **Table 3.4** presents the data for a ten-interval example patient and healthy control DNA digested with MspI before and after 5-hmC conversion. Since the interval DNA size is a range between two values, it would be inappropriate to directly multiply the AUC by the weight. This would result categorical data due to "binning". Furthermore, doing so would hyper-inflate by the higher weights. Instead, best approximation of the weighted AUC is produced using the mean

interval range. This is done by again applying the AUC calculation but including the mean weight value. The total positive AUC of each interval was calculated for

$$f(x)AUC(w) = \sqrt{\int_a^b f(x)\,\mathrm{d}x(w_2 - w_1)}^2 .$$

Figure 3.7 illustrates the AUC of each. The total AUC $f(x)_1$ and AUC $f(x)_2$ are listed for example patient and healthy control in **Table 3.4**. The percentage difference, or similarity (for clarity in the case of MspI and MspI+T4), between patient and control is calculated.

Table 3.4: Calculation of weighted AUC in example healthy control and patient for $f(x)_1$ and $f(x)_2$ using a ten point interval. The % maximum measurement possible can also be though of as "similarity" between curves; the theoretical minimum level of 5-hmC is discussed at the beginning of this section. A ratio of $wf(x)_1$ to $wf(x)_2$ is also suitable usage instead of a percentage. The first calculated value (Control AUC $wf(x)_1$) is produced by $\sqrt{(50-60)[\frac{(200+350)}{2}]}^2 = 2750$, and so forth.

Density weighting calculations								
Mean interval size kb	Control AUC $f(x)_1$	Control AUC $wf(x)_1$	Control AUC $f(x)_2$	Control AUC $wf(x)_2$	Patient AUC $f(x)_1$	Patient AUC $wf(x)_1$	Patient AUC $f(x)_2$	Patient AUC $wf(x)_2$
60	200	2750	0	750	200	2750	0	0
50	350	15000	150	10312.50	350	15000	0	3750
12.5	450	4143.75	400	5206.25	450	4143.75	200	2337.50
4	525	806.25	825	1237.50	525	806.25	350	600
2.5	550	550	825	762.50	550	550	450	487.50
1.5	550	403.13	700	468.75	550	403.13	525	403.13
0.75	525	170.63	550	157.50	525	170.63	550	192.5
0.4	450	60	350	41.25	450	60	550	80.63
0.25	350	68.75	200	37.50	350	68.75	525	121.88
0	200		100		200		450	
AUC(w)	-	23952.50	-	18973.75	-	23952.50	-	7973.13
% max			79.21%				33.29%	

Figure 3.7 illustrates the case of 5-mC hypermethylation due to a loss of 5-hmC where low % *similarity* between curves (AUC(w)) signifies low 5-ghmC (or 5-hmC). A reduction of 5-hmC results in more genomic DNA digestion by MspI after T4-BGT treatment

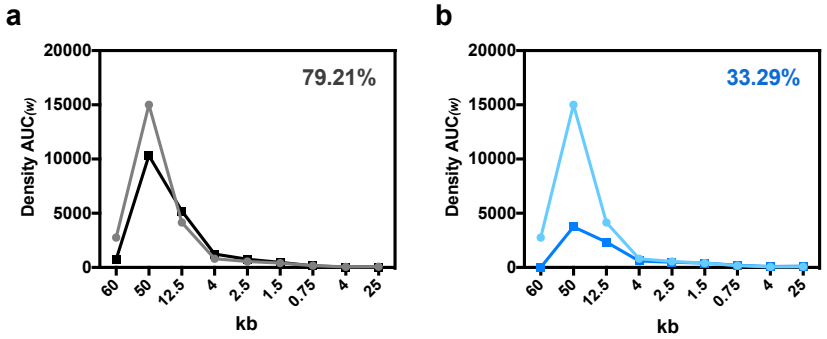


Figure 3.7: **Example weighted AUC of $f(x)_1$ (light) and $f(x)_2$ (dark) from (a) normal 5-hmC in healthy control versus (b) reduced 5-hmC in patient.** Extended legend.

and therefore, a smaller peak in high molecular weight DNA. For patients with DNA hypermethylation a global increase in 5-mC is found during calculation of methylation difference comparing HpaII-treated DNA (aliquot 3) to MspI-treated DNA (aliquot 2). An increase in % *difference* between curves in this case represents hypermethylation.

3.6 Statistical analysis

To correctly handle the weighted serial measurement data generated, several methods could be employed. Each method is based on calculating a summary statistic for each subject [13]. This can be the mean value, slope of a line, or max/min values if appropriate. The measurement summary data is then used as the raw data, in a second step, representing each individual. The simplest method of summarising this data type is usually done by using the summary statistic for each subject, averaging within the group, and performing a comparison of group means using a t test of Mann Whitney U [14].

A similar method usage can be seen in Badger et al. [15], using the area under the curve as the summary measurement for each study participant for swelling due to lymphedema of the limb. Similar approaches are used by Hay et al. [16] in the study of paracetamol plus ibuprofen for the treatment of fever, and Peacock et al. [17] for studying the acute effects of winter air pollution on respiratory function. Matthews [18] and Altman [19] also provide robust reading on this topic.

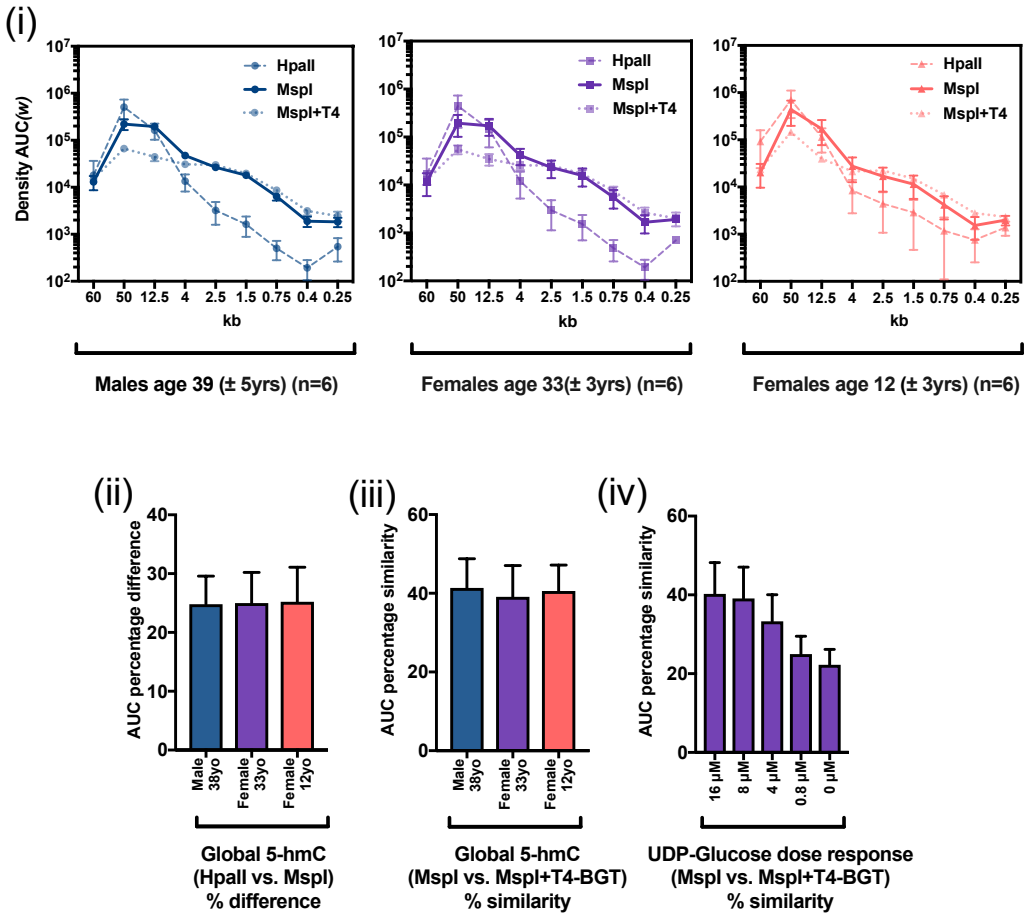


Figure 3.8: **Relative quantification of methylation in adult male and adult/adolescent female groups.** (i) Curves are shown for individual digestions. (ii) Differences between global methylation are expected based on age and sex. However, the % difference between *types* of methylation were relatively small in our tests. (iii) The maximum and minimum effective UDP-Glucose dose was found by testing a range of concentrations.

An alternative but valid method of comparing case and control data in this case is possible using the two sample Kolmogorov-Smirnov test. This is a nonparametric test that compares the cumulative distributions of two data sets. Data are not required to be sampled from Gaussian distributions and results will not change if the data is transformed to logarithms, reciprocals, or any transformation. This method would be appropriate to report the maximum difference between two distributions and return a D statistic and P-value. This method was not included here because it is less intuitive to interpret a comparison of maximum differences, compared to viewing a bar plot of the relative quantification. chapter 5 illustrates and discusses the cumulative distribution of two

groups of protein pathway network data. The Kolmogorov-Smirnov test could also be applied for this comparison (if it was needed) to allow for the automation of detecting the optimal distribution of protein networks.

3.7 Results

Relative quantification of genomic 5-mC and 5-hmC was applied to compare three age-sex matched groups. Genomic DNA was sourced from PBMCs via whole blood from six individuals for each group of males aged 38 (± 4 years) females aged 33 (± 3 years), and females ages 12 (± 2 years). **Figure 3.8** (i) plots the density curves after digestion with HpaII, MspI, and MspI with T4-BGT treatment for each sample group. **Figure 3.8** (ii) graphs the relative levels of global 5-mC as % difference between curves, 100% representing the maximum detectable level of methylation genome-wide. Similarly, (iii) shows the relative level of 5-hmC. Although significantly different, comparisons between healthy control age-sex matched groups cannot be inferred. Hall et al. [20] measure no significant difference in global methylation levels between healthy males and females using the Infinium HumanMethylation450 BeadChip. However, age-sex match differences are expected when comparing density curves by enzyme digestion. To mimic a case of 5-hmC hypomethylation a dose response to decreasing UDP-glucose was performed. **Figure 3.8** (iv) shows an optimum concentration of UDP-glucose at 8 µM to measure 5-hmC in healthy female adults (39%). Complete lack of UPD-glucose supplement results in a reduction in 5-ghmC measurement to 22% and mimicking 5-hmC hypomethylation. This might be considered baseline for normalisation. UDP-glucose saturation of 5-hmC at 16 µM produces a maximum measurement in healthy adult females of 40%. A measurement window for age-sex matched groups can be established for patient queries in a similar fashion to determining the local reference range for blood counts at haematology laboratories. Parametric paired t-test is used for establishing significant changes.

3.8 Discussion

Measuring methylation status provides a valuable source of genetic data [8]. Cytosine methylation provides significant control of gene expression and chromatin remodelling; 5-mC and 5-hmC status affects spatial and temporal gene expression. Embryonic development and cellular differentiation processes are dependent on these mechanisms. CpG methylation levels are non-random throughout the genome [21]. Methylation of promoter regions often reduces gene expression. In mammals the majority of active promoters reportedly associate with unmethylated CpGs.

Hydroxymethylation often correlates with increased gene expression [22, 23]. As the intermediate substrate during 5-mC demethylation, a reduction in 5-hmC promotes hypermethylation [24–26]. Changes in chromatin structure are influenced by hypermethylation as it initiates heterochromatin formation [27, 28]. Histone-modifying and chromatin-remodelling proteins are recruited to 5-mC by methyl binding domain proteins [28]. Selective X chromosome inactivation is controlled by changes in chromatin structure. A change in methylation status typically drives uni-directional cell differentiation [29]. A number of diseases are directly related to disruption of any of these processes. Perturbation of methylation is reported for a number of diseases including Fragile X, Rett syndrome, Prader-Willi, Angelman, and Beckwith-Wiedemann Syndromes. Hypomethylation of heterochromatin can cause genomic instability while hypermethylation of euchromatin drives transcriptional repression.

While relative-quantification of genome-wide methylation cannot provide gene-specific information, it can be applied for a very low cost and requires minimal DNA. This application may be ideal for many cases as a first step before committing to costly methods such as whole-genome bisulfite sequencing or methylation arrays [8]. Furthermore, we have developed this application for cases where patient material is too limited to pursue other commercially available methods.

3.9 Conclusion

A cost-effective and practical method for measuring differences in genome-wide methylation status with limited DNA material is present here.

Bibliography

[1] Silvanus Thompson. *Calculus Made Easy, Second Edition.* MacMillan and Co. Ltd., 1914.

[2] Zachary D Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

[3] En Li and Yi Zhang. Dna methylation in mammals. *Cold Spring Harbor perspectives in biology*, 6(5):a019133, 2014.

[4] Aimée M Deaton and Adrian Bird. Cpg islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022, 2011.

[5] Susan C Wu and Yi Zhang. Active dna demethylation: many roads lead to rome. *Nature reviews Molecular cell biology*, 11(9):607–620, 2010.

[6] Matthias Bochtler, Agnieszka Kolano, and Guo-Liang Xu. Dna demethylation pathways: Additional players and regulators. *Bioessays*, 39(1), 2017.

[7] Hao Wu and Yi Zhang. Reversing dna methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1):45–68, 2014.

[8] Xiaoji Wu and Yi Zhang. Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 2017.

[9] Andrew E Teschendorff and Caroline L Relton. Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, pages nrg–2017, 2017.

[10] Johannes Schindelin, Curtis T Rueden, Mark C Hiner, and Kevin W Eliceiri. The imagej ecosystem: An open platform for biomedical image analysis. *Molecular reproduction and development*, 82(7-8):518–529, 2015.

[11] The Document Foundation. Libre office 2017, 2017. URL https://www.libreoffice.org.

[12] Kyoko Ito, Joun Lee, Stephanie Chrysanthou, Yilin Zhao, Katherine Josephs, Hiroyo

Sato, Julie Teruya-Feldstein, Deyou Zheng, Meelad M Dawlaty, and Keisuke Ito. Non-catalytic roles of tet2 are essential to regulate hematopoietic stem and progenitor cell homeostasis. *Cell Reports*, 28(10):2480–2490, 2019.

[13] JN1 Matthews, Douglas G Altman, MJ Campbell, and Patrick Royston. Analysis of serial measurements in medical research. *Bmj*, 300(6719):230–235, 1990.

[14] Janet Peacock and Philip Peacock. *Oxford handbook of medical statistics*. Oxford University Press, 2011.

[15] Caroline MA Badger, Janet L Peacock, and Peter S Mortimer. A randomized, controlled, parallel-group clinical trial comparing multilayer bandaging followed by hosiery versus hosiery alone in the treatment of patients with lymphedema of the limb. *Cancer*, 88(12):2832–2837, 2000.

[16] Alastair D Hay, Céire Costelloe, Niamh M Redmond, Alan A Montgomery, Margaret Fletcher, Sandra Hollinghurst, and Tim J Peters. Paracetamol plus ibuprofen for the treatment of fever in children (pitch): randomised controlled trial. *Bmj*, 337:a1302, 2008.

[17] JL Peacock, P Symonds, P Jackson, SA Bremner, JF Scarlett, DP Strachan, and HR Anderson. Acute effects of winter air pollution on respiratory function in schoolchildren in southern england. *Occupational and Environmental Medicine*, 60 (2):82–89, 2003.

[18] JNS Matthews. A refinement to the analysis of serial data using summary measures. *Statistics in medicine*, 12(1):27–37, 1993.

[19] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990.

[20] Elin Hall, Petr Volkov, Tasnim Dayeh, Jonathan Lou S Esguerra, Sofia Salö, Lena Eliasson, Tina Rönn, Karl Bacos, and Charlotte Ling. Sex differences in the genome-wide dna methylation pattern and impact on gene expression, microrna levels and insulin secretion in human pancreatic islets. *Genome biology*, 15(12):522, 2014.

[21] Michael Stevens, Jeffrey B Cheng, Daofeng Li, Mingchao Xie, Chibo Hong, Cécile L Maire, Keith L Ligon, Martin Hirst, Marco A Marra, Joseph F Costello, et al. Estimating absolute methylation levels at single-cpg resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome research*, 23(9): 1541–1553, 2013.

[22] Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[23] Elisabet Pujadas and Andrew P Feinberg. Regulated noise in the epigenetic landscape of development and disease. *Cell*, 148(6):1123–1131, 2012.

[24] Skirmantas Kriaucionis and Nathaniel Heintz. The nuclear dna base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science*, 324 (5929):929–930, 2009.

[25] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[26] Cornelia G. Spruijt, Felix Gnerlich, Arne H. Smits, Toni Pfaffeneder, Pascal W.T.C. Jansen, Christina Bauer, Martin Münzel, Mirko Wagner, Markus Müller, Fariha Khan, H. Christian Eberl, Anneloes Mensinga, Arie B. Brinkman, Konstantin Lephikov, Udo Müller, Jörn Walter, Rolf Boelens, Hugo van Ingen, Heinrich Leonhardt, Thomas Carell, and Michiel Vermeulen. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, 152(5):1146 – 1159, 2013. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2013.02.004. URL http://www.sciencedirect.com/science/article/pii/S0092867413001529.

[27] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone h3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, 41(2):246, 2009.

[28] Andrew P Feinberg. The key role of epigenetics in human disease prevention and mitigation. *New England journal of Medicine*, 378(14):1323–1334, 2018.

[29] John B Gurdon, Tom R Elsdale, and Michel Fischberg. Sexually mature individuals of xenopus laevis from the transplantation of single somatic nuclei. *Nature*, 182 (4627):64, 1958.

4 Germline *TET2* deficiency

4.1 Introduction

4.1.1 The epigenetic landscape

Conrad H Waddington, in *The strategy of the genes* [1], described a hypothetical landscape in which the fate of a pluripotent cell is guided in the same way that a marble rolling down a hill encounters many peaks and valleys; each path encountered along the way may lead to unique destinations. As it progresses, the number of possible outcomes for the journey diminish. This simple metaphor from 1957 illustrates the differentiation potential of a cell. Mutation and epigenetic alterations modify the itinerary outlined by germline DNA. Shortly after the theory of the epigenetic landscape was published, Gurdon et al. [2] wrote a modest letter in Nature, reporting that differentiated cells could not only give rise to alternative, differentiated cells but could also give rise to any of the cell types in a sexually mature adult animal. The genetically identical products later became well known as clones (the term clone was only first applied to animals five years later, by JBS Haldane). This work was succinctly described in it's 2012 Nobel Prize title summary "for the discovery that mature cells can be reprogrammed to become pluripotent." Before this, it was undetermined whether the specialist roles seen in each cell after differentiation was caused by gene silencing or if cells lost the genetic material altogether. In context, the epigenetic landscape described by Waddington [1] appears correct; a cell can differentiate

by being guided along a furrow that provides genetic modifications to limit the cell's potential while also maintaining the source code.

DNA methylation had been introduced before any discussion of epigenetics, by Rollin D Hotchkiss [3] in 1948 while he was reporting on the separation of nucleotides. It was almost thirty years until methylation was linked with gene expression [4].

It was apparent from the work of Gurdon et al. [2], that even after a cell had become differentiated it would be possible for it to return to it's earlier form. The explanation of how was only shown more recently. Takahashi and Yamanaka [5] showed that induction of pluripotency was capable with four transcription factors, Oct3/4, Sox2, Klf4, and c-Myc (sometimes collectively referred to as the *Yamanaka factors*) [This paper also termed the designation of iPS (induced pluripotent stem *cells*)]. Their important finding earned a share of the 2012 Nobel Prize with Gurdon. Waddington's metaphor described a marble running down a mountain, Takahashi and Yamanaka [5] showed a marble defying gravity and rolling back uphill to chose a new path.

The definition of epigenetics has changed (and become more specific) from Waddington's original meaning. We now know that modifications exist beyond the level of DNA sequence that can control gene expression at stable levels, which are conserved after cell division, and are susceptible to environmental modification [6]. Environmental signals have a strong influence on cell plasticity and are well known in the context of ageing and disease susceptibility [7].

4.1.2 Epigenetic categories

Recently, Feinberg [8] summarised the forms of epigenetic information into three categories. **DNA methylation** is the first, and most widely recognised form. CpG sites are the stage for methylation; a cytosine, that precedes a guanosine in the 5' to 3' orientation, can be modified by covalent addition of a methyl group. The modification can be replicated in newly divided cells. The process is carried out by the enzyme DNA methyltransferase 1 (DNMT1) through recognition of a hemimethylated sites on newly synthesized DNA to imprint the modification to the complementary, daughter-strand, CpG.

New methylation sites are introduced by the activity of de novo methyltransferases. These modifications generally result in preventing access for transcription factor or enhancer binding. Methylation can be removed through either active or passive mechanisms. Briefly, passive demethylation occurs though dilution during cell division when methylation is not replicated by DNMT1. Active demethylation occurs through the enzymatic action of ten-eleven translocation (TET) methylocytosine dioxygenases (which will be discussed in detail in the following sections) [9].

Post-translational modification of nucleosomal histones is the second form of epigenetic regulation. Histones act as a scaffold around which DNA coils. Modifications consist of acetylation, methylation, phosphorylation, ubiquitylation, and sumoylation. As with methylation, the post-translational modifications can be reversed though enzymatic processes including that of lysine demethylases and deacetylases, although not as well defined. While necleosomal modification are ATP-independent, another form of epigenetic controls occurs though ATP-dependent nucleosome density remodeling where condensed regions are unavailable for transcription or vice-versa. Histone-dependent epigenetics has been well reviewed by Soshnev et al. [10].

High-order chromatic structures form the third category of epigenetic regulation. This process involves heterochromatin (nucleosomal compaction often near the nuclear membrane) and euchromatin (nucleosomal accessibility) The nuclear membrane/lamina, distal to the nuclear core, is said to be a repressive environment, though it may contain regions that are transcriptionally permissive [8]. Nucleosomal compaction occurs towards this outer region of the nucleus where heterochromatin contains compact genomic sequences which might be seen with increased methylation, and coiled around histones that are tagged with repressive post-translational modification markers. Long regions of heterochromatin often consist of large epigenomic domains (lamina-associated domains). These large condensed chromosomal regions may be separated by stretches of DNA that are relieved of condensation and available for transcription. These active regions are more likely to harbour histones showing activation-promoting post-translational modifications and reduced DNA methylation. The multi-tiered structure of condensation allows for promoter-enhancer interactions, topological associations, multi-gene expression

correlations, or chromosomal domain silencing [11–14].

The effect of mutations which disrupt any of these systems have been described as damaging toward (i) epigenetic modifiers, (ii) epigenetic mediators, and (iii) epigenetic modulators [15]. *Epigenetic modifiers* are genes whose products have direct involvement; protein coding genes that control DNA methylation, post-translational modification, or higher-order chromatin structure (e.g. *TET2*, *IDH*, *ARID*). Mutation in these genes can have a cascade effect since they control the regulation of many downstream genes. The downstream genes which are targeted by the modifiers are *epigenetic mediators*. Damaging mutations in mediators might only be linked with epigenetics when they occur in, for example, tumour suppressor genes resulting in the same disease as mutation in one of the upstream epigenetic modifiers. *Epigenetic modulators* are upstream of modifiers and have the potential to induce or suppress differentiation-specific epigenetic states. Pro-inflammatory NF-κB signalling has been reported as an example of a modulator that links the environment and epigenome [15]. Excessive signalling in this pathway can trigger an epigenetically-dependent interleukin-6 positive feed-back loop.

Epigenetic modulators		*Epigenetic modifiers*		*Epigenetic mediators*
Secondary effect		Primary effect		Primary effect
on multiple pathways	$\rightarrow$	on multiple genes	$\rightarrow$	on single genes
(*NF-κB*, *IL-6*)		(*TET2*, *IDH1*)		(*IGF-2*, *OCT4*, *WNT*)

4.1.3 DNA methyltransferase

The methylation process occurs when the 5' carbon of cytosine is modified with the addition of a methyl group. S-Adenosyl methionine acts as a donor and the reaction is carried out by DNA methyltransferases (DNMTs) enzymes [16]. However, the control of methylation processes involves many factors. Damaging variants in several methylation-regulating genes have been found to frequently occur in haematopoietic malignancy. This includes DNA methyltransferase 3A (*DNMT3A*), ten-eleven translocation 2 (*TET2*), and isocitrate dehydrogenase 1 (*IDH1*) and IDH2 [17]. For these methyltransferases, the

effect of defective protein results in a increased capacity for self-renewal and blockage of differentiation for haematopoietic stem/progenitor cells. Most notably, this drives clonal expansion into a pre-leukemic stem cell state. While differentiation is restrained, the ability to acquire further, proliferative, driver variants accelerates transformation to clonality. [17].

DNMT1 maintains the methylation pattern for newly synthesized DNA strands by copying that of the mother strand. Conversely, methylation can occur at specific loci independently of replication through the activity of DNMT3, a de novo methyltransferase. DNMT2 is said to perform primarily as an RNA methyltransferase through its roll in t-RNAAsp methylation [18]. DNMT2 has also been seen to methylate DNA in vivo [19] Unlike DNMT1, DNMT3 (A and B) are defined as as de novo methyltransferases since they can act upon unmethylated DNA substrates [20].

AML due to somatic variants in *DNMT3A* was first identified by Ley et al. [21] using whole genome sequencing. Furthermore, they found that 22 % of AML also had variants in the same gene. Variants have also been found at about half that rate in myelodysplastic syndrome (MDS) and myeloproliferative neoplasms (MPN) [22, 23]. Disease-causing variants in this gene have been shown to have reduced DNA methylation activity [24]. *IDH1* and *IDH2* are another set of genes that often harbour variants in AML, MDS, and MPN, and frequently in low-grade malignant gliomas. [25–29]. [30, 31].

As genetic screening become more robust, aberrant protein function were also found in other cancer types, including T-cell lymphoma [32–34]. These genes produce enzymes that perform as part of the citric acid cycle. *IDH2* is the mitochondrial homolog of *IDH1*. They function to catalyze the oxidative decarboxylation of isocitrate. As a result, this produces α-ketoglutarate (α-KG) and NADPH. Most cases of IDH-dependent disease occur as heterozygous variants and, due to the enrichment for pathogenic mutation occurring at conserved residues, the mechanism is thought to be a gain-of-function [35].

Interestingly, mutually exclusive mutations in IDH and TET2 were identified in many cases of disease [36, 37]. This observation suggested a shared involvement, mechanistically, through the same pathways. The neomorphic activity of mutant IDH was shown to

perform the catalysis of a-KG to D-2-hydroxyglutarate (D-2-HG), consuming NADPH in the process [27, 28, 38]. Normally produced at a low level, D-2-HG accumulates in cells harbouring pathogenic IDH variants. Since D-2-HG has a similar structure to a-KG, it competes for the a-KG-dependent dioxygenases (including TET2), thus preventing the hydroxylation of 5-methylcytosine (5mC) by TET2. Furthermore, the hypermethylation and gene expression signature of TET2-dependent disease was mirrored both in IDH-dependent disease and in vitro [36, 39]. In an indirect manner, mutant IDH can interfere with the normal function of TET2, resulting in the same phenotype as TET2 deficiency.

4.1.4 TET-dependent DNA demethylation

In plants 5mC can be excised by Repressor of silencing 1 (ROS1)/Demeter (DME) family of DNA glycosylases and replaced with unmodified cytosine through base excision repair (BER) [40]. The ROS1/DME family of proteins has not been identified in mammals, however TET proteins carry out this role as part of the TET-TDG pathway followed by BER. Mayer et al. [41] in the year 2000 identified genome wide loss of 5mC in mouse zygotes. They showed (using staining methods) that for paternal DNA demethylation occurred in the first 8 hours after fertilisation, before DNA replication begins. The maternal genome was demethylated after several cleavage divisions. This was followed up by Oswald et al. [42] in the same year, showing active demethylation of the paternal genome in the mouse zygote using bisulfite sequencing. The first mechanistic reports showed tissue-specific accumulation of 5-hydroxymethylcytosine (5hmC) and the conversion of 5mC to 5hmC by TET1 in humans in 2009 [43, 44]. In these two papers, Kriaucionis and Heintz [43] had provided evidence that a high abundance of 5hmC can be found in the brain and Tahiliani et al. [44] demonstrated the TET1-dependent conversion of 5mC to 5hmC.

A role for TET1 in cancer was reported in 2003 showing that it acted as a complex with MLL (myeloid/lymphoid or mixed-lineage leukaemia 1) (KMT2A) [45, 46], a positive global regulator of gene transcription that is named after its role cancer regulation. These publications, if pursued further, may have arrived at the same conclusion that is currently known for TET protein function (a hypothetical time-line might have resulted in a slightly

boring recap here, at least to non-cancer biologists). Instead, and more interestingly, in 2009 Tahiliani et al. [44] used a computational search for enzymes that could modify 5mC. Methylation was known to be crucial for gene silencing, mammalian development, and retrotransposon silencing. The mammalian TET proteins were found to be orthologues of *Trypanosoma brucei* base J-binding protein 1 (JBP1) and JBP2.

Trypanosomes are a parasitic protozoa, best known to non-parasitologists as the cause of sleeping sickness and Chagas disease. Base J (*or β-d-glucopyranosyloxymethyluracil*) had been found in *Trypanosoma brucei* DNA in the early 1990s [47], although the evidence of an unusual form of DNA modification goes back to at least the mid 1980s [48]. It was the first hypermodified base that was known in eukaryotic DNA. At least by 2008, base J was known to be always found in telomeric repeats of any organism tested [49], as well as other genomic regions. JBP1 and JBP2 had been proposed around this time to oxidize the 5-methyl group of thymine in the conversion to base J. Base J is formed when first, a specific thymidine in DNA is converted into hydroxymethyldeoxyuridine (HOMedU), and then HOMedU is glycosylated to form base J (**Figure 4.1**). In simpler terms, JBP1 and JBP2 catalyse this oxidation of thymine to 5-hydroxymethyluracil (5hmU) during the biosynthesis of trypanosome base J [44, 49–51].

Brief reference definitions

HOMeU: hydroxymethyluracil.
HOMedU: hydroxymethyldeoxyuridine.
Uracil: demethylated form of thymine.
Uridine: glycosylated form of uracil.
Thymine combined with deoxyribose creates the nucleoside deoxythymidine.

The biochemistry was bolstered by further functional investigations supporting the base modification process: in 2007 by Yu et al. [50], in 2009 by Cliffe et al. [51], and bridging the gap for 2008 was a collaborative review by the senior authors of each paper; Piet Borst and Robert Sabatini [49].

The biosynthesis of base J is strikingly similar to the TET1-dependent mechanism of demethylation. While JBP1 and JBP2 catalyse oxidation of thymine to 5hmU, TET1 was responsible for 5mC to 5hmC conversion [44]. Together with TET1, TET2 and TET3 were also demonstrated to carry out the oxidation of 5mC to 5hmC [52]. TET proteins

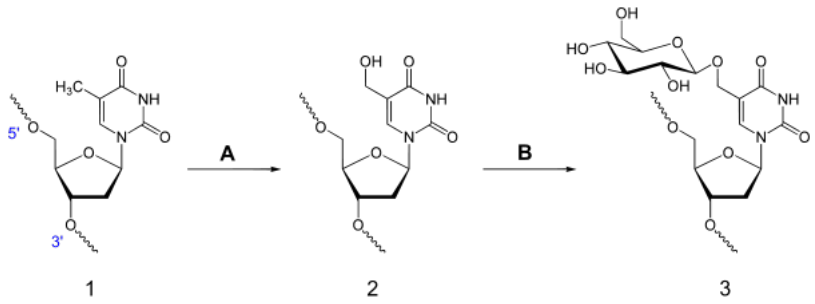


Figure 4.1: **Biosynthesis of base J.** Conversion A: thymidine hydroxylase, Conversion B: β-glucosyltranferase; 1: dT (deoxythymidine); 2: HOMedU; 3: β-D-glucosyl-HOMedU (dJ, or "base J"). Public domain re-illustration the figure by Borst and Sabatini [49].

have also been reported as catalysing the oxidation of 5hmC to 5fC and 5caC. In two back-to-back Science articles He et al. [53] demonstrated that (i) TET converts 5mC to 5fC and 5caC, and (ii) 5fC and 5caC are both present in mouse ESCs and organs, while Ito et al. [54] show that (i) TET converts 5mC and 5hmC to 5caC, (ii) the 5caC can then be excised by thymine DNA glycosylase (TDG), and (iii) depleting TDG causes 5caC accumulation in mouse embryonic stem cells (ESCs) [this paper also shows the use of two-dimensional thin-layer chromatography for separation of cytosine and its modified forms as 5mC, 5hmC, etc., which is unlikely to be seen today with more advanced available such as methylation-specific antibodies].

The oxidations of 5hmC to 5fC and 5caC described in those papers is also similar to oxidation of thymine to 5-hydroxyuracil, 5-formyluracil and 5-carboxyluracil, which is carried out by thymine hydroxylase as part of the thymidine salvage pathway [55].

This dull inventory of biochemical modifications becomes more interesting in the next step biologically. In general terms, DNA methylation causes specific sequences to become inaccessible for expression. The process of demethylation is initiated through modification of the 5mC to 5hmC, 5fC, etc. To return to the unmodified form of cytosine (C), the site is targeted for TDG-dependent base excision repair [53, 56, 57]. The "thymine" in TDG (*thymine* DNA glycosylase) might be considered a misnomer; TDG was previously known for removing thymine moieties from G/T mismatches. The process involves hydrolysing the carbon-nitrogen bond between the sugar-phosphate DNA backbone and

the mismatched thymine. Only in 2011 had He et al. [53] and Ito et al. [54] published the activity for TDG as also excising the oxidation products of 5-methylcytosine. Furthermore, in the same year Maiti and Drohat [56] show that TDG excises both 5fC and 5caC. The site left behind remains abasic until it is repaired by the base excision repair system. The biochemical process was mostly rounded out when, very recently (2016), Weber et al. [57] published the evidence of base excision repair coupled with TET and TDG. There are reports of 5hmC-specific binding proteins which might suggest a role greater than just an intermediate of the demethylation process [58–60]. Indeed, the honeybee has received some spotlight as a model of the "social repertoire" of the epigenome via in-depth analyses of TET dioxygenase [61, 62].

In simple terms, TET–TDG–BER produces demethylation; TET proteins oxidise 5mC to create the substrate for TDG-dependent excision. Base excision repair then replaces 5mC with C.

4.1.5 5-Formylcytosine and 5-carboxylcytosine

5fC and 5caC have not received as much attention as 5hmC. This is partially due to more difficult detection. 5caC is said to occur at levels that are 10-1000x less than 5hmC as well as having fewer tools for measurement [63]. 5hmC is detected at roughly 2-100x less than 5mC, with variance among tissues reported [43, 64]. Bisulfite sequencing is typically the most widespread methylation assay (outlined briefly in **Box 4.1.5**). This method can only differentiate between methylated and unmethylated DNA sequences. 5mC and 5hmC are both registered as methylated while 5fC and 5caC are affected by bisulfite sequencing as though unmethylated [65]. The low levels of 5caC also support the idea that it acts as a late intermediate in the demethylation process [63]. 5fC has also been implicated in regulation through stalling RNA pol II [66].

Box 4.1.5. Bisulfite converts cytosine residues to uracil. 5-methylcytosine is protected from the reaction. After conversion only methylated cytosines remain unconverted. The chemical reaction that occurs during conversion of cytosine to uracil consists of (i) sulphonation, (ii) hydrolytic deamination, and (iii) desulphonation.

4.1.6 TET-dependent DNA demethylation in cancer

Much of the mechanism of TET function has been addressed through embryonic stem cell research [65] , although 5hmC is found in different tissues throughout the lifecycle [43, 67]. The most striking outcome of aberrant TET activity is its association with the development of cancer. Damaging variants in *TET2* were attributed as the cause of several myeloid malignancies around the same time as the protein's function was reported for TET-dependent oxidation [68–74]. Not only were damaging TET2 mutations found in disease, but the levels of 5hmC were also affected, linking the molecular mechanism of impaired demethylation with disease [75]. In mice the depletion of TET2 skewed the differentiation of haematopoietic precursors [75], as well as amplifying the rate of haematopoietic or progenitor cell renewal [76–79] (*as an aside*, [79] has quantifications of 5hmC and 5mC, measured by dot blot in heterozygous and homozygous knock out mouse DNA, showing results that are reminiscent of what I find in my assay on humans, not identical but consistent). This increased proliferation of cells may promote malignancy, particularly in a hypermethylated genomic landscape.

Within the last five years there has been enough evidence to reasonably state that in the absence of other obvious causes, mutations in TET2 have potential to initiate the sequence of events leading to haematological malignancy. TET2 mutations have been identified from pre-malignant haematopoietic stem cells in patients who had later developed myelodysplastic syndrome or acute myeloid leukaemia [17, 80, 81]. TET2 mutations were said to be preserved in tumours along with further somatic cancer-promoting variants. Rasmussen and Helin [82] reiterates that TET2 variants have also been found for "aged healthy individuals with clonal haematopoiesis" and who also have an "increase (for their) propensity to develop haematopoietic malignancies" [83–86]. (*Note:* the paper by Genovese et al. [84] has a beautiful illustration and data presentation in Fig 4 showing haematopoietic clones undergoing progression toward myeloid cancer).

TET2 inhibition in mice has been shown to increase cell proliferation. Transformation to a cancerous state has been said to occur at a slow rate, and occurring with low penetrance [82]. However, in humans the same time scales would not be considered as

having a long latency.

Identical variants which arise in different cell types can result in malignancies that might be described on the opposite ends of the phenotypic spectrum. In that case one might argue that phenotypically-driving categorisation might impede scientific discovery compare to a genetically-driven categorisation. In the case of TET2-dependent disease the literature seems to report a bias for myeloid differentiation [82]. It is not clear if there is a true bias for malignancy to arise though the myeloid progenitor (as opposed to the lymphoid progenitor) cell lineage, or if the reports to date and still too limited. TET-dependent disease in humans often reported as developing myelodysplastic syndrome and acute myeloid leukaemia, while mouse studies may be limited in their classification methods. The questions evoked by these reports may include: (i) Are the cells of myeloid lineage the only cell types that can bear damaging *TET2* variants and continue to propagate? (ii) Is TET expression the same in other cells types as in myeloid cells? (iii) Would humans with germline *TET2* variants only produce (or favour) malignancy of myeloid/lymphoid lineage?

Haematopoietic malignancies were also seen with mutation in TET1 and TET3 [70]. TET1 or TET1/TET2 double knock out mice were found to produce lymphoid, B cell malignancy, [87, 88] though that is not to say that TET1 necessarily has a greater role in either lymphoid or myeloid cell types. Second degree evidence of TET involvement is shown where mutation or loss of regulators of TET (IDH1, IDH2 and WT1) also results in methylation-dependent disease [36, 89–91]. A final point of importance is that each individual TET proteins (TET1-3) are not redundant. Deficiency in one individual protein can not be fully compensated by the other family members despite their seemingly parallel function [82].

4.2 Aims and objectives

To investigate the validity of the *TET2* gene candidate as the cause of a closely shared phenotype in two families. To explore the population genetics data of both local and global allele frequencies of rare variants. To investigate potential secondary germline or

somatic variants that could contribute to disease. To assay methylation profiles in all patients and relatives. To functionally assess the mechanism of TET2 loss of enzyme activity.

4.3 Methods

4.3.1 PBMC purification

Fresh blood samples were collected on several occasions from the proband and all family members, in 6mL EDTA coated collection tubes. From this human peripheral blood mononuclear cells (PBMCs) were isolated by density centrifugation using Lymphoprep. Blood was diluted (2:1) with sterile 1x Phosphate-buffered Saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10mM Na2HPO4, 2mM NaH2PO4). 45mL was transferred into a 50mL Falcon tube and centrifuged at 4000rpm for 20 min. The layer of white blood cells formed at the interface between red blood cells and serum was carefully transferred onto a new 50mL Falcon tube containing 15mL Lymphoprep. The total volume was brought to 50mL with sterile 1x PBS if required. Centrifugation was applied at 1200rpm for 20 min. The top layer was carefully removed and PBMCs were collected from the subsequent layer taking care to avoid the remaining layer of Lymphopred/red blood cells. Isolated PBMCs were washed in sterile 1x PBS. Cells were stored in freezing medium (fetal bovine serum *BFS) and ETDA) and stored at -80° C. If cells are to be used immediately for most typical cell culture experiments they should be counted and seeded at a density of $1x10^6$ cells/mL in complete RPMI in appropriate experiment vessels. Alternative methods to Lymphoprep PMBC purification can include seperation using ficol or simply red blood cell lysis using an appropriate buffer.

4.3.2 Whole exome sequencing

Germ-line variant analysis was carried out for patients 1-3; DNA was purified from whole blood and prepared using SureSelectXT with All Exon v6 capture library and sequenced on Illumina HiSeq 3000 for 2 × 150-bp paired-end sequencing. Reads were aligned with

BWA-MEM to GRCh37/hg19 and variant calling was performed according to GATK-best practices. A detailed protocol section is provided in chapter 5. Somatic variant analysis was carried out on exome data from Patient 1 and 2; genomic DNA, extracted from the early passage primary dermal fibroblasts and lymphoma tissue, was submitted for whole exome sequencing using Nextera Rapid Capture Exomes kit (Illumina) coupled with massively parallel sequencing by the Illumina NovaSeq Sequencing system. The DNA sequences were mapped to the hg19 human genome by NovoAlign (http://novocraft.com/main). In parallel, homozygosity mapping was performed using the Affymetrix Genome-Wide Human SNP 5.0 microarray. Homozygous regions were identified using Homozygosity Mapper (http://homozygositymapper.org) and further analysed using microsatellite markers.

4.3.3 PCR and Sanger sequencing

Amplification of genomic DNA for Sanger sequencing was performed by the standard PCR method. PCR clean-up was performed with ExoSAP-IT (Affymetrix, Santa Clara, USA). Sanger sequencing was then provided using the same primers, primers sequences are listed in Supplementary Table 3. Sanger sequencing using BigDye Terminator Cycle Sequencing Kit, version 3.1 (Applied Biosystems, MA, USA) and analysis on an ABI 3130XL DNA analyzer (Applied Biosystems, MA, USA).

Table 4.1: Oligonucleotide primers

Oligo Name	**Sequence**
SDHA f	TGGGAACAAGAGGGCATCTG
SDHA r	CCACCACTGCATCAAATTCATG
UBE4A f	GGATGGACGTTCCTATTCCCC
UBE4A r	AGGTCTGCAAGAGACTTGATTC
TET1 f	TCTGTTGTTGTGCCTCTGGA
TET1 r	GCCTTTAAAACTTTGGGCTTC
TET2 f	AAAGATGAAGGTCCTTTTTATACCC
TET2 r	ATAGCTTTACCCTTCTGTCCAAAC
TET3 f	CACTCCGGAGAAGATCAAGC
TET3 r	GGACAATCCACCCTTCAGAG
TET2 f seq f1 jar	CTTTCGCATTCACACACACTTT
TET2 r seq f2 jar	GAGTTCCCCTGCACATGTTC
TET2 ORF f	ATGGAACAGGATAGAACCAAC
TET2 ORF r	TCATATATATCTGTTGTAAGGCCC

4.3.4 DNA Methylation

Purified DNA was quantified with the Qubit dsDNA BR Assay Kit and normalized. In duplicate, three aliquots of DNA were prepared for each sample according to the EpiMark analysis kit (E3317S, NEB) manufacturer instructions. One aliquot per sample was treated with T4 β-glucosyltransferase (T4- BGT) (10 units per sample) (M0357S) to convert 5hmC to glucosylated 5hmC (5ghmC) using uridine diphosphate glucose (UPD) (1.24 uL). T4 β-glucosyltransferase (T4-BGT)-dependent glucosylation of 5hmC to form 5ghmC was facilitated by heating at 37oC for 6 hours. After glucosylation, enzyme restriction was performed on all samples. MspI (R0106S) and HpaII (R0171S) recognize the same DNA sequence ('5 CCGG 3') but are differentially sensitive to methylation status. MspI cleaves both 5mC and 5hmC. However, MspI cleavage is blocked by 5ghmC. HpaII cannot cleave modified sites; any modification with 5mC, 5hmC, or 5ghmC at either cytosine will prevent cleavage. All samples were heated at 37oC for 6-12 hours to allow for complete digestion. After digestion the separation and analysis of DNA samples up to greater than 60,000 base pairs was performed using the Agilent Genomic DNA ScreenTape assay (cat. 5067- 5365/6) with the Agilent 2200 TapeStation system according to manufacturer specifications, with quantification sensitive down to 5 pg/uL. High-resolution assay reports for each Genomic DNA ScreenTape were analysed with the use of ImageJ software to plot lane densities [92]. DNA runs from high to low molecular weight on the gel from top to bottom; the density plots present this data from left to right. The area under the curve (AUC) at each decile interval was recorded and used for calculation of differences between 5mC and 5hmC concentration. The minimum level of 5hmC genome wide is found when the difference before and after T4-BGT treatment approaches zero for MspI-restricted DNA. Similarly, the difference between MspI and HpaII restriction indicates non-specific methylation.

4.3.5 Western blotting

Protein was purified by lysis of purified PBMCs using sodium orthanovanadate, Complete Protease Inhibitor Cocktail and PMSF (Sigma), with RIPA buffer (50mM Tris HCL

pH 7.5, 150mM Sodium Chloride, 0.5% Sodium Deoxycholate, 0.1% Sodium Dodecyl sulfate (SDS). Protein concentration was determined using Pierce Bicinchoninic Acid Assay (Pierce, Thermo Scientific) according manufacturer's instructions.

Cell lysates were denaturated at 90°C for 10 minutes, equal amount of lysates were loaded. The Novex Mini Gel Tank and blot module, Bolt 4–12% Bis–Tris Plus Gels, was used for electrophoresis in 1x SDS NuPAGE MOPS Running Buffer (Novonex, Life Technologies, USA). Proteins were transferred to a polyvinylidene fluoride (PVDF) membrane (Thermo Scientific Pierce, Life Technologies) in NuPAGE Tris - Glycine Buffer (Life Technologies) and subsequently blocked with 5% bovine serum albumin (BSA) or 5% non-fat milk in Tris Buffered Saline/0.1% Tween 20 (TBS/T) for 1hour at RT, followed by incubation with anti-human primary antibodies: mouse TET2 (Active Motif, 61389, clone 21F11, 1:1000, USA) and rabbit GAPDH (Cell Signaling Technology, 5174, clone D16H11, 1:2000, USA) for overnight at 4°C. The blots were then washed three times with TBS/T and incubated with appropriate Horseradish Peroxidase (HRP)-conjugated secondary antibodies: anti-mouse (Cell Signaling Technology, 7076, 1:5000, USA) and anti-rabbit (Cell Signaling Technology, 7074S, 1:2000, USA) in 5% non-fat milk in TBS/T for 1 hour at room temperature (RT). The blots were developed with ImmobilonTM Western Chemiluminescent HRP substrate (Millipore) according to the manufacturer's instruction or used Super Signal West Femto/Pico (Thermo Fisher Scientific).The Chemiluminescent images were captured on a G:BOX Chemi using GeneSnap Software (Syngene, India).

4.3.6 Gene expression by PCR and quantitative RT-PCR

The clearance of Sendai virus vectors and endogenous expression of pluripotent markers were validated by vector- and marker-specific primers, respectively. RNA was extracted by ReliaPrep RNA Miniprep systems (Promega, USA), followed by reverse transcription using GoScript Reverse Transcription System (Promega, USA) according manufacturer's instructions. Quantitative RT-PCR (qPCR) was provided by GoTaq Green Master Mix and product was detected by standard 2% agarose gel electrophoresis. qRT-PCR reaction was run on the Applied Biosystems QuantStudio 7 Flex Real-Time PCR System (Thermo Fisher Scientific, USA) with the GoTaq qPCR Master Mix (Promega, USA). The data

were analysed using the QuantStudio software (Thermo Fisher Scientific, USA) and relative gene expression was B determined using the 2-delta-delta Ct method using SDHA and UBE4A as a housekeeping genes. Primers sequences are listed in table 4.1.

4.4 Results

4.4.1 Family summary

Table 4.2: Two families where TET2 deficiency was identified.

Family	Treatment centre	Parents	Probands	Siblings
1 (TET2 p.H1382R)	Newcastle	Mother *Het* Father *Het*	P1 *Hom* P2 *Hom*	Brother *Het* Brother *WT* Sister *WT*
2 (TET2 p.Q1632*)	Leeds	Mother *Het* Father *Het*	P3 *Hom*	Sister *Het* Sister *Het*

Patient 1 (P1) and P2 were chronologically the first cases where disease was ultimately attributed to homozygous TET2 deficiency. This family, (family 1) were under the care of Prof Hambleton in Great North Children's Hospital, Newcastle upon Tyne. The probands had also attended Department of Paediatrics, Leeds General Infirmary, and St James's University Hospital at earlier ages. Prof Hambleton's research group and clinical collaborators had been investigating this case when family 2 (P3) came into our study as a patient with PID. P3 was found to have a homozygous non-sense variant in *TET2*. The diagnosis of lymphoma in P3 supported this germ-line variant as the genetic determinant of disease since there is consensus that somatic mutations in this gene result in similar disease features. Since the patients from family 1 and 2 had such similar conditions and genetic diagnosis, a collaboration was established to complete this study. The family pedigrees are shown in **Figure 4.2** and **4.3**.

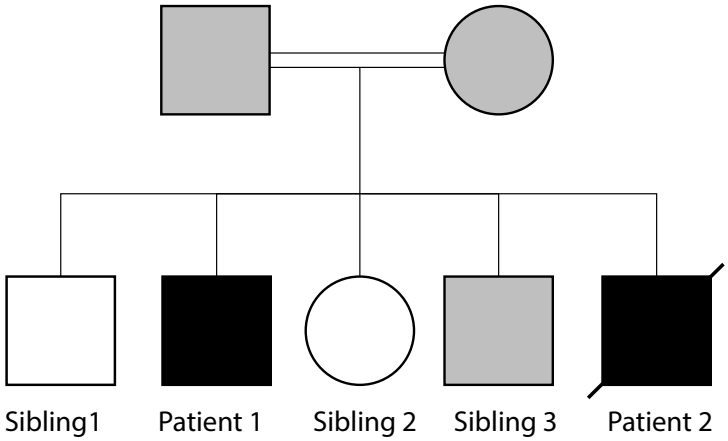


Figure 4.2: **Pedigree of family one.**

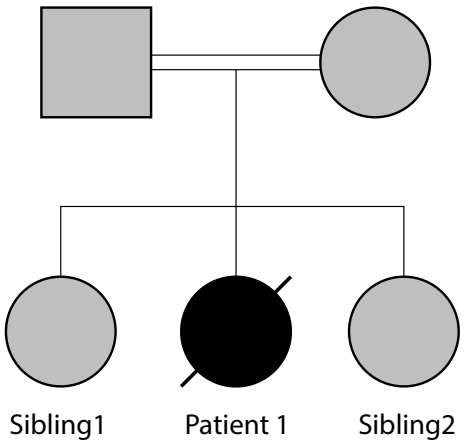


Figure 4.3: **Pedigree of family two.**

4.4.2 Clinical histories

The information collected in this section has been gathered from many physicians and clinical scientists. Prof. Hambleton and Dr. Savic are primarily responsible for this information.

This section provides detailed clinical histories to adequately interpret the finding of this chapter. However, a briefer summary of the key clinical factors are summarised in the subsequent section (subsection 4.4.3). The main clinical features and treatment strategies of patients 1-3 are described individually in the following subsections. Summaries of these descriptions are illustrated as a timeline of events in **Figure 4.4** and **4.5**.

4.4.2.1 Patient 1 history

P1 (**Figure 4.4**) received a clinical diagnosis of autoimmune lymphoproliferative syndrome (ALPS) but lacked a genetic determinant. He presented to hospital at 4 weeks of age due to pneumonia associated with Respiratory Syncytial Virus (RSV) and Cytomegalovirus (CMV) infection and was treated with Ganciclovir . Subsequently, P1 showed failure to thrive and developmental delay, and had frequent admissions due to recurrent lower respiratory tract infections. From 18 months of age, he developed hepatosplenomegaly, chronic lymphadenopathy, persistent EBV viremia, recurrent respiratory infections progressing to bronchiectasis and autoimmune complications such as immune thrombocytopenia and autoimmune hemolytic anaemia, requiring frequent transfusions. An autoimmune screen showed positive Coomb's test, anti-nuclear antibody (ANA) and rheumatoid factor (RF). A lymph node biopsy showed EBV-associated follicular hyperplasia (**Figure 4.6**).

At the age of 3 years, a diagnosis of ALPS was suspected on the basis of this combination of lymphoproliferative disease and autoimmunity, and confirmed with the demonstration of defective Fas-mediated apoptosis (subsection 4.4.16) and raised double negative (CD4-CD8-) TCR$\alpha\beta$ T-cells (20%) (Table 4.3). Throughout this illness, standard lymphocyte subsets were grossly normal. His IgM and IgA levels were normal, but IgG and IgG1 were intermittently high (Table 4.4).

P1 was treated with high dose (2g/kg) intravenous immunoglobulin, Rituximab (anti-CD20 antibody) and corticosteroid. Despite some initial response by the age of 4 years, P1's condition deteriorated markedly, with the development of massive hepatosplenomegaly, liver dysfunction and evolution of his lymphadenopathy to become hard and 'knobbly'. Further investigations including lymph node biopsy showed an EBV-positive Hodgkin-like polymorphic B-cell lymphoproliferative disorder (**Figure 4.6**). He was started on low intensity chemotherapy (vincristine and rituximab), but despite hyperhydration developed the life-threatening complication of tumour lysis syndrome with acute renal failure, requiring prolonged intensive care, and further complicated by Stenotrophomonas pneumonia and sepsis. He received four doses of adoptive EBV-specific cytotoxic T-cells with therapeutic benefit.

Since his lymphoproliferative disorder apparently arose in the context of an inborn error of immunity, P1 was prepared for haematopoietic stem cell transplant (HSCT) as a potentially lifesaving procedure. Because of his gross hepatosplenomegaly, he required a splenectomy prior to transplant. Histological examination of the explanted spleen showed granulomatous chronic inflammation, extramedullary haematopoiesis, and the effects of Rituximab therapy (**Figure 4.6**); both the spleen and a liver biopsy were free of the lymphoproliferative disorder.

At 4 years 4 months old, he received a matched sibling donor bone marrow HSCT after reduced intensity conditioning using fludarabine 150mg/m2, melphalan 140mg/m2 and alemtuzumab 1mg/kg (given days -14 to -10). Exceptionally, he showed evidence of autologous T-lymphocyte reconstitution in the periphery before completion of his conditioning (2513/ul) so received additional serotherapy on days -2 and -1 in the form of anti-thymocyte globulin, 2 x 2mg/kg. The T-replete graft contained 10.2×10^6 CD34+ stem cells. He engrafted rapidly but already by day +18 he showed mixed chimerism in whole blood, 91% donor. One month after BMT, he was at risk of rejecting the graft altogether with evidence of a large population of recipient T-cells (only 22% donor, compared with 100% donor myeloid cells). However, after receiving further alemtuzumab 0.9 mg/kg and an unconditioned stem cell top-up from the same sibling donor, the proportion of donor cells improved to reach 100% T-cells but only 18% B-cells and

negligible donor myeloid chimerism by 12 months post-transplant.

P1 was monitored frequently and his condition was clinically stable for the first two years post-transplant. An important and interesting finding was that starting from 3 months after transplant, he developed leucocytosis, monocytosis, neutrophilia and lymphocytosis affecting all subsets (subsection 4.4.16). At 26 months post-transplant (6 years 6 months old), he was noted to have recurrent lymphadenopathy, hepatomegaly and hypercalcemia of unknown cause. Lymph node and bone marrow trephine biopsies showed non-caseating granulomata with no evidence of malignancy (**Figure 4.6**) and transiently increased serum angiotensin converting enzyme.

At 36 months post-transplant (7 years 4 months old), he developed thrombocytopenia with positive ANA and ongoing widespread lymphadenopathy. At the same time, his IgG levels, which were previously normal, were found to be elevated (29.7g/L), whereas IgM and IgA remained within normal limits. His β2-microglobulin was also increased (14.6mg/ml, normal range $<$2.7mg/ml). Based on the recurrence of lymphadenopathy, hepatomegaly and autoimmune phenomena, he was diagnosed with relapse of ALPS. The Fas-mediated apoptosis assay was repeated, and found to be normal (subsection 4.4.16); this was expected given that the patient had 78% donor T-cells. Nonetheless, soluble FasL, vitamin B12 and IL10 were all raised. P1 then received a course of rituximab, following which his IgG and β2-microglobulin levels normalized. As anticipated, the absolute CD19 count dropped to 0, whilst at the same time the absolute CD3 count was also reduced.

At 48 months post-transplant (8 years 4 months old), P1 was admitted to hospital with central cyanosis and poor lung function secondary to severe bronchiectasis. The lymphadenopathy and hepatomegaly were reduced in size but new lumps were noted in his scrotum and tongue. Excision biopsy of these lesions demonstrated two granular cell tumours that were completely excised (**Figure 4.6**). A liver biopsy at the time showed granulomatous inflammation with cirrhosis (**Figure 4.6**). Subsequently, he developed severe immune thrombocytopenia at 60 months (9 years 4 months old) and anaemia at 62 months (9 years 6 months old) post-transplant, requiring blood product support and

a further course of rituximab for presumed autoimmune etiology. During this time, he also presented with headache and hypertension and later had a seizure, with brain CT showing a small right frontal bleed. In addition, he developed pleural effusions and ascites of unknown cause which were managed with fluid restriction and diuretics and gradually resolved.

At 63 months post-transplant (9 years 7 months old), another lump was noted at his left upper arm and excision biopsy showed another granular cell tumour. At 84 months post-transplant (11 years 4 months old), he developed another episode of thrombocytopenia and increased lymphadenopathy. At 98 months post-transplant (12 years 6 months old), he developed respiratory and gut failure secondary to *E.coli* sepsis complicating a severe febrile diarrheal illness acquired in Pakistan. His stool was positive for cryptosporidium, norovirus and sapovirus. He also required blood transfusion due to thrombocytopenia and anaemia.

Care was shifted to a palliative footing as he was severely debilitated and not expected to survive, however he showed a remarkable recovery in the home environment. Currently, the patient is 17 years old and 13 years post transplantation. His general condition is poor and he is no longer in full time education. He requires supplemental oxygen and his exercise capacity is severely limited. He was evaluated for short stature and failure of pubertal development. He has become transfusion-dependent for chronic anaemia, the cause of which is uncertain but the patient has declined further bone marrow investigation.

4.4.2.2 Patient 2 history

Patient 2 (P2) (**Figure 4.5**) was the younger brother of P1. His health problems began with haematuria and nephrotic range proteinuria in the first 4 weeks of life. Subsequently he was noted to have hypothyroidism and hypogammaglobulinaemia that were attributed to renal losses and treated with thyroxine and immunoglobulin supplementation; lymphocyte numbers were normal. He briefly required pediatric intensive care for presumed aspiration pneumonia complicated by pneumothorax. He developed CMV viremia around 8 weeks of age with evidence of respiratory involvement, treated with ganciclovir.

At around the same age he developed hepatosplenomegaly and lymphadenopathy. A lymph node biopsy showed a nodal peripheral T-cell lymphoma of T follicular helper phenotype and clonal TCRG gene rearrangement (**Figure 4.6**). He was thrombocytopenic but no autoantibody tests were documented and Coomb's test was negative. Renal biopsy showed granulomatous nephritis (**Figure 4.6**) and a diagnosis of possible BCGosis was made; he received anti-mycobacterial therapy and supportive care.

Based on the presence of lymphadenopathy, lymphoma, hepatosplenomegaly and autoimmunity in the context of a positive family history, P2 was clinically diagnosed with ALPS. This diagnosis was confirmed by defective Fas-mediated apoptosis (subsection 4.4.16), increased DN TCR$\alpha\beta$ T-cells (1.9 %) and raised soluble Fas ligand, 0.96 ng/ml (Table 4.3). In addition there was a very low fraction of IgM memory (0.33%) and class-switched memory B-cells (0.03%). He was treated with cyclophosphamide and methylprednisolone for his lymphoma; only one dose of vincristine was given due to deranged liver function. A maternal CD3/CD19-depleted haploidentical peripheral blood HSCT was performed at 9 months of age, following modified intensity conditioning (alemtuzumab 1mg/kg, treosulfan 42 g/m2, Fludarabine 150 mg/m2; 2 x 106 CD34+ stem cells). The HSCT was unsuccessful as he developed early graft rejection accompanied by early autologous lymphoid reconstitution, with evidence of 100% of T-cells being recipient in origin. ATG was ineffective in rescuing donor chimerism. Subsequently, 3 months post-transplant, his condition deteriorated when he developed respiratory failure and required intubation for presumed sepsis. The patient died at the age of 13 months, 4 months post-transplant.

To summarize, both brothers presented with clinical features of ALPS including lymphadenopathy, hepatosplenomegaly, lymphoma and autoimmune phenomena such as immune-mediated thrombocytopenia and anaemia. Both patients also bore laboratory features of ALPS including defective Fas-mediated apoptosis, increased DN TCR$\alpha\beta$ T-cells and increased soluble Fas ligand. In addition, they both showed developmental delay, susceptibility to infection including CMV and EBV, together with granulomatous inflammation, which are not commonly associated features of ALPS. Since neither patient bore mutations in genes already associated with inborn errors of immunity, and parents

were related, a novel autosomal recessive disorder was suspected.

4.4.2.3 Patient 3 history

Patient 3 (P3, family 2) (**Figure 4.5**) was the second child born to related parents from the same community as P1 and P2. Her problems with infection emerged at around 18 months of age, after which she suffered from frequent respiratory tract infections requiring multiple courses of antibiotics. On at least two occasions, pneumonias resulted in acute respiratory failure and the need for ventilatory support. There was also a history of loose stools and relatively poor weight gain.

Immunologic investigations suggested impaired humoral immunity: She was found to be IgA deficient, had reduced levels of IgM, normal levels of total IgG, but reduced IgG2 subclass. She also had impaired response to pneumococcal challenge. Subsequent investigations showed essentially normal lymphocyte subsets, but absent class-switched memory B-cells.

Despite the institution of immunoglobulin replacement, antibiotic prophylaxis and physiotherapy, ultimately this progressed to bronchiectasis with an overnight oxygen requirement. Furthermore there was longstanding clinical and laboratory evidence of pathological lymphoproliferation in the form of hepatosplenomegaly and lymphadenopathy as well as increased double negative T cells (DNTs) in peripheral blood (9% aged 8 years). However, Fas-dependent apoptosis as well as T-proliferative response were normal (Table 4.3, and subsection 4.4.16). A lymph node biopsy showed EBV-associated follicular hyperplasia (**Figure 4.6**) and no evidence of malignancy at that time. Although there was no definite evidence of autoimmunity either clinically or serologically, moderate thrombocytopenia was evident over several years. During this time she developed two benign skin tumours: a cellular dermatofibroma and a pilomatrixoma (**Figure 4.6**).

These immunological features occurred against a background of significant global developmental delay, for example P3 walked at 4 years of age and was not able to attend mainstream school. She had feeding problems and there was a suspicion of recurrent aspiration, managed by fundoplication and creation of a gastrostomy for enteral feeding.

At the age of 12 years, P3 presented with a mediastinal mass and pericardial effusion and investigations revealed a primary mediastinal large B-cell lymphoma (**Figure 4.6**). She tolerated R-CHOP chemotherapy and went into remission. At around this time she developed worsening headaches and idiopathic intracranial hypertension was detected. Imaging revealed skull thickening consistent with extramedullary haematopoiesis.

After extensive discussions with her family and in view of the ongoing risks to her health and quality of life, P3 went forward to allogeneic HSCT. After conditioning consisting of alemtuzumab (1mg/kg), fludarabine (150 mg/m2) and treosulfan (42 g/m2), she received PBSC containing 9×10^6 CD34+ stem cells from an 11/12 matched unrelated donor (single antigen mismatch in host vs graft direction). She tolerated myeloablative conditioning poorly and developed multi-organ failure requiring intensive care (respiratory, renal, circulatory and gut). Although she survived this phase, she showed very early reconstitution of autologous T-cells with progressive loss of her graft despite clinical evidence of graft-versus-host disease of skin and liver (grade III). Care was shifted to a palliative footing at home, where she died.

Table 4.3: Major clinical features of 3 patients with immunodeficiency and immune dysregulation. DNT: double negative T-cells, HSCT: haematopoietic stem cell transplantation, ND: not determined.

	Patient 1	**Patient 2**	**Patient 3**
Immunodeficiency			
Recurrent respiratory tract infections	++	+	++
Bronchiectasis	++	+	++
Herpes viral infection	++	+	+
Lymphoproliferation			
Lymphadenopathy	+	+	+
Hepatosplenomegaly	+	+	+
Lymphoma	+	+	+
Autoimmunity			
Autoimmune cytopenias	+	+	-
Autoantibodies	+	+	-

Laboratory Values			
Class-switched memory B-cells	ND	ND	Low
FasL-mediated apoptosis	Impaired	Impaired	Normal
Soluble Fas Ligand	Increased	Increased	Normal
DNT cells	High		High
Specific antibodies	Normal	ND	Low
Developmental delay			
Developmental delay	+	+	+
Outcome of HSCT			
Autologous T-cell reconstitution	+	+	+
	Split mixed chimerism	Rejected and died	Rejected and died

Table 4.4: Immunoglobulin levels of patients before transplantation. Age matched reference values are displayed in brackets. Bold values: abnormal laboratory values. For patient P2, and P3 aged 12.8 years, the values were taken on immunoglobulin supplementation

	Patient 1	**Patient 2**	**Patient 3**	
[g/l]	**age 3.5 years**	**age 5 months**	**age 2.5years**	**age 12.8 years**
IgG	22.4* (4.9 - 16.1)	12.9* (2.4 - 8.8)	13.5 (4.9 - 16.1)	8.6 (5.4 - 16.1)
IgA	1.16 (0.4 - 2.0)	0.98* (0.1 - 0.5)	<0.06* (0.4 - 2.0)	0.06* (0.8 - 2.8)
IgM	1.24 (0.5 - 2.0)	0.22 (0.2 - 1.0)	0.24* (0.5 - 2.0)	0.05* (0.5 - 1.9)
IgG1	20.78* (3.5 - 9.4)		13.4* (3.2 - 9.0)	
IgG2	2.47 (0.6 - 3.0)		0.17* (0.5 - 2.8)	
IgG3	2.76* (0.1-1.3)		2.39* (0.1 - 1.2)	

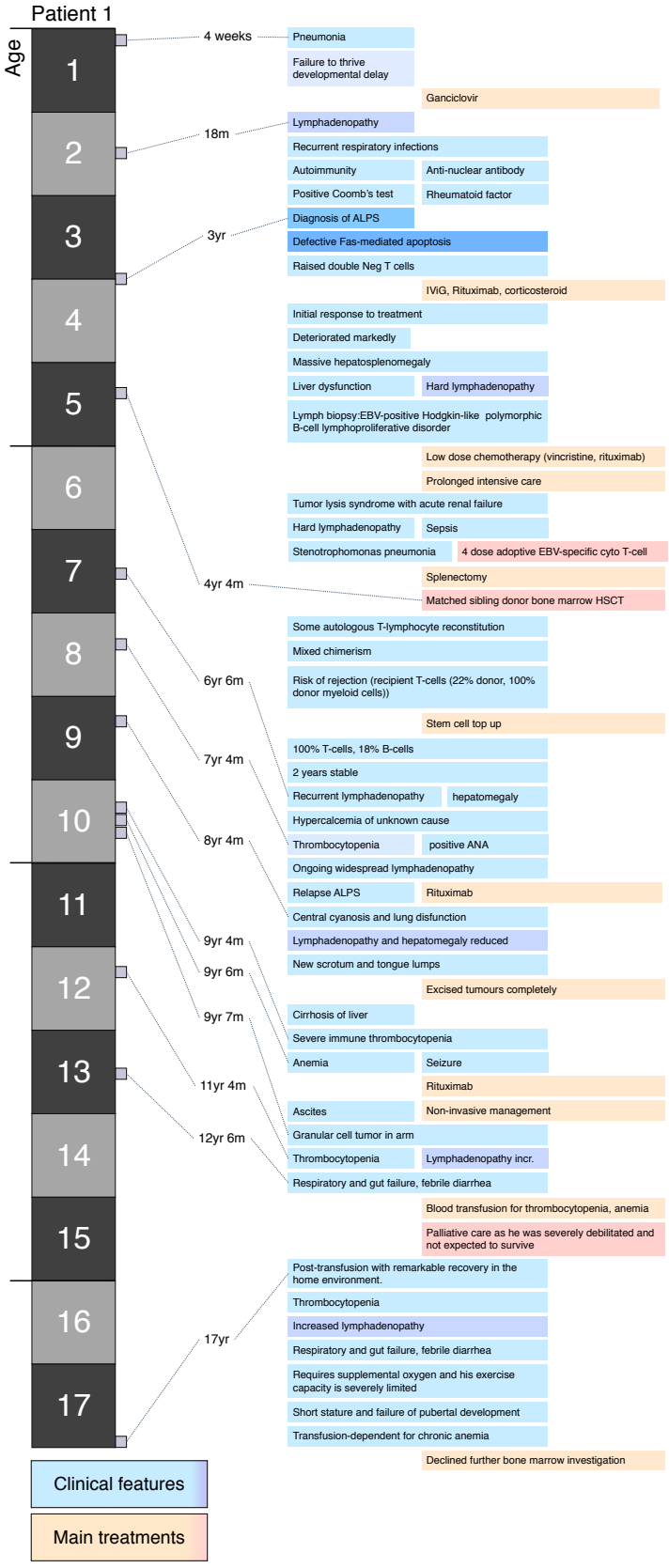


Figure 4.4: **Clinical history timeline of patient 1.** Blue tone colours are indicative of clinical features, while yellow-red tone boxes indicate treatments. Key events are colour matched between patients 1-3 (on the opposite page). The numbered age boxes represent each year of life. Small purple time-point boxes indicate specifically recorded events or assays. The onset and features of disease that are common to all patients can be seen, supporting TET2 deficiency as the common cause of disease.

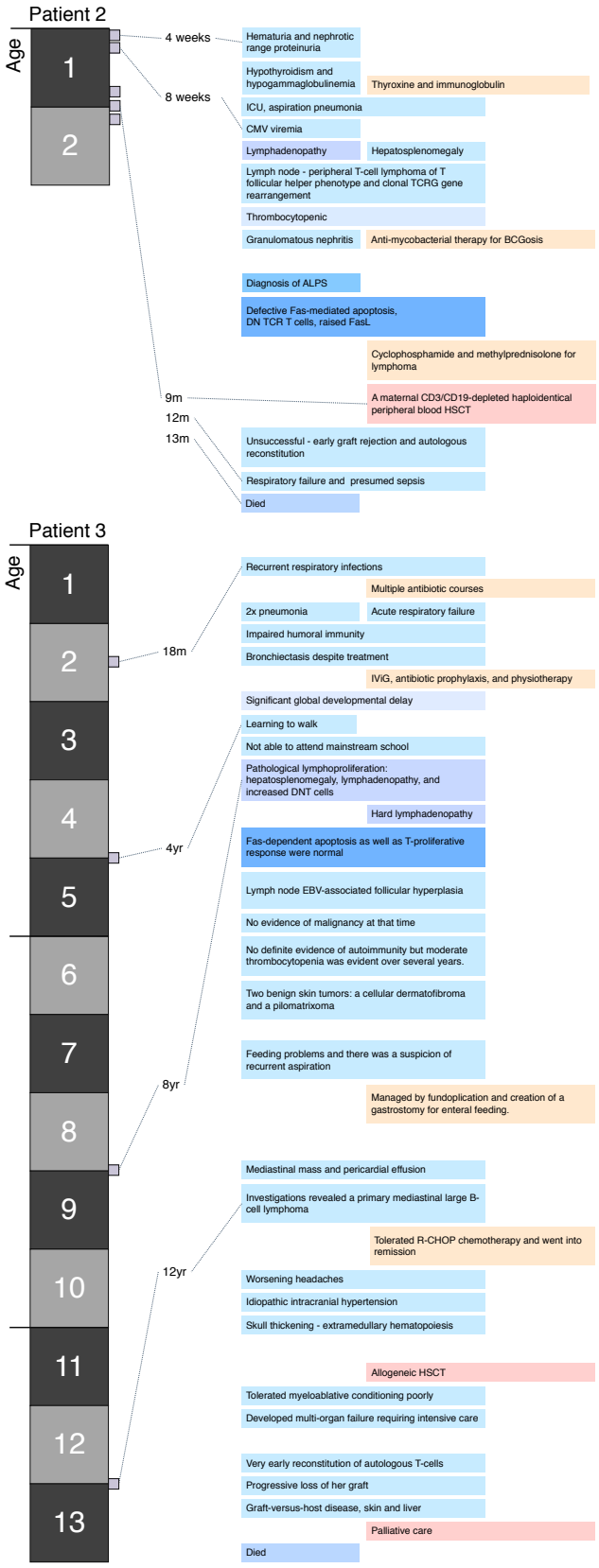


Figure 4.5: **Clinical history time line of patient 2 and 3.**

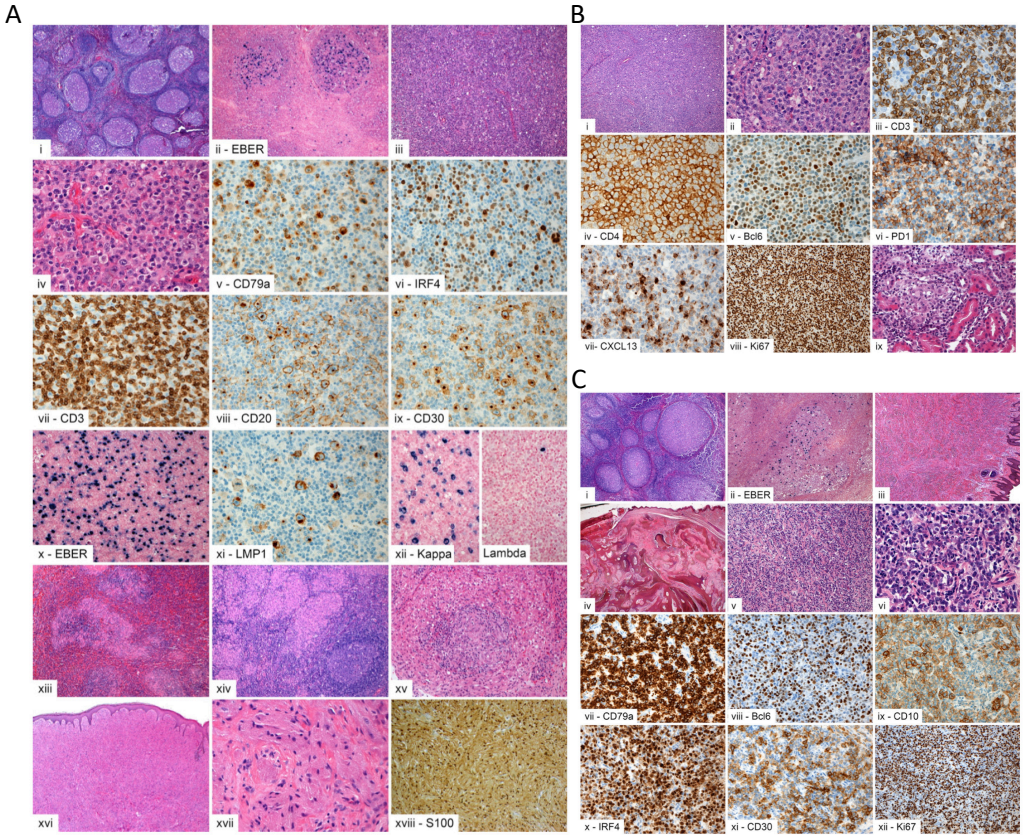


Figure 4.6: **Histopathology of lymphoid tumors and other significant pathology.** Histology performed by Dr Chris Bacon, Newcastle University.
(A) P1. **(i-ii) Lymph node biopsy showing EBV-associated follicular hyperplasia**: *i*, H&E (x40); *ii,* EBV EBER (x100). **(iii-xii) Lymph node biopsy showing EBV-positive polymorphic B-cell lymphoproliferative syndrome**: *iii*, H&E (x100); *iv*, H&E (x600); *v*, CD79a (x400); *vi*, IRF4 (x400); *vii*, CD3 (x400); *viii*, CD20 (x400); *ix*, CD30 (x400); *x*, EBV EBER (x200); *xi*, EBV LMP1 (x400); *xii*, Kappa/Lambda immunoglobulin light chains (x400). **(xiii-xv) Spleen, lymph node and liver showing granulomatous inflammation**: *xiii*, H&E spleen (x100); *xiv*, H&E lymph node biopsy (x100); *xv*, H&E liver biopsy (x200). **(xvi-xviii) Scrotal skin showing granular cell tumor**: *xvi*, H&E (x40); *xvii*, H&E (x400); *xviii*, S100 (x200).
(B) P2. **(i-viii) Lymph node biopsy showing nodal peripheral T-cell lymphoma with T follicular helper phenotype**: *i*, H&E (x100); *ii*, H&E (x600); *iii*, CD3 (x600); *iv*, CD4 (x600); *v*, Bcl6 (x600); *vi*, PD1 (x600); *vii*, CXCL13 (x600); *viii*, Ki67 (x200). **(ix) Renal biopsy showing granulomatous inflammation**: H&E (x400).
(C) P3. **(i-ii) Lymph node biopsy showing EBV-associated follicular hyperplasia**: *i*, H&E (x40); *ii*, EBV EBER (x100). **(iii) Skin showing a cellular dermatofibroma**: H&E (x40). **(iv) Skin showing a pilomatrixoma**: H&E (x20). **(v-xii) Mediastinal mass biopsy showing primary mediastinal large B-cell lymphoma**: *v*, H&E (x200); *vi*, H&E (x600); *vii*, CD79a (x400); *viii*, Bcl6 (x400); *ix*, CD10 (x400); *x*, IRF4 (x400); *xi,* CD30 (X400); *xii*, Ki67 (x200).

4.4.3 Clinical presentation of patients with immunodeficiency

Immunological features

P1 & 2 - Two siblings from a consanguineous British Pakistani background were found with marked predisposition to herpesviral disease and early onset of ALPS, manifesting as lymphadenopathy, hepatosplenomegaly, autoimmune cytopenias and impaired Fas-dependent apoptosis together with raised serum markers of ALPS (subsection 4.4.16).
P3 - A third, unrelated child from the same community presented with recurrent respiratory and viral infections progressing to bronchiectasis in the context of humoral immunodeficiency. She also had problems that included hepatosplenomegaly and moderate thrombocytopenia. She had a notable absence of IgA, reduced IgG2 subclass, impaired specific antibody responses to vaccine antigens, an excess of DNT cells (9%) and a complete lack of class switched memory B-cells (0%).

Developmental delay

P1 & 2 - were noted to have mild-moderate developmental delay.
P3 - showed significant global developmental delay.

Lymphoma

P1 - At 4 years of age, P1 developed an EBV-positive Hodgkin-like polymorphic B-cell lymphoproliferative disorder.
P2 - Within the first year of life, P2 developed a nodal peripheral T-cell lymphoma of T follicular helper phenotype (**Figure 4.6**).
P3 - At the age of 12, P3 developed a mediastinal mass, which proved to be an EBV-negative primary mediastinal large B-cell lymphoma (**Figure 4.6**).

Treatment and outcome

P1 & 2 - Modified chemotherapy was followed by HSCT for both.

P1 - developed mixed chimerism after a matched sibling donor transplant, followed by transient relapse of ALPS and disordered haematopoiesis.

P2 - rejected his haplo-identical maternal transplant and died of sepsis.

P3 - obtained remission with R-CHOP chemotherapy and proceeded towards HSCT. She tolerated myeloablative conditioning poorly and developed multi-organ failure requiring intensive care (respiratory, renal, circulatory, gut). Although she survived this phase, she showed very early reconstitution of autologous T-cells with progressive loss of her graft and clinical evidence of graft-versus-host disease of skin and liver (grade III). Care was shifted to a palliative footing at home, where she died. The clinical and laboratory phenotype of all three patients is outlined in Table 4.3 and 4.4

4.4.4 Initial genetic investigation

This study began with genomic investigation for patient 3 as she was suspected to have a genetic cause for primary immunodeficiency. Whole exome sequencing was performed and routine analysis was carried out. Tailored investigation of rare functional variants identified a strong candidate in the gene TET2; Homozygous GRCh37, ENST00000540549, c.4894C>T, p.Q1632*, exon 11/11. Figure 4.7 shows the Integrative genomics viewer (IGV) image of an aligned and sorted bam file from proband exome sequence on the right. The indicated variant appears across all sequence reads and is therefore homozygous. The same sequence is shown in the confirmatory Sanger sequence on the left. This gene has been shown to be have a clear link with the development of lymphoma. Therefore, upon identifying this variant the patients' medical history was checked. Between the time of acquiring a blood sample for DNA extraction and exome sequencing analysis, the patient had been diagnosed with lymphoma.

Patients P1 and P2 were initially exome sequenced in Newcastle and bore the homozygous missense mutation c.4145A>G, p.H1382R in exon 9 of *TET2*. This was predicted to be highly damaging since it affects the Fe(II) binding motif, known to be critical for

TET2 enzyme activity [93].

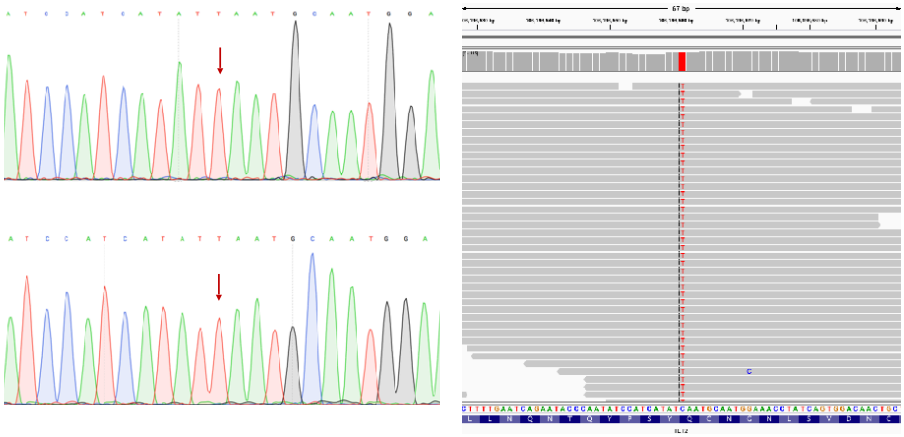


Figure 4.7: **Exome data and Sanger confirmation.** (Left) Sanger sequence confirmation of homozygous variants in forward (top) and reverse complement (bottom) sequence. (Right) Integrative genomics viewer image of an aligned and sorted bam file from proband whole exome sequence. Homozygous C>T is seen in red.

4.4.5 Sanger sequencing

The confirmatory Sanger sequencing results are shown in summary for both families in **Figure 4.8**. Sanger sequencing confirmed the inheritance of the homozygous SNV in the two proband siblings of family 1. and the inheritance of the homozygous stop variant in the proband of family 2. DNA for the entirety of family 2 was sequenced and the full results are shown in **Figure 4.9**. Sanger sequencing was carried for family 1 in Newcastle University to confirm the inheritance pattern shown on their pedigree.

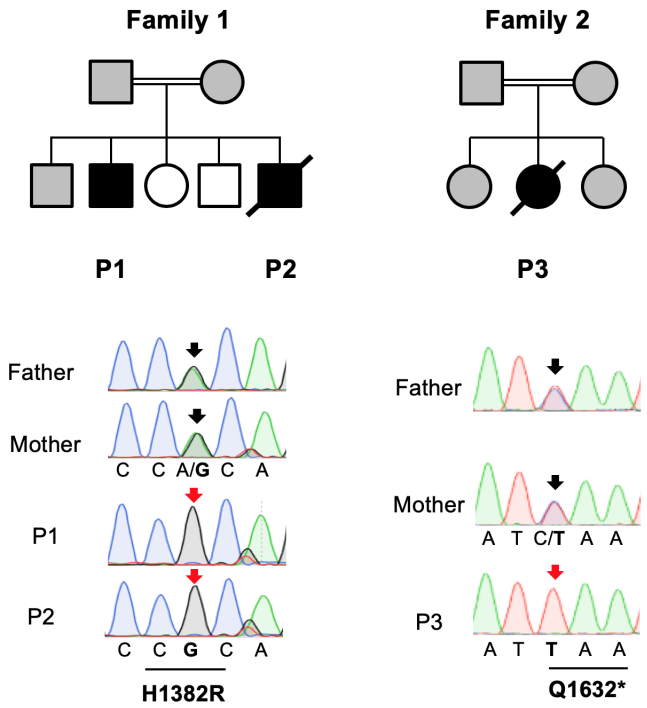


Figure 4.8: **Sanger sequencing of both families.**

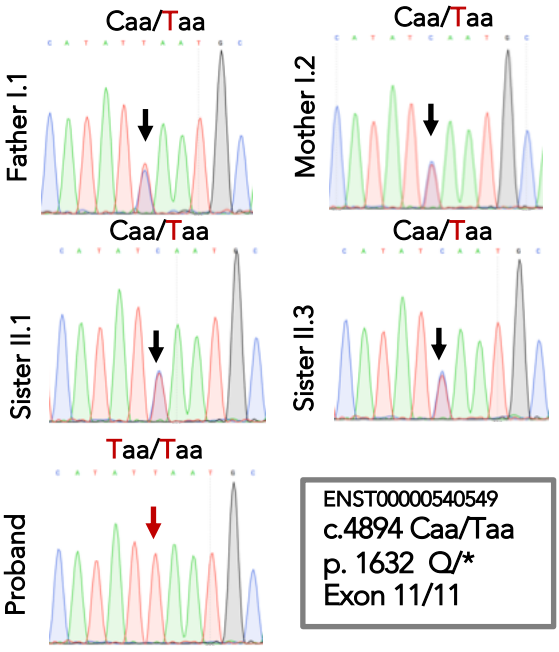


Figure 4.9: **Sanger sequencing in family two.** The alternate variant is shown in red. Only the proband was found to have homozygous inheritance.

4.4.6 Somatic mutation screening

The proband from family 1 (P3) was investigated for somatic variants by the Leeds Haematological Malignancy Diagnostic Service (HMDS), within the Department of Clinical Haematology, Leeds Teaching Hospitals NHS Trust. This screening intended to identify somatic mutations due to hypermutagenicity with loss of *TET2* (ostensibly frequent in haematopoietic stem/progenitor cells). Deep sequencing with a restricted panel of genes was performed on bone marrow-derived DNA after lymphoma was diagnosed. A bone marrow sample from pre-lymphoma was also available. This was intended for sequencing to allow a time-line of genetic variability. However, the initial post-lymphoma sample showed no significant somatic variation to justify further investigation.

Somatic *TET2* variants were observed in other lymphoma patients from Leeds. Somatic variant p.Q1632* was observed the local cohort of patients with myeloid malignancy (2/1221 acute myeloid leukaemia (AML), 2/286 chronic myelomonocytic leukaemia (CMML); table 4.5). The second variant, p.H1382R, was seen in heterozygosity in patients with myeloid malignancy (3/1221 patients with AML and 1/286 patients with CMML; table 4.5). It will be interesting to follow the future of TET2 deficiency to link the association in cell type versus malignancy.

Table 4.5: Instances of TET2 p.H1382R and p.Q1632* identified in a cohort of 4324 patients with suspected or confirmed myeloid malignancy, investigated as described by Cargo et al. [94]. The cohort included 1221 cases of acute myeloid leukaemia (AML) and 285 cases of chronic myelomonocytic leukaemia (CMML). VAF, variant allele frequency; LOH, loss of heterozygosity.

Patient	Age	Position	cDNA	Protein	VAF	Depth	Diagnosis
1	48	4:106190867	4145A>G	H1382R	0.426	2947	AML
2	80	4:106190867	4145A>G	H1382R	0.951	2817	AML
3	59	4:106190867	4145A>G	H1382R	0.946	2605	AML
4	83	4:106190867	4145A>G	H1382R	0.201	3742	CMML
5	65	4:106196561	4894C>T	Q1632*	0.4	1352	CMML
6	71	4:106196561	4894C>T	Q1632*	0.436	2090	AML
7	81	4:106196561	4894C>T	Q1632*	0.943	1550	CMML

8	70	4:106196561	4894C>T	Q1632*	0.879	2062	AML

4.4.7 Known *TET2* variants in childhood malignancy

Data from the St. Jude Children's Research Hospital was downloaded and analysed to query known causes of TET2-dependent malignancy in children (https://pecan.stjude.cloud/). This data consisted of 4,469 samples with 55,874 variants. Variants in *TET2* were found in two of the largest pediatric studies; The Pediatric Cancer Genome Project (PCGP 2,050 subjects) and Therapeutically applicable research to generate effective treatments (TARGET 1,719 subjects). Twenty-two of 32 variants were predicted as damaging and subjects had a diagnosis of 9 cancer subtypes. The variants identified are shown in the summary of this data in **Figure 4.10**. Variant R1383S (which may have a similar effects as H1382R) and other non-sense variants (similar to the effect of Q1632*) are present in this cohort, but occurring only in a heterozygous state.

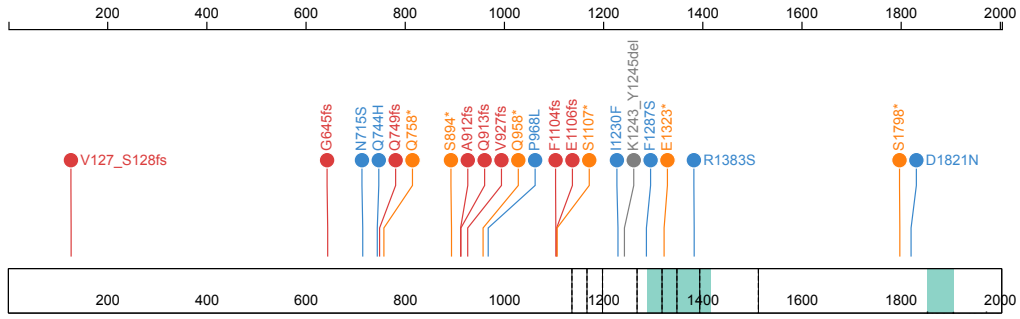


Figure 4.10: **St Jude data** The dataset used in this search for TET2-dependent determinants cancer used of 4,469 samples with 55,874 variants. TET2 genomic region was selected and damaging germline and somatic variants were found. Only heterozygous variants were found in childhood cancer cases. The protein oxygenase domain regions are highlighter in turquoise colour, which contain the substrate and DNA binding regions and 2-oxoglutarate binding regions, as shown in more detail on figure 4.12. Variants are colour coded as red; frameshift, blue; missense, orange;stop, grey; deletion.

4.4.8 Known *TET2* variants in a similar genetic background

To determine if this mutation is shared in individuals from a similar genetic population, the Born in Bradford (BiB) study data was acquired and analysed for variants in *TET2*. This

dataset has been described in detail by Narasimhan et al. [95]. Briefly, the dataset consists of exome sequencing for 3,222 British adults of Pakistani heritage with high parental relatedness. This dataset has been used previously in discovery of 1,111 rare-variant homozygous genotypes with predicted loss-of-function in 781 genes. TET2 p.Q1632* was not present. The variant identified in family 2, p.H1382R, was also not present in the BiB dataset. We suspect that both mutations arose privately in small, non-overlapping ancestral communities. Figure 4.11 shows the coding variants in the BiB dataset for *TET2*.

To confirm the absence of both variants presents in family 1 and 2 was sufficient. The probands could be mapped on a genetic ancestry plot against the BiB dataset. This test would confirm if their ancestral genetics happens to overlay those of BiB, and therefore confirm the novelty of the pathogenic variants to a very small group. This could be easily illustrated by a principal component analysis (PCA) plotting the genetics of BiB and the 1000 Genomes public dataset of mixed ancestry. However, the information from test would not provide any exceptionally positive outcome and therefore I did not consider it ethically valuable to perform.

4.4.9 Known *TET2* variants in the general population

To investigate whether or not the first identified variant from family 1 was a rare variant, the Genome Aggregation Database (gnomAD) of 138,632 exomes and genomes was queried. This mutation was not reported.

To assess the frequency of variants in *TET2*, gnomAD allele frequencies were used to map the highly-conserved coding regions in humans [96]. The frequency of conserve coding regions in *TET2* within the background population is illustrated in **Figure 4.12**. This figures presents the overall conservation and variant load in the general population. The top line in blue indicates conserved amino acids in humans by a density in blue lines, scaled across cDNA positions. Clusters of conserved amino acids are seen especially in the C-terminal domain which contains the protein catalytic domain. Notable features of conservation occur around major functional domains; specific amino acid residues with

known functions in these domains include sites for interaction with DNA (p.1290–1303), 2-oxoglutarate binding region (p.1896-1898), and part of the substrate binding domain (p.1902–1904). Clusters are also seen around selected amino modification sites; asymmetric dimethylarginine and phosphoserines. No definite conservation occurs around areas of compositional bias such as Pro-rich or Gln-rich. In particular, four large conserved regions of similar size are evident; p.20-136, p.1071-1247, p.1265-1512, p.1550-1745, p.1771-1959.

The middle line separates variants by their effect; deletions (red), frameshifts (green), nonsense (blue), splice variants (pink). The frequency of alternative variants occurring on the same amino acid residue are illustrated by colour intensity. At the bottom, a combined illustration of all four pathogenic variant types (potential damaging) are shown. Comparing the top and bottom bars, regions of high conservation versus damaging mutations are seen. Functional domains are annotated. The coding gene regions that are highly conserved and also found to be damaged in cancer are marked with black horizontal bars.

4.4.10 Mutations in Cancer-related genes

While the following section did not uncover a significant result, it is useful to include in light of the patients' development of lymphoma. All rare variants found by exome sequencing were compared to the database of Catalogue of Somatic Mutations in Cancer (COSMIC), the most comprehensive database of cancer-related variation. This was performed to (i) identify any potential secondary germline variants (that are known to contribute to disease in the somatic state) and (ii) potentially identify any known pathogenic variants that explain features seen in the patient more accurately than that of TET2 LoF. Association investigation is commonly done via GWAS; identifying a small number of disease-associated SNPs from a large database in a single patient would not provide any statistically-relevant result. The patient in this case had a strong immune phenotype, a feature that is usually due to only one or a small number of damaging variants. However, it is useful to visualise the genome-wide spread of disease-associated genes and to investigate the genomic mutation load compared to controls.

Figure 4.13 shows (A) firstly the difficulty in identifying "cancer-related genes" that are not from a strict statistical definition of association; the majority of functional variants occurred in genes that can be labelled as lymphoma-related. (B) Very few genes harboured more than one rare functional variant. (C-E) Variants were unequally distributed and particularity chromosome 5, 8, and 12 carry the highest burden of "cancer-association". While (A-E) provides little empirical data, (F) the comparison of germline mutational load compared to controls (12 unrelated technical controls from the same library preparation and sequencing batch) shows that P3 had a significantly high frequency of functional rare variants. Runs of homozygosity were also found in chromosomes 12, 17, and 14. Consanguinuity is potentially responsible for this occurrence. Confirming this within the community is potentially valuable. However, this was not done since additional ethical approval is desirable (discussed in subsection 4.4.8).

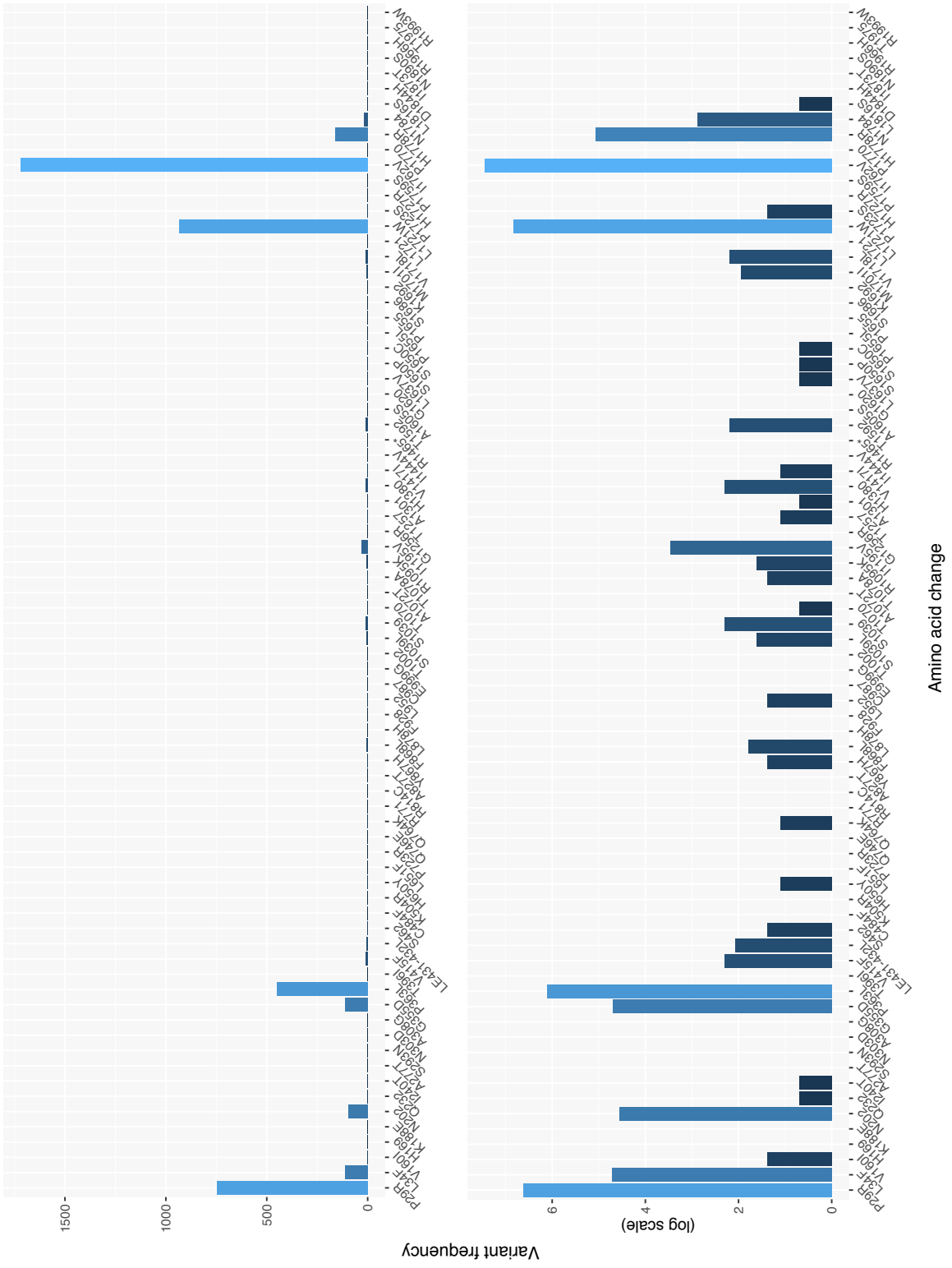


Figure 4.11: **BiB cohort summary statistics for *TET2*.**Variant frequency within the cohort is shown as bars and by increasing colour lightness. For scale, the **top** shows the variant frequency on a normal y-axis, while the **bottom** frame uses a log scale to view the frequency of rare variants clearly.

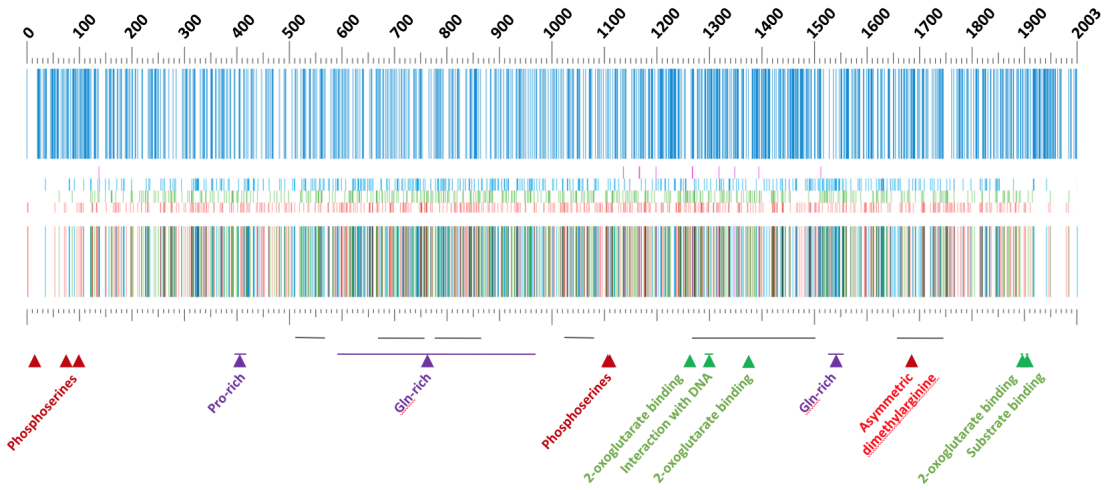


Figure 4.12: **Variant frequency and consequence in population genetics.** [Top] Residues that are conserved in humans are shown by blue density with equal thickness for each residue. [Middle] Potential LoF variants found in gnomAD (deletions; red, frameshifts; green, nonsense; blue, splice variants; pink). [Bottom] All variants present in gnomAD overlayed. Black horizontal bars are conserved and, based on cancer genetics databases, potentially important for cancer suppression.

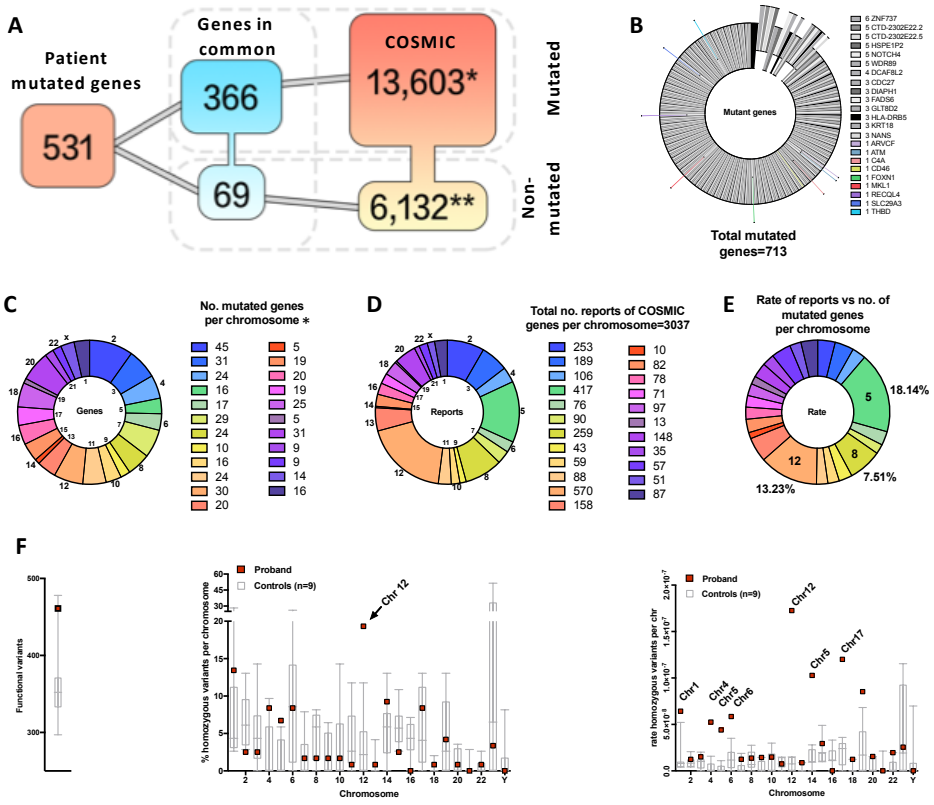


Figure 4.13: **Known genetic foctors in lymphoma.** A. The proband has 531 functional variants found by exome sequencing. Genes on COSMIC database that have been implicated for lymphoma were compared to all gene variants for P3. Since COSMIC is extremely comprehensive approximate 68% of genes are linked with disease in some way. 366 variants occur in lymphoma related genes on COSMIC; only 69 genes are mutated in the proband which are not reported as associated with lymphoma. B. Only 14 genes contain multiple variants. C. Number of genes that are mutated per chromosome in proband. D. The frequency at which these are reported in lymphoma via COSMIC. E. The rate of reports plus the number of genes mutated in proband; chr 5, 12, and 8 share the most. F. The proband has far higher functional variants than average (despite normal sequence quality). Chr 12 has the highest % of homozygous variants. Regions of homozygosity occur particularly on chromosomes 12, 17, and 14.

4.4.11 Acquired somatic mutations in genes within the RAS signalling pathway in patients' lymphoma tissue

Somatic mutations of TET2 are prevalent within many types of malignancy, including myeloid and lymphoid neoplasms, where they are believed to represent an initiating event [97–100]; therefore, we hypothesized that lymphomagenesis would require a "second hit". To identify putative cooperating mutations, we performed high depth WES of patients' lymphoid tumour samples and made a pairwise comparison with germline data, confirming hits by Sanger sequencing. In the EBV-positive Hodgkin-like polymorphic B-cell lymphoproliferative disorder of P1, we detected a single point mutation p.K117N in KRAS, an oncogene that frequently harbors somatic variants in various solid tumour types, as well as haematological malignancies (table 4.6). Our observation supports the previously suggested collaboration of TET2 loss-of-function and KRAS gain-of-function mutations, reported in myeloid neoplasia [81, 101, 102]. Another somatic variant, again within the RAS signalling pathway, was found in the peripheral T-cell lymphoma of patient P2, affecting the gene ERBIN (Erbb2 Interacting Protein). ERBIN acts within the RAS signalling pathway by disrupting RAS-RAF interaction [103]. p.R1194H has not previously been reported in the context of neoplasia and is rare in the population (table 4.6). The acquisition of somatic variants could not be assayed in the primary mediastinal large B-cell lymphoma of patient P3, due to lack of material.

Table 4.6: **Somatic missense mutations in RAS signalling pathway-related genes**, was performed by Dr Mikulasova. KRAS (P1) and ERBIN (P2), identified by pairwise comparison of lymphoma tissue and germline high depth WES for the EBV-positive Hodgkin-like polymorphic B-cell lymphoproliferative disorder of patient P1 and the peripheral T-cell lymphoma of patient P2, respectively. These somatic variants were confirmed by Sanger sequencing in Newcastle by Prof Hambleton's group. Next generation sequencing allele fractions (NGS AF) correspond to the approximate proportion of tumor cells present in lymphoma tissue sample. Due to lack of sufficient lymphoma tissue DNA from patient P3, we were unable to perform a WES analysis on lymphoma sample.

	KRAS c.351A>T, p.K117N	ERBIN c.3581G>A, p.R1194H
NGS %cells:	22%*	46%*
NGS AF:	11%	23%
Sanger:	17%	40%

100G:	0	0
Gnomad:	3.98E-06	1.20E-05
ExAc:	0	2 alleles
Variant ID	rs770248150	rs760950077
CADD:	21.2	15.92
PolyPhen:	0.998 probably damaging	0 benign
SIFT:	0.011 damaging	0.251 tolerated
PROVEAN:	-4.56 deleterious No homozygous	1.8 neutral No homozygous
Mutation Taster:	disease-causing	polymorphism
COSM ID	COSM6854421 COSM4696721	
Associated phenotypes	Colorectal neoplasms Hepatocellular carcinoma Malignant melanoma Multiple myeloma Carcinoma of oesophagus Adenocarcinoma of stomach	No associated phenotypes

4.4.12 Interactions in damaged protein pathways

We were interested to see if there was any obvious polygenic effect of germline rare variants in affected patients. One way to do this is by selecting all genes that harbour any functional (potentially damaging) variant and annotating that geneset with a database of functional protein association. STRING database is a high quality source of known and predicted protein-protein interactions. Applying the same database to any query removes our interpretive bias. Although, the database itself will be enriched for inherently

biased protein function information since highly studied proteins will have the strongest evidence for associations. Therefore, it is mostly useful to tell us about interactions that we can readily interpret. This is in contrast to a geneset enrichment or burden test that would instead tell us about any statistical gene enrichment. The latter would be useless in a cohort of N=3 patients where we have no statistical power. Therefore, producing interpretable protein pathway information (from STRING db) is a justifiable exploration. Chapter 5 includes a novel statistical method for network-based analysis in section 5.5.4 for larger cohorts.

The known pathway interactions between genes which contain functional variants are shown for each patient separately first and then combined. **Figure 4.14** shows the PPI for all genes with functional variants for family 2 patient 1. The same method is used again for family 2 patient 2 (**Figure 4.15**), and family 1 patient 1 (**Figure 4.16**). Lastly, the combined data shows the PPI for variant genes from all 3 patients **Figure 4.17**.

Although no significantly enriched pathways were found to be common among all patients, this may be considered a useful negative result. Since this is likely the first report of germline homozygous TET2 deficiency, the effect size should be considered. From this analysis, we see no evidence of other contributing germline variants, nor did we find any other candidate variants through routine best-practice exome analysis of individual variants. This network analysis could potentially be repeated in the future on a large cohort of unrelated individuals with a shared phenotype to uncover a significant pathway enrichment.

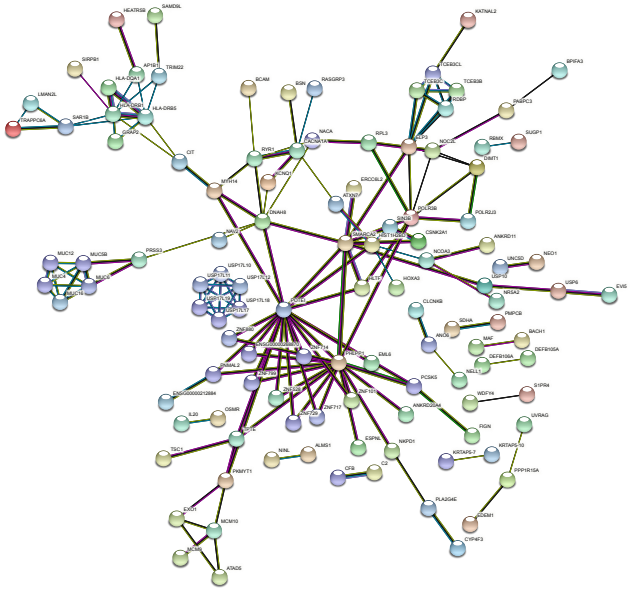


Figure 4.14: **PPI for proteins with gene variants in F1 P1.** The STRING database was used to query all for all genes harbouring a functional coding variant. Known protein-protein interactions are indicated by weighted lines connecting each protein. Protein pathways group into tight clusters.

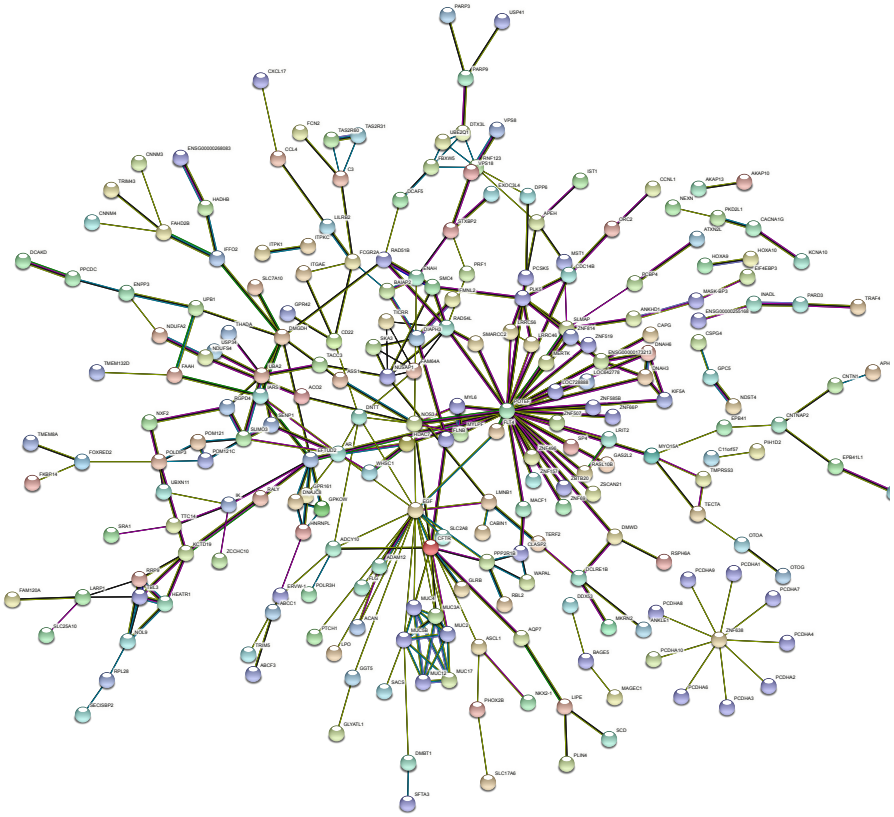


Figure 4.15: **PPI for proteins with gene variants in F1 P2.** The STRING database was used to query all for all genes harbouring a functional coding variant.

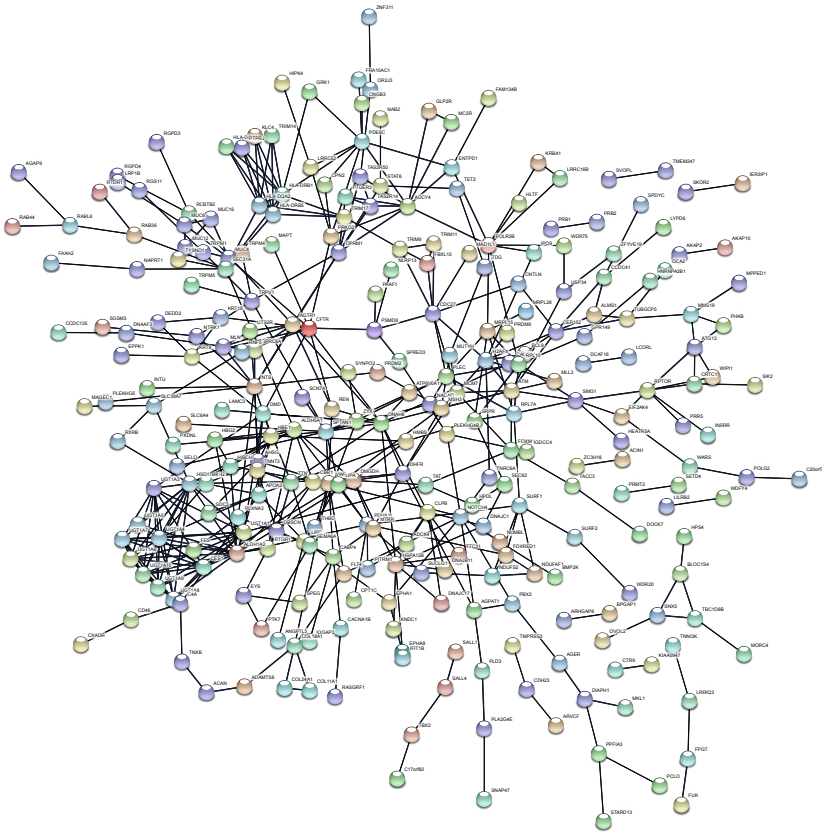


Figure 4.16: **PPI for proteins with gene variants in F2 P3.** The STRING database was used to query all for all genes harbouring a functional coding variant.

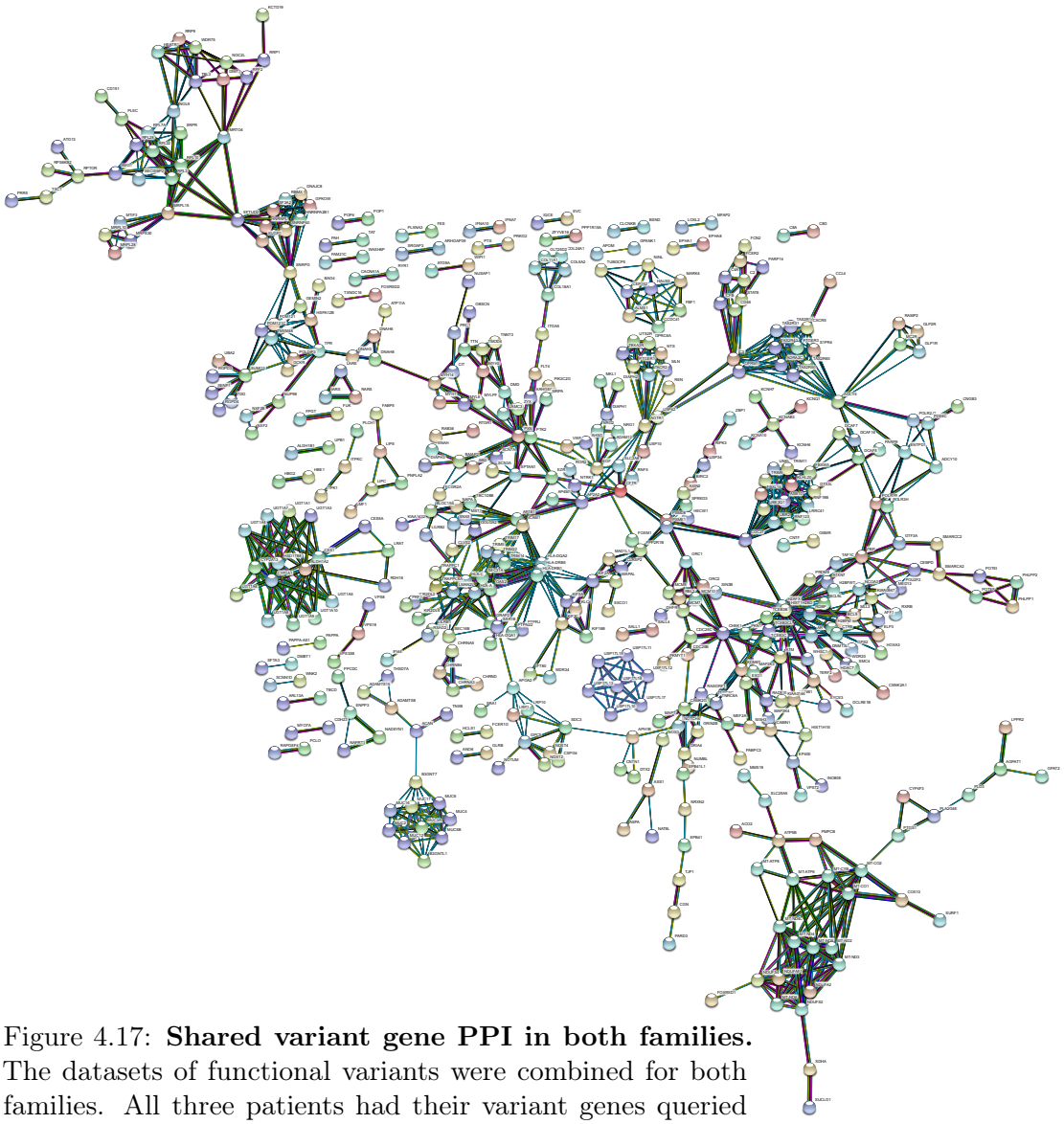


Figure 4.17: **Shared variant gene PPI in both families.** The datasets of functional variants were combined for both families. All three patients had their variant genes queried to identify potential shared protein pathways including or in addition to the TET2-dependent methylation pathway.

4.4.13 Mutant protein expression

To quantify protein expression in proband and first degree relatives, western blotting was performed on family 1 in Newcastle and on family 2 in Leeds. In family 2, the stop variant carried by the proband was expected to result in either a truncated protein or be lost through non-sense mediated decay. Several experiments were performed to confirm that no truncated protein was produced. Two independent experiments show the expression of protein in a healthy control, haploinsufficiency for all heterozygous relatives and complete loss of protein in the proband who carries a homozygous variant (**Figure 4.18**).

Figure 4.19 presents this data again along side that of family 1 from Newcastle, where the expression of TET2 H1382R protein was not impaired relative to TET2 wt in primary cells. It was also not impaired in a recombinant system (subsection 4.4.14).

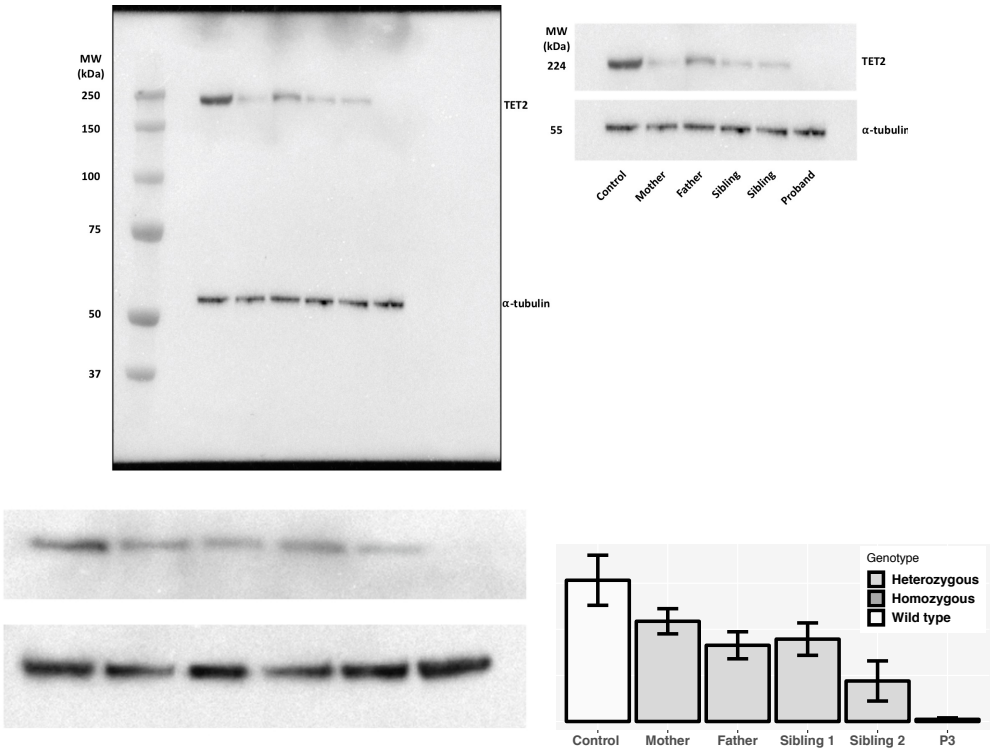


Figure 4.18: **TET2 protein expression in PBMC.** Western blot Protein purified from frozen PBMCs using sodium orthanovanadate, Complete Protease Inhibitor Cocktail and PMSF (Sigma), with RIPA buffer. The Novex Mini Gel Tank and blot module, Bolt 4–12% Bis–Tris Plus Gels, and PVDF Transfer Membrane (Thermo Fisher Scientific). Imaging used Super Signal West Femto/Pico (Thermo Fisher Scientific).

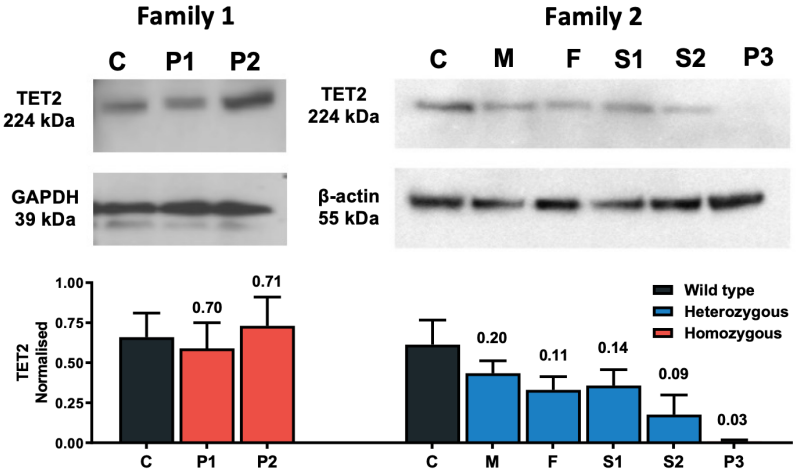


Figure 4.19: **TET2 protein expression in PBMC.** For context, a single Westernblot image and data from figure 4.18 (family 2) are duplicated for comparison with family 1 (completed in Newcastle by Prof Hambleton's group).

4.4.14 Enzymatic activity immunofluorescence

While family 2 carried a loss of protein expression, family 1 carried a variant that was expected to result in an expressed protein that is catalytically dead. The catalytically important Fe(II) binding motif has been shown on the protein structure [93]. Prof. Hambleton's group carried out the following work to compare the enzymatic activity of TET2wt and TET2H1382R by immunofluorescence microscopy analysis of transfected HEK293T cells stained for 5-hmC [75]. This revealed intense staining in TET2wt cells, but no increase of 5-hmC signal in cells expressing TET2H1382R, thus providing evidence agreeing with the predicted loss of its 5-hydroxymethylating enzymatic activity (**Figure 4.20**).

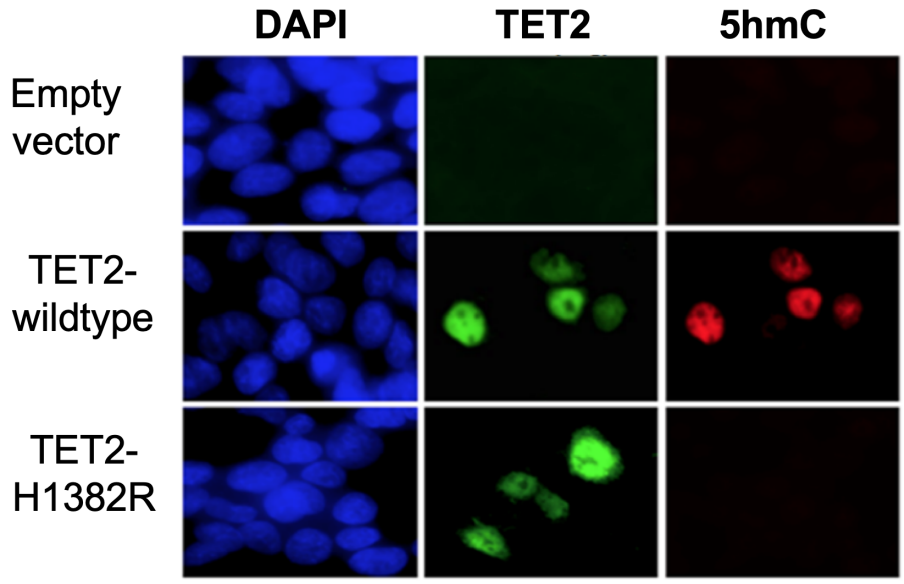


Figure 4.20: **Immunofluorescence showing impaired TET2 hydroxymethylating activity.** (Completed in Newcastle by Prof Hambleton's group). 5hmC immunofluorescence staining in HEK293T cells transfected with either empty lentiviral vector, Flag-tagged wild type TET2 or mutant TET2H1382R. Blue: DAPI stain, green: Flag, red: 5hmC staining. Result is representative of three independent experiments.

4.4.15 Effect of loss of TET2 function on total blood DNA methylation

The relative quantification of methylation status was carried out as demonstrated in section chapter 3. **Figure 4.21**. The methylation profile ratios are shown in Figure 4.22, while the methylation profile curves are shown in **Figures 4.23** and **4.24**. While the method used here is limited to viewing global methylation, we see a dosage effect that may not be seen with more sophisticated methods such as methylation sequencing. "Healthy controls" who also carried *TET2* coding variants were found in our sequence database of several hundred unrelated individuals. The first three of these variants are considered benign and we see no effect on methylation status for these individuals. The fourth one of which carried the same variant as family 1 (heterozygous) and shows a methylation status equivalent to the other heterozygous carriers of this damaging variant. *Note*; this variant has been reported in the heterozygous state previously it but does not appear in GnomAD is should be considered rare. Therefore, this "healthy control" has been flagged to confirm that mix-up has not occurred with their DNA sample or sample ID. Methylation blotting was also attempted but ultimately unsuccessful due to the limited material available. We expect that methylation blotting would provide the same results observed in **Figure 4.21**, although lacking the finer details seen in **Figures 4.23** and **4.24**.

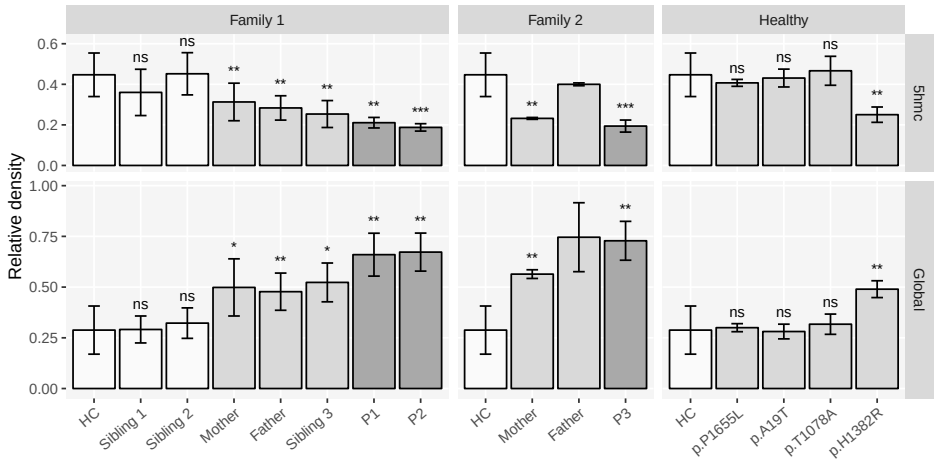


Figure 4.21: **Effect of loss of TET2 function on total blood DNA methylation status.** Bar plots of global 5mC and 5hmC methylation in patients showed increased 5mC levels and decreased 5hmC, with intermediate values in their relatives who were heterozygous carriers. "Healthy controls" were unrelated individuals in our sequence database who also carried *TET2* coding variants, one of which also carried (heterozygous) the same variants as seen family 1. Data shown are mean±SD from 2 independent experiments and seven healthy controls. P-values are shown for unpaired t-tests compared to healthy controls.

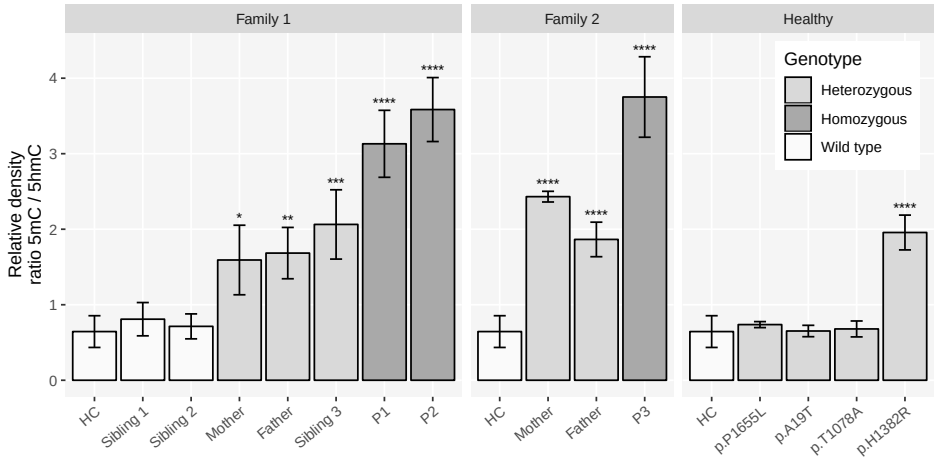


Figure 4.22: **Consequences of TET2 loss-of-function on DNA methylation.** Increased ratio of 5mC to 5hmC, as determined by DNA methylation assay of total blood DNA, in patients bearing homozygous H1382R and Q1632* mutations compared to homozygous wild type controls or siblings. Heterozygous relatives showed significantly increased, intermediate levels. Data shown as mean ± SD from 2 independent experiments and seven healthy controls. P-values are shown for unpaired t-tests compared to healthy controls.

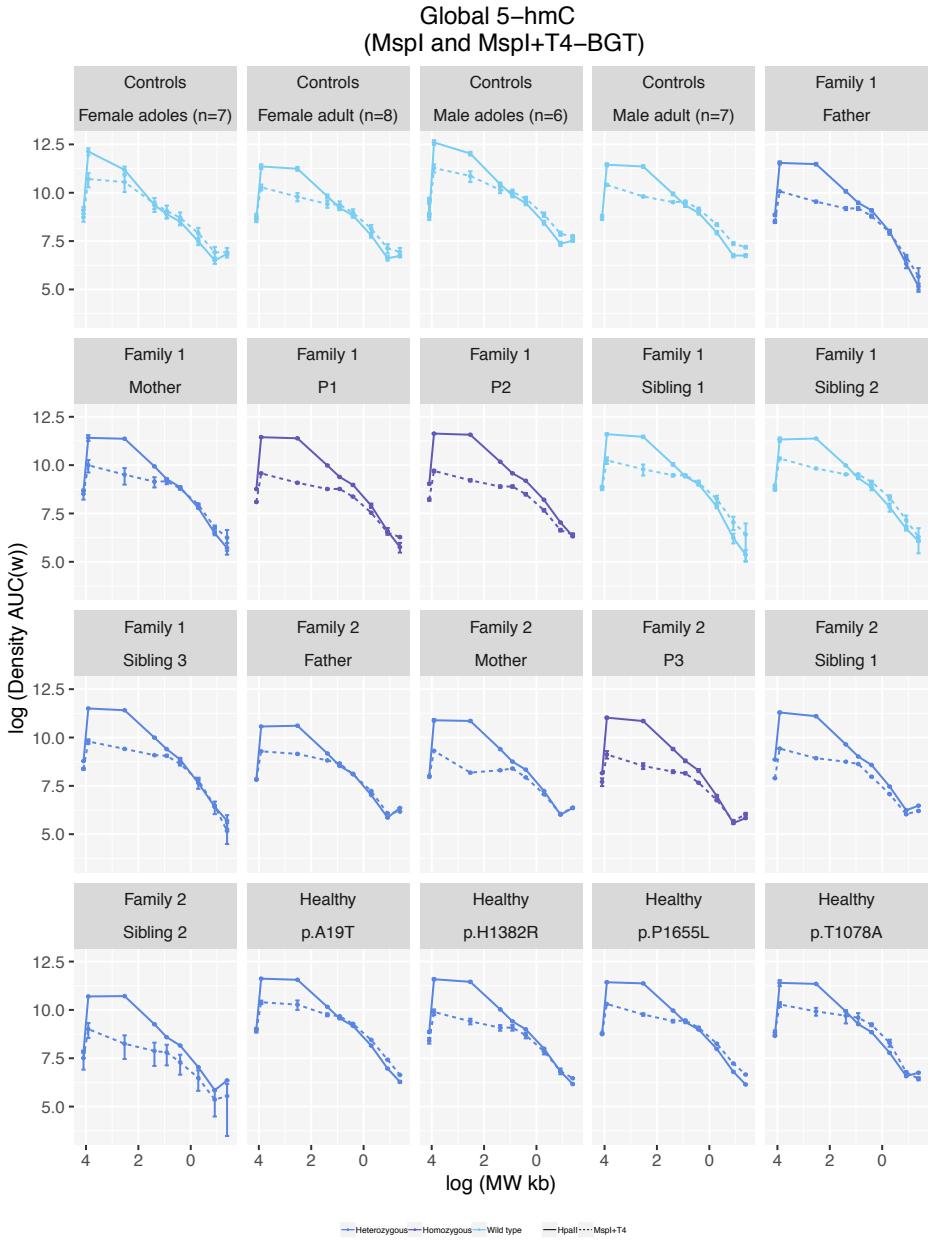


Figure 4.23: **Methylation profile curves.** Representative curves of global 5mC and 5hmC methylation in healthy controls, patients and their relatives. Density AUC(w) after MspI and HpaII digestion and T4-BGT pre-treatment for production of 5-ghmC.

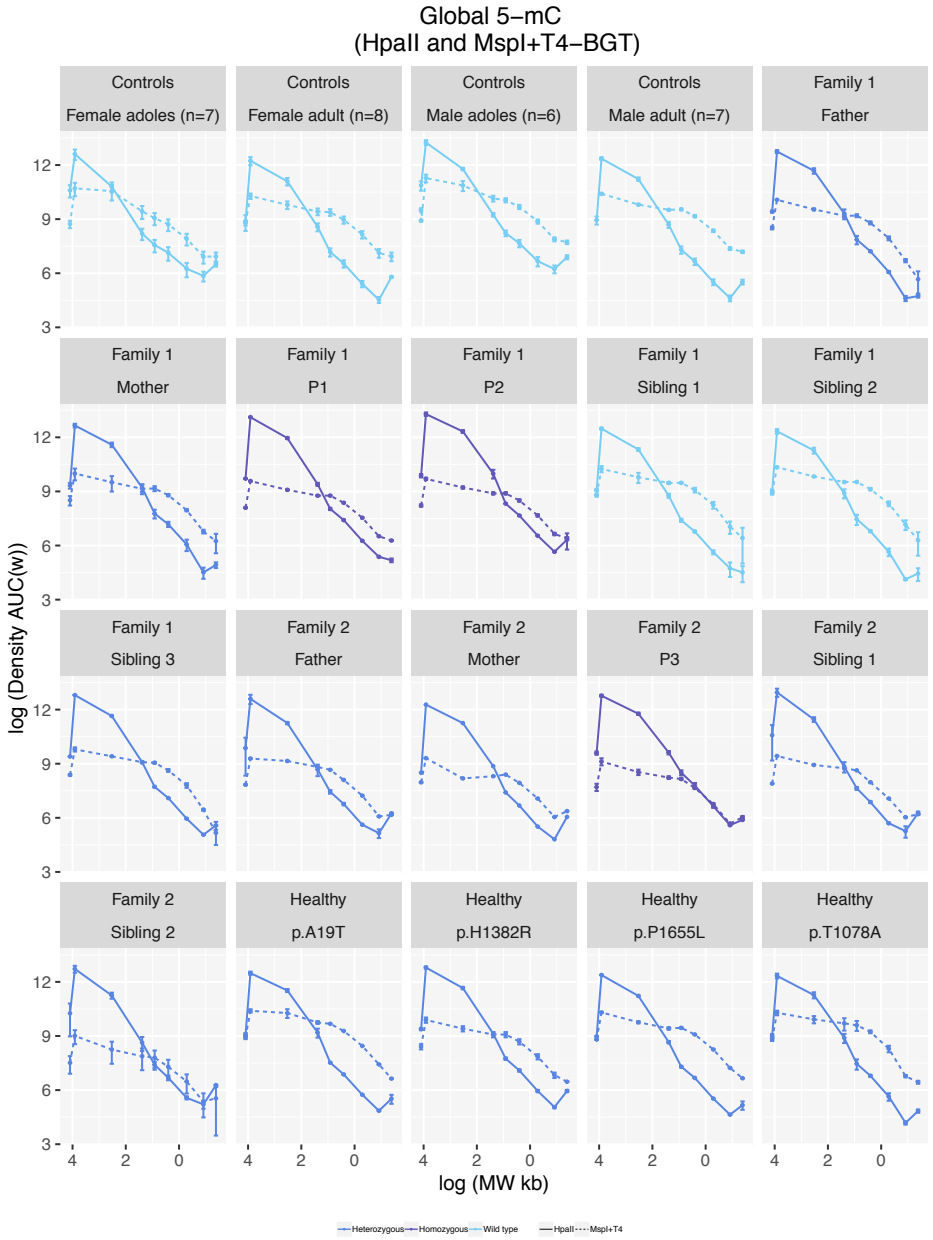


Figure 4.24: **Methylation profile curves continued.**

4.4.16 Effect of TET2-deficiency on T-cell homeostasis

The timeline of patient histories (**Figure 4.4-4.5**) shows that all three patients had features of ALPS. These features were variable combinations of hepatosplenomegaly and lymphadenopathy which progressed to lymphoma/lymphoproliferative disorder (n=3), raised proportion of DNT-cells (n=3), and clinically significant autoimmunity (n=2). In P1 and P2 this suggestive clinical picture was accompanied by clearly impaired T-lymphoblast apoptosis in vitro (**Figure 4.25**). It was also accompanied together with serologic markers of ALPS, such as elevated levels of sCD25, FasL and IL-10. P3 did not show serum markers of ALPS or disordered apoptosis in vitro, contrary to expectation since she had elevated DNTs. Known ALPS disease genes were screened for relevant mutations but none were found in any patient and Fas expression was normal (methods for screening a list of genes based on coordinates can be read in **chapter 5**). The amount of available patient material was limited and therefore detailed studies of T-cell function and transcriptome in vitro could not be done. However, all three patients had impaired responses to human herpesviruses, which highly suggests an impaired T-cell immunity in vivo. We also observed that TET2-deficient T-cells post-HSCT behaved with a strong pro-proliferative phenotype. Lastly, despite receiving a T cell-depleting conditioning regimen containing full dose serotherapy all three patients showed autologous T-lymphoid reconstitution in less than two months after HSCT (**Figure 4.4-4.5** and **Figure 4.26**).

4.4.17 TET2 deficiency impairs human B-cell terminal differentiation

Patients 1-3 received exogenous immunosuppression therapy which obscures the detection of abnormalities of humoral immunity. However, each showed clear evidence of a deficiency in class-switched memory B-cells; a primary defect of humoral immunity. Healthy B-cells execute an appropriate differentiation programme in response to antigenic stimulation. We hypothesized that TET2-deficiency inherently impairs this ability for patients (one the questions posed in subsection 4.1.6 based on the literature to date).

To test this, the research group of Dr Doody provided their in vitro culture system with a T cell-dependent stimulus to enable the generation of long-lived plasma cells

from primary B-cells as described in detail by **Cocco M, Stephenson S, Care MA, et al. In vitro generation of long-lived human plasma cells. J Immunol 2012;189:5773-85.** A schematic of this B cell differentiation assay is shown illustrated in **Figure 4.27** (A). For TET2Q1632* B-cells, the appearance of short-lived plasmablasts at day 6 indicated that they were capable of initiating plasma cell differentiation. However, these cells failed to progress to phenotypically mature plasma cells, which emerge at day 13 in healthy donors and persist during the time-frame of the assay (**Figure 4.27**(B)). The same cultures showed high levels of IgM from TET2Q1632* plasmablasts. This declined as the cells died. There was also a complete failure to generate IgG (**Figure 4.27** (C)), consistent with the murine model of TET2 deficiency [100].

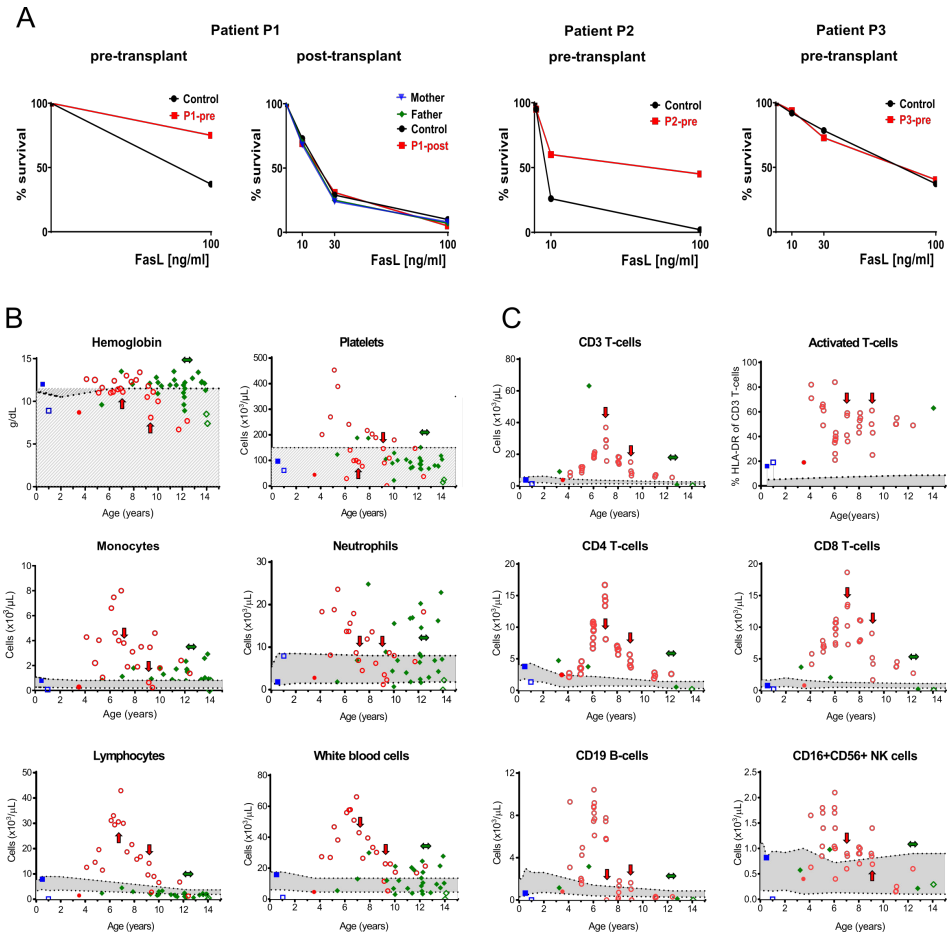


Figure 4.25: **Fas ligand-mediated apoptosis and peripheral blood cell counts in patients before and after haematopoietic stem cell transplantation.** Fas ligand-mediated apoptosis was assayed by Frederic Rieux-Laucat and Anne Rensing-Ehl. (A) Fas Ligand-induced apoptosis in patients' and healthy controls' PHA and IL-2 stimulated T-blasts determined by flow cytometry using Annexin-V/PI staining showed impaired apoptosis in patient P1 and P2 before transplantation, normal response of patient P3 before transplantation, and repaired response of patient P1 after transplantation compared to healthy control cells. (B) Hemoglobin and total blood cell counts, and (C) absolute numbers of lymphocyte subsets, B-cells, NK-cells and percentages of activated T-cells in peripheral blood of all patients. Hatched area: sub-normal range; grey area: normal range. Red arrow: treatment with Rituximab in patient P1 (age 9 years), green double-ended arrow: R-CHOP in patient P3 (at age 12-12.5 years), red circles: P1; blue squares: P2; green diamonds: P3. Filled symbols: pre-transplantation; open symbols: post-transplantation.

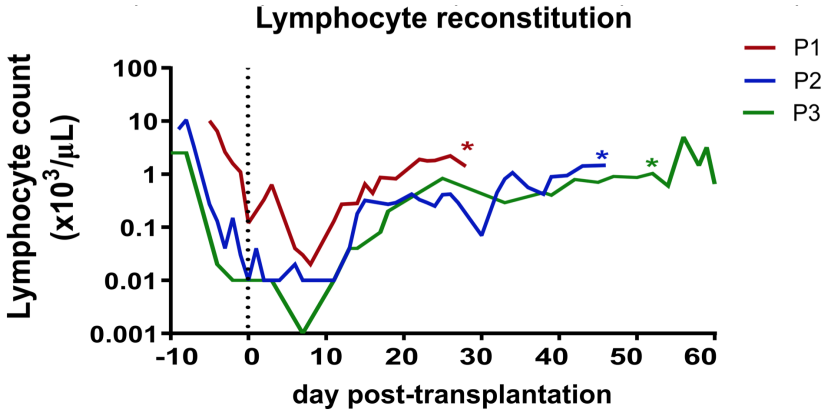


Figure 4.26: **Autologous lymphocyte reconstitution post-HSCT.** Data compiled from all three patients in Newcastle and Leeds. Figure compiled by Dr Spegarova. Rapid autologous lymphocyte reconstitution after HSCT in patients with homozygous TET2 loss-of-function, despite full T-cell depleting serotherapy. Asterisks indicate the first measurement of T cell chimerism in each patient: P1 (D+28) 78% recipient, P2 (D+46) 100% recipient, P3 (D+52) 91% recipient.

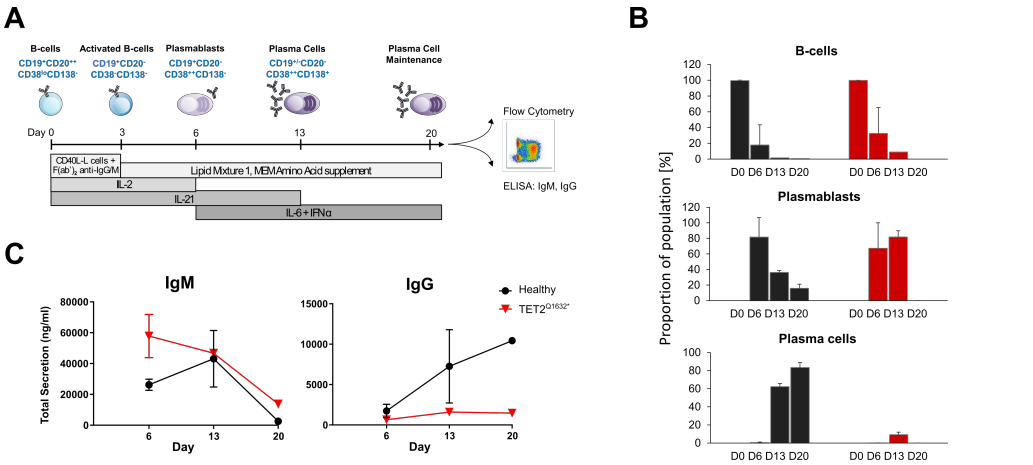


Figure 4.27: **Failure of TET2-deficient B-cells to generate mature plasma cells and produce IgG in vitro.** This assay has been carried out by the group of Dr Doody. (A) Illustration showing in vitro B-cell differentiation methodology. T-cell dependent immune stimulus is mimicked and cytokine stimulation applied. (B) Flow cytometry was performed for patient P3 samples and healthy control from in vitro differentiated primary B-cells. A defect in B-cell maturation is seen with impaired cell survival. (C) Secreted IgM and IgG quantified by ELISA during B-cell differentiation. A block in class-switch recombination is evident for patient cells. The data are presented as mean ± SD from two independent experiments.

4.4.18 TET2-deficiency skews in vitro haematopoietic differentiation towards the myeloid lineage

The effect of TET2 loss-of-function on haematopoiesis has been investigated by Dr Spegarova, recounted here. An in vitro disease model was produced using patient-derived induced pluripotent stem cells (iPSC). Primary fibroblasts from patients P1 and P2 and healthy volunteers were reprogrammed into iPSC that were fully characterized as shown in **Figure 4.28**, and differentiated into haematopoietic precursors as described by Olivier et al. [104] (**Figure 4.29**). TET2-deficient cultures had a higher proportion of erythro-megakaryocytic progenitors, and persistently lower fraction of myeloid progenitors, as detected by flow cytometry (**Figure 4.29 B**).

While the low fraction of myeloid progenitors appears to contradict the ultimate conclusion of this section, we see several features that indicate a skew towards the myeloid lineage. A colony forming unit assay revealed a skewed and boosted clonogenic potential of TET2 H1382R haematopoietic progenitors towards the myeloid lineage, at the expense of clonogenically impaired erythroid and megakaryocytic lineages (**Figure 4.29 C, D**). This observation is in agreement with a previous study showing hyper-proliferation and impaired differentiation of TET2-deficient erythroid cells in mice [105]. This could be correlated with in vivo findings of marked monocytosis and variable neutrophilia in the face of chronic non-immune thrombocytopenia in patients P1 and P3 (**Figure 4.25**). Moreover, the surviving patient P1 has become transfusion-dependent over time, albeit we cannot rule out a late effect of chemotherapy on his marrow reserve.

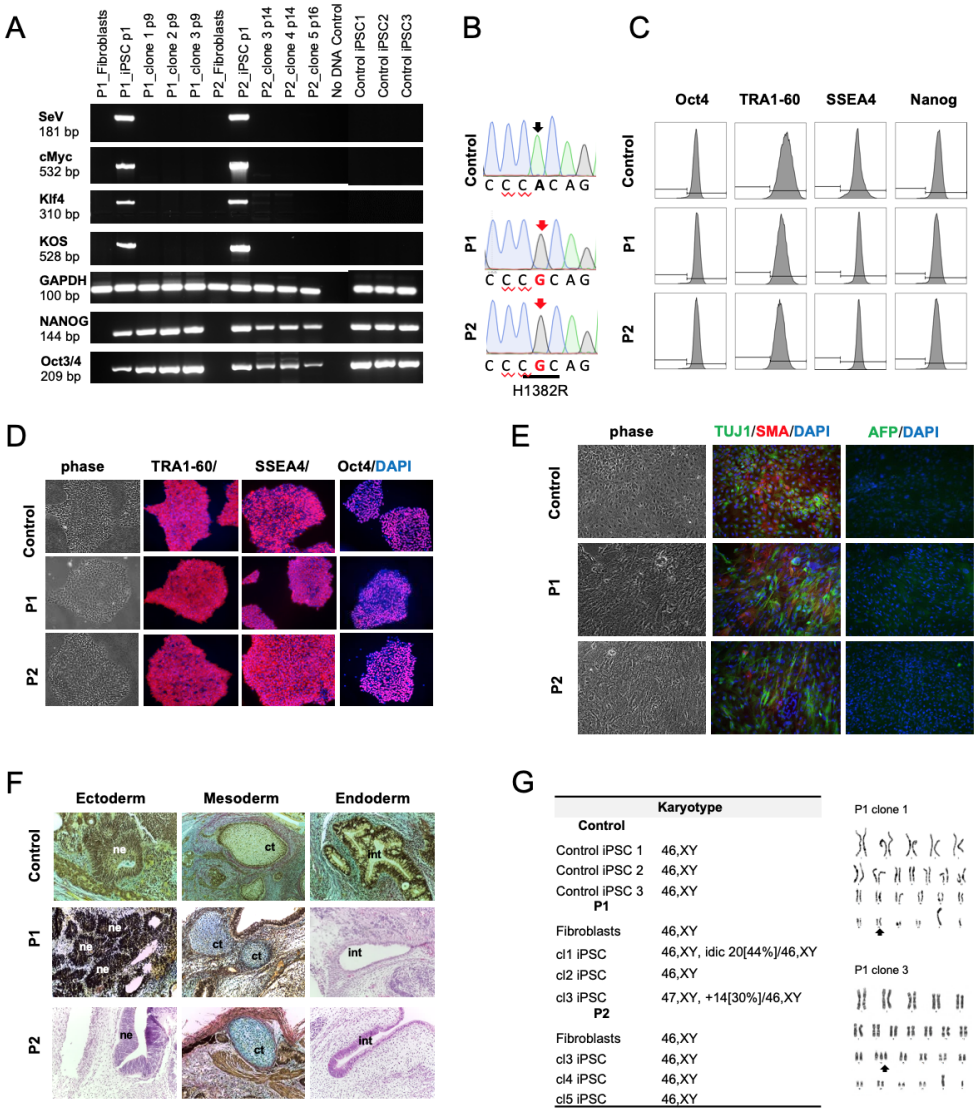


Figure 4.28: **Characterisation of derived iPSC from patient P1 and P2 and healthy individuals.** This assay has been carried out by Dr Spegarova. A) Clearance of Sendai virus vectors and expression of pluripotency markers Nanog and Oct3/4 detected by PCR in TET2-deficient patient-derived iPSC P1 and P2 and three healthy control iPSC lines in various passage number p1, p9, p14 or p16. B) Sanger sequencing result of TET2 gene in patients and non-affected iPSC confirming TET2H1382R mutation. C) Representative pictures of expression of pluripotency markers TRA1-60, SSEA4, Oct4 and Nanog detected by flow cytometry and D) by immunofluorescence in iPSC lines. Blue: DAPI, Red: pluripotency markers. E) Differentiation of iPSC into 3-germ layers in vitro and F) in vivo. Ectoderm - beta-III tubulin (TUJ1, green), Mesoderm - smooth muscle actin (SMA, red) and Endoderm - alpha-fetoprotein (AFP, red). Ne - neuroepithelium, ct - cartilage, int - intestinal epithelium. Blue: DAPI. G) Karyotyping of fibroblasts and iPSC lines, number of positive metaphases in brackets. Isodicentric chromosome 20 in 44% of P1 clone1 cells, and trisomy of chromosome 14 in 30% of P1 clone 3 cells, highlighted by arrow.

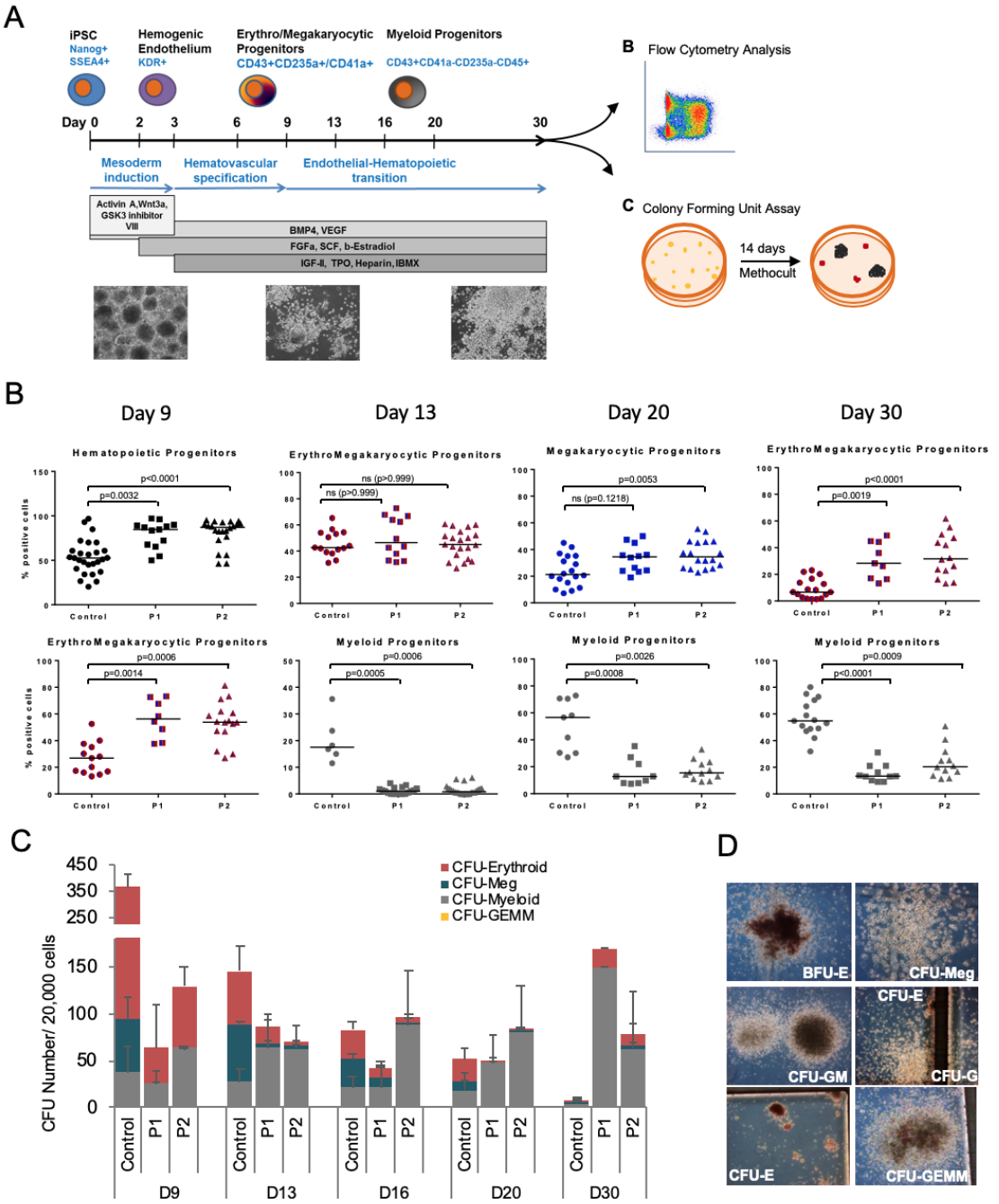


Figure 4.29: **Impaired in vitro haematopoietic differentiation by TET2-deficient iPSC.**(Continued on the following page.)

Figure 4.29: **Impaired in vitro haematopoietic differentiation by TET2-deficient iPSC.** This assay has been carried out by Dr Spegarova. A) Schematic presentation of experimental strategy to assess the haematopoietic differentiation capacity from iPSC in vitro24 with representative pictures of cell culture at major differentiation stages; starting with embryoid bodies at Day 0-3, followed by presence of haematopoietic precursors and their budding at Day 5-12, culminating in proliferation and maturation at Day 13-30. B) Scatter box plots demonstrating the percentage of positive cells detected by flow cytometry at individual time points during differentiation; showing haematopoietic progenitors (CD34+/-CD43+), erythro-megakaryocytic progenitors (CD43+CD235a+CD41a+), erythroid progenitors (CD43+CD235a+CD41a-), megakaryocytic progenitors (CD43+CD235a-CD41a+) and myeloid progenitors (CD43+CD235a-CD41a-CD45+). Bar represents median value from minimum 6 independent experiments. Statistical significance was calculated using non-parametric Kruskal-Wallis test. C) Quantification and classification of individual colony forming unit (CFU) types showed skewed differentiation towards myeloid lineage at the expense of erythroid and megakaryocytic colonies. Data are presented as mean ± SD from minimum 3 independent experiments. D) Representative pictures of individual CFU classified according to characteristic morphologic features. Data from three healthy control lines are presented in one Control group. CFU types: Erythroid (E), megakaryocytic (Meg), granulocytic (G), monocytic (M), and erythroid burst forming unit (BFU-E).

4.5 Discussion

In their recent work, Fraietta et al. [106] show that while developing therapeutic CD19-targeted T cells (CAR-T anti-tumour therapy), the integration of the CD19-CAR sequence happened to interrupt the TET2 coding sequence. This fortuitous integration event had a combined effect; disrupting TET2 promoted the therapeutic efficacy of CD19-targeted T cells in the patient. Inhibiting TET2 normally would not be considered ameliorative in cancer immunology. However, in this case the pathogenic effect of TET2 deficiency allowed for more successful CAR-T cell therapy. It is likely that this type of combined genetic alteration will improve CART function. This anecdotal example demonstrates the lymphoproliferative effect of TET2 deficiency.

Lymphoma, or most cancers, can often be subtyped based on genotype associations. It is not surprising that diverse contributing factors may result in a broadly similar phenotype, which can then be divided into categories based on the founding gene mutations; this effect is generally due to umbrella definitions based on morphological features. Defining diseases by phenotype has been required (i) when genotyping methodes were not available, (ii) when treatment strategies could be applied for disease progression that follows a phenotypically common route despite having a widely different genetic causes, and (iii) when disease management is aided by having general categorisation rather than endless genotype-based definitions for each individual disease. In the case of diffuse large B-cell lymphomas, Schmitz et al. [107] report an instance when genotype subgrouping is beneficial for both treatment and prediction. Gene-expression signatures and responses to immunochemotherapy were defined based on the co-occurrence of genetic alterations. They identified different mutant-gene-dependent sources of "chronic active" B-cell receptor signalling. This report exemplifies a middle ground for disease categorisation by defining subgroups based on the signalling pathway of importance. This approach incorporates genetic findings while also implying the mechanism of disease based on the pathway. For example based on the genetics, Schmitz et al. [107] could define the subtypes of diffuse large B-cell lymphomas into four aberrant pathways: PI3 kinase pathway, BCR-dependent NF-κB activation, other NF-κB, and antiapoptotic BCL2 family. For the

individuals described in this chapter, development of lymphoma may be the most difficult challenge, both physically and psychologically. However, from a research perspective it may be productive to briefly consider this an incidental result of the aberrant immune response found in TET2. Understanding the donwstream effect of TET2 deficiency does not necessarily explain the mechanism of protein function. For example, the outcome RAG deficiency is directly related to the protein function; non-functional protein cannot bind to recombination signal sequences to allow for recombination of T cell receptor, B cell receptor, or antibody coding genes therefore resulting in a lack of their expression. However, in TET2 deficiency there is a widespread effect on methylation. The genes ultimately controlled by methylation and the levels of proteins expressed as a result can produce very different phenotypic characteristics. The discussion of TET2 deficiency can therefore be divided into that of (i) mechanism and (ii) meta-scale outcome or phenotype.

When published, this chapter may be the first report of germline homozygous TET2 loss-of-function, identified in association with combined immunodeficiency, autoimmunity and childhood lymphoma in two unrelated kindreds. As a severe autosomal recessive trait, homozygous null mutations of this gene are absent from any databases of human genomic variation that we could find. The extensive literature on human TET2-deficiency focuses exclusively on somatic variation in the context of CHIP, myeloid and lymphoid malignancies. However, Kaasinen et al. [108] report a heterozygous germline frameshift mutation resulting in adult lymphoma. At present we must consider healthy heterozygous family members as being at some increased risk of lymphoid malignancy but this is difficult to quantify. However, the families may be monitored as being potentially pre-lymphomic.

Our findings confirm the strong link between TET2 loss-of-function and lymphomagenesis, with early onset lymphoid tumors of diverse types in all three affected children. These patients' phenotype is also consistent with observations of TET2-knockout mice that are viable, fertile and develop normally, but demonstrate myeloproliferation, splenomegaly and lymphomagenesis [77, 109–111]. The TET2 defect elevates blood DNA methylation levels, especially at active enhancers and cell-type specific regulatory regions with binding sequences of master transcription factors involved in haematopoiesis, endorsing the importance of TET2 in regulation of haematopoietic differentiation [108, 112, 113].

Furthermore, 5hmC level in DNA was reduced dramatically in homozygous TET2 mutant mice compared to heterozygous ones, as likewise noticed in peripheral blood of our cohort [113, 114].

We found no impairment in the reprogramming efficiency or pluripotent potential of TET2-deficient iPSC, just as mouse embryonic stem cells (mESC) deleted for TET proteins retain pluripotency [114, 115]. Increased haematopoietic repopulating capacity with skewing of cell differentiation toward monocytic/granulocytic lineages was described in Tet2-knockout(KO) mice who died by 1 year of age because of the development of myeloid malignancies [79].

Our results from in vitro haematopoietic differentiation of TET2-deficient iPSC confirm a similar effect in humans, showing boosted clonogenic potential of myeloid progenitors at the expense of impaired erythroid and megakaryocytic progenitors. There were echoes of this in vivo in the two patients who survived infancy, both of whom showed coexistent monocytosis and frequent neutrophilia along with thrombocytopenia. Our patients' manifest immunodeficiency and immune dysregulation emphasizes a broader role of TET2 in homeostasis and function of the human adaptive immune system. Whereas an immunodeficiency phenotype has not been reported for Tet2-KO mice to date, recent studies imply a crucial role for TET2 in maintaining T-cell homeostasis and B-cell development as recently reviewed by Lio and Rao [109], Feng et al. [116]. Thus TET2-deficient B-cells in both species inefficiently generate mature plasma cells and show impaired class-switch recombination [100, 117].

That this should cause an immunodeficiency phenotype in human beings in the natural environment is perhaps not surprising. It is noteworthy that clinically relevant autoimmunity and impaired T-cell apoptosis in our TET2H1382R patients co-segregated and that these abnormalities were absent from both P3 and the knockout mouse model. One attractive hypothesis is that the hypomorphic nature of the H1382R mutation dissociates the enzymatic and non-enzymatic epigenetic activities of TET2 potentially modulating disease phenotype [116, 118]. However, it is by no means uncommon for individual inborn errors of immunity to produce a broad disease spectrum, ranging from

lymphoproliferation to immunodeficiency [119].

The Swiss Institute of Bioinformatics STRING database was used to quantify the relationship between functional variants identified by whole exome sequencing. The relationships are grouped into validated protein pathways and the strength of interactions is based on several criteria. While the majority of gene variants are likely to have a benign effect on protein function, it is valuable to visualise potential polygenic contributions. The data presented here consists of variants identified by screening for germline functional variants. Somatic variants are likely to exist at allele frequencies below the filtering thresholds used and less likely to be included in this figure. However, if one were to specifically target low-frequency (potential somatic) variants, this same technique can be used to identify their common pathway involvements. Specifically, in the case of tumour sequencing experiments for example, filtering all tumour-specific variants (in DNA derived for tumour sample) against germline whole blood-derived DNA reveals the clonally amplified somatic variants. Human genomics in general has not produced major examples of polygenic disease profiles, although mutually exclusive mutation in TET and IDH proteins have been shown as the cause of leukaemia [36].

We can conclude that impaired T-cell apoptosis shows variable expressivity in TET2-deficiency, but at present we cannot confidently ascribe this to a genotype-phenotype effect. Indeed, all of the patients reconstituted autologous T-cells strikingly early after conditioned haematopoietic stem cell transplantation, suggesting a cell-autonomous pro-proliferative phenotype. Had suitable material been available, it would have been ideal to explore our observations further at a transcriptomic and epigenomic level alongside with more detailed profiling of immune cells [120–122]. Such analysis awaits the identification of future cases, which might now be achieved by targeted screening among children with lymphoid malignancy, especially on a background of consanguinity and immunodeficiency.

4.6 Conclusion

The present findings expand understanding of the critical role of TET2 within the human haematopoietic system and define a new inborn error of immunity.

Bibliography

[1] Conrad Hal Waddington et al. The strategy of the genes. a discussion of some aspects of theoretical biology. with an appendix by h. kacser. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, 1957.

[2] John B Gurdon, Tom R Elsdale, and Michel Fischberg. Sexually mature individuals of xenopus laevis from the transplantation of single somatic nuclei. *Nature*, 182 (4627):64, 1958.

[3] Rollin D Hotchkiss. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175(1): 315–332, 1948.

[4] Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[5] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4): 663–676, 2006.

[6] Elisabet Pujadas and Andrew P Feinberg. Regulated noise in the epigenetic landscape of development and disease. *Cell*, 148(6):1123–1131, 2012.

[7] Andrew E Teschendorff, James West, and Stephan Beck. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Human molecular genetics*, 22 (R1):R7–R15, 2013.

[8] Andrew P Feinberg. The key role of epigenetics in human disease prevention and mitigation. *New England journal of Medicine*, 378(14):1323–1334, 2018.

[9] Zhao-xia Chen and Arthur D Riggs. Dna methylation and demethylation in mammals. *Journal of Biological Chemistry*, 286(21):18347–18353, 2011.

[10] Alexey A Soshnev, Steven Z Josefowicz, and C David Allis. Greater than the sum of parts: complexity of the dynamic epigenome. *Molecular cell*, 62(5):681–694, 2016.

[11] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone h3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, 41(2):246, 2009.

[12] Jennifer C Harr, Teresa Romeo Luperchio, Xianrong Wong, Erez Cohen, Sarah J Wheelan, and Karen L Reddy. Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and a-type lamins. *J cell biol*, 208(1):33–52, 2015.

[13] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[14] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[15] Andrew P Feinberg, Michael A Koldobskiy, and Anita Göndör. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, 17(5):284, 2016.

[16] Julie A. Law and Steven E. Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3): 204–220, Feb 2010. ISSN 1471-0064. doi: 10.1038/nrg2719. URL http://dx.doi.org/10.1038/nrg2719.

[17] Steven M. Chan and Ravindra Majeti. Role of dnmt3a, tet2, and idh1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia. *International journal of Hematology*, 98(6):648–657, Dec 2013. ISSN 1865-3774. doi: 10.1007/s12185-013-1407-8. URL https://doi.org/10.1007/s12185-013-1407-8.

[18] Mary Grace Goll, Finn Kirpekar, Keith A. Maggert, Jeffrey A. Yoder, Chih-Lin Hsieh, Xiaoyu Zhang, Kent G. Golic, Steven E. Jacobsen, and Timothy H. Bestor. Methylation of trnaasp by the dna methyltransferase homolog dnmt2. *Science*, 311(5759):395–398, 2006. ISSN 0036-8075. doi: 10.1126/science.1120976. URL http://science.sciencemag.org/content/311/5759/395.

[19] Sameer Phalke, Olaf Nickel, Diana Walluscheck, Frank Hortig, Maria Cristina Onorati, and Gunter Reuter. Retrotransposon silencing and telomere integrity in somatic cells of drosophila depends on the cytosine-5 methyltransferase dnmt2.

Nature Genetics, 41(6):696–702, May 2009. ISSN 1546-1718. doi: 10.1038/ng.360. URL http://dx.doi.org/10.1038/ng.360.

[20] Jill S Butler and Sharon YR Dent. The role of chromatin modifiers in normal and malignant hematopoiesis. *Blood*, 121(16):3076–3084, 2013.

[21] Timothy J Ley, Li Ding, Matthew J Walter, Michael D McLellan, Tamara Lamprecht, David E Larson, Cyriac Kandoth, Jacqueline E Payton, Jack Baty, John Welch, et al. Dnmt3a mutations in acute myeloid leukemia. *New England Journal of Medicine*, 363(25):2424–2433, 2010.

[22] Matthew J Walter, Li Ding, Dong Shen, Jin Shao, Marcus Grillot, Michael McLellan, Robert Fulton, Heather Schmidt, Joelle Kalicki-Veizer, Michelle O'Laughlin, et al. Recurrent dnmt3a mutations in patients with myelodysplastic syndromes. *Leukemia*, 25(7):1153, 2011.

[23] F Stegelmann, L Bullinger, RF Schlenk, P Paschka, M Griesshammer, C Blersch, S Kuhn, S Schauer, Hartmut Döhner, and K Döhner. Dnmt3a mutations in myeloproliferative neoplasms. *Leukemia*, 25(7):1217, 2011.

[24] Xiao-Jing Yan, Jie Xu, Zhao-Hui Gu, Chun-Ming Pan, Gang Lu, Yang Shen, Jing-Yi Shi, Yong-Mei Zhu, Lin Tang, Xiao-Wei Zhang, et al. Exome sequencing identifies somatic mutations of dna methyltransferase gene dnmt3a in acute monocytic leukemia. *Nature genetics*, 43(4):309, 2011.

[25] Elaine R Mardis, Li Ding, David J Dooling, David E Larson, Michael D McLellan, Ken Chen, Daniel C Koboldt, Robert S Fulton, Kim D Delehaunty, Sean D McGrath, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, 361(11):1058–1066, 2009.

[26] Hai Yan, D Williams Parsons, Genglin Jin, Roger McLendon, B Ahmed Rasheed, Weishi Yuan, Ivan Kos, Ines Batinic-Haberle, Siân Jones, Gregory J Riggins, et al. Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, 360(8): 765–773, 2009.

[27] Stefan Gross, Rob A Cairns, Mark D Minden, Edward M Driggers, Mark A Bittinger, Hyun Gyung Jang, Masato Sasaki, Shengfang Jin, David P Schenkein, Shinsan M Su, et al. Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *Journal of*

Experimental Medicine, 207(2):339–344, 2010.

[28] Patrick S Ward, Jay Patel, David R Wise, Omar Abdel-Wahab, Bryson D Bennett, Hilary A Coller, Justin R Cross, Valeria R Fantin, Cyrus V Hedvat, Alexander E Perl, et al. The common feature of leukemia-associated idh1 and idh2 mutations is a neomorphic enzyme activity converting α-ketoglutarate to 2-hydroxyglutarate. *Cancer cell*, 17(3):225–234, 2010.

[29] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L Gallia, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 2008.

[30] MM Patnaik, CA Hanson, JM Hodnefield, TL Lasho, CM Finke, RA Knudson, RP Ketterling, A Pardanani, and A Tefferi. Differential prognostic effect of idh1 versus idh2 mutations in myelodysplastic syndromes: a mayo clinic study of 277 patients. *Leukemia*, 26(1):101, 2012.

[31] A Tefferi, TL Lasho, O Abdel-Wahab, P Guglielmelli, J Patel, D Caramazza, L Pieri, CM Finke, O Kilpivaara, M Wadleigh, et al. Idh1 and idh2 mutation studies in 1473 patients with chronic-, fibrotic-or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. *Leukemia*, 24(7):1302, 2010.

[32] Darrell R Borger, Kenneth K Tanabe, Kenneth C Fan, Hector U Lopez, Valeria R Fantin, Kimberly S Straley, David P Schenkein, Aram F Hezel, Marek Ancukiewicz, Hannah M Liebman, et al. Frequent mutation of isocitrate dehydrogenase (idh) 1 and idh2 in cholangiocarcinoma identified through broad-based tumor genotyping. *The oncologist*, 17(1):72–79, 2012.

[33] M Fernanda Amary, Krisztian Bacsi, Francesca Maggiani, Stephen Damato, Dina Halai, Fitim Berisha, Robin Pollock, Paul O'donnell, Anita Grigoriadis, Tim Diss, et al. Idh1 and idh2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. *The Journal of pathology*, 224(3):334–343, 2011.

[34] Rob A Cairns, Javeed Iqbal, François Lemonnier, Can Kucuk, Laurence De Leval, Jean-Philippe Jais, Marie Parrens, Antoine Martin, Luc Xerri, Pierre Brousset, et al. Idh2 mutations are frequent in angioimmunoblastic t-cell lymphoma. *Blood*,

pages blood–2011, 2012.

[35] Julie-Aurore Losman and William G Kaelin. What a difference a hydroxyl makes: mutant idh,(r)-2-hydroxyglutarate, and cancer. *Genes & development*, 27(8):836–852, 2013.

[36] Maria E. Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S. Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F. Fernandez, Martin S. Tallman, Zhuoxin Sun, Kristy Wolniak, Justine K. Peeters, Wei Liu, Sung E. Choe, Valeria R. Fantin, Elisabeth Paietta, Bob LÃ¶wenberg, Jonathan D. Licht, Lucy A. Godley, Ruud Delwel, Peter J.M. Valk, Craig B. Thompson, Ross L. Levine, and Ari Melnick. Leukemic idh1 and idh2 mutations result inâ a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer Cell*, 18(6):553 – 567, 2010. ISSN 1535-6108. doi: https://doi.org/10.1016/j.ccr.2010.11.015. URL http://www.sciencedirect.com/science/article/pii/S1535610810004836.

[37] Verena I Gaidzik, Peter Paschka, D Spath, Marianne Habdank, CH Kohne, Ulrich Germing, Marie von Lilienfeld-Toal, Gerhard Held, Heinz-August Horst, Detlef Haase, et al. Tet2 mutations in acute myeloid leukemia (aml): results from a comprehensive genetic and clinical analysis of the aml study group. *J Clin Oncol*, 30(12):1350–1357, 2012.

[38] Lenny Dang, David W White, Stefan Gross, Bryson D Bennett, Mark A Bittinger, Edward M Driggers, Valeria R Fantin, Hyun Gyung Jang, Shengfang Jin, Marie C Keenan, et al. Cancer-associated idh1 mutations produce 2-hydroxyglutarate. *Nature*, 462(7274):739, 2009.

[39] Julie-Aurore Losman, Ryan E Looper, Peppi Koivunen, Sungwoo Lee, Rebekka K Schneider, Christine McMahon, Glenn S Cowley, David E Root, Benjamin L Ebert, and William G Kaelin. (r)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. *Science*, 339(6127):1621–1625, 2013.

[40] Jian-Kang Zhu. Active dna demethylation mediated by dna glycosylases. *Annual review of genetics*, 43:143–166, 2009.

[41] Wolfgang Mayer, Alain Niveleau, Jörn Walter, Reinald Fundele, and Thomas Haaf. Embryogenesis: demethylation of the zygotic paternal genome. *nature*, 403(6769):501, 2000.

[42] J Oswald, S Engemann, N Lane, W Mayer, A Olek, R Fundele, W Dean, W Reik, and J Walter. Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478, 2000.

[43] Skirmantas Kriaucionis and Nathaniel Heintz. The nuclear dna base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science*, 324(5929):929–930, 2009.

[44] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[45] RB Lorsbach, J Moore, S Mathew, SC Raimondi, ST Mukatira, and JR Downing. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia*, 17(3):637, 2003.

[46] Ryoichi Ono, Tomohiko Taki, Takeshi Taketani, Masafumi Taniwaki, Hajime Kobayashi, and Yasuhide Hayashi. Lcx, leukemia-associated protein with a cxxc domain, is fused to mll in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). *Cancer research*, 62(14):4075–4080, 2002.

[47] Janet H Gommers-Ampt, Fred Van Leeuwen, Antonius LJ de Beer, Johannes FG Vliegenthart, Miral Dizdaroglu, Jeffrey A Kowalak, Pamela F Crain, and Piet Borst. β-d-glucosyl-hydroxymethyluracil: a novel modified base present in the dna of the parasitic protozoan t. brucei. *Cell*, 75(6):1129–1136, 1993.

[48] Andre Bernards, Nel van Harten-Loosbroek, and Piet Borst. Modification of telomeric dna in trypanosoma brucei; a role in antigenic variation? *Nucleic Acids Research*, 12(10):4153–4170, 1984.

[49] Piet Borst and Robert Sabatini. Base j: discovery, biosynthesis, and possible functions. *Annu. Rev. Microbiol.*, 62:235–251, 2008.

[50] Zhong Yu, Paul-Andre Genest, Bas ter Riet, Kate Sweeney, Courtney DiPaolo, Rudo Kieft, Evangelos Christodoulou, Anastassis Perrakis, Jana M Simmons, Robert P Hausinger, et al. The protein that binds to dna base j in trypanosomatids has features of a thymidine hydroxylase. *Nucleic acids research*, 35(7):2107–2115, 2007.

[51] Laura J Cliffe, Rudo Kieft, Timothy Southern, Shanda R Birkeland, Marion Mar-

shall, Kate Sweeney, and Robert Sabatini. Jbp1 and jbp2 are two distinct thymidine hydroxylases involved in j biosynthesis in genomic dna of african trypanosomes. *Nucleic acids research*, 37(5):1452–1462, 2009.

[52] Shinsuke Ito, Ana C D'Alessio, Olena V Taranova, Kwonho Hong, Lawrence C Sowers, and Yi Zhang. Role of tet proteins in 5mc to 5hmc conversion, es-cell self-renewal and inner cell mass specification. *nature*, 466(7310):1129, 2010.

[53] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047): 1303–1307, 2011.

[54] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[55] Jeffrey A Smiley, Melisa Kundracik, Daniel A Landfried, Vincient R Barnes Sr, and Armend A Axhemi. Genes of the thymidine salvage pathway: thymine-7-hydroxylase from a rhodotorula glutinis cdna library and iso-orotate decarboxylase from neurospora crassa. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1723(1-3):256–264, 2005.

[56] Atanu Maiti and Alexander C Drohat. Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011.

[57] Alain R Weber, Claudia Krawczyk, Adam B Robertson, Anna Kuśnierczyk, Cathrine B Vågbø, David Schuermann, Arne Klungland, and Primo Schär. Biochemical reconstitution of tet1–tdg–ber-dependent active dna demethylation reveals a highly coordinated mechanism. *Nature communications*, 7:10806, 2016.

[58] Carina Frauer, Thomas Hoffmann, Sebastian Bultmann, Valentina Casa, M. Cristina Cardoso, Iris Antes, and Heinrich Leonhardt. Recognition of 5-hydroxymethylcytosine by the uhrf1 sra domain. *PLoS ONE*, 6(6):e21306, Jun 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0021306. URL http://dx.doi.org/10.1371/journal.pone.0021306.

[59] Marian Mellén, Pinar Ayata, Scott Dewell, Skirmantas Kriaucionis, and Nathaniel Heintz. Mecp2 binds to 5hmc enriched within active genes and accessible chromatin in the nervous system. *Cell*, 151(7):1417 – 1430, 2012. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2012.11.022. URL http://www.sciencedirect.com/science/article/pii/S0092867412014079.

[60] Cornelia G. Spruijt, Felix Gnerlich, Arne H. Smits, Toni Pfaffeneder, Pascal W.T.C. Jansen, Christina Bauer, Martin Münzel, Mirko Wagner, Markus Müller, Fariha Khan, H. Christian Eberl, Anneloes Mensinga, Arie B. Brinkman, Konstantin Lephikov, Udo Müller, Jörn Walter, Rolf Boelens, Hugo van Ingen, Heinrich Leonhardt, Thomas Carell, and Michiel Vermeulen. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, 152(5):1146 – 1159, 2013. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2013.02.004. URL http://www.sciencedirect.com/science/article/pii/S0092867413001529.

[61] Erik M. K. Rasmussen and Gro V. Amdam. Cytosine modifications in the honey bee (apis mellifera) worker genome. *Frontiers in Genetics*, 6, Feb 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00008. URL http://dx.doi.org/10.3389/fgene.2015.00008.

[62] Marek Wojciechowski, Dominik Rafalski, Robert Kucharski, Katarzyna Misztal, Joanna Maleszka, Matthias Bochtler, and Ryszard Maleszka. Insights into dna hydroxymethylation in the honeybee from in-depth analyses of tet dioxygenase. *Open Biology*, 4(8), 2014. doi: 10.1098/rsob.140110. URL http://rsob.royalsocietypublishing.org/content/4/8/140110.

[63] Chun-Xiao Song and Chuan He. Potential functional roles of dna demethylation intermediates. *Trends in Biochemical Sciences*, 38(10):480 – 484, 2013. ISSN 0968-0004. doi: https://doi.org/10.1016/j.tibs.2013.07.003. URL http://www.sciencedirect.com/science/article/pii/S0968000413001187.

[64] Chun-Xiao Song, Chengqi Yi, and Chuan He. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature Biotechnology*, 30(11):1107–1116, Nov 2012. ISSN 1546-1696. doi: 10.1038/nbt.2398. URL http://dx.doi.org/10.1038/nbt.2398.

[65] William A Pastor, L Aravind, and Anjana Rao. Tetonic shift: biological roles of tet proteins in dna demethylation and transcription. *Nature reviews Molecular cell*

biology, 14(6):341, 2013.

[66] Matthew W Kellinger, Chun-Xiao Song, Jenny Chong, Xing-Yu Lu, Chuan He, and Dong Wang. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of rna polymerase ii transcription. *Nature Structural and Molecular Biology*, 19(8):831–833, Jul 2012. ISSN 1545-9985. doi: 10.1038/nsmb.2346. URL http://dx.doi.org/10.1038/nsmb.2346.

[67] Maxim Ivanov, Mart Kals, Marina Kacevska, Isabel Barragan, Kie Kasuga, Anders Rane, Andres Metspalu, Lili Milani, and Magnus Ingelman-Sundberg. Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. *Genome Biology*, 14(8):R83, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-8-r83. URL http://dx.doi.org/10.1186/gb-2013-14-8-r83.

[68] François Delhommeau, Sabrina Dupont, Véronique Della Valle, Chloé James, Severine Trannoy, Aline Massé, Olivier Kosmider, Jean-Pierre Le Couedic, Fabienne Robert, Antonio Alberdi, Yann Lécluse, Isabelle Plo, François J. Dreyfus, Christophe Marzac, Nicole Casadevall, Catherine Lacombe, Serge P. Romana, Philippe Dessen, Jean Soulier, Franck Viguié, Michaela Fontenay, William Vainchenker, and Olivier A. Bernard. Mutation in tet2 in myeloid cancers. *New England journal of Medicine*, 360(22):2289–2301, 2009. doi: 10.1056/NEJMoa0810069. URL https://doi.org/10.1056/NEJMoa0810069. PMID: 19474426.

[69] Saskia M C Langemeijer, Roland P Kuiper, Marieke Berends, Ruth Knops, Mariam G Aslanyan, Marion Massop, Ellen Stevens-Linders, Patricia van Hoogen, Ad Geurts van Kessel, Reinier A P Raymakers, Eveline J Kamping, Gregor E Verhoef, Estelle Verburgh, Anne Hagemeijer, Peter Vandenberghe, Theo de Witte, Bert A van der Reijden, and Joop H Jansen. Acquired mutations in tet2 are common in myelodysplastic syndromes. *Nature Genetics*, 41:838 EP –, 05 2009. URL http://dx.doi.org/10.1038/ng.391.

[70] Omar Abdel-Wahab, Ann Mullally, Cyrus Hedvat, Guillermo Garcia-Manero, Jay Patel, Martha Wadleigh, Sebastien Malinge, JinJuan Yao, Outi Kilpivaara, Rukhmi Bhat, et al. Genetic characterization of tet1, tet2, and tet3 alterations in myeloid malignancies. *Blood*, 114(1):144–147, 2009.

[71] Anna M Jankowska, Hadrian Szpurka, Ramon V Tiu, Hideki Makishima, Manuel

Afable, Jungwon Huh, Christine L O'Keefe, Rebecca Ganetzky, Michael A McDevitt, and Jaroslaw P Maciejewski. Loss of heterozygosity 4q24 and tet2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood*, 113(25): 6403–6410, 2009.

[72] Ayalew Tefferi, A Pardanani, KH Lim, O Abdel-Wahab, TL Lasho, J Patel, N Gangat, CM Finke, S Schwager, A Mullally, et al. Tet2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia*, 23(5):905, 2009.

[73] Ayalew Tefferi, RL Levine, KH Lim, O Abdel-Wahab, TL Lasho, J Patel, CM Finke, A Mullally, CY Li, A Pardanani, et al. Frequent tet2 mutations in systemic mastocytosis: clinical, kitd816v and fip1l1-pdgfra correlates. *Leukemia*, 23(5):900, 2009.

[74] Ayalew Tefferi, KH Lim, O Abdel-Wahab, TL Lasho, J Patel, MM Patnaik, Curtis A Hanson, A Pardanani, DG Gilliland, and RL Levine. Detection of mutant tet2 in myeloid malignancies other than myeloproliferative neoplasms: Cmml, mds, mds/mpn and aml. *Leukemia*, 23(7):1343, 2009.

[75] Myunggon Ko, Yun Huang, Anna M. Jankowska, Utz J. Pape, Mamta Tahiliani, Hozefa S. Bandukwala, Jungeun An, Edward D. Lamperti, Kian Peng Koh, Rebecca Ganetzky, X. Shirley Liu, L. Aravind, Suneet Agarwal, Jaroslaw P. Maciejewski, and Anjana Rao. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant tet2. *Nature*, 468:839 EP –, 11 2010. URL http://dx.doi.org/10.1038/nature09586.

[76] Kelly Moran-Crusio, Linsey Reavie, Alan Shih, Omar Abdel-Wahab, Delphine Ndiaye-Lobry, Camille Lobry, Maria E. Figueroa, Aparna Vasanthakumar, Jay Patel, Xinyang Zhao, Fabiana Perna, Suveg Pandey, Jozef Madzo, Chunxiao Song, Qing Dai, Chuan He, Sherif Ibrahim, Miloslav Beran, Jiri Zavadil, Stephen D. Nimer, Ari Melnick, Lucy A. Godley, Iannis Aifantis, and Ross L. Levine. tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*, 20(1):11–24, 2018/06/27 2011. doi: 10.1016/j.ccr.2011.06.001. URL http://dx.doi.org/10.1016/j.ccr.2011.06.001.

[77] Cyril Quivoron, Lucile Couronné, Véronique Della Valle, Cécile K. Lopez, Isabelle

Plo, Orianne Wagner-Ballon, Marcio Do Cruzeiro, Francois Delhommeau, Bertrand Arnulf, Marc-Henri Stern, Lucy Godley, Paule Opolon, Hervé Tilly, Eric Solary, Yannis Duffourd, Philippe Dessen, Hélène Merle-Beral, Florence Nguyen-Khac, Michaëla Fontenay, William Vainchenker, Christian Bastard, Thomas Mercher, and Olivier A. Bernard. Tet2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell*, 20(1):25–38, 2011. doi: https://doi.org/10.1016/j.ccr.2011.06.003. URL http://www.sciencedirect.com/science/article/pii/S153561081100225X.

[78] Myunggon Ko, Hozefa S. Bandukwala, Jungeun An, Edward D. Lamperti, Elizabeth C. Thompson, Ryan Hastie, Angeliki Tsangaratou, Klaus Rajewsky, Sergei B. Koralov, and Anjana Rao. Ten-eleven-translocation 2 (tet2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proceedings of the National Academy of Sciences*, 108(35):14566–14571, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1112317108. URL http://www.pnas.org/content/108/35/14566.

[79] Zhe Li, Xiaoqiang Cai, Chen-Leng Cai, Jiapeng Wang, Wenyong Zhang, Bruce E. Petersen, Feng-Chun Yang, and Mingjiang Xu. Deletion of tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. *Blood*, 118(17):4509–4518, 2011. ISSN 0006-4971. doi: 10.1182/blood-2010-12-325241. URL http://www.bloodjournal.org/content/118/17/4509.

[80] Max Jan, Thomas M. Snyder, M. Ryan Corces-Zimmerman, Paresh Vyas, Irving L. Weissman, Stephen R. Quake, and Ravindra Majeti. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Science Translational Medicine*, 4(149):149ra118–149ra118, 2012. ISSN 1946-6234. doi: 10.1126/scitranslmed.3004315. URL http://stm.sciencemag.org/content/4/149/149ra118.

[81] Raphaël Itzykson, Olivier Kosmider, Aline Renneville, Margot Morabito, Claude Preudhomme, Céline Berthon, Lionel Adès, Pierre Fenaux, Uwe Platzbecker, Olivier Gagey, Philippe Rameau, Guillaume Meurice, Cédric Oréar, François Delhommeau, Olivier A. Bernard, Michaela Fontenay, William Vainchenker, Nathalie Droin, and Eric Solary. Clonal architecture of chronic myelomonocytic leukemias. *Blood*, 121 (12):2186–2198, 2013. ISSN 0006-4971. doi: 10.1182/blood-2012-06-440347. URL http://www.bloodjournal.org/content/121/12/2186.

[82] Kasper Dindler Rasmussen and Kristian Helin. Role of tet enzymes in dna methylation, development, and cancer. *Genes & development*, 30(7):733–750, 2016. doi: 10.1101/gad.276568.115. URL http://genesdev.cshlp.org/content/30/7/733.abstract.

[83] Lambert Busque, Jay P Patel, Maria E Figueroa, Aparna Vasanthakumar, Sylvie Provost, Zineb Hamilou, Luigina Mollica, Juan Li, Agnes Viale, Adriana Heguy, Maryam Hassimi, Nicholas Socci, Parva K Bhatt, Mithat Gonen, Christopher E Mason, Ari Melnick, Lucy A Godley, Cameron W Brennan, Omar Abdel-Wahab, and Ross L Levine. Recurrent somatic tet2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature Genetics*, 44:1179 EP –, 09 2012. URL http://dx.doi.org/10.1038/ng.2413.

[84] Giulio Genovese, Anna K Kähler, Robert E Handsaker, Johan Lindberg, Samuel A Rose, Samuel F Bakhoum, Kimberly Chambert, Eran Mick, Benjamin M Neale, Menachem Fromer, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England journal of Medicine*, 371(26):2477–2487, 2014. doi: 10.1056/NEJMoa1409405. URL https://doi.org/10.1056/NEJMoa1409405. PMID: 25426838.

[85] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, Craig H. Mermel, Noel Burtt, Alejandro Chavez, John M. Higgins, Vladislav Moltchanov, Frank C. Kuo, Michael J. Kluk, Brian Henderson, Leena Kinnunen, Heikki A. Koistinen, Claes Ladenvall, Gad Getz, Adolfo Correa, Benjamin F. Banahan, Stacey Gabriel, Sekar Kathiresan, Heather M. Stringham, Mark I. McCarthy, Michael Boehnke, Jaakko Tuomilehto, Christopher Haiman, Leif Groop, Gil Atzmon, James G. Wilson, Donna Neuberg, David Altshuler, and Benjamin L. Ebert. Age-related clonal hematopoiesis associated with adverse outcomes. *New England journal of Medicine*, 371(26):2488–2498, 2014. doi: 10.1056/NEJMoa1408617. URL https://doi.org/10.1056/NEJMoa1408617. PMID: 25426837.

[86] Mingchao Xie, Charles Lu, Jiayin Wang, Michael D McLellan, Kimberly J Johnson, Michael C Wendl, Joshua F McMichael, Heather K Schmidt, Venkata Yellapantula, Christopher A Miller, Bradley A Ozenberger, John S Welch, Daniel C Link, Matthew J Walter, Elaine R Mardis, John F Dipersio, Feng Chen, Richard K

Wilson, Timothy J Ley, and Li Ding. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine*, 20:1472 EP –, 10 2014. URL http://dx.doi.org/10.1038/nm.3733.

[87] Luisa Cimmino, Meelad M Dawlaty, Delphine Ndiaye-Lobry, Yoon Sing Yap, Sofia Bakogianni, Yiting Yu, Sanchari Bhattacharyya, Rita Shaknovich, Huimin Geng, Camille Lobry, Jasper Mullenders, Bryan King, Thomas Trimarchi, Beatriz Aranda-Orgilles, Cynthia Liu, Steven Shen, Amit K Verma, Rudolf Jaenisch, and Iannis Aifantis. Tet1 is a tumor suppressor of hematopoietic malignancy. *Nature Immunology*, 16:653 EP –, 04 2015. URL http://dx.doi.org/10.1038/ni.3148.

[88] Zhigang Zhao, Li Chen, Meelad M. Dawlaty, Feng Pan, Ophelia Weeks, Yuan Zhou, Zeng Cao, Hui Shi, Jiapeng Wang, Li Lin, Shi Chen, Weiping Yuan, Zhaohui Qin, Hongyu Ni, Stephen D. Nimer, Feng-Chun Yang, Rudolf Jaenisch, Peng Jin, and Mingjiang Xu. Combined loss of tet1 and tet2 promotes b cell, but not myeloid malignancies, in mice. *Cell Reports*, 13(8):1692 – 1704, 2015. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2015.10.037. URL http://www.sciencedirect.com/science/article/pii/S2211124715012073.

[89] ChristineÂ Guo Lian, Yufei Xu, Craig Ceol, Feizhen Wu, Allison Larson, Karen Dresser, Wenqi Xu, Li Tan, Yeguang Hu, Qian Zhan, Chung wei Lee, Di Hu, BillÂ Q. Lian, Sonja Kleffel, Yijun Yang, James Neiswender, AbrahamÂ J. Khorasani, Rui Fang, Cecilia Lezcano, LynÂ M. Duncan, RichardÂ A. Scolyer, JohnÂ F. Thompson, Hojabr Kakavand, Yariv Houvras, LeonardÂ I. Zon, MartinÂ C. Mihm, UrsulaÂ B. Kaiser, Tobias Schatton, BruceÂ A. Woda, GeorgeÂ F. Murphy, and YujiangÂ G. Shi. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell*, 150(6):1135 – 1146, 2012. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2012.07.033. URL http://www.sciencedirect.com/science/article/pii/S0092867412010124.

[90] Raajit Rampal, Altuna Alkalin, Jozef Madzo, Aparna Vasanthakumar, Elodie Pronier, Jay Patel, Yushan Li, Jihae Ahn, Omar Abdel-Wahab, Alan Shih, Chao Lu, PatrickÂ S. Ward, JenniferÂ J. Tsai, Todd Hricik, Valeria Tosello, JacobÂ E. Tallman, Xinyang Zhao, Danette Daniels, Qing Dai, Luisa Ciminio, Iannis Aifantis, Chuan He, Francois Fuks, MartinÂ S. Tallman, Adolfo Ferrando, Stephen Nimer, Elisabeth Paietta, CraigÂ B. Thompson, JonathanÂ D. Licht, ChristopherÂ E.

Mason, LucyÂ A. Godley, Ari Melnick, MariaÂ E. Figueroa, and RossÂ L. Levine. Dna hydroxymethylation profiling reveals that wt1 mutations result in loss of tet2 function in acute myeloid leukemia. *Cell Reports*, 9(5):1841 – 1855, 2014. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2014.11.004. URL http://www.sciencedirect.com/science/article/pii/S2211124714009589.

[91] Yiping Wang, Mengtao Xiao, Xiufei Chen, Leilei Chen, Yanping Xu, Lei Lv, Pu Wang, Hui Yang, Shenghong Ma, Huaipeng Lin, Bo Jiao, Ruibao Ren, Dan Ye, Kun-Liang Guan, and Yue Xiong. Wt1 recruits tet2 to regulate its target gene expression and suppress leukemia cell proliferation. *Molecular Cell*, 57(4):662 – 673, 2015. ISSN 1097-2765. doi: https://doi.org/10.1016/j.molcel.2014.12.023. URL http://www.sciencedirect.com/science/article/pii/S1097276514010016.

[92] Johannes Schindelin, Curtis T Rueden, Mark C Hiner, and Kevin W Eliceiri. The imagej ecosystem: An open platform for biomedical image analysis. *Molecular reproduction and development*, 82(7-8):518–529, 2015.

[93] Lulu Hu, Ze Li, Jingdong Cheng, Qinhui Rao, Wei Gong, Mengjie Liu, Yujiang Geno Shi, Jiayu Zhu, Ping Wang, and Yanhui Xu. Crystal structure of tet2-dna complex: insight into tet-mediated 5mc oxidation. *Cell*, 155(7):1545–1555, 2013.

[94] Catherine Cargo, Matthew Cullen, Jan Taylor, Mike Short, Paul Glover, Suzan Van Hoppe, Alex Smith, Paul Evans, and Simon Crouch. The use of targeted sequencing and flow cytometry to identify patients with a clinically significant monocytosis. *Blood*, 133(12):1325–1334, 2019.

[95] Vagheesh M. Narasimhan, Karen A. Hunt, Dan Mason, Christopher L. Baker, Konrad J. Karczewski, Michael R. Barnes, Anthony H. Barnett, Chris Bates, Srikanth Bellary, Nicholas A. Bockett, Kristina Giorda, Christopher J. Griffiths, Harry Hemingway, Zhilong Jia, M. Ann Kelly, Hajrah A. Khawaja, Monkol Lek, Shane McCarthy, Rosie McEachan, Anne O'Donnell-Luria, Kenneth Paigen, Constantinos A. Parisinos, Eamonn Sheridan, Laura Southgate, Louise Tee, Mark Thomas, Yali Xue, Michael Schnall-Levin, Petko M. Petkov, Chris Tyler-Smith, Eamonn R. Maher, Richard C. Trembath, Daniel G. MacArthur, John Wright, Richard Durbin, and David A. van Heel. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*,

352(6284):474–477, 2016. ISSN 0036-8075. doi: 10.1126/science.aac8624. URL https://science.sciencemag.org/content/352/6284/474.

[96] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[97] François Lemonnier, Lucile Couronné, Marie Parrens, Jean-Philippe Jaïs, Marion Travert, Laurence Lamant, Olivier Tournillac, Thérèse Rousset, Bettina Fabiani, Rob A Cairns, et al. Recurrent tet2 mutations in peripheral t-cell lymphomas correlate with tfh-like features and adverse clinical parameters. *blood*, 120(7): 1466–1469, 2012.

[98] Fazila Asmar, Vasu Punj, Jesper Christensen, Marianne T Pedersen, Anja Pedersen, Anders B Nielsen, Christoffer Hother, Ulrik Ralfkiaer, Peter Brown, Elisabeth Ralfkiaer, et al. Genome-wide profiling identifies a dna methylation signature that associates with tet2 mutations in diffuse large b-cell lymphoma. *Haematologica*, 98 (12):1912–1920, 2013.

[99] Kazuya Shimoda, Kotaro Shide, Takuro Kameda, Tomonori Hidaka, Yoko Kubuki, Ayako Kamiunten, Masaaki Sekine, Keiichi Akizuki, Haruko Shimoda, Takumi Yamaji, et al. Tet2 mutation in adult t-cell leukemia/lymphoma. *Journal of Clinical and Experimental Hematopathology*, 55(3):145–149, 2015.

[100] Pilar M Dominguez, Hussein Ghamlouch, Wojciech Rosikiewicz, Parveen Kumar, Wendy Béguelin, Lorena Fontan, Martín A Rivas, Patrycja Pawlikowska, Marine Armand, Enguerran Mouly, et al. Tet2 deficiency causes germinal center hyperplasia, impairs plasma cell differentiation, and promotes b-cell lymphomagenesis. *Cancer discovery*, 8(12):1632–1653, 2018.

[101] Xi Jin, Tingting Qin, Meiling Zhao, Nathanael Bailey, Lu Liu, Kevin Yang, Victor Ng, Tomoyasu Higashimoto, Rosemary Coolon, Gina Ney, et al. Oncogenic n-ras and tet2 haploinsufficiency collaborate to dysregulate hematopoietic stem and progenitor cells. *Blood advances*, 2(11):1259–1271, 2018.

[102] Hassan Awada, Yasunobu Nagata, Abhinav Goyal, Mohammad F Asad, Bhumika Patel, Cassandra M Hirsch, Teodora Kuzmanovic, Yihong Guan, Bartlomiej P

Przychodzen, Mai Aly, et al. Invariant phenotype and molecular association of biallelic tet2 mutant myeloid neoplasia. *Blood advances*, 3(3):339–349, 2019.

[103] Penggao Dai, Wen C Xiong, and Lin Mei. Erbin inhibits raf activation by disrupting the sur-8-ras-raf complex. *Journal of Biological Chemistry*, 281(2):927–933, 2006.

[104] Emmanuel N Olivier, Lamin Marenah, Angela McCahill, Alison Condie, Scott Cowan, and Joanne C Mountford. High-efficiency serum-free feeder-free erythroid differentiation of human pluripotent stem cells using small molecules. *Stem cells translational medicine*, 5(10):1394–1405, 2016.

[105] Xiaoli Qu, Shijie Zhang, Shihui Wang, Yaomei Wang, Wei Li, Yumin Huang, Huizhi Zhao, Xiuyun Wu, Chao An, Xinhua Guo, et al. Tet2 deficiency leads to stem cell factor–dependent clonal expansion of dysfunctional erythroid progenitors. *Blood*, 132(22):2406–2417, 2018.

[106] Joseph A Fraietta, Christopher L Nobles, Morgan A Sammons, Stefan Lundh, Shannon A Carty, Tyler J Reich, Alexandria P Cogdill, Jennifer JD Morrissette, Jamie E DeNizio, Shantan Reddy, et al. Disruption of tet2 promotes the therapeutic efficacy of cd19-targeted t cells. *Nature*, 558(7709):307, 2018.

[107] Roland Schmitz, George W Wright, Da Wei Huang, Calvin A Johnson, James D Phelan, James Q Wang, Sandrine Roulland, Monica Kasbekar, Ryan M Young, Arthur L Shaffer, et al. Genetics and pathogenesis of diffuse large b-cell lymphoma. *New England journal of Medicine*, 378(15):1396–1407, 2018.

[108] Eevi Kaasinen, Outi Kuismin, Kristiina Rajamäki, Heikki Ristolainen, Mervi Aavikko, Johanna Kondelin, Silva Saarinen, Davide G Berta, Riku Katainen, Elina AM Hirvonen, et al. Impact of constitutional tet2 haploinsufficiency on molecular and clinical phenotype in humans. *Nature communications*, 10(1):1252, 2019.

[109] Chan-Wang J Lio and Anjana Rao. Tet enzymes and 5hmc in adaptive and innate immune systems. *Frontiers in immunology*, 10, 2019.

[110] E Solary, OA Bernard, Ayalew Tefferi, François Fuks, and William Vainchenker. The ten-eleven translocation-2 (tet2) gene in hematopoiesis and hematopoietic diseases. *Leukemia*, 28(3):485, 2014.

[111] Enguerran Mouly, Hussein Ghamlouch, Veronique Della-Valle, Laurianne Scourzic,

Cyril Quivoron, Damien Roos-Weil, Patrycja Pawlikowska, Véronique Saada, K Diop M'Boyba, Cécile K Lopez, et al. B-cell tumor development in tet2-deficient mice. *Blood advances*, 2(6):703–714, 2018.

[112] Kasper D Rasmussen, Ivan Berest, Sandra Keβler, Koutarou Nishimura, Lucía Simón-Carrasco, George S Vassiliou, Marianne T Pedersen, Jesper Christensen, Judith B Zaugg, and Kristian Helin. Tet2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. *Genome research*, 29(4):564–575, 2019.

[113] Gary C Hon, Chun-Xiao Song, Tingting Du, Fulai Jin, Siddarth Selvaraj, Ah Young Lee, Chia-an Yen, Zhen Ye, Shi-Qing Mao, Bang-An Wang, et al. 5mc oxidation by tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Molecular cell*, 56(2):286–297, 2014.

[114] Michael Reimer, Kirthi Pulakanti, Linzheng Shi, Alex Abel, Mingyu Liang, Subramaniam Malarkannan, and Sridhar Rao. Deletion of tet proteins results in quantitative disparities during esc differentiation partially attributable to alterations in gene expression. *BMC developmental biology*, 19(1):16, 2019.

[115] Thierry Langlois, Barbara da Costa Reis Monte-Mor, Gaëlle Lenglet, Nathalie Droin, Caroline Marty, Jean-Pierre Le Couédic, Carole Almire, Nathalie Auger, Thomas Mercher, François Delhommeau, et al. Tet2 deficiency inhibits mesoderm and hematopoietic differentiation in human embryonic stem cells. *Stem Cells*, 32 (8):2084–2097, 2014.

[116] Yimei Feng, Xiaoping Li, Kaniel Cassady, Zhongmin Zou, and Xi Zhang. Tet2 function in hematopoietic malignancies, immune regulation, and dna repair. *Frontiers in oncology*, 9, 2019.

[117] Katia Schoeler, Andreas Aufschnaiter, Simon Messner, Emmanuel Derudder, Sebastian Herzog, Andreas Villunger, Klaus Rajewsky, and Verena Labi. Tet enzymes control antibody production and shape the mutational landscape in germinal center b cells. *The FEBS journal*, 2019.

[118] Hao Lian, Wen-Bin Li, and Wei-Lin Jin. The emerging insights into catalytic or non-catalytic roles of tet proteins in tumors and neural development. *Oncotarget*, 7 (39):64512, 2016.

[119] Capucine Picard, H Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean-Laurent Casanova, Talal Chatila, Yanick J Crow, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, et al. International union of immunological societies: 2017 primary immunodeficiency diseases committee report on inborn errors of immunity. *Journal of clinical immunology*, 38(1):96–128, 2018.

[120] Yan-ping Xu, Lei Lv, Ying Liu, Matthew D Smith, Wen-Cai Li, Xian-ming Tan, Meng Cheng, Zhijun Li, Michael Bovino, Jeffrey Aubé, et al. Tumor suppressor tet2 promotes cancer immunity and immunotherapy efficacy. *The Journal of clinical investigation*, 2019.

[121] Hiroko Nakatsukasa, Mayumi Oda, Jinghua Yin, Shunsuke Chikuma, Minako Ito, Mana Koga-Iizuka, Kazue Someya, Yohko Kitagawa, Naganari Ohkura, Shimon Sakaguchi, et al. Loss of tet proteins in regulatory t cells promotes abnormal proliferation, foxp3 destabilization and il-17 expression. *International immunology*, 31(5):335–347, 2019.

[122] Xiaojing Yue, Chan-Wang J Lio, Daniela Samaniego-Castruita, Xiang Li, and Anjana Rao. Loss of tet2 and tet3 in regulatory t cells unleashes effector function. *Nature communications*, 10(1):2011, 2019.

5 Genomic analysis for primary immune disorders

Abbreviations

BWA (Burrows-Wheeler transformation aligner), FDR (False discovery rate), GO (Gene Ontology), GrCh38 (Genome Reference Consortium Human Build 38), GVCF (Genomic Variant Call Format), KEGG (Kyoto Encyclopedia of Genes and Genomes), LoF (loss-of-function), NCBI (National Center for Biotechnology), Pfam (Protein families database), PPI (Protein-protein interaction), VCF (variant call format).

5.1 Introduction

This chapter contains theory and examples for the investigation of rare disease by exome sequencing used throughout this thesis. Each section is generally self-contained with a brief introduction. A specific section is devoted to a novel method of rare disease cohort network analysis in Sec 5.5. A separate introduction is also included to begin that section in context. This procedure was developed to provide a statistical method for the detection of damaged protein pathways that drive disease. The method is based on measuring variant enrichment and clustering by protein-protein interactions (PPI).

A detailed overall analysis plan is illustrated in subsection 5.3.1. A accompanying

data storage plan is also provided in the same section that directly maps to the analysis plan. A rough overview "infographic" of a next generation sequencing study is shown **Figure 5.1**. The general requirements, personnel responsibilities, and cost-breakdown is shown.

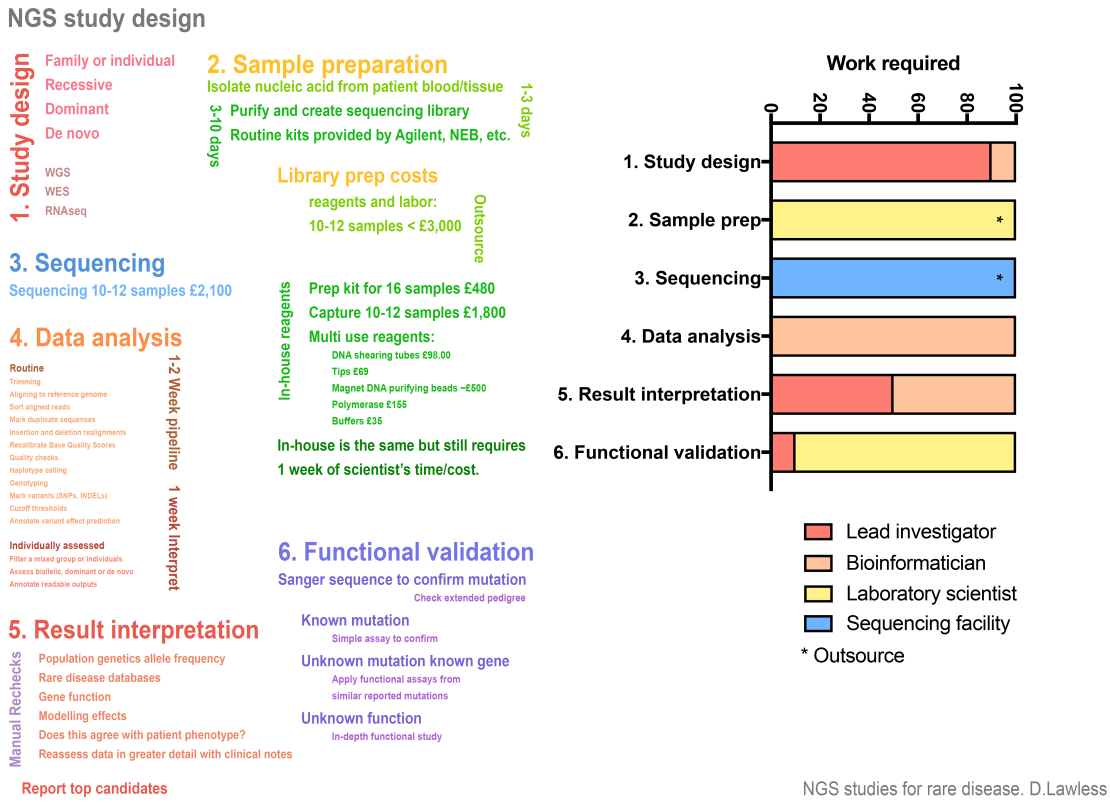


Figure 5.1: **Whole exome sequencing experiment design.** The general requirements, personnel responsibilities, and cost-breakdown is shown for a small NGS study of approximately ten participants. If library preparation and sequencing is performed at a dedicated facility then scaling up to very large cohorts (>1,000) potential only differs in one critical feature; implementing the bioinformatic methods used in this chapter also requires a critical expertise in high-performance computing. No methods have been included to demonstrate job scheduling and parallelisation across large computer clusters.

5.2 Exome sequencing

5.2.1 Sample preparation

For genomic investigations, a patient generally donates a small blood sample (2-6mL) along with signed consent to use their biological material and data in genetic and functional research. Patient DNA is purified from peripheral blood monocytes. In most cases, the purification is done using a commercial kit such as that from Qiagen (51104 QIAamp DNA Blood Mini Kit). This protocol takes about 1 hour to purify 1-10 patient samples. Sometimes patient DNA is provided from an external source such as a local hospital where blood samples are processed routinely by dedicated staff. In this case, the purification method may be unknown so extra care should be taken when checking suitability for sequencing experiments. Consideration should be given to the possibility of sample mix up, that contamination could have occurred, etc.

High-throughput sequencing experiments benefit from consistency during sequencing library preparation. While there are several commercial options available, the protocol used in this study was the SureSelect XT target enrichment system for Illumina paired-end multiplexed sequencing library. A detailed protocol is available from the manufacturer. However, the process can be summarised in four main steps. After DNA quality has been checked, the basic protocol consists of:
(1) DNA fragmentation into 100-300 base pair strands, either (i) by using an enzyme that digests the DNA or (ii) by breaking by sonication; the DNA is suspended inside a small glass tube containing a glass rod which is vibrated by sonic waves inside a water bath.
(2) Another round of quality control checks to ensure that the DNA is fragmented into the correct size range.
(3) These fragments are bound by probes that specifically recognise the coding sequences which collectively make up the exome.
(4) The DNA that has been selectively purified is then tagged by adding a tail of nucleotides in specific sequences that label each of the individual samples with a unique code. When the sequencing step is performed later, all of the samples will get mixed together. The unique tag allows us to later re-identify which sequences belong to every person included

in the study.

While it is important that library preparation is performed accurately, the individual steps could be replaced by alternative methods. The crucial component is an end product of targeted DNA fragments that have been tagged appropriately to allow the sequencing chemistry on the chosen system and that fragment lengths are in the correct range. A more detailed summary of the procedure is outlined;

Preparation of sample

1. DNA is sheared, the most frequently used methods are by enzymatic digestion and sonication.
2. Fragmented DNA is purified using AMPure XP beads.
3. Quality assessment.
4. End repair.
5. Purify using AMPure XP beads.
6. Adenylation at 3' end.
7. Purify using AMPure XP beads.
8. Paired-end adaptor ligation.
9. Purify using AMPure XP beads.
10. Amplification.
11. Purify using AMPure XP beads.
12. Assess quality.

Hybridisation and capture

1. Hybridize capture library probes to DNA.
2. Capture the hybridized DNA using streptavidin-coated beads.

Note: at this step, custom gene target libraries can be used.

Indexing and multiplexing

1. Captured libraries are amplified with indexing primers.
2. Purify using AMPure XP beads.
3. Assess quality and concentration of indexed library DNA.
4. Pool samples at equal concentrations.

5.2.2 Capture library

For targeted sequencing experiments, the most important step in library preparation is the hybridisation of capture library probes. Libraries of probes that are complementary to exome coding sequences can be ordered from a number of commercial suppliers. For a whole exome, this consists of hundreds of thousands of short RNA oligonucleotide strands bound to biotin. When the capture library hybridisation mix is added to the DNA, most of the short probes bind to their complementary DNA sequences over 12-16 hours. To separate these selected fragments from the remaining bulk of unwanted DNA, streptavidin-coated magnetic beads are added. The streptavidin attaches to the biotin and therefore the DNA-bound probe can be pulled out using a strong magnet. Unbound DNA can then be washed away. Experiments in this study have been performed using Agilent capture library SureSelect Human All Exon V4-6, although several other options are available.

Targeted panels can also be used to focus on smaller sets of genes. For example, in some immunological conditions a panel of 50 genes might be targeted rather than a library for all known genes (exome). Cancer genetics screening services sometimes use a small panel of 40-100 genes. These small panels cut down on cost and focus only on genes where interpretation of variants would be possible. For the same price as whole exome, less capture library is needed and more samples can be sequenced.

As of 2018, all-exon capture library costs roughly £16,000 for enough reagent to prepare 96 DNA samples. This accounts for about 50% of the cost of the total library preparation materials. In total, the library preparation costs about £200 per sample. Once the samples have been prepared it cost about another £200 to sequence; approximately £400 total.

5.2.3 Sequencing

The sequencing carried out in this study was performed on Illumina platforms. These include the MiSeq for very small runs of a select set of genes, HiSeq 3000, 4000, and HiSeq

X for whole exome or whole genome sequencing. The prepared libraries of patient DNA are pooled to contain 5-12 samples per pool. Since each sample has a unique identifier tag, it is OK to pool them together and later separate out all the individual data per person. On the HiSeq 3000 approximately 12 samples can be run per lane with acceptable coverage. This provides about 30-50X reads per nucleotide, sufficiently deep to confidently identify true germline mutations. There are 8 lanes per sequencing flow cell. Therefore, a single sequencing run can contain anything from 50-100 patient samples. Depending on the sequencing platform the run can take up to 5 days to complete.

5.2.4 Ultra-deep sequencing

Mendelian disorders can be successfully explained using exome and whole genome sequencing. Both the interpretability and cost per sample are improved in cases where a gene sequencing panel can be used. Some conditions, particularly autoinflammatory disorders, can arise from low frequency somatic variants that are capable of driving disease through potent gain-of-function mechanisms. It is worth noting that a "gain-of-function" can also be considered as a succinct description for systems where a loss of inhibitory activity occurs that directly results in increased signaling cascade activity that would otherwise rest in an inactive state; a homeostatic pathway. E.g. loss of an autoinhibitory feature for a single protein or loss of an inhibitory mechanism that is responsible for direct repression in the absence of stimulation or specific agonist. In such cases, a low frequency de novo variant will escape detection with typical sequencing methods, but ultra-deep sequencing offers a method for detection. This option uses a high concentration of capture reagent to prepare a highly enriched library and sequence at high-density on a flow cell to produce ultra-deep sequencing reads (e.g. >5,000x versus 50x, as typical for whole exome sequencing). In this case, PCR-free preparation is ideal for somatic variant detection, naturally.

5.3 Genomic analysis

Like any data science, bioinformatics is a discipline of data manipulation. The majority of jobs could be accomplished simply with a method for sequence alignment and data mining using grep, sed, and awk. However, the development of specialised genomics-based tools allows us to standardise procedures and expand the avenues of exploration. One of the greatest single, collaborative, sources of genomics analysis tool is the Genome Analysis Toolkit developed by The Broad Institute.

While not every tool was used in this study, a synopsis of analysis options is worthwhile; an overview of GATK provides a good example of the current trends. The software provided by GATK includes methods for data manipulation. As of writing, there are 291 packages in this software suite. These are divided into major topics of genomic data handling that include:

- Tools dedicated to managing read data in SAM, BAM or CRAM formats.
- Diagnostics and QC to collect sequencing quality and comparative metrics;
- Interval manipulation to process genomic intervals in various formats. For example, converting a BED file to a Picard interval list;
- Metagenomics. For example, microbial community composition and pathogen detection using read filtering, microbe reference alignment, and abundance scoring;
- Tools that manipulate FASTA format references. For example, creating a custom capture library relies on oligonucleotide baits for hybrid selection reactions, or making BWA-MEM index image files, or a sequence dictionary to accompany a reference.
- Variant calling and genotyping for variants such as SNVs, SNPs, and Indels. For example, haplotype calling of germline SNPs and indels by performing a local re-assembly of haplotypes, such HaplotypeCaller gVCF files are generally merged into batches of single gVCFs to manage databases, and joint genotyping is a common

approach on these databases. Some tools also specialise in calling somatic SNVs and indels also by local assembly of haplotypes.

- Variant manipulation software for handling variant call format (VCF) data.
- Base calling. This is software that is used at the early stage of sequence data interpretation to process the raw data, i.e. base calls, and other attributes such as the adapters used.
- Read filters which can be applied by the engine to select reads for analysis.
- Variant annotations is a software that can be used during critical stages of analysis by other tools, i.e. HaplotypeCaller, Mutect2, VariantAnnotator and GenotypeGVCFs.
- Copy number variant discovery using read coverage to detect copy number variants.
- Coverage analysis using allele depths as the metric.
- Structural variant discovery.
- Variant evaluation and refinement. For example, variant calls can be further detailed using annotations which are not offered by the base software.
- Variant filtering that allows annotation of the FILTER column in a dataset.

5.3.1 Routine analysis

Routine analysis can be summarised in order as raw sequence data quality control, read trimming, reference alignment, subsequently followed by the GATK best practices for SNV and indels. **Figure 5.8** illustrates the basic analysis workflow structure. Proceeding top to bottom, the procedure making up the left side of fig. 5.8 contains the procedures for routine analysis. Each rectangle box labels a program function, key input and output data are shown with light slanted boxes. The most important data retention steps are indicated with a "data store" symbol. The right-hand side of the same figure illustrates the second phase of analysis used in this study; tailored analysis, or cohort-specific analysis. The annotation, filtering, and segregation of data here depends on the project. A generally useful strategy will output gene candidate data based on inheritance type

to produce individual datasets for each (i) functional heterozygous variants (including de novo, somatic, known dominant genes, etc.), (ii) homozygous only variants, and (iii) potential compound heterozygous variants, and (iv) a master version of all variants that have completed the filtering pipeline. These datasets are generally small (<1MB per individual) and combined into all individuals per sequence run or cohort.

Genome and exome analysis is an iterative process. Although there are routine steps, different methods will be used depending on each experiment. Data storage is a major factor in genetic analysis. Not only are the initial files large in size, many intermediate files are produced and may themselves be important to retain for a certain period. Key output files are shown by light slanted boxes. As shown in **Figure 5.3**, storage structure is divided between long-term and short-term storage. A /work/ directory is used for long-term storage and is backed up routinely. Short-term storage is used for intermediate files which are held in /scratch/ directories and not backed up. File sizes are represented by colour, dark orange indicating large and light yellow indicating small sizes.

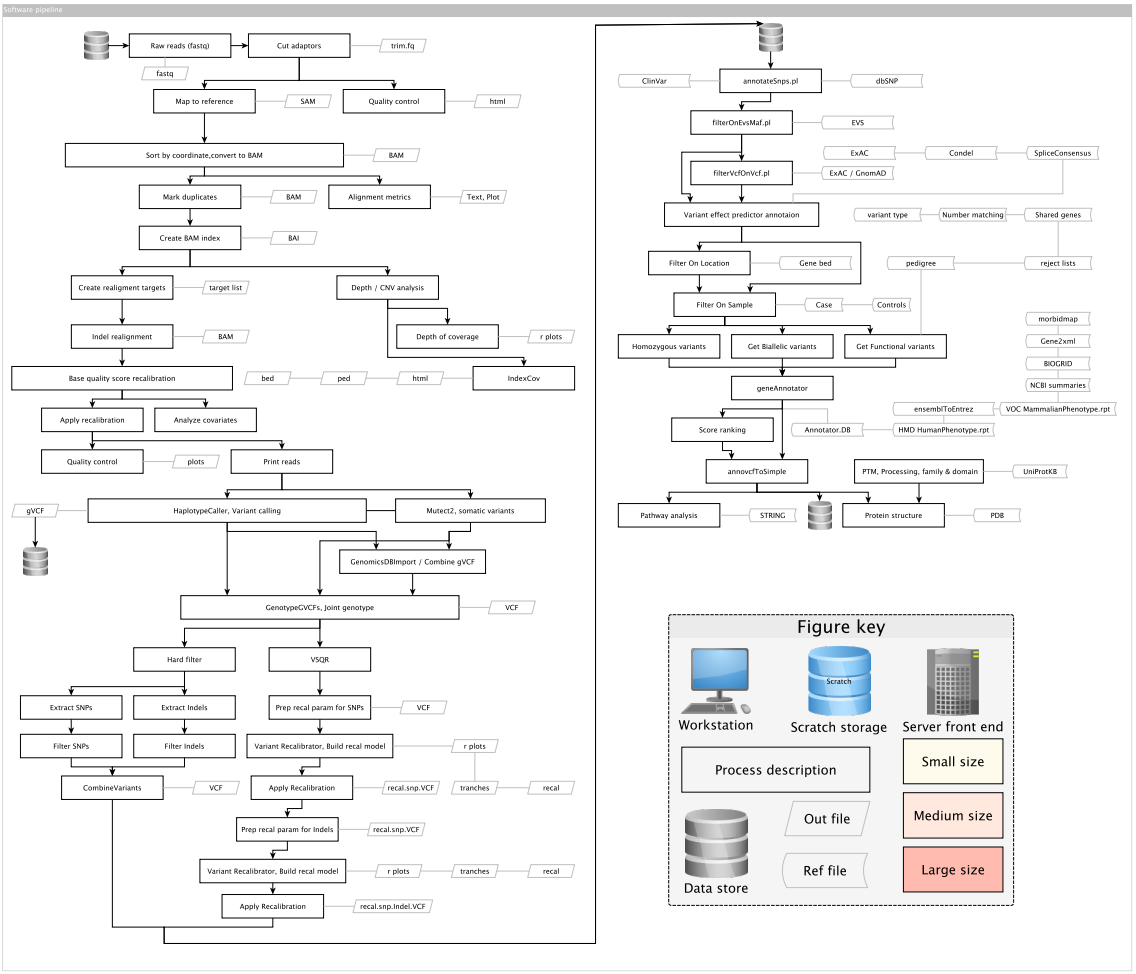


Figure 5.2: **Analysis workflow structure.** Tools used are shown in square boxes. Reference data used secondary to inputs are shown as light boxes with curved sides. Key output files are shown by light slanted boxes. Storage structure is divided between long-term and short-term storage. The same figure key is applied to Fig. 5.3.

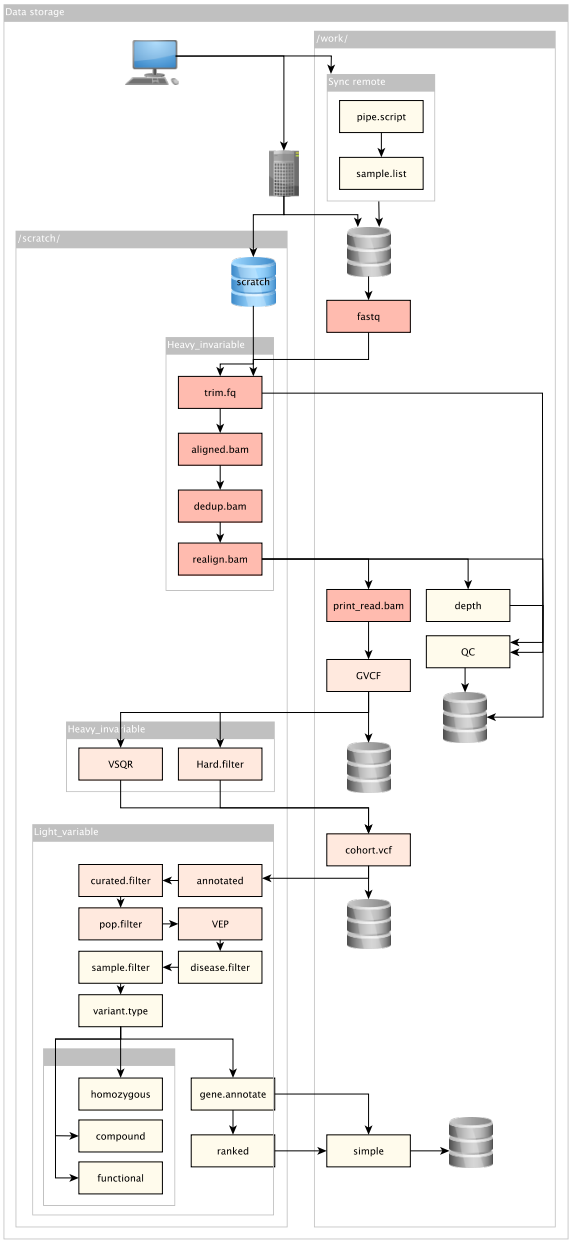


Figure 5.3: **Analysis workflow storage structure.** Storage structure is divided between long-term and short-term storage. A /work/ directory is used for long-term storage and is backed up routinely. Short-term storage is used for intermediate files which are held in /scratch/ directories and not backed up. File sizes are represented by colour, dark orange indicating large and light yellow indicating small sizes. Figure key is shown in Fig. 5.8.

5.3.2 Sequence alignment to reference genome

The analysis methods are normally run as a pipeline workflow. The basic methods do not have major changes in theory, although there are usually several methods or software options available for each step. Once a working pipeline is established, most of a researcher's time can be spent on the tailored analysis at the end of the pipeline, which requires more specialised steps. Each individuals' exome sequence data contains approximately 3-8 GB of raw data. This is output as 150bp raw unmapped sequence fragments that must be aligned to the reference human genome. The raw sequence data is normally collected into a fasta format file called a "fastq" file (pronounced "fast" "q").

An important consideration for sequence analysis is the reference genome used for comparison. The coordinates for individual nucleotides vary between reference versions. For example, aligning with one reference version will produce a file that contains chromosome, position, and variants specific to that genome reference. Annotation will be required to interpret results, but if databases based on coordinates from different reference versions are used during this step the results will be incorrect.

The current human genome reference is a version of Genome Reference Consortium Human Build 38 patch release 13 (GrCh38)
(https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39).
Because of the timing when next generation sequencing became popular, many researchers tend to use genome build GrCh37 in their analysis
(https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/)
However, it is preferable to use the more recent GrCh38. A lot of the best standardised methods that are used in the field were developed while genome build GrCh37 was the most recent version. Thousands of database samples will be in storage which have been aligned with this reference. Bioinformatic analysis is extremely more powerful when comparing many samples than when looking at one sample individually. Therefore, many people still tend to align their data to GrCh37 so that they can use their reference databases without going back and realigning all of their old samples again to GrCh38.

The most popular method for aligning short read data to the reference human genome

is "BWA-MEM" (a Burrows-Wheeler transformation aligner) [1]. BWA-MEM was used to align sequencing data in this study to GrCh37 (for an example usage see page 268).

5.3.3 Read adaptor trimming

Since Illumina-based sequencing technology relies on duplexed samples, identification sequence tags were added to all sequence libraries. During analysis these tag sequences can affect alignment and are therefore removed from each read . The command line usage is shown on page 267.

5.3.4 Read sorting

To allow downstream analyses to run efficiently, the sequences within files are rearranged based on their coordinate position after alignment with the reference genome. This process is carried out using SamTools This software is part of the The Broad Institute-maintained Genome Analysis Toolkit (GATK). Their standardised pipeline is illustrated here in 5.4; a protocol familiar to most bioinformaticians. And example of usage can be seen on page 268.

5.3.5 Read deduplication

Sequence library preparation may contain a PCR amplification step. Individual fragments of genomic DNA will be amplified. If a read contains a variant then, after amplification, we only want to count this occurrence once so that we do not interpret an inflated allele depth. Therefore, identical reads are marked as duplicates. Alternative overlapping reads that also contain the same variant will result in detection of a true germline variant. When no other overlapping reads contain the variant then the allele depth will remain low and be filtered out later by a frequency threshold, or flagged as potentially somatic. For command line usage examples of this step, see page 268.

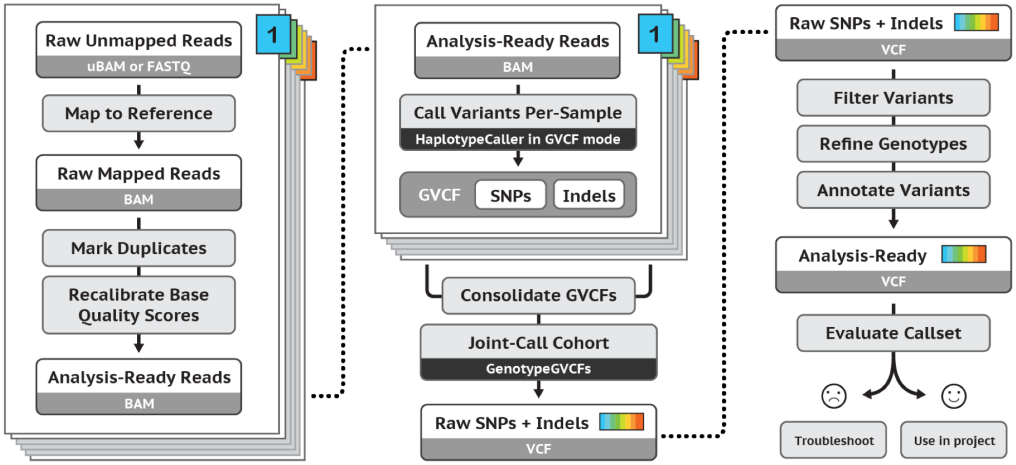


Figure 5.4: **GATK best practices.** Illustration from software.broadinstitute.org. Per-sample variant calling is used to produce a file in GVCF format. GVCFs are consolidated from multiple samples into a GenomicsDB datastore. Joint genotyping is carried out, and finally, variant quality score recalibration filtering is used to produce the final multi-sample callset with the desired balance of precision and sensitivity. Further downstream analysis, including annotation is not shown.

5.3.6 Read realignment and targets

After sequence alignment, regions of misalignments will inevitably exist. To deal with this feature, a local realignment process is used such that the number of mismatching bases is minimized across all the reads. This main source of misalignments corrected in this step are due insertions and deletions. Current versions of the GATK suite no longer require this step as it is integrated into the downstream process of haplotype assembly (via HaplotypeCaller or MuTect2). However, the step is included here since it is a well known legacy feature and is a very useful concept to understand for new users. As usage example is provided on page 268.

5.3.7 Base quality score recalibration

The alignment steps are difficult and computationally intensive. There are methods to double check the alignment and see if more appropriate corrections can be made. Once the quality control is all done, we are left with a Bam file format which is ready for variant analysis. Most of the bioinformatic community agrees on some best practices

using the tools maintained by the Broad Institute. The GATK is widely used for the QC and variant analysis of genomic data.

Joint analysis of multiple samples increases the accuracy of our methods. Not only are the algorithms checking for consistencies in the data, but sometimes the sequence library preparation induces errors in the sequences produced. For example, sometimes a particular nucleotide position can be sequenced incorrectly. In isolation we would expect that this patient has a true mutation in the gene, but when we compare the whole cohort we see that it is just a common sequencing artefact.

When we look at the number of variants compared to the reference genome there can be hundreds of thousands. The vast majority of these can be ignored by [1] comparing the in-house database of false positive, [2] comparing the unrelated samples sequenced on the same run to remove library preparation errors, [3] compare to databases of common polymorphisms.

In genome wide association studies, researchers are generally looking at the mild effects of common polymorphisms which occur in the general population and may associate with a particular phenotype. In rare disease analysis we are focusing on the very rare variants that have a strong effect to produce a severe phenotype. Therefore, another step for pruning out the data is to compare to large cohorts of "healthy" populations to leave only the very rare variants in our dataset. The command line arguments can be see on page 269.

5.3.8 Haplotype calling

The final output, illustrated in the GATK best practices figure above, is stored in a Genomic Variant Call Format (GVCF). The GVCF file type that now presents our data has one row for each nucleotide along the genome. The row contains the DNA position, the nucleotide (either wild-type (ref) or mutation (alt)) and lots of quality and metrics information. We analyse variants against curated databases of known mutations. We also analyse again separately for indels, since a shift in the sequence position due to a indel could affect the alignment accuracy. For an example see page 270.

5.3.9 Cohort joint genotyping

We can merge 10-100s of samples together by combining the files to simplify how we handle the data. Tracking hundreds of files is exponentially more difficult than tracking 1. The GVCF contains a row of data for every single nucleotide. We can condense the information by converting to a VCF which instead only keeps information for every variant but not every wild type nucleotide (since wild type is healthy and of no interest to us). The GATK documentation provides a great explanation of the shared features and differences between gVCF and VCF files.

As our dataset becomes smaller we can double check to focus on just the most likely disease-causing mutations. Often times, a research group or clinical research team will collect genetic material from patients who they would like to diagnose genetically, or even collect a great database of patients with a shared phenotype. There are many of facilities that will sequence the samples commercially. When one orders exome or whole genome sequencing commercially, most facilities will also provide data analysis.

The output of their analysis is usually this VCF file (mostly contain the chromosome, nucleotide position, and a selection of quality control information). This file is usually the end-point of routine analysis. However, it does not really put one in a position for a genetic diagnosis. Very good services will also provide lists of top candidate genetic determinants along with information on each of the genes and possible mechanisms of pathogenicity (although the number of companies doing high-level tailored analysis is small but growing). There are usually more hurdles in determining candidate variants of unknown significance. An example of the command line arguments used can be found on page 270.

5.3.10 Tailored analysis

Routine analysis typically takes up to a week, although it is usually performed on a standardised pipeline that can run automatically on a high-performance computing platform. A large part of custom filtering begins when the routine analysis steps have

been completed; downstream analysis is adapted for each particular challenge. The discussion in chapter 2 explains some foundational steps towards a fully automatic system that relies only on some input features, such as clinical information. While many software packages exist that claim to output tailored analysis, these tend to either only tackle a specific niche or require lots of curated auxiliary input data.

The output of non-routine analysis (outlined in this chapter) sometimes takes only five minutes to interpret a cause of disease. In other cases, data that has been sequenced years previously has not yet relented an explanation for phenotypes that almost certainly should be explained by coding variants present within the sequence data. For example, for a dozen patients who share a similar and severe phenotype, it is often likely that the same gene or related genes could cause their disease. Unrelated patients with a rare dominant disease are not to all carry the same disease-causing variant; they may have different variants in shared gene, or variants among different genes which all contribute to a shared pathway that would result in the same end-point phenotype.

For example, in **Figure 5.5** above we see that from a group of unrelated people, all of the candidate genes carrying functional variants are joined by their shared functional interactions. For an autoinflammatory phenotype, genes like *NLRP3*, *NOD2*, *TNFAIP3*, *MyD88*, *IKBKB*, *FASLG*, or *TMEM173* might all have different functions but damaging mutations in any of these could result in phenotypes that, on the surface, appear related.

Another circumstance might be seen in a small cohort of patients with a shared autoinflammatory phenotype. For example, the gene *NLRP7* has relatively few publications examining it's role in autoinflammatory disease. One would not consider this a strong candidate gene if faced with a variant of unknown significance in this gene from a single patient. However, three or four very rare or novel mutations in unrelated patients should be given consideration as producing an autoinflammatory disease. Single case, or small cohorts lack the power to measure significant associations. Therefore in the situation proposed here, manual interpretation is required (biased as it may be).

NLRP7 variants not reported as producing disease, like *MEFV*, or *TNFAIP3*. However, we must consider that genes plausibly responsible for causing disease in a dominant manner

and that are highly conserved are generally under purifying selective pressure. Damaging mutations may be not be compatible with viability and therefore we never see cases of disease. Variants which are damaging to protein function but that do not completely destroy all of the normal structure may produce a phenotype that is pathogenic but viable with modern medical intervention.

In the example of NLRP7, the protein is known to Inhibit CASP1/caspase-1-dependent IL-1β secretion. The functional domains of this protein are shared in other pro-inflammatory processes. Pyrin, NACHT, and LRR, domain variants are all studied for autoinflammatory diseases. The related gene, *NLRP3*, is probably the most widely recognised gene where damaging variants in these functional domains produce severe immune disorders. In cases where we have protein structures, we can also model the effect.

In our example, *NLRP7* variants have been reported as the genetic determinant of a condition that causes early neonatal death and ectopic pregnancy. Many of the reported variants are stop mutations that will either produce a truncated protein or prevent expression of the allele altogether through nonsense-mediated decay. It is difficult predict the mechanism of disease in cases like this where the two outcomes have opposing paths. That is to say, a truncated protein may have an active functional domain which can no longer be inhibited since the C-terminal domains are missing, while haploinsufficiency would mean that cells cannot perform their normal function for the pathway since 50% of the protein is depleted (in heterozygous cases). Haploinsufficiency can result in a disease that phenotypically resembles a gain-of-function when the responsible protein normally acts as an inhibitor for an inflammatory pathway [2]. This is not expected with *NLRP7* and therefore heterozygous loss-of-function does not explain disease.

For a candidate gene like this, we have some plausible evidence but cannot really progress any further without new functional studies. The first step involved confirmatory Sanger sequencing for all patients identified through exome sequencing. Next, any close relatives that are available might be also sequenced for the same variants. If the mutations are disease-causing then other carriers would also be expected to have some shared phenotype features. The possibilities in functional experiments vary widely and

are highly dependent on the candidate genes. The procedure outlined in this hypothetical example is generally applicable in for the majority of single-case studies and illustrates the importance of tailored analysis. The initial findings of genomic analysis may produce more follow up questions, including whether other probable gene candidates can be ruled out, for which the patient carries only the "normal" reference alleles (e.g. *CFTR* screening for cystic fibrosis/lung disease).

5.4 Integrating databases

5.4.1 Population genetics

GnomAD (version r2.0.2) [3] was used in these studies as the best source of population genetics data. The reference genome is GRCh37. Offline local database mirrors were used in most cases. Input sets used GnomAD variant allele frequencies and reference sequences processed as VCF and CSV files. chapter 2 outlines a specific data transformation using the gnomAD database, but in general, gnomAD was used as a filtering threshold for determining the expected population frequency of each variant. A strict threshold for rare variants could be set to ignore and candidate variants that are more frequent than 0.001. However, in most cases a more lenient level is used and any remaining benign or common variants are removed by "technical control" (filter on cohort to remove common variants between individuals that do not share a phenotype). A more modest cut-off threshold allows us to sometimes identify variant that are present in the general population, which are responsible for a recessive disease with no predictable heterozygous loss-of-function intolerance.

Other sources of population genetics data comes from resources such as ClinVar and dbSNP, which as they grow in size become an annotated and curated for of population data. These resources allow us to calculate the expected frequencies for disease-causing variants. However, since these are manually curated database and predominantly European based, they are inherently biased and not reliable for statistical applications.

5.4.2 Phenotype, genotype, and function

Population genetics database gnomAD has been individually addressed in section 5.4.1, as this is the most important type of annotation and filtering criteria for genetic determinants of rare disease. Additionally, in these studies many phenotype and genotype databases have been used for annotation and interpretation. Specifically, the most frequently used data came from MGI Phenotype, MorbidMap, VOC MammalianPhenotype, Gencode symbol, UniProtKB, Enterez ID, ENSGene ID, GO ID, Description, OMIM, BIOGRID interactions, HGMD human phenotype, ClinVar, and dbSNP. In most cases, every candidate variant was annotated with the main information per gene from a local database containing the information from each of the listed resources.

These are the "basic" information databases that we used to annotate variants. In a cohort study, data mining can find correlations and was therefore included for posterity as it does not significantly increase the data storage. Even if an obvious cause of disease was found we may later return to the data to find other cofactors or genetic modifiers. Or for example, in a single case study, a variant of unknown significance may have no statistical basis to be selected or ignored. We use this information to decide if that mutation is worth consideration: Is it in a protein domain of known function? Are there other cases reported with the same phenotype? What is the gene function, ontology, etc.?

We have also used some gene lists that are specific to disease, druggability, etc. A major contributor for collecting these gene lists has been the Mac Arthur et al. [4]. These gene lists can be used is special cases. For example, a study looking at (1) dominant pathogenic mutations, and (2) in known immune genes might filter to included only those known observables. We could decide to only study SNPs in FDA-approved drug targets.

Table 5.1: **List of gene lists.** An example of genelists that are used for tailored analysis. Originally compiled by [4] (*CRISPR screening studies).

Gene List	Gene Count	Reference
Universe	19,194	HUGO 2018 [5]
FDA-approved drug targets	385	Wishart 2018 [6]
Drug targets	201	Nelson 2012 [7]

Autosomal dominant genes	307	Blekhman 2008 [8]
Autosomal dominant genes	631	Berg 2013 [9]
Autosomal recessive genes	527	Blekhman 2008 [8]
Autosomal recessive genes	1073	Berg 2013 [9]
X-linked genes	66	Blekhman 2008 [8]
X-linked recessive genes	102	Berg 2013 [9]
X-linked dominant genes	34	Berg 2013 [9]
X-linked ClinVar genes	61	Landrum 2014 [10]
All dominant genes	709	Blekhman 2008, Berg 2013 [8, 9]
All recessive genes	1183	Blekhman 2008, Berg 2013 [8, 9]
Homozygous LoF tolerant	330	Lek 2016 [3]
Essential in culture	283	Hart 2014 [11]
Essential in culture*	683	Hart 2017 [12]
Non-essential in culture*	913	Hart 2017 [12]
Essential in mice	2,454	Blake '11, Georgi '13, Liu '13 [13–15]
Genes nearest to GWAS peaks	6,336	MacArthur 2017 [16]
DNA Repair Genes	178	Wood 2005 [17]
DNA Repair Genes	151	Kang 2012 [18]
ClinGen haploinsufficient genes	294	Rehm 2015 [19]
Olfactory receptors	371	Mainland 2015 [20]
Reported in ClinVar	3078	Landrum 2014 [10]
Kinases	347	UniProt 2016 [21]
GPCRs from guide to pharmacology	391	Alexander 2017, Harding 2018. [22, 23]
GPCRs from Uniprot	756	UniProt 2016 [21]
Natural product targets	37	Dancik 2010 [24]
BROCA - Cancer Risk Panel	66	BROCA Cancer Risk Panel [25]
ACMG V2.0	59	Kalia 2017 [26]

GPI-anchored proteins	135	UniProt 2016 [21]

Verma et al. [27] take an interesting approach to comparing druggable targets with population genetics data. DrugBank is a database for over 800 genes with over 950 unique drugs. Genetic data can be filtered for these genes and targeted for LoF variants. Association analysis consists of logistic regression using the ICD-9 codes, and linear regression using quantitative variables. This gene binding and regression analysis steps are done using BioBin.

The International Statistical Classification of Diseases and Related Health Problems (commonly known as the ICD) provides alpha-numeric codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or disease. Nearly every health condition can be assigned to a unique category and given a code, up to six characters long. Such categories usually include a set of similar diseases

BioBin relies on the Library of Knowledge Integration (LOKI), which integrates multiple databases providing a comprehensive biological knowledge platform for variant binning [28]. The LOKI database consolidates biological information from several sources, most notably the National Center for Biotechnology (NCBI) dbSNP and Entrez Gene, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Gene Ontology (GO), Protein families database (Pfam), NetPath-signal transduction pathways, amongst others [29–34].

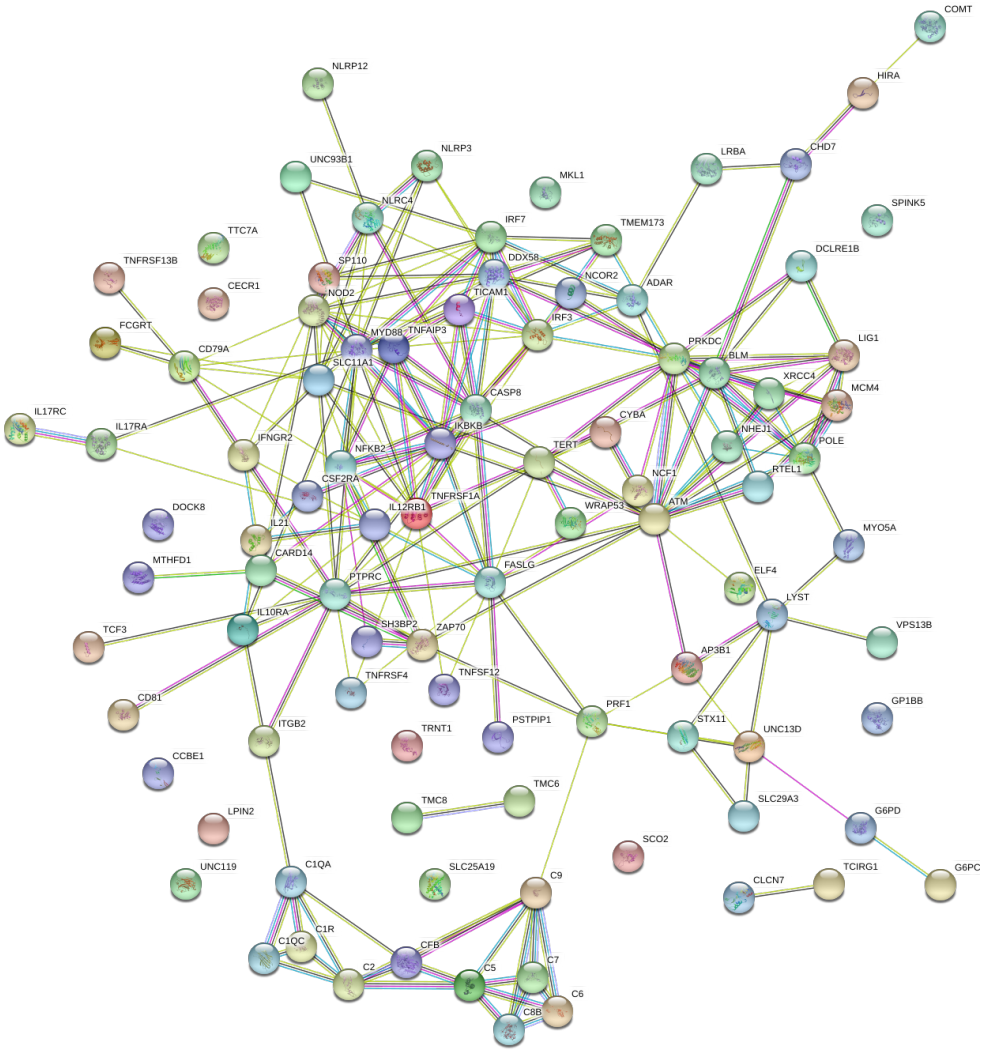


Figure 5.5: **Shared pathways in candidate genes.** Output of a STRING database query for known and predicted protein-protein interactions. These interactions include physical and functional associations. (STRING is an SIB-run service of the ELIXIR core data open resources for publicly funded research).

5.5 Rare disease cohort network analysis

5.5.1 Introduction

The exome sequencing is most commonly used for genetic diagnosis in single use cases. Over the next decade exome and genome sequencing will become very commonplace for the average person at least in high-GDP countries. A massive expansion in population genetics data will provide the information that GWAS studies have always sought. We will still be left with large genomic black holes; that is invariant coding and non-coding regions where rare variants act as genetic determinants of disease and where carriers are either rarely found or non-viable with life. To uncover the function of these genetic loci we will still be at the mercy of cohort size in rare disorder studies. For true rare genetic disorders, a disease frequency of 0.01% equates to approximately 800,000 cases worldwide for diseases that are not embryonic lethal and where lifespan is about normal. If we consider high income populations where genomic sequencing would be common, then we may have 100,000 cases. However, even with a potentially large pool of candidate cases there are multiple reasons why genetic studies can fail to find the cause of a Mendelian disease. Organisation of large rare disease studies is a complex task. Adequate recruitment may not be possible. For candidate cases, it may be impossible to clearly separate overlapping phenotypes. Therefore, now and well into the future, rare disease studies will generally be limited to a maximum number of living participants on the scale of hundreds.

Current best practices in genomic analysis will first identify "low hanging fruit"; single cases in a cohort with a clear genetic determinant (e.g. haploinsufficiency of a well-defined dominant gene). The second order will identify commonly mutated genes or loci based on burden testing comparing cases to controls or background population genetics. Many disorders have a phenotype that can be derived from mutations in several different genes. The encoded genes generally are a part of the same protein pathway, even directly upstream and downstream of each other in some cases. For example, chapter 1 covers this topic with individual cases of RAG1 and RAG2 deficiency.

Proposed here is a statistically robust and unbiased method to find variants in protein-

coding genes that share a common functional protein pathway for a disease cohort. **Figure 5.6** provides a high-level graphical summary of the concept. **Figure 5.7** conveys the theory of the procedures for this method in more detail with the major datasets explicitly shown.

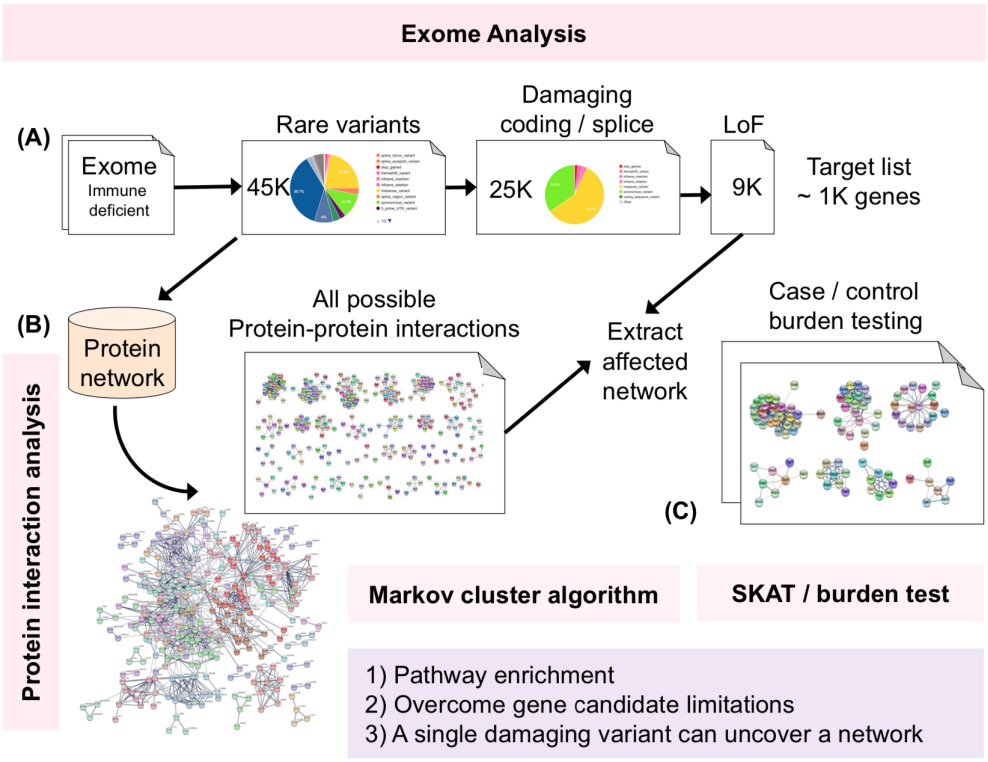


Figure 5.6: **Deleterious rare variants in damaged protein pathways in rare disease.** A. GATK best practices were used for whole exome analysis with joint genotyping for cases and controls; 200 in total. Custom filtering [35] extracted variants of high impact consequence (ostensibly loss-of-function (LoF)), present only in cohort cases. B. Genes harboring rare predicted LoF variants were grouped based on protein-protein interactions [36] using a Markov cluster algorithm [37]. C. Case-control testing was performed on each protein pathway cluster.

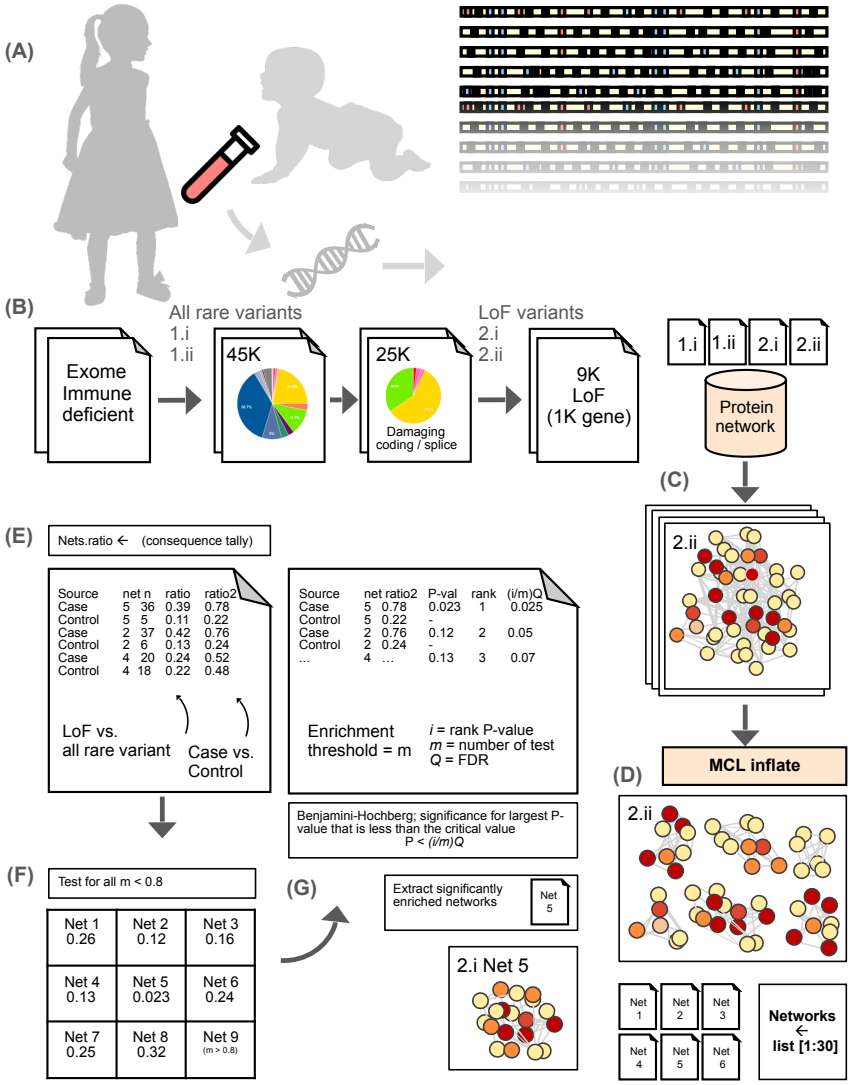


Figure 5.7: **Rare variant analysis and protein pathway significant enrichment.** A. DNA is collected and sequenced. B. Routine genomic analysis is carried out according to best practices, for both (i) control and (ii) case groups of patients. First, all rare variants are output, followed by a smaller subset of loss-of-function (LoF) variants. C. Genes harbouring functional variants were clustered based on their respective protein-protein interactions according to STRING DB, including function and ontology. D. A clustering method is applied to break a large highly connected network into smaller individual ones. E. The number of tests can be reduced by, for example, testing only networks that carry a threshold level of LoFs and are therefore biologically relevant to disease. F. Deleterious variant load per network was tested for enrichment in cases, controls, or random sampling. G. Multiple testing correction is applied to identify the critical significant threshold.

5.5.2 Exome analysis

Exome sequencing analyses has been discussed in detail. The rare disease cohort network analysis requires less tailored analysis steps than traditional variant interpretation. Therefore, the data preparation is briefly outlined here.

Sequences were trimmed and quality controlled using FastQC via Trim-galore. Reads were aligned to GrCh37 using BWA-MEM. GATK "best practices" were used for marking duplicate reads, recalibration, and whole cohort variant quality score recalibration before generating genomic VCFs with HaplotypeCaller and joint genotyping. Filtering and prediction of functional consequences was performed using Variant Effect Predictor (http://www.ensembl.org/info/docs/tools/vep/index.html), Exome Variant Server (http://evs.gs.washington.edu/EVS/), The Single Nucleotide Polymorphism database (https://www.ncbi.nlm.nih.gov/projects/SNP/) and ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), The Exome Aggregation Consortium and The Genome Aggregation Database (http://gnomad.broadinstitute.org). Filtering of common variations and annotation was performed using VCFhacks (https://github.com/gantzgraf/vcfhack). Candidate variants were required to pass the following filtering conditions: frequency (count/coverage) between 20-100%, according to VEP-annotation at least one canonical transcript is affected with one of the following consequence: variants of the coding sequence, frameshift, missense, protein altering, splice acceptor, splice donor, or splice region; an inframe insertion or deletion; a start lost, stop gained, or stop retained, or according to VEP an GnomAD frequency unknown, <=0.01, or with clinical significance 'path'. VCFhacks [35] was used for cohort-specific filtering retained functional variants that were present in at least one case but absent in controls (for case-driven PPI clustering). The same criteria were used to also collect functional variants that were present in at least one control but absent in controls (for control-driven PPI clustering).

5.5.3 Cluster list preparation

Group-specific variant data was extracted from the joint cohort. Specifically, the datasets came from the routine analysis pipeline show in **Figure 5.8** as the output of the process "filter on Sample" and converted from VCF to tsv format using the process "annovcftoSimple" using the tool VCFhacks [35]. Four gene lists were prepared consisting of the following groups; (1) variants present in controls and (2) variants present in cases and further divided for genes that harboured either (i) all rare variants or (ii) only potential loss-of-function variants. Specifically, the datasets for (i) all rare variants came from the output of the process "filter on sample" via "get functional variants". The datasets for (ii) potential loss-of-function variants is a subset of (i), processed by the R script at the step where "damaging variants" are written out to file.

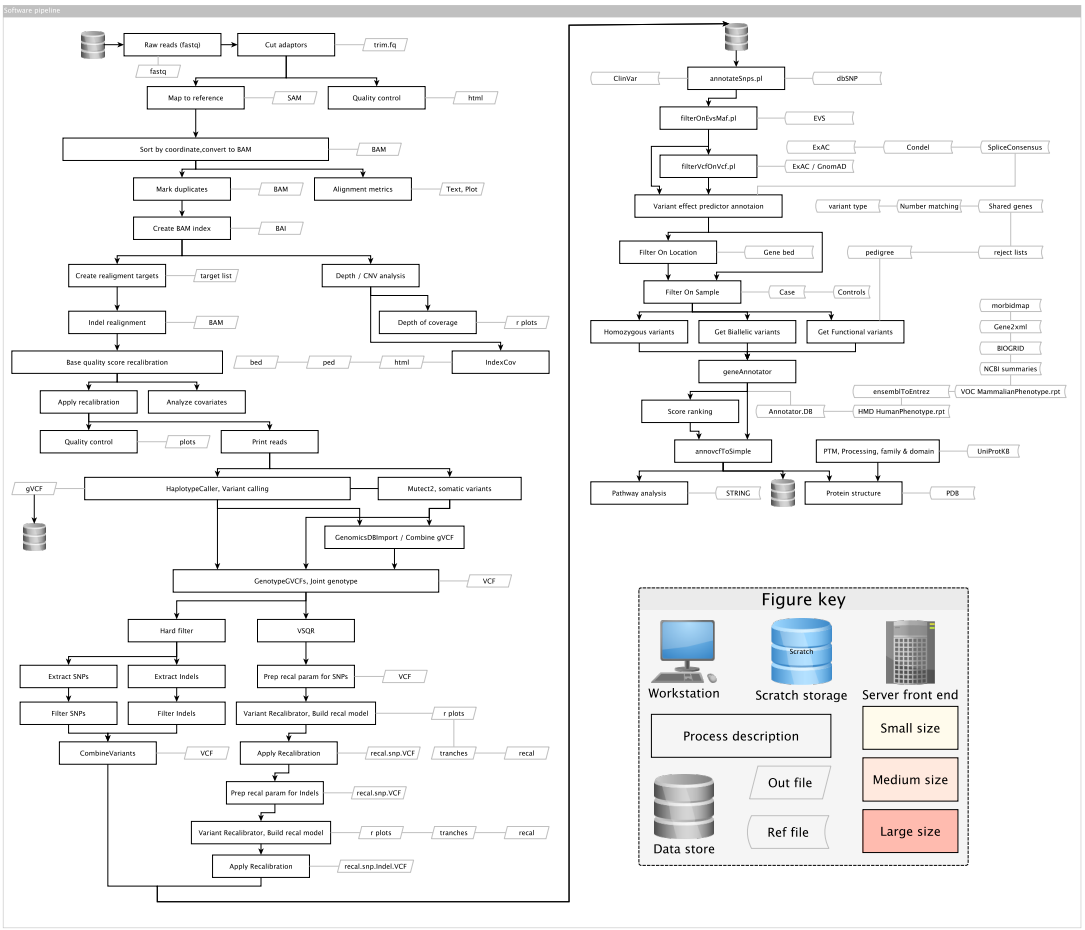


Figure 5.8: Analysis workflow structure. Tools used are shown in square boxes. Reference data used secondary to inputs are shown as light boxes with curved sides. Key output files are shown by light slanted boxes. Storage structure is divided between long-term and short-term storage.

5.5.4 Network construction

Group-specific gene lists [1 (i-ii) and 2 (i-ii)] were assessed for PPI using the STRING database [36] via Cytoscape [38]. An initial PPI network was generated for each of the 4 dataset groups. The STRINGdb default confidence score cut-off (0.4) was used for these tests. This score is the measure of evidence required to create an interaction between two nodes. A stricter value can be set if networks are too large. Query genes were defined as nodes, PPI were defined as edges, and networks of proteins linked through PPI were defined as clusters. Clusters or networks can also be generally considered as making up a part of a protein pathway.

Table 5.2: PPI for protein-coding genes harbouring potential LoF rare variants prior to clustering into pathway-specific networks. Query genes were defined as nodes, PPI were defined as edges, and networks of proteins liked through PPI were defined as clusters. The majority of query proteins group into a single large, weakly connected network cluster. PPI; Protein-protein interactions.

	Network cluster	Number of nodes	Number of edges	Number of clusters
Cases	Total	1956	9559	114
	No edges	1	0	107
	One edge	2	1	6
	Large multi-edge	1837	9553	1
Controls	Total	2305	14139	102
	No edges	1	0	77
	One edge	2	1	3
	Two edges	3	2	1
	Large multi-edge	2219	14134	1

Table 5.2 lists the characteristics of PPIs for genes found to harbour functional, potential LoF rare variants in cases and controls (i.e. gene lists 2 [i-ii]). Most query

proteins were seen to cluster into one large multi-edge node which contained many weak interactions. The data used in this table is represented again visually in **Figure 5.9**. Each dot, or node represents a protein-coding gene that has at least one potentially damaging variant. The edges, or lines connecting nodes, represent known PPI data that link proteins. This visual information clearly illustrates the body of functional protein data that can be included in variant analysis. However, since nearly every protein has some potential evidence of effect on many other proteins, then no clear definable protein pathway can be seen.

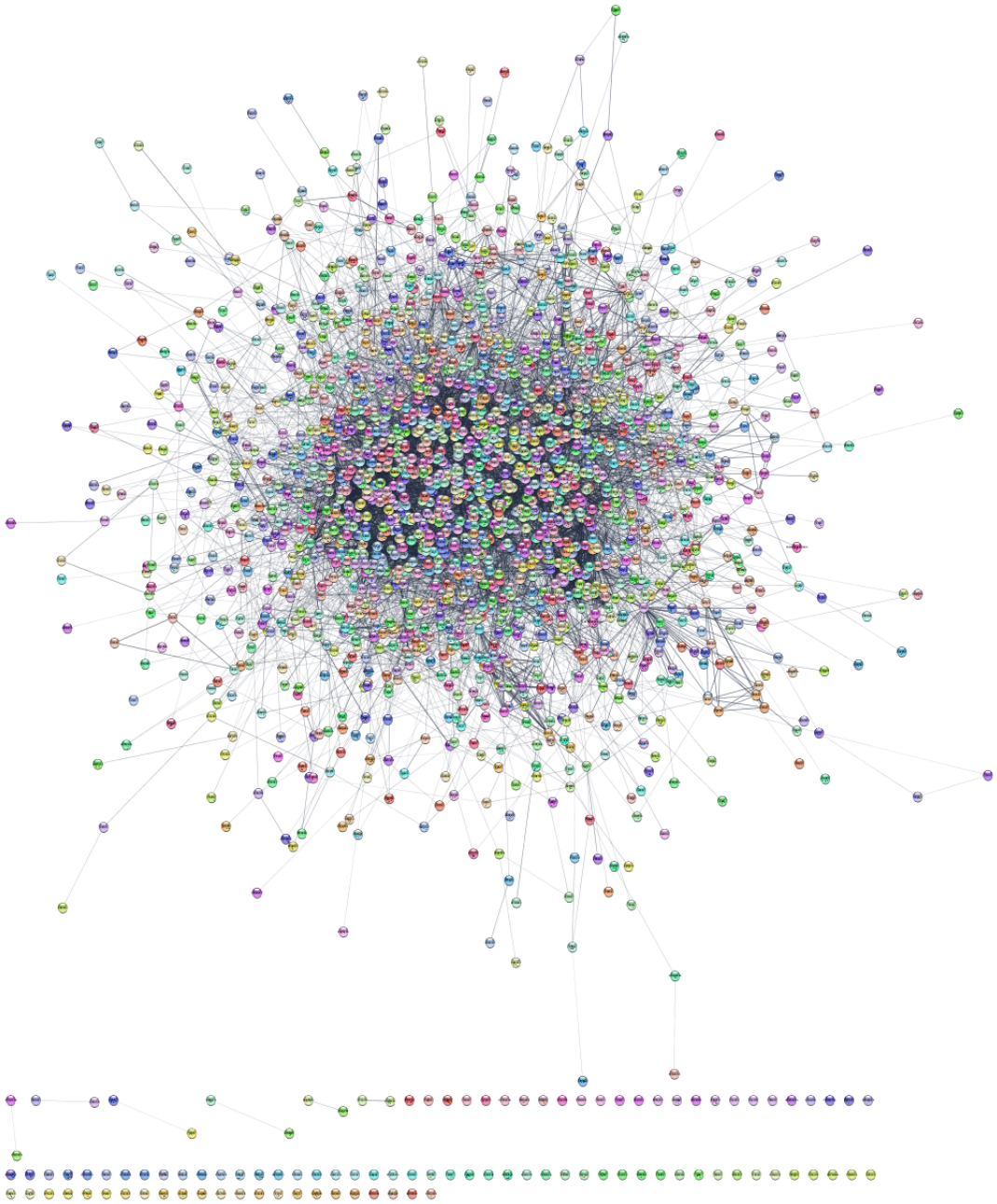


Figure 5.9: **Genes harbouring potentially damaging variants in a disease cohort.** A visual representation of PPI occurring in all genes that harbour potentially damaging functional variants in a typical disease cohort.

To segregate protein pathways and refine the number of genes (nodes) in each cluster, the Markov cluster algorithm (MCL) was used [37, 39]. The principal data-specific adjustment required for using MCL is the inflation operator, which regulates cluster granularity or tightness. The optimum inflation parameter for separating protein pathways was found to be 2.5, using a measure of uniform distribution across datasets. **Figure 5.10** illustrates an optimal inflation of a large PPI network into smaller, clearly defined protein pathway clusters.

As a reference example, **table 5.3** lists three inflation parameters tested for most consistent separation (2.5, 3, 4) and shows the effect of adjustment on the total number of edges (protein interactions). The median number of nodes (query proteins) are shown for cases and controls (also shown as total number of nodes in **table 5.2**).

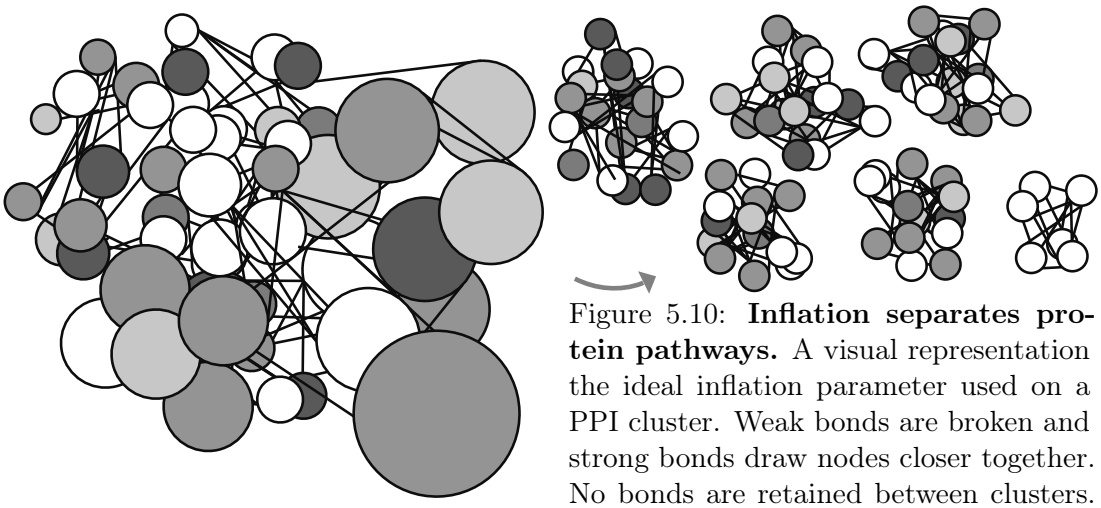


Figure 5.10: **Inflation separates protein pathways.** A visual representation the ideal inflation parameter used on a PPI cluster. Weak bonds are broken and strong bonds draw nodes closer together. No bonds are retained between clusters. With this type of inflation each protein network cluster can be investigated without considering overlaps.

Table 5.3: **PPI cluster size and Markov cluster algorithm inflation.** The result of using three different inflation parameters are shown. Horizontally first the number of nodes (proteins queried for their potential interactions) is shown. Vertically the number of edges (PPI or connections between protein nodes) are shown. As inflation parameters change, first poor cluster separation is seen, then an excess of very small networks is made (i.e. inflation 4 creates many networks consisting of only 1-2 nodes with only 1-2 edges). * The PPI only before inflation are illustrated in figure 5.9.

		Total count median Case/control ± S.D.	Node/Edge ratio Case/control ± S.D.
Number of nodes		2130.5 ± 246.78	
Number of edges	PPI only *	11849 ± 3238.55	0.18 ± 0.03
	Inflation 2.5	2787.5 ± 740.34	0.78 ± 0.12
	Inflation 3	4229.5 ± 3669.18	0.77 ± 0.61
	Inflation 4	1199.5 ± 146.37	1.78 ± 0.01

Figure 5.11 illustrates the effect of adjusting the inflation parameter for MCL clustering on protein networks. After MCL clustering, cases and controls were found to group into 928 and 1034 networks clusters respectively. Of these, 494 and 568 were single-node (single-protein) "clusters" which shared no interaction with another protein while 434 and 466 clusters had at least one interaction between proteins. The cumulative probability plot (figure 5.12) shows the cumulative sum of proteins per network against network rank size. **Figure 5.13** shows qqplot for the same data for distribution compared between groups after inflation at 2.5.

Figure 5.14 shows the number of proteins per network. For example, 235 clusters (470 protein nodes) were seen for cases where only one interaction was shared between two proteins. A median of 0.78 nodes-per-edge (proteins-per-interaction) was found in the cases group; naturally the majority of edges appear in large network clusters and therefore the frequency of nodes-per-edge increases as network sizes decrease.

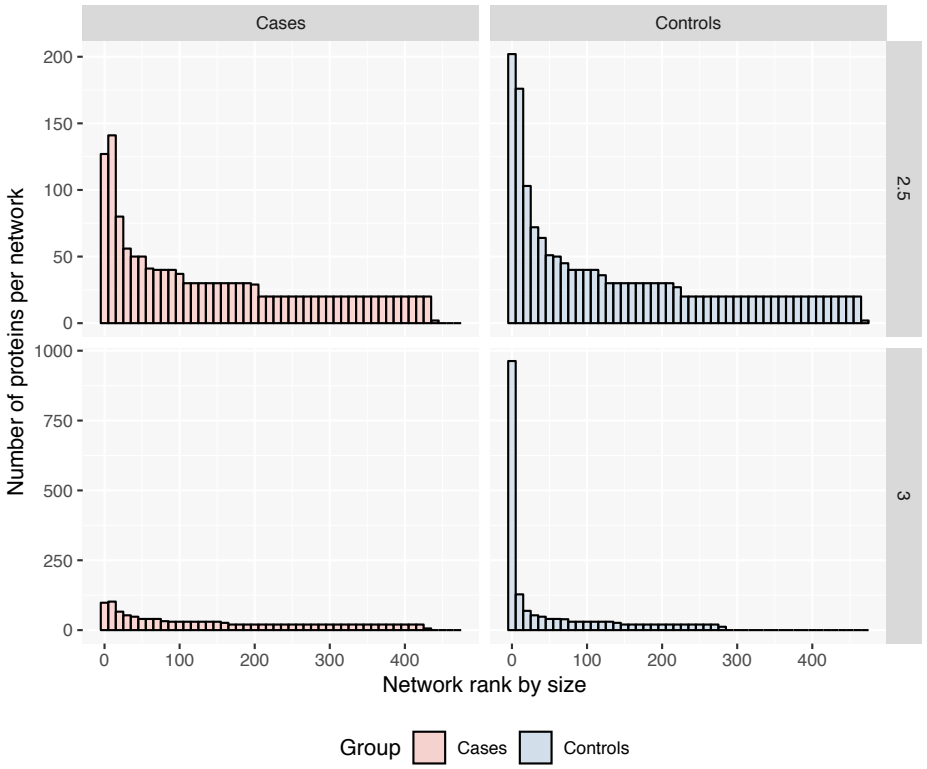


Figure 5.11: **Effect of inflation on network size distribution.** The outcome on network size is demonstrated to compare effect of two inflation parameters. An ideal separation of networks should provide an geometric decrease in the number of proteins per network regardless of the sample group. Inflation parameter 2.5 produced the ideal distribution while inflation parameter 3 produced one large, poorly separated network and a large increase in single-protein nodes on one group. Binwidth of 10.

5.5.5 Random sampling

With our group-specific gene lists [1 (i-ii) and 2 (i-ii)], prepared in section 5.5.4, we found the distribution of genes per networks and output the list of genes in each network for all 4 datasets. The mean number of genes per network rank was found between cases and controls, again for (i) all rare variants and (ii) only potential loss-of-function variants. A third gene cluster list was produced by random sampling gene symbols in artificial networks equal to the same median size as case-driven and control-driven clusters in from datasets (i) and (ii). The resulting dataset [3 (i-ii)] mirrors those of cases and controls but instead of true PPI networks, the networks contained randomly assigned genes.

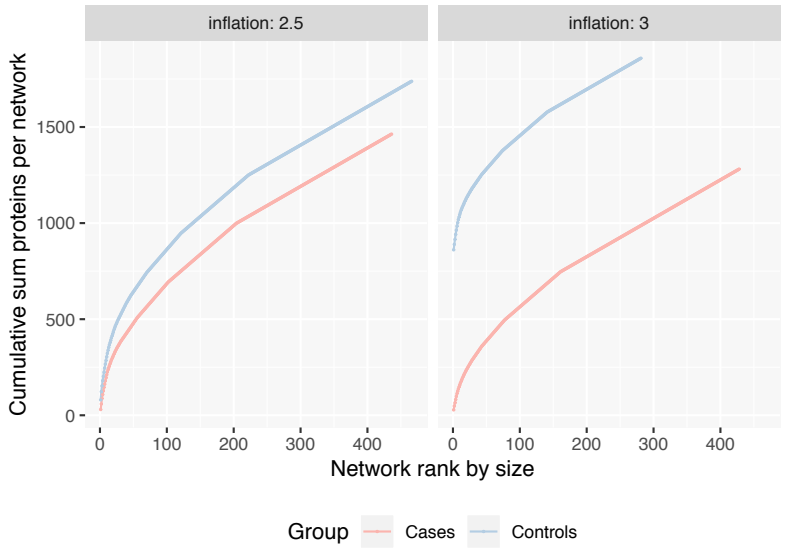


Figure 5.12: **Cumulative sum of network rank by size**. The effect of inflation on network size distribution could be potentially measured automatically by quantifying the cumulative sum of network rank by size and determining the best inflation parameter to use. This process would reduce user bias.

5.5.6 Expanding damaged gene MCL clusters

For each of the 4 MCL-clustered datasets, cases and controls [1 (i-ii) and 2 (i-ii)], the cluster ID and list of gene symbols was extracted. The gene lists of network clusters made from datasets (ii) (potential LoF) were used to find the network clusters in (i), all-variant gene clusters, the contained the same overlapping genes. This occurs where list (ii) is a subset of list (i). The clusters that contained gene overlaps were extracted since they contained at least one potential LoF per network. Using this output, the total variant load in "damaged pathways" could be compared. For clarity, this procedure is summarised again in Box 5.5.6; items **A-B**. Item **C** outlines the remaining steps. **Figure 5.9** illustrates the effect of inflation with an ideal inflation parameter. The large network of PPI were separated into individually contained protein networks.

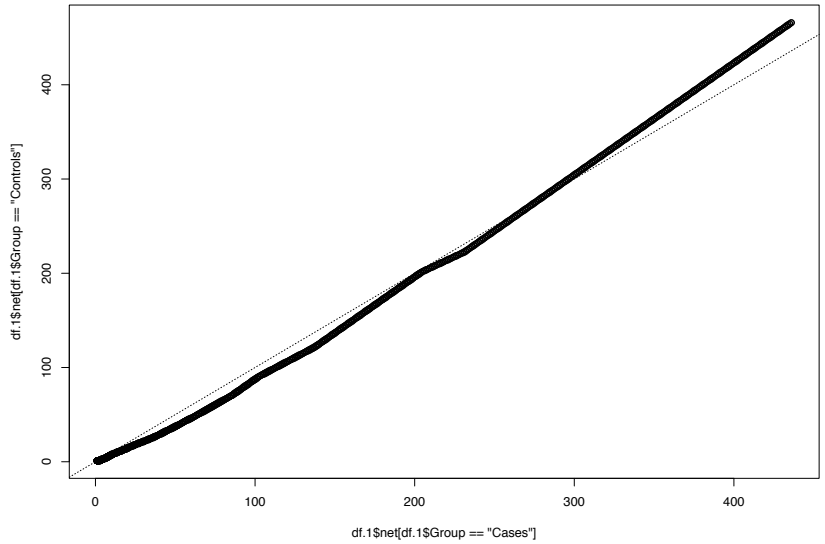


Figure 5.13: **QQ plot illustrating uniform inflation.** The data presented in figure 5.12 is used to produce the quantile-quantile plot for the most uniform distribution between the case and control groups after all inflation parameters were tested.

Box 5.5.6.

- **A.** Get (1) all variant genes in (i) cases and (ii) controls; (2) get only "LoF"-type genes in (i) cases and (ii) controls. Starting with [2 (i-ii))] import via Cytoscape with STRINGdb, cluster with MCL, and export table. Extract network ID and gene symbol. Repeat again with [1 (i-ii)].
- **B.** Then select only the networks from the output of [1 (i-ii)] where those networks are also present in [2 (i-ii)] (overlapping genes) and therefore harbour several LoF genes.
- **C.** Lastly, perform a means test for total variant load on the selected networks (1). Burden rank and test number is defined in the following sections. Multiple testing correction is also subsequently applied.

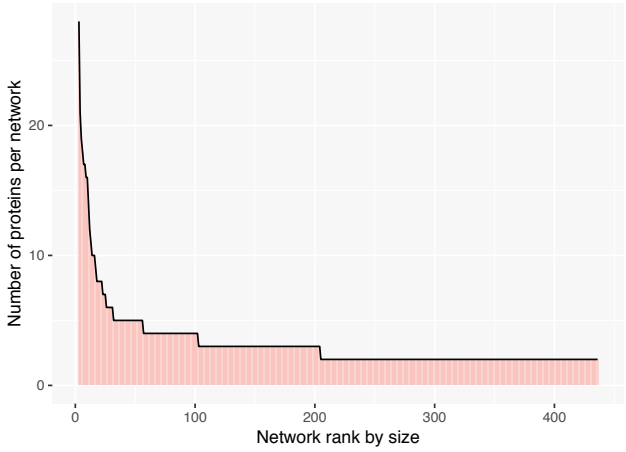


Figure 5.14: Number of proteins per network for case-driven clustering. The size of protein networks has a geometric distribution that decreases until protein (nodes) with no interactions remain; in this cases approximately 200 out of 400 proteins did not play a major role in a single pathway.

5.5.7 Burden rank

Our downstream case-control testing compares the mean total variant load per network. To prevent dilution of our significance threshold due to multiple testing an arbitrarily high number of networks we assumed that protein networks harbouring loss-of-function variants at a consistent frequency in all groups were unlikely to contain genes of interest. To remove these networks, we firstly found (p) the ratio of LoF to all variants within the group per network, and secondly found (q); the ratio of p between groups per network. Networks were ranked by value q. Values passing a threshold of 0.7 were included for total variant load means testing (i.e. 70% of ostensibly damaging variants occurred in cases regardless of the proportion of total variants). This also has the effect that even if there is no significant difference in a case/control total-variant means test downstream, potential false negatives may be rescued by checking for LoF enrichment. This method is applied to real data in section 5.5.10 and **table 5.5**.

5.5.8 Determining the number of tests m

The number of tests should be determined by the predefine LoF ratio per network, q. This value is arbitrarily set and has the problem that an investigator can decide to use a

higher threshold to nude the critical significant threshold in a desired direction. However, testing roughly the top 20-30% of networks is suggested. In our experiments we set our test number as the top 25% of burden-ranked networks. This will be approximately 10 networks to test (the asymptote of network numbers peak when the study size increases over approx. 400 samples as all of the possible PPIs are saturated once the maximum queryable genes are included). Study sizes that are much larger than this will likely only (1) be for disease that are not very rare and (2) be large enough to start expecting single gene significance levels without requiring network analysis. However, some very strict filtering rules could allow larger studies with this method.

5.5.9 Significance testing

We hypothesised firstly that no variant enrichment would be seen in random sampling or control-driven gene clusters, and secondly enrichment would only be seen in case-driven clusters for protein-pathways that provide susceptibility to viral infection. For measuring a significant enrichment of functional variants in a protein network, there are three factors to consider.

1. Our aim is to do a comparison of means between case and control, for total variant load per network.

2. This is done in three iterations; [1] control-driven, [2] case-driven, and [3] random sample-driven.

3. We correct our significance threshold to account for multiple testing using the Benjamini-Hochberg procedure.

With our group-specific gene lists [1 (i-ii), 2 (i-ii), and [3 (i-ii)], prepared in sections 5.5.4 and 5.5.5, we found the distribution of genes per networks and output the list of genes in each network for all 6 datasets. In each of the 3 "all variant" datasets we simply do a comparison of means for total variant load per network comparing case to control, or random.

While the test is not complicated, the significance threshold deserves an in-depth explanation; this is a novel method and most people replicating this study will not have experience with the statistical procedures required. The statistical significance also only allows a narrow margin for successful discovery. When a large number of tests are performed, one is likely to produce P-values that are "statistically significant" by chance ($P < 0.05$), even if the null hypothesis is true. The null hypothesis would state that "random controls and people with disease have the same average frequency of potentially pathogenic variants in some protein pathway". The alternate hypothesis would state that "people with disease have an increased frequency of potentially pathogenic variants in some protein pathway than random controls".

Traditionally, Bonferroni correction has been used in cases like this. For each "family" (network means test) being tested one must correct the critical P-value. For example, for one test a significant P-value might be 0.05 and below this we consider the result to be significant. The chance of getting this result if the null hypothesis was true would be 5%. That does not mean that there is 5% chance that it is true. The following examples are reiterated summary of the topic found in the Handbook of biological statistics [40].

For multiple tests of "families" then we need to adjust the P-value since we are more likely to get false positives by chance. In a published example, García-Arenzana et al. [41] tested 25 associations with mammographic density, which is an important risk factor for breast cancer. The 25 "families" tested were dietary variables including "Total calories", "Olive oil", "whole milk", "white meat", etc. For each variable a P-value was given for its association with mammographic density, i.e. total calories $P < 0.001$, olive oil $P = 0.008$, whole milk $P = 0.039$.

To perform a Bonferroni correction, the critical P-value (or significant threshold) should be divided by the number of tests, $0.05/25 = 0.002$. Therefore, only "total calories" would be significantly associated with the risk factor. If 75 more variables were measured (100 total) then the critical P-value would have to be $0.05/100 = 0.0005$. However, it may not be reasonable to invalidate the significance of the original findings. Using Bonferroni correction for family-wise error rate can mean extremely small P-values. So instead we use

Table 5.4: Benjamini-Hochberg procedure example. The BH-critical value ($(i/m)Q$) is produced on the first row by i; rank =1, m; number of tests =25, Q; FDR = 0.25, to give $(1/25) * 0.25 = 0.01$. The P-value for proteins is below the BH-critical value and therefore the first five tests are significantly associated. Table replicated from García-Arenzana et al. [41] and based on the example used by McDonald [40].

Dietary variable	P value
Total calories	<0.001
Olive oil	0.008
Whole milk	0.039
White meat	0.041
Proteins	0.042
Nuts	0.06
Cereals and pasta	0.074
White fish	0.205
Butter	0.212
Vegetables	0.216
Skimmed milk	0.222
Red meat	0.251
Fruit	0.269
Eggs	0.275
Blue fish	0.34
Legumes	0.341
Carbohydrates	0.384
Potatoes	0.569
Bread	0.594
Fats	0.696
Sweets	0.762
Dairy products	0.94
Semi-skimmed milk	0.942
Total meat	0.975
Processed meat	0.986

a more powerful method for controlling the false discovery rate; the Benjamini–Hochberg procedure [42, 43].

In this procedure, we compare each individual P-value to its Benjamini-Hochberg critical value, $(i/m)Q$, where i is the rank, m is the total number of tests, and Q is the chosen false discovery rate. The largest P-value that has $P < (i/m)Q$ (i.e. P less than BH-critical value) is significant, and all of the P-values smaller than it are also significant, even the ones that aren't less than their own Benjamini-Hochberg critical value.

So in the same example, with 25 tests and Benjamini-Hochberg critical value for a false discovery rate set to 0.25, table 5.4 shows the outcome. The largest P-value that is less than its $(i/m)Q$ values is 0.042 for protein. Therefore, the first 5 variables are significantly associated, including whole milk and white meat despite the fact that their BH-critical value is higher than their P-value. If we were to never have measured protein in this example, m the number of tests would be 24, slightly increasing the BH-critical value, and again identify a significant association for the first 4 tests. Someone interested can recalculate this table to see this effect.

The choice of a false discover rate depends on the application. False positives can waste time, resources, and pollute future work. Minimising false negatives could result in missing a very important finding, that is, when there is a real effect but it is not deemed statistically significant. Allowing a pre-determined level false negatives is often reasonable. As in our application, finding enriched protein networks is the main goal, and downstream work will also be done such as clinical interpretation or functional studies which will catch false negatives. Therefore, the false discovery rate does not have to be very small; consider that our input dataset is already filtered down to ostensibly damaging rare variants. Furthermore, the input dataset is essentially the result of traditional best practices in exome or genome sequencing analysis.

5.5.10 Enrichment testing

For all networks, the top 30 networks in size (largest to smallest; 1-30) were ordered using the burden rank (sec 5.5.7). From these, the number of tests was set (according to the rules defined in sec 5.5.8, so that only the top 8 burden-ranked networks were means tested for their total variant load. **Figure 5.15** shows the test of means for the top 8 protein pathway networks. Table 5.5 lists the P-values assessed for significance using the BH-procedure. We found that only one of the networks was significantly associated with a pathogen-specific immunodeficiency. The variant load was significantly higher than for controls. The total potential LoF variants only accounted for 30.5% of total variants in the network but was ranked high during the burden rank (see sec 5.5.7) because no controls harboured potential LoF variants in this network and therefore 100% occurred in cases. This protein network contained genes responsible for pathogen detection; some genes *might have been* identified as candidates using the routine exome analysis pipeline such as the antiviral receptors and antiviral interferon regulatory factors. However, most of the other genes that are integral to this pathway would not have been identified by standard best practices. The protein network is shown in **Figure 5.16** where potential LoF variants-harbouring genes are coloured in red. Gene candidates with variants of unknown significance are coloured in red and, anecdotally, the colouring thereafter becomes lighter (orange to yellow) based on the likelihood of candidates being

identified by manual interpretation of unknown candidates.

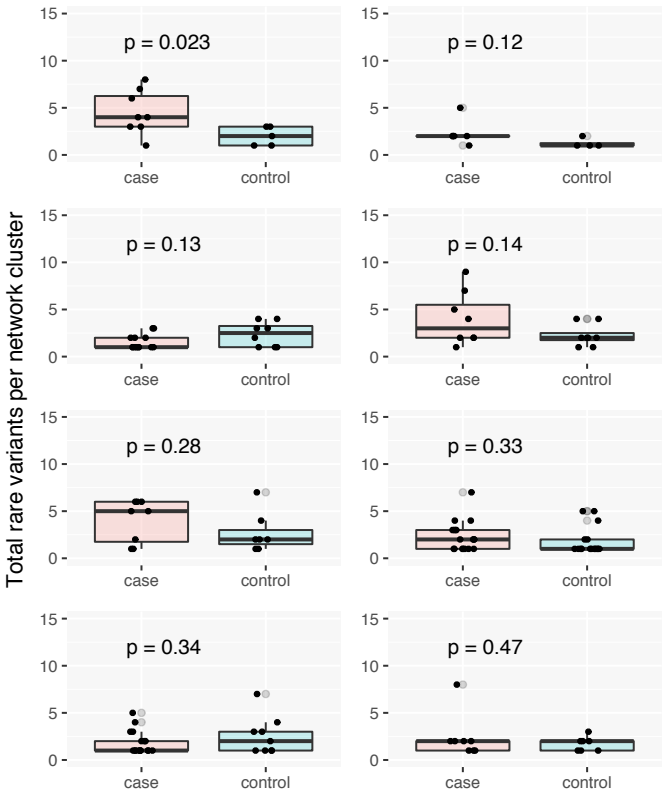


Figure 5.15: **Case and control means test.** The total rare variants per network are shown, comparing groups. A test of means was conducted in this test dataset and P-values are shown.

Table 5.5: Benjamini-Hochberg procedure for real data. The top 30 networks in size (largest to smallest; 1-30) were ordered using the burden rank (sec 5.5.7). From these, the number of tests was set (sec 5.5.8, so only the top 8 burden-ranked networks were means tested for their total variant load. The BH-critical value ($(i/m)Q$) was produced on the first row by i; rank =1, m; number of tests =8, Q; FDR = 0.2, to give $(1/8) * 0.2 = 0.025$. In this case the first P-value is the only one that falls below the BH-critical value and therefore is the only significantly associated protein network with a pathogen-specific immunodeficiency.

Net ID by size	**LoF freq in cases**	**LoF freq due to cases per network**	**P-value**	**rank**	$(i/m)Q$
22	0.306	1	0.023	1	0.025
27	0.429	1	0.12	2	0.05

16	0.6	0.919	0.13	3	0.075
19	0.281	0.835	0.14	4	0.1
25	0.25	1	0.28	5	0.125
11	0.357	0.838	0.33	6	0.15
10	0.516	0.856	0.34	7	0.175
18	0.474	0.85	0.47	8	0.2

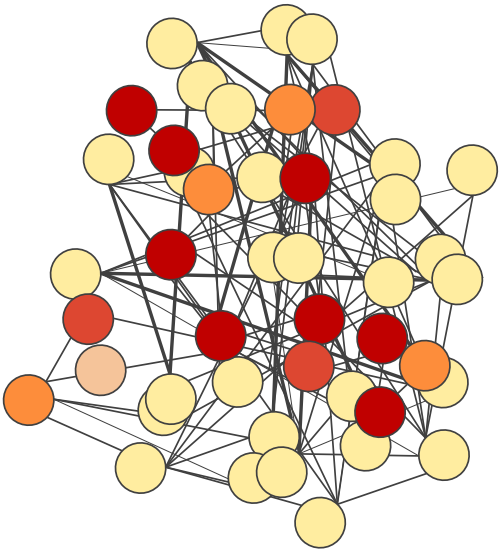


Figure 5.16: **Protein network with significantly enriched variant load.** From the example data, network 22 was significantly enriched for rare variants. The same clustering method was again used on all variants with a less stringent variant frequency (<1% in general population and present in any cohort sample). With the resulting, more common variants, the full protein network can be seen (about double in size compared to only very rare variants). Gene candidates with variants of unknown significance are coloured in red and, anecdotally, the colouring thereafter becomes lighter (orange to yellow) based on the likelihood of candidates being identified by manual interpretation of unknown candidates.

5.6 Discussion

Exome sequence data is usually about 4 GB of information per person. Whole genomes are approximately 50GB of data. The analysis of whole genome sequencing is almost identical to the exome pipeline outlined here. While there is much more information (for not much of a higher cost), a lot of the non-coding sequence contains information that we can't yet interpret. For Mendelian disease the whole exome often uncovers the coding variants that explain disease. We may not understand anything else outside the exome (and the surrounding splice regions) in relation to a patients' disease. Mutations in the promoters or enhancers that prevent transcription may not be as readily interpretable as the majority of coding variant effects. Therefore, whole genome is often not required. This excuse for performing exome sequencing rather than whole genome mostly depends

on value for money. Performing all the different kinds of analysis, including non-coding genome analysis, requires many people with expertise in each topic. Even if whole genome data was available to smaller research teams, it is often the case that they cannot carry out all the work required to interpret it. For national level genomics, there is no question that whole genome sequencing is preferential. We can retain the data for decades with hundreds of experts to share the work-load, while the cost is essentially a political factor. An important question to address is the right for a person to agree to genetic forfeiture. We are at the brink of preventative medicine using genome sequencing in newborns. Regardless of the popular ethical consensus, any preventative non-consenting genomic analysis can be considered coercion.

Figure 5.16 illustrates how not only can very rare or damaging variants be clustered, but the same network can be expanded to include peripherally interacting genes. This modification may be used for downstream functional work such as looking at pathway-level expression data. An important consideration for protein network cohort analysis is evident in **Figure 5.14**. About 50% of genes with a functional variant are do not get clustered into a PPI network (protein pathway). However, some of these genes could still harbour a potential loss-of-function or damaging variant. If we found 3 significantly enriched protein networks, a potential 4th missed network (false negative), because of unclustered genes, would not detract from the significant findings. Therefore, the singleton genes remaining from MCL clustering should be listed and reassessed based on traditional interpretations; variant effect, loss-of-function intolerance, etc. The converse, a false positive because of over clumping weakly related proteins, would be negative.

The analysis of genomic data is an iterative process. Therefore, access to raw unprocessed genetic information is often required to utilize cutting edge methods [44, 45]. Furthermore, genetic analysis is a complex, multi-stage procedure. Due to the inherent complexity, there is a number of output streams which consist of different data types. To provide seamless integration with current best practices in precision medicine, it is valuable to adhere to standard genomic data types, including CRAM, SAM/BAM, FASTQ, and VCF [46–49]. There is benefit to creating new data formats that increase efficiency. However, by focusing on key data types in genomics, one can enable integration

with most current software [44, 50].

An interesting caveat to genetic data is that at pre-processing stages, several data types cannot be currently provided with protection through the use of modern cryptographic methods [51, 52]. There is currently a severe lack of tools that complement current methods required to interrogate genetic data at different stages while protecting individual personal genetic records. Furthermore, despite the attempts to promote data privacy and integrity through global initiatives, such as Global Aliance 4 Genome Health, little has been done to produce queryable data that protects the genetic identity of a subject.

The privacy concerns at the early stages of data processing are often overlooked. Almost every method offered for data security relies on protecting only fully process data (e.g. already variant called VCF format data) or summary statistics. In worst cases, privacy concerned genomics falls back to "trust-based" systems where data generators or researchers are required to accept responsibility for preventing any re-anonymisation. Of course, researcher trust is an important factor, however, relying on this method for protecting subject information is immoral. Unlike nearly all clinical data, genetic data is inherently identifiable and is not readily anonymised. The information that makes up the data is itself the identity or commodity. In nearly every other type of clinical data, it is only a commodity when there is an identity to which it is paired or if it is part of a large dataset-of-normals. The lack of strong methods of genetic data protection is not an apparent risk generally. Extrapolating the risk which differentiates other types of data that requires informed consent is a difficult task for many experts. Relying on patient consent and trust in data protection is not sufficient for the future of global genomics. Successfully overcoming these challenges will allow for the use of analysis methods that otherwise provide vulnerabilities against the protection of private data [1], [GA4GH (https://www.ga4gh.org)].

5.7 Conclusion

A pipeline of routine exome analysis was outlined. Important points on tailored analysis are demonstrated. A new method was developed for the unbiased detection of a protein

network, driving disease, based on potential loss of function variants.

5.8 Command line example code

5.8.1 Whole exome analysis

```
$ #!/bin/bash
##############################################
#### The basic protocol for analysis.
#### It is best to set up a loop that
#### can run the protocol on all samples.
##############################################
# Make project organisation folders
mkdir ~/1.fastq/ && \
mkdir ~/2.trim/ && \
mkdir ~/3.sort/ && \
mkdir ~/4.dedup/ && \
mkdir ~/5.realtar/ && \
mkdir ~/6.indelrealn/ && \
mkdir ~/7.baserecal/ && \
mkdir ~/9.printbam/ && \
mkdir ~/10.gvcf/ && \
mkdir ~/geno/ && \


##############################################
#### Typical workflow
##############################################


##############################################
#### Cut read adaptors and run FastQC (see page 233)
##############################################
trim_galore -q 20 -fastqc_args \
"-outdir ~/2.trim/QC_reports"
-illumina -gzip \
-o ~/2.trim/ -length 20 -paired \
~/1.fastq/Sequencing_ID_L001_R1_001.fastq.gz \
~/1.fastq/Sequencing_ID_L001_R2_001.fastq.gz && \
```

```
##############################################
#### Align reads to reference genome (see page 233)
##############################################
bwa mem -t 12 -M ~/ref/human_g1k_v37.fasta \
~/2.trim/Sequencing_ID_L001_R1_001_val_1.fq.gz \
~/2.trim/Sequencing_ID_L001_R2_001_val_2.fq.gz \
-v 1 -R '@RG\tID:Sample_ID\tSM:Sample_ID \
\tPL:ILLUMINA\tLB:Sample_ID' \
-M | samtools view -Sb - > ~/2.trim/Sample_ID.bam && \


##############################################
#### Sort reads (see page 233)
##############################################
java -Xmx8g -jar ~/picard/picard-tools-2.5.0/picard.jar
SortSam \
I= ~/2.trim/Sample_ID.bam \
O= ~/3.sort/Sample_ID.sort.bam \
SO=coordinate CREATE_INDEX=TRUE && \


##############################################
#### Mark duplicate reads (see page 233)
##############################################
java -Xmx8g -jar ~/picard/picard-tools-2.5.0/picard.jar
MarkDuplicates \
I= ~/3.sort/Sample_ID.sort.bam \
O= ~/4.dedup/Sample_ID.sort.dedup.bam \
M= ~/4.dedup/Sample_ID.sort.dedup.metrics \
CREATE_INDEX=TRUE && \


##############################################
#### Create indel realigner targets (see page 234)
##############################################
java -Xmx6g -jar ~/GATK/GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R ~/ref/human_g1k_v37.fasta \
-known ~/ref/1000G_phase1.indels.b37.vcf \
-known ~/ref/Mills_and_1000G_gold_standard.\
indels.b37.sites.vcf \
-I ~/4.dedup/Sample_ID.sort.dedup.bam \
-o ~/5.realtar/Sample_ID.sort.dedup.bam.intervals && \
```

```
##############################################
#### Indel realignment (see page 234)
##############################################
java -Xmx6g -jar ~/GATK/GenomeAnalysisTK.jar \
-T IndelRealigner \
-R ~/ref/human_g1k_v37.fasta \
-known ~/ref/1000G_phase1.indels.b37.vcf \
-known ~/ref/Mills_and_1000G_gold_standard.\
indels.b37.sites.vcf \
-I ~/4.dedup/Sample_ID.sort.dedup.bam \
-targetIntervals \
~/5.realtar/Sample_ID.sort.dedup.bam.intervals \
-o ~/6.indelrealn/Sample_ID.sort.dedup.indelrealn.bam && \


##############################################
#### Recalibrate base quality scores using a recalibration model
(BQSR) (see page 234)
##############################################
java -Xmx8g -jar ~/GATK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R ~/ref/human_g1k_v37.fasta \
-knownSites ~/ref/dbSnp146.b37.vcf.gz \
-knownSites ~/ref/1000G_phase1.indels.b37.vcf \
-knownSites ~/ref/Mills_and_1000G_gold_standard.\
indels.b37.sites.vcf \
-o ~/7.baserecal/Sample_ID.sort.dedup.indelrealn.recal.grp \
-I ~/6.indelrealn/Sample_ID.sort.dedup.indelrealn.bam \
-nct 6 && \


##############################################
#### Optional check for base recalibration
##############################################


##############################################
#### Print final reads after applying BQSR (see page 234)
##############################################
java -Xmx12g -jar ~/GATK/GenomeAnalysisTK.jar \
-T PrintReads \
```

```
-R ~/ref/human_g1k_v37.fasta \
-I ~/6.indelrealn/Sample_ID.sort.dedup.indelrealn.bam \
-BQSR ~/7.baserecal/Sample_ID.sort.dedup.indelrealn.recal.grp \
-o ~/9.printbam/Sample_ID.sort.dedup.indelrealn.recal.bam \
-disable_indel_quals && \


##############################################
#### Haplotype variant calling (see page 235)
##############################################
java -Xmx8g -jar ~/GATK/GenomeAnalysisTK.jar \
-T HaplotypeCaller -emitRefConfidence GVCF \
-R ~/ref/human_g1k_v37.fasta -D ~/ref/dbSnp146.b37.vcf.gz \
-stand_call_conf 30 \
-stand_emit_conf 10 \ # deprecated
-I ~/9.printbam/Sample_ID.sort.dedup.indelrealn.recal.bam \
-o ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-L ~/ref/SureSelectAllExonV6/S07604514_Regions_b37.bed \
-ip 30 && \


##############################################
#### Joint genotyping (see page 236)
##############################################
java -Xmx12g -jar ~/GATK/GenomeAnalysisTK.jar \
-T GenotypeGVCFs \
-R ~/ref/human_g1k_v37.fasta \
-D ~/ref/dbSnp146.b37.vcf.gz -stand_call_conf 30 \
-stand_emit_conf 10 \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-V ~/10.gvcf/Sample_ID.sort.dedup.indelrealn.recal.HC.g.vcf \
-o ~/geno/genotype.vcf -nda -showFullBamList -nt 12 && \


##############################################
#### Hard filter selecting SNVs
##############################################
```

```
java -Xmx12g -jar ~/GATK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R ~/ref/human_g1k_v37.fasta \
-selectType SNP \
-variant ~/geno/genotype.vcf \
-o ~/geno/genotype.raw-snps.vcf && \


##############################################
#### Hard filter selecting INDELs
##############################################
java -Xmx12g -jar ~/GATK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R ~/ref/human_g1k_v37.fasta \
-variant ~/geno/genotype.vcf \
-selectType INDEL -selectType MNP \
-o ~/geno/genotype.raw-indels.vcf && \


##############################################
#### Applying hard filter for SNVs
##############################################
java -Xmx8g -jar ~/GATK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R ~/ref/human_g1k_v37.fasta \
-V ~/geno/genotype.raw-snps.vcf \
-filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 ||\
MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0" \
-filterName "snp_hard_filter" \
-o ~/geno/genotype.raw-snps.filtered.snvs.vcf && \


##############################################
#### Applying hard filter for INDELs
##############################################
java -Xmx8g -jar ~/GATK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R ~/ref/human_g1k_v37.fasta \
-V ~/geno/genotype.raw-indels.vcf \
-filterExpression "QD < 2.0 || FS > 200.0 ||\
ReadPosRankSum < -20.0" \
-filterName "indel_hard_filter" \
-o ~/geno/genotype.raw-indels.filtered.indels.vcf && \
```

```
##############################################
#### Combine filtered results
##############################################
java -Xmx8g -jar ~/GATK/GenomeAnalysisTK.jar \
-T CombineVariants -R ~/ref/human_g1k_v37.fasta \
-variant ~/geno/genotype.raw-snps.filtered.snvs.vcf \
-variant ~/geno/genotype.raw-indels.filtered.indels.vcf \
-o ~/geno/genotype.fltd-combinedvars.vcf \
-genotypemergeoption UNSORTED && \


##############################################
#### Filter variants in EdbSNP >/= 1% \
#### and not listed as pathogenic by ClinVar
##############################################
perl ~/vcfhacks-v0.2.0/annotateSnps.pl \
-d ~/ref/dbSnp146.b37.vcf.gz ~/ref/clinvar_20160531.vcf.gz \
-f 1 -pathogenic \
-i ~/geno/genotype.fltd-combinedvars.vcf \
-o ~/geno/genotype.fltd-combinedvars.1pcdbsnp.vcf \
-t 12 && \


##############################################
#### Filter variants in EVS greater >/= 1%
##############################################
perl ~/vcfhacks-v0.2.0/filterOnEvsMaf.pl -d ~/ref/evs/ \
-f 1 -progress \
-i ~/geno/genotype.fltd-combinedvars.1pcdbsnp.vcf \
-o ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.vcf \
-t 12 && \


##############################################
#### Exac filter for population frequency
##############################################
perl ~/vcfhacks-v0.2.0/filterVcfOnVcf.pl \
-i ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.vcf \
-f ~/ref/ExAC/ExAC.r0.3.sites.vep.vcf.gz \
-o ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.exac.vcf \
-w -y 0.01 \
```

```
-b # progress bar \
-t # number of threads && \


##############################################
#### Annotate with variant effect predictor
##############################################
perl ~/variant_effect_predictor/variant_effect_predictor.pl \
-offline -vcf -everything \
-dir_cache ~/variant_effect_predictor/vep_cache \
-dir_plugins ~/variant_effect_predictor/vep_cache/Plugins \
-plugin Condel,~/variant_effect_predictor/vep_cache/\
Plugins/config/Condel/config/ \
-plugin ExAC,~/ref/ExAC/ExAC.r0.3.sites.vep.vcf.gz \
-plugin SpliceConsensus \
-fasta ~/variant_effect_predictor/fasta/\
Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz \
-i ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.exac.vcf \
-o ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.exac.vep.vcf
\
-fork 12 && \


##############################################
#### Confirm samples names
##############################################
perl ~/vcfhacks-v0.2.0/getSampleNames.pl \
-i ~/geno/genotype.fltd-combinedvars.1pcdbsnp.1pcEVS.exac.vep.vcf
&& \
```

5.8.2 Data extraction

```
##############################################
#### Extract columns
##############################################
#### A list of files for
#### where the data from
#### one column is compiled into
#### a master table
##############################################
```

```
##########################################
#### This takes column 3 from every file
#### and appends to the output file.
#### Space delimited.
##########################################
awk 'FNR==1{f++}{a[f,FNR]=$3}END{for(x=1;x<=FNR;x++)\
{for(y=1;y<ARGC;y++)printf("%s ",a[y,x]);print ""}}' \
./../*/pheno.txt > master.txt


##########################################
#### The spacer method can be changed; tab, space, comma, etc.
#### Another way to convert later is with the following command.
#### [The tab character (after "s/") must be removed
#### and printed to the command line using "ctrl+v" then "tab".]
##########################################
sed 's/ /,/g' input.tsv > output.csv
```

5.8.3 Candidate filter

```
##########################################
#### Filter a VCF on a candidate gene list.
##########################################
#### List format as "X:1-2000",
#### or -b for a bed file or
#### a list file with 1 per line.
##########################################
for f in ~/immune.panel/vep/*.vcf
do
perl ~/vcfhacks-v0.2.0/filterVcfOnLocation.pl \
-i ~/immune.panel/vep/$f \
-b ~/deep.panel.bed \
-o ~/immune.panel/filter/$( basename -s .vcf $f ).panel.vcf \
done


##########################################
#### Post-routine analysis candidate filtering.
#### Similar filtering can be done without going back
#### to analysis stages to create a virtual panel.
##########################################
```

```
#### Export all gene names and give the count.
sort list.txt | uniq -c > InflammatoryDisorderCohortHitCount.txt
#### Format to csv.
#### Cross against a "master" list of immune genes.

###############################################
#### In R, import data
###############################################
master <- read.csv("./master.csv", stringsAsFactors=FALSE)
InflammatoryDisorderCohortHitCount <-
read.csv("./ InflammatoryDisorderCohortHitCount.csv",
stringsAsFactors=FALSE)

###############################################
#### Merge the master immune gene list
#### with the Inflammatory disorder cohort hits.
###############################################
combine <-
merge(master, InflammatoryDisorderCohortHitCount,
by = "Gene", all = TRUE)

###############################################
#### Remove the genes that happen to overlap
#### gene of interest and remove anything from
#### the master list that is not in the cohort list.
###############################################
clean <- na.omit(combine)

###############################################
#### Write out the table.
###############################################
write.csv(clean, './GenesOfInterest.csv', row.names = FALSE)

###############################################
#### The output can be sorted as of interest
#### e.g.  autosomal dominant autoinflammatory gene.
###############################################
```

5.8.4 Tailored filtering

```
###############################################
#### Filter on sample.
```

```
#### May need use a "-freq" option
#### to account for index hopping.
#### Filter on sample removes anything shared
#### with cases (-s) that are not listed but not others (-x).
###############################################
perl /home/vcfhacks-v0.2.0/filterOnSample.pl \
-i ~/samples.vep.vcf \
-s case.1 case.2 case.3 -x \
-o ~/samples.getFunctionalVariantsVep.vcf


###############################################
#### Get variants.
###############################################
#### Getting functional variants.  The -n flag allows
#### selections only when >2 samples
#### have variants in a shared gene.
perl /home/vcfhacks-v0.2.0/getFunctionalVariants.pl \
-s case.1 case.2 case.3 \
-i ~/samples.vep.vcf \
-f -n 2 \
-o ~/samples.getFunctionalVariantsVep.SharedGenes.vcf


#### Candidate compound heterozygous.
#### Only variants that are common in ALL -s are considered.
#### Flag -n specifies the number of cases
#### required to return a genotype.
perl /home/vcfhacks-v0.2.0/findBiallelic.pl \
-i ~/samples.vep.vcf \
-s case.1 case.2 case.3 \
-n 1 \
-o ~/samples.findBiallelic.all.vcf


###############################################
#### Rank, annontate, and simplify
###############################################
perl /home/vcfhacks-v0.2.0/rankOnCaddScore.pl \
-c /data/shared/cadd/v1.3/*.gz \
-i ~/samples.getFunctionalVariantsVep.SharedGenes.vcf \
-o ~/samples.getFunctionalVariantsVep.SharedGenes.cadd.ranked.vcf
\
```

```
-progress


perl /home/vcfhacks-v0.2.0/geneAnnotator.pl \
-d /home/vcfhacks-v0.2.0/data/geneAnnotatorDb \
-i ~/samples.findBiallelic.all.vcf \
-o ~/samples.findBiallelic.all.gene.anno


perl /home/vcfhacks-v0.2.0/annovcfToSimple.pl \
-i ~/samples.findBiallelic.all.gene.anno \
-vep -gene_anno \
-canonical_only \
-u -contains_variant \
-o ~/samples.findBiallelic.all.gene.anno.simple.xlsx
```

Bibliography

[1] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 05 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://doi.org/10.1093/bioinformatics/btp324.

[2] Dylan Lawless, Shelly Pathak, Thomas Edward Scambler, Lylia Ouboussad, Rashida Anwar, and Sinisa Savic. A case of adult-onset still's disease caused by a novel splicing mutation in tnfaip3 successfully treated with tocilizumab. *Frontiers in Immunology*, 9, Jul 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.01527. URL http://dx.doi.org/10.3389/fimmu.2018.01527.

[3] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[4] Mac Arthur Lab. List of gene lists for genomic analysis. *GitHub*, 2018. URL https://github.com/macarthur-lab/gene_lists.

[5] HUGO Gene Nomenclature Committee at the European Bioinformatics Institute.

Genenames.org. *European Bioinformatics Institute*, 2018.

[6] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1037. URL https://doi.org/10.1093/nar/gkx1037.

[7] Matthew R. Nelson, Daniel Wegmann, Margaret G. Ehm, Darren Kessner, Pamela St. Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, Liling Warren, Jennifer Aponte, Matthew Zawistowski, Xiao Liu, Hao Zhang, Yong Zhang, Jun Li, Yun Li, Li Li, Peter Woollard, Simon Topp, Matthew D. Hall, Keith Nangle, Jun Wang, Gonçalo Abecasis, Lon R. Cardon, Sebastian Zöllner, John C. Whittaker, Stephanie L. Chissoe, John Novembre, and Vincent Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012. ISSN 0036-8075. doi: 10.1126/science.1217876. URL https://science.sciencemag.org/content/337/6090/100.

[8] Ran Blekhman, Orna Man, Leslie Herrmann, Adam R. Boyko, Amit Indap, Carolin Kosiol, Carlos D. Bustamante, Kosuke M. Teshima, and Molly Przeworski. Natural selection on genes that underlie human disease susceptibility. *Current Biology*, 18(12): 883 – 889, 2008. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2008.04.074. URL http://www.sciencedirect.com/science/article/pii/S0960982208006015.

[9] Jonathan S Berg, Michael Adams, Nassib Nassar, Chris Bizon, Kristy Lee, Charles P Schmitt, Kirk C Wilhelmsen, and James P Evans. An informatics approach to analyzing the incidentalome. *Genetics in Medicine*, 15(1):36, 2013.

[10] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1113. URL https://doi.org/10.1093/nar/gkt1113.

[11] Traver Hart, Kevin R Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason

Moffat. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology*, 10(7), 2014.

[12] Traver Hart, Amy Hin Yan Tong, Katie Chan, Jolanda Van Leeuwen, Ashwin Seetharaman, Michael Aregger, Megha Chandrashekhar, Nicole Hustedt, Sahil Seth, Avery Noonan, et al. Evaluation and design of genome-wide crispr/spcas9 knockout screens. *G3: Genes, Genomes, Genetics*, 7(8):2719–2727, 2017.

[13] Judith A. Blake, Carol J. Bult, James A. Kadin, Joel E. Richardson, Janan T. Eppig, and the Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Research*, 39(suppl_1):D842–D848, 11 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq1008. URL https://doi.org/10.1093/nar/gkq1008.

[14] Benjamin Georgi, Benjamin F. Voight, and Maja Bućan. From mouse to human: Evolutionary genomics analysis of human orthologs of essential genes. *PLOS Genetics*, 9(5):1–10, 05 2013. doi: 10.1371/journal.pgen.1003484. URL https://doi.org/10.1371/journal.pgen.1003484.

[15] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp v2.0: A database of human non-synonymous snvs and their functional predictions and annotations. *Human Mutation*, 34(9):E2393–E2402, 2013. doi: 10.1002/humu.22376. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22376.

[16] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45 (D1):D896–D901, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1133. URL https://doi.org/10.1093/nar/gkw1133.

[17] Richard D. Wood, Michael Mitchell, and Tomas Lindahl. Human dna repair genes, 2005. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 577(1):275 – 283, 2005. ISSN 0027-5107. doi: https://doi.org/10.1016/j.mrfmmm.2005.03.007. URL http://www.sciencedirect.com/science/article/pii/S0027510705001636. Mechanisms of DNA Repair.

[18] Josephine Kang, Alan D. D'Andrea, and David Kozono. A DNA Repair Pathway Focused Score for Prediction of Outcomes in Ovarian Cancer Treated With Platinum-Based Chemotherapy. *JNCI: Journal of the National Cancer Institute*, 104(9): 670–681, 04 2012. ISSN 0027-8874. doi: 10.1093/jnci/djs177. URL https://doi.org/10.1093/jnci/djs177.

[19] Heidi L. Rehm, Jonathan S. Berg, Lisa D. Brooks, Carlos D. Bustamante, James P. Evans, Melissa J. Landrum, David H. Ledbetter, Donna R. Maglott, Christa Lese Martin, Robert L. Nussbaum, Sharon E. Plon, Erin M. Ramos, Stephen T. Sherry, and Michael S. Watson. Clingen — the clinical genome resource. *New England Journal of Medicine*, 372(23):2235–2242, 2015. doi: 10.1056/NEJMsr1406261. URL https://doi.org/10.1056/NEJMsr1406261. PMID: 26014595.

[20] Joel D Mainland, Yun R Li, Ting Zhou, Wen Ling L Liu, and Hiroaki Matsunami. Human olfactory receptor responses to odorants. *Scientific Data*, 2:150002 EP –, 02 2015. URL https://doi.org/10.1038/sdata.2015.2.

[21] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1099. URL https://doi.org/10.1093/nar/gkw1099.

[22] Stephen PH Alexander, Arthur Christopoulos, Anthony P Davenport, Eamonn Kelly, Neil V Marrion, John A Peters, Elena Faccenda, Simon D Harding, Adam J Pawson, Joanna L Sharman, Christopher Southan, Jamie A Davies, and CGTP Collaborators. The concise guide to pharmacology 2017-18, g protein-coupled receptors. *British Journal of Pharmacology*, 174(S1):S17–S129, 2017. doi: 10.1111/bph.13878.

[23] Simon D Harding, Joanna L Sharman, Elena Faccenda, Chris Southan, Adam J Pawson, Sam Ireland, Alasdair J G Gray, Liam Bruce, Stephen P H Alexander, Stephen Anderton, Clare Bryant, Anthony P Davenport, Christian Doerig, Doriano Fabbro, Francesca Levi-Schaffer, Michael Spedding, Jamie A Davies, and NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Research*, 46(D1):D1091–D1106, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1121. URL https://doi.org/10.1093/nar/gkx1121.

[24] Vlado Dančík, Kathleen Petri Seiler, Damian W. Young, Stuart L. Schreiber, and

Paul A. Clemons. Distinct biological network properties between the targets of natural products and disease genes. *Journal of the American Chemical Society*, 132(27): 9259–9261, 07 2010. doi: 10.1021/ja102798t. URL https://doi.org/10.1021/ja102798t.

[25] University of Washington Deptartment of Laboratory Medicine. Broca-cancerriskpanel. *BROCA Web portal*, 2018. URL http://depts.washington.edu/labweb/Divisions/MolDiag/MolDiagGen/index.htm.

[26] Sarah S. Kalia, Kathy Adelman, Sherri J. Bale, Wendy K. Chung, Christine Eng, James P. Evans, Gail E. Herman, Sophia B. Hufnagel, Teri E. Klein, Bruce R. Korf, Kent D. McKelvey, Kelly E. Ormond, C. Sue Richards, Christopher N. Vlangos, Michael Watson, Christa L. Martin, and David T. Miller. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (acmg sf v2.0): a policy statement of the american college of medical genetics and genomics. *Genetics In Medicine*, 19:249 EP –, 11 2016. URL https://doi.org/10.1038/gim.2016.190.

[27] Shefali Setia Verma, Navya Josyula, Anurag Verma, Xinyuan Zhang, Yogasudha Veturi, Frederick E Dewey, Dustin N Hartzel, Joe Leader, Marylyn D Ritchie, and Sarah A Pendergrass. Rare variants in drug target genes contributing to complex diseases, phenome-wide. *Scientific reports*, 8(1):4624, 2018.

[28] Sarah A Pendergrass, Alex Frase, John Wallace, Daniel Wolfe, Neerja Katiyar, Carrie Moore, and Marylyn D Ritchie. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData mining*, 6(1):25, 2013.

[29] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 45(Database issue):D12, 2017.

[30] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2011.

[31] Marija Milacic, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, 4(4):1180–1211, 2012.

[32] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T

Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25 (1):25, 2000.

[33] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2013.

[34] Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Mathivanan, Christian Pecquet, et al. Netpath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):R3, 2010.

[35] David A. Parry. Vcfhacks; simple to use commandline programs for vcf filtering and manipulation. *GitHub repository*, 2015. URL https://github.com/david-a-parry/vcfhacks.

[36] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian vonÂ Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, 10 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw937. URL https://doi.org/10.1093/nar/gkw937.

[37] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 04 2002. ISSN 0305-1048. doi: 10.1093/nar/30.7.1575. URL https://doi.org/10.1093/nar/30.7.1575.

[38] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[39] Stijn Marinus Van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2000.

[40] John H McDonald. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

[41] Nicolás García-Arenzana, Eva María Navarrete-Muñoz, Virginia Lope, Pilar Moreo,

Carmen Vidal, Soledad Laso-Pablos, Nieves Ascunce, Francisco Casanova-Gómez, Carmen Sánchez-Contador, Carmen Santamariña, et al. Calorie intake, olive oil consumption and mammographic density among spanish women. *International journal of cancer*, 134(8):1916–1925, 2014.

[42] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[43] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[44] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(1):11.10.1–11.10.33, 2013. doi: 10.1002/0471250953.bi1110s43. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43.

[45] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 2018. doi: 10.1101/201178. URL https://www.biorxiv.org/content/early/2018/07/24/201178.

[46] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 06 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://doi.org/10.1093/bioinformatics/btp352.

[47] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome research*, 21(5):734—740, May 2011. ISSN 1088-9051. doi: 10.1101/gr.

114819.110. URL http://europepmc.org/articles/PMC3083090.

[48] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2009.

[49] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 06 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr330. URL https://doi.org/10.1093/bioinformatics/btr330.

[50] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278, 2014.

[51] David Froelicher, Patricia Egger, João Sá Sousa, Jean Louis Raisaro, Zhicong Huang, Christian Mouchet, Bryan Ford, and Jean-Pierre Hubaux. Unlynx: a decentralized system for privacy-conscious data sharing. *Proceedings on Privacy Enhancing Technologies*, 2017(4):232–250, 2017.

[52] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. {GAZELLE}: A low latency framework for secure neural network inference. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1651–1669, 2018.

6 Appendix data

Table 6.1: **Percentage of variants per gene.** Percentage of mutated versus non-mutated amino acids in RAG1 and RAG2 based on GnomAD population genetics data.

Protein	Status	Value	Total residues	%
RAG1	Invariant	584	1043	55.99
RAG1	Mutated	459	1043	44.01
RAG2	Invariant	295	527	55.98
RAG2	Mutated	232	527	44.02

Table 6.2: **Residue frequencies and mutation per gene.** Basic statistics for RAG1/2 were produced using SMS2. Results are shown for the canonical sequences, a 1043-residue sequence of RAG1-201 peptide ENSP00000299440 and for a 527-residue sequence of RAG2-201 peptide ENSP00000308620. Both count and frequency of residue usage are provided. The number of times each residue is mutated was found in population genetic data. The number of mutant versus wild-type is shown, from which the rate of each is derived. The values for mutation and residue frequency sum to form the raw MRF, used in the main analysis dataframe.

RAG1						
Residue	Mutant	WT	Total	Rate mutated	Frequency	MRF
Ala	28	40	68	0.412	0.065	0.027
Arg	45	21	66	0.682	0.063	0.043
Asn	13	25	38	0.342	0.036	0.012
Asp	23	25	48	0.479	0.046	0.022
Cys	13	21	34	0.382	0.033	0.012
Gln	12	22	34	0.353	0.033	0.012
Glu	33	50	83	0.398	0.080	0.032
Gly	15	31	46	0.326	0.044	0.014
His	27	16	43	0.628	0.041	0.026
Ile	24	24	48	0.500	0.046	0.023
Leu	26	71	97	0.268	0.093	0.025
Lys	31	60	91	0.341	0.087	0.030

Met	20	8	28	0.714	0.027	0.019
Phe	14	32	46	0.304	0.044	0.013
Pro	29	23	52	0.558	0.050	0.028
Ser	43	44	87	0.494	0.083	0.041
Thr	21	19	40	0.525	0.038	0.020
Trp	5	5	10	0.500	0.010	0.005
Tyr	8	17	25	0.320	0.024	0.008
Val	29	30	59	0.492	0.057	0.028
RAG2						
Ala	9	11	20	0.450	0.038	0.017
Arg	12	6	18	0.667	0.034	0.023
Asn	18	13	31	0.581	0.059	0.034
Asp	15	18	33	0.455	0.063	0.028
Cys	6	12	18	0.333	0.034	0.011
Gln	3	11	14	0.214	0.027	0.006
Glu	13	20	33	0.394	0.063	0.025
Gly	20	17	37	0.541	0.070	0.038
His	9	7	16	0.563	0.030	0.017
Ile	14	17	31	0.452	0.059	0.027
Leu	15	22	37	0.405	0.070	0.028
Lys	9	24	33	0.273	0.063	0.017
Met	7	6	13	0.538	0.025	0.013
Phe	9	19	28	0.321	0.053	0.017
Pro	12	17	29	0.414	0.055	0.023
Ser	16	26	42	0.381	0.080	0.030
Thr	18	17	35	0.514	0.066	0.034
Trp	2	5	7	0.286	0.013	0.004
Tyr	10	9	19	0.526	0.036	0.019
Val	16	17	33	0.485	0.063	0.030

Table 6.3: **MRF data tables. The complete scores are listed for each protein.** Both the wild type and alternative variants reported on GnomAD are shown (*WT* and *Gnomad*). Multiple alternative variants are reported for some residues which are annotated as *Multiallelic*. A column listing the 1% average of MRF scores using a sliding window is present (*Av 1%*); this is used in Figure 2.4 (iii) with a cut-off threshold at the 75th percentile to clearly visualise high scoring clusters. The Boolean conservation score is based on population genetics data, where 1 represents no known variant at a residue site (*Boolean C*). *MRFC* (MRF with conservation C score applied). Variants known to cause RAG deficiency are underlined in columns *Residue* and *WT*.

RAG1								RAG2							
MRFC	Av. 1 %	MRF	Residue	WT	GnomAD	Multiallelic	Boolean C	MRFC	Av. 1 %	MRF	Residue	WT	GnomAD	Multiallelic	Boolean C
0	0.0158	0.0192	1	*Met*	?	NA	0	0.0133	0.024	0.0133	1	Met	NA	NA	1
0.0268	0.0175	0.0268	2	Ala	NA	NA	1	0.0304	0.0194	0.0304	2	Ser	NA	NA	1
0.0268	0.0204	0.0268	3	Ala	NA	NA	1	0.0285	0.0156	0.0285	3	Leu	NA	NA	1
0.0412	0.0159	0.0412	4	Ser	NA	NA	1	0.0057	0.019	0.0057	4	Gln	NA	NA	1
0	0.0157	0.0134	5	Phe	Leu	NA	0	0	0.0197	0.0133	5	Met	Lys	NA	0
0	0.0182	0.0278	6	Pro	Ser	NA	0	0.0304	0.014	0.0304	6	Val	NA	NA	1
0.0278	0.0155	0.0278	7	Pro	NA	NA	1	0.0342	0.019	0.0342	7	Thr	NA	NA	1
0.0201	0.0128	0.0201	8	Thr	NA	NA	1	0	0.0247	0.0304	8	*Val*	Ile	NA	0
0	0.0087	0.0249	9	Leu	Val	NA	0	0.0304	0.0197	0.0304	9	Ser	NA	NA	1

0.0144	0.0087	0.0144	10	Gly	NA	NA	1	0.0342	0.0129	0.0342	10	Asn	NA	NA	1
0.0249	0.0087	0.0249	11	Leu	NA	NA	1	0	0.0163	0.0342	11	Asn	Ser	NA	0
0	0.0059	0.0412	12	Ser	Asn	NA	0	0	0.0159	0.0266	12	Ile	Thr	NA	0
0	0.0062	0.0412	13	Ser	Ala	NA	0	0.0171	0.0144	0.0171	13	Ala	NA	NA	1
0	0.0062	0.0268	14	Ala	Thr	1	0	0.0285	0.0156	0.0285	14	Leu	NA	NA	1
0	0.0048	0.0278	15	Pro	Ser	NA	0	0.0266	0.0201	0.0266	15	Ile	NA	NA	1
0	0.0023	0.0221	16	Asp	His	1	0	0.0057	0.0243	0.0057	16	Gln	NA	NA	1
0	0.0049	0.0316	17	Glu	Asp	NA	0	0.0228	0.0186	0.0228	17	Pro	NA	NA	1
0.023	0.0072	0.023	18	Ile	NA	NA	1	0.038	0.0194	0.038	18	Gly	NA	NA	1
0	0.0102	0.0115	19	Gln	Arg	NA	0	0	0.0239	0.0171	19	Phe	Cys	NA	0
0	0.0115	0.0259	20	His	Tyr	1	0	0.0304	0.022	0.0304	20	Ser	NA	NA	1
0	0.0156	0.0278	21	Pro	Ser	1	0	0.0285	0.0144	0.0285	21	Leu	NA	NA	1
0.0259	0.0188	0.0259	22	His	NA	NA	1	0.0133	0.0178	0.0133	22	Met	NA	NA	1
0.023	0.017	0.023	23	Ile	NA	NA	1	0	0.0175	0.0342	23	Asn	Ser	NA	0
0.0297	0.017	0.0297	24	Lys	NA	NA	1	0.0171	0.0194	0.0171	24	Phe	NA	NA	1
0.0134	0.0183	0.0134	25	Phe	NA	NA	1	0.0285	0.0178	0.0285	25	Asp	NA	NA	1
0.0412	0.0213	0.0412	26	Ser	NA	NA	1	0.038	0.0239	0.038	26	Gly	NA	NA	1
0.0316	0.0187	0.0316	27	Glu	NA	NA	1	0.0057	0.0239	0.0057	27	Gln	NA	NA	1
0.0048	0.0164	0.0048	28	Trp	NA	NA	1	0.0304	0.0216	0.0304	28	Val	NA	NA	1
0	0.0134	0.0297	29	Lys	Arg	NA	0	0.0171	0.0175	0.0171	29	Phe	NA	NA	1
0.0134	0.0121	0.0134	30	Phe	NA	NA	1	0.0171	0.0163	0.0171	30	Phe	NA	NA	1
0.0297	0.0123	0.0297	31	Lys	NA	NA	1	0.0171	0.0114	0.0171	31	Phe	NA	NA	1
0	0.0091	0.0249	32	Leu	Val	NA	0	0	0.0114	0.038	32	Gly	Glu	NA	0
0	0.01	0.0134	33	Phe	Ile	NA	0	0.0057	0.008	0.0057	33	Gln	NA	NA	1
0	0.0131	0.0431	34	Arg	Trp	1	0	0.0171	0.0053	0.0171	34	Lys	NA	NA	1
0	0.0118	0.0278	35	Val	Met	1	0	0	0.0099	0.038	35	*Gly*	AlafsTer	1	0
0.0431	0.0088	0.0431	36	Arg	NA	NA	1	0.0038	0.0121	0.0038	36	Trp	NA	NA	1
0	0.0088	0.0412	37	Ser	Thr	1	0	0.0228	0.0133	0.0228	37	Pro	NA	NA	1
0.0134	0.012	0.0134	38	Phe	NA	NA	1	0.0171	0.0133	0.0171	38	Lys	NA	NA	1
0.0316	0.0151	0.0316	39	Glu	NA	NA	1	0.0228	0.0148	0.0228	39	Arg	NA	NA	1
0	0.0178	0.0297	40	Lys	Asn	NA	0	0	0.0148	0.0304	40	Ser	Tyr	NA	0
0	0.0147	0.0201	41	Thr	AspfsTer	NA	0	0.0114	0.0182	0.0114	41	Cys	NA	NA	1
0	0.0147	0.0278	42	Pro	Thr	1	0	0.0228	0.0137	0.0228	42	Pro	NA	NA	1
0.0316	0.0165	0.0316	43	Glu	NA	NA	1	0.0342	0.0137	0.0342	43	Thr	NA	NA	1
0.0316	0.0163	0.0316	44	Glu	NA	NA	1	0	0.0148	0.038	44	Gly	Ter	NA	0
0.0268	0.0163	0.0268	45	Ala	NA	NA	1	0	0.0102	0.0304	45	Val	Ile	NA	0
0.0115	0.0163	0.0115	46	Gln	NA	NA	1	0.0171	0.0034	0.0171	46	Phe	NA	NA	1
0	0.0163	0.0297	47	Lys	Glu	NA	0	0	0.0034	0.0171	47	His	Tyr	1	0
0.0316	0.0131	0.0316	48	Glu	NA	NA	1	0	0.0095	0.0285	48	Leu	Pro	NA	0
0.0297	0.0131	0.0297	49	Lys	NA	NA	1	0	0.0095	0.0285	49	Asp	Ala	NA	0
0	0.0105	0.0297	50	Lys	del	1	0	0.0304	0.0129	0.0304	50	Val	NA	NA	1
0	0.0093	0.0221	51	Asp	Tyr	1	0	0.0171	0.0129	0.0171	51	Lys	NA	NA	1
0	0.0121	0.0412	52	Ser	Phe	NA	0	0.0171	0.0129	0.0171	52	His	NA	NA	1
0	0.0089	0.0134	53	Phe	Ser	1	0	0	0.0068	0.0342	53	Asn	His	NA	0
0.0316	0.0059	0.0316	54	Glu	NA	NA	1	0	0.0068	0.0171	54	His	Arg	NA	0
0	0.0059	0.0144	55	Gly	Glu	NA	0	0	0.0091	0.0304	55	Val	Ile	NA	0
0	0.0059	0.0297	56	Lys	Glu	NA	0	0.0171	0.0125	0.0171	56	Lys	NA	NA	1
0.0278	0.0101	0.0278	57	Pro	NA	NA	1	0.0285	0.0171	0.0285	57	Leu	NA	NA	1
0	0.0101	0.0412	58	Ser	Pro	NA	0	0.0171	0.0171	0.0171	58	Lys	NA	NA	1
0	0.0096	0.0249	59	Leu	Arg	NA	0	0.0228	0.0137	0.0228	59	Pro	NA	NA	1
0	0.0096	0.0316	60	Glu	Val	NA	0	0	0.0114	0.0342	60	Thr	Ala	NA	0
0	0.0096	0.0115	61	Gln	Arg	NA	0	0	0.008	0.0266	61	Ile	Val	NA	0
0.0412	0.0068	0.0412	62	Ser	NA	NA	1	0.0171	0.0034	0.0171	62	Phe	NA	NA	1
0	0.0098	0.0278	63	Pro	Leu	NA	0	0	0.0034	0.0304	63	Ser	Tyr	NA	0
0.0268	0.0098	0.0268	64	Ala	NA	NA	1	0	0.0095	0.0171	64	Lys	Glu	NA	0
0	0.0098	0.0278	65	Val	Phe	1	0	0	0.0083	0.0285	65	Asp	Tyr	NA	0
0	0.0098	0.0249	66	Leu	Val	1	0	0.0304	0.0121	0.0304	66	Ser	NA	NA	1
0	0.0057	0.0221	67	Asp	His	NA	0	0.0114	0.0121	0.0114	67	Cys	NA	NA	1
0.0297	0.0086	0.0297	68	Lys	NA	NA	1	0.019	0.0167	0.019	68	Tyr	NA	NA	1
0	0.0087	0.0268	69	Ala	Gly	NA	0	0	0.0152	0.0285	69	Leu	Val	NA	0

0	0.0115	0.0221	70	Asp	Glu	NA	0	0.0228	0.0129	0.0228	70	Pro	NA	NA	1
0	0.0143	0.0144	71	Gly	Cys	NA	0	0.0228	0.0091	0.0228	71	Pro	NA	NA	1
0	0.0143	0.0115	72	Gln	Lys	1	0	0	0.0091	0.0285	72	Leu	Phe	NA	0
0.0297	0.0113	0.0297	73	Lys	NA	NA	1	0	0.0046	0.0228	73	Arg	Cys	1	0
0.0278	0.0113	0.0278	74	Pro	NA	NA	1	0	0.0034	0.019	74	Tyr	Ser	NA	0
0.0278	0.0138	0.0278	75	Val	NA	NA	1	0	0.0102	0.0228	75	Pro	Ala	NA	0
0.0278	0.0163	0.0278	76	Pro	NA	NA	1	0.0171	0.0102	0.0171	76	Ala	NA	NA	1
0	0.0193	0.0201	77	Thr	Asn	NA	0	0.0342	0.0171	0.0342	77	Thr	NA	NA	1
0	0.019	0.0115	78	Gln	Glu	1	0	0	0.0171	0.0114	78	Cys	Tyr	NA	0
0	0.0162	0.0278	79	Pro	Ala	1	0	0.0342	0.0137	0.0342	79	Thr	NA	NA	1
0.0249	0.0134	0.0249	80	Leu	NA	NA	1	0	0.0068	0.0171	80	Phe	Ile	1	0
0.0249	0.0106	0.0249	81	Leu	NA	NA	1	0	0.0129	0.0171	81	Lys	Arg	NA	0
0.0297	0.012	0.0297	82	Lys	NA	NA	1	0	0.0061	0.038	82	Gly	Ser	NA	0
0.0268	0.0161	0.0268	83	Ala	NA	NA	1	0.0304	0.0061	0.0304	83	Ser	NA	NA	1
0	0.0161	0.0259	84	His	Arg	NA	0	0	0.0121	0.0285	84	Leu	Phe	NA	0
0	0.0166	0.0278	85	Pro	Thr	1	0	0	0.0171	0.0247	85	Glu	Gln	NA	0
0	0.0122	0.0297	86	*Lys*	ValfsTer	1	0	0.0304	0.011	0.0304	86	Ser	NA	NA	1
0.0134	0.0125	0.0134	87	Phe	NA	NA	1	0.0247	0.011	0.0247	87	Glu	NA	NA	1
0.0412	0.0098	0.0412	88	Ser	NA	NA	1	0	0.011	0.0171	88	Lys	Glu	1	0
0	0.0098	0.0297	89	Lys	Asn	1	0	0	0.0087	0.0171	89	His	Arg	1	0
0.0297	0.0098	0.0297	90	Lys	NA	NA	1	0	0.0038	0.0057	90	Gln	Glu	NA	0
0.0134	0.0128	0.0134	91	Phe	NA	NA	1	0.019	0.0038	0.019	91	Tyr	NA	NA	1
0	0.0114	0.0259	92	His	Gln	NA	0	0	0.0072	0.0266	92	Ile	Phe	NA	0
0	0.0073	0.0221	93	Asp	Asn	NA	0	0	0.0072	0.0266	93	Ile	Thr	NA	0
0	0.0073	0.0125	94	Asn	del	NA	0	0.0171	0.011	0.0171	94	His	NA	NA	1
0	0.0073	0.0316	95	Glu	Lys	1	0	0	0.0144	0.038	95	*Gly*	Arg	NA	0
0.0297	0.0059	0.0297	96	Lys	NA	NA	1	0.038	0.0144	0.038	96	Gly	NA	NA	1
0	0.0059	0.0268	97	Ala	Val	NA	0	0.0171	0.011	0.0171	97	Lys	NA	NA	1
0	0.0059	0.0431	98	Arg	Ser	NA	0	0	0.011	0.0342	98	Thr	Ala	NA	0
0	0.0071	0.0144	99	Gly	Ser	1	0	0	0.0102	0.0228	99	Pro	Arg	NA	0
0.0297	0.0071	0.0297	100	Lys	NA	NA	1	0	0.0118	0.0342	100	Asn	Ser	NA	0
0	0.0041	0.0268	101	Ala	Glu	1	0	0.0342	0.0118	0.0342	101	Asn	NA	NA	1
0	0.0041	0.023	102	Ile	SerfsTer	NA	0	0.0247	0.0178	0.0247	102	Glu	NA	NA	1
0	0.0041	0.0259	103	His	Arg	NA	0	0	0.0178	0.0304	103	Val	Ile	NA	0
0.0115	0.0067	0.0115	104	Gln	NA	NA	1	0.0304	0.0144	0.0304	104	Ser	NA	NA	1
0	0.0062	0.0268	105	Ala	Val	NA	0	0	0.0095	0.0285	105	Asp	Tyr	NA	0
0	0.0075	0.0125	106	Asn	Ser	1	0	0.0171	0.0133	0.0171	106	Lys	NA	NA	1
0	0.0075	0.0249	107	Leu	Val	1	0	0	0.0072	0.0266	107	Ile	Phe	NA	0
0	0.007	0.0431	108	*Arg*	Gly	1	0	0.019	0.0072	0.019	108	Tyr	NA	NA	1
0.0259	0.0076	0.0259	109	His	NA	NA	1	0	0.0099	0.0304	109	Val	Asp	NA	0
0.0249	0.009	0.0249	110	Leu	NA	NA	1	0	0.0099	0.0133	110	Met	Val	1	0
0.0125	0.0103	0.0125	111	Cys	NA	NA	1	0.0304	0.0121	0.0304	111	Ser	NA	NA	1
0	0.0103	0.0431	112	Arg	Cys	1	0	0	0.0144	0.0266	112	Ile	Val	NA	0
0	0.0103	0.023	113	Ile	HisfsTer	NA	0	0.0304	0.0144	0.0304	113	Val	NA	NA	1
0.0125	0.012	0.0125	114	Cys	NA	NA	1	0.0114	0.0083	0.0114	114	Cys	NA	NA	1
0.0144	0.0095	0.0144	115	Gly	NA	NA	1	0	0.0152	0.0171	115	Lys	Glu	1	0
0.0125	0.0082	0.0125	116	Asn	NA	NA	1	0	0.0125	0.0342	116	Asn	Ile	1	0
0	0.0082	0.0412	117	Ser	Tyr	NA	0	0.0342	0.0102	0.0342	117	Asn	NA	NA	1
0	0.0108	0.0134	118	Phe	LeufsTer	1	0	0.0171	0.0102	0.0171	118	Lys	NA	NA	1
0.0431	0.0096	0.0431	119	Arg	NA	NA	1	0	0.0102	0.0171	119	Lys	ArgfsTer	NA	0
0	0.0081	0.0268	120	Ala	Val	NA	0	0	0.0068	0.0304	120	Val	LeufsTer	NA	0
0	0.0069	0.0221	121	Asp	Val	NA	0	0	0.0034	0.0342	121	Thr	Ile	NA	0
0	0.0077	0.0316	122	Glu	Gln	1	0	0.0171	0.0057	0.0171	122	Phe	NA	NA	1
0.0259	0.0105	0.0259	123	His	NA	NA	1	0	0.0057	0.0228	123	Arg	Cys	1	0
0	0.0061	0.0125	124	Asn	Asp	1	0	0.0114	0.0106	0.0114	124	Cys	NA	NA	1
0	0.0061	0.0431	125	Arg	Gly	NA	0	0	0.0106	0.0342	125	Thr	Ala	NA	0
0	0.0076	0.0431	126	Arg	Ser	1	0	0.0247	0.0163	0.0247	126	Glu	NA	NA	1
0.0077	0.0076	0.0077	127	Tyr	NA	NA	1	0.0171	0.014	0.0171	127	Lys	NA	NA	1
0.0278	0.005	0.0278	128	Pro	NA	NA	1	0.0285	0.0201	0.0285	128	Asp	NA	NA	1
0	0.0072	0.0278	129	Val	Gly	NA	0	0	0.0152	0.0285	129	Leu	ArgfsTer	NA	0

0	0.0086	0.0259	130	His	Tyr	1	0	0.0304	0.0118	0.0304	130	Val	NA	NA	1
0.0144	0.0116	0.0144	131	Gly	NA	NA	1	0	0.0061	0.038	131	Gly	Ala	NA	0
0	0.0128	0.0278	132	Pro	Ser	NA	0	0	0.0106	0.0285	132	Asp	Asn	NA	0
0	0.0126	0.0278	133	Val	Met	NA	0	0	0.0095	0.0304	133	Val	Ile	NA	0
0.0221	0.0126	0.0221	134	Asp	NA	NA	1	0.0228	0.0129	0.0228	134	Pro	NA	NA	1
0.0144	0.0151	0.0144	135	Gly	NA	NA	1	0.0247	0.0175	0.0247	135	Glu	NA	NA	1
0.0297	0.0161	0.0297	136	Lys	NA	NA	1	0.0171	0.0175	0.0171	136	Ala	NA	NA	1
0.0201	0.0161	0.0201	137	Thr	NA	NA	1	0.0228	0.0129	0.0228	137	Arg	NA	NA	1
0.0249	0.0191	0.0249	138	Leu	NA	NA	1	0	0.0114	0.019	138	Tyr	Ser	NA	0
0	0.0169	0.0144	139	Gly	Val	NA	0	0	0.014	0.038	139	Gly	Ser	1	0
0.0249	0.0154	0.0249	140	Leu	NA	NA	1	0.0171	0.0148	0.0171	140	His	NA	NA	1
0.0249	0.0154	0.0249	141	Leu	NA	NA	1	0.0304	0.0216	0.0304	141	Ser	NA	NA	1
0	0.0134	0.0431	142	Arg	Ter	1	0	0.0266	0.0277	0.0266	142	Ile	NA	NA	1
0.0297	0.0136	0.0297	143	Lys	NA	NA	1	0.0342	0.0243	0.0342	143	Asn	NA	NA	1
0	0.0156	0.0297	144	Lys	Glu	1	0	0.0304	0.0182	0.0304	144	Val	NA	NA	1
0	0.0173	0.0316	145	Glu	LysfsTer	1	0	0	0.0129	0.0304	145	Val	Leu	NA	0
0.0297	0.0152	0.0297	146	Lys	NA	NA	1	0	0.0061	0.019	146	Tyr	Cys	NA	0
0	0.0152	0.0431	147	Arg	SerfsTer	NA	0	0	0	0.0304	147	Ser	Cys	NA	0
0.0268	0.0123	0.0268	148	Ala	NA	NA	1	0	0	0.0228	148	Arg	Gln	NA	0
0.0201	0.0148	0.0201	149	Thr	NA	NA	1	0	0	0.038	149	Gly	Trp	NA	0
0.0412	0.0171	0.0412	150	Ser	NA	NA	1	0	0.0027	0.0171	150	Lys	Asn	NA	0
0.0048	0.0141	0.0048	151	Trp	NA	NA	1	0	0.0027	0.0304	151	Ser	Arg	NA	0
0	0.0141	0.0278	152	Pro	Leu	NA	0	0.0133	0.0027	0.0133	152	Met	NA	NA	1
0	0.0142	0.0221	153	Asp	Glu	NA	0	0	0.0027	0.038	153	Gly	Ala	NA	0
0.0249	0.0135	0.0249	154	Leu	NA	NA	1	0	0.0027	0.0304	154	Val	Ala	NA	0
0.023	0.0094	0.023	155	Ile	NA	NA	1	0	0	0.0285	155	Leu	Phe	1	0
0	0.0112	0.0268	156	Ala	Val	NA	0	0	0.0076	0.0171	156	Phe	Ser	NA	0
0	0.0112	0.0297	157	Lys	Arg	1	0	0	0.0076	0.038	157	Gly	Ter	NA	0
0.0278	0.0112	0.0278	158	Val	NA	NA	1	0.038	0.0076	0.038	158	Gly	NA	NA	1
0.0134	0.0117	0.0134	159	Phe	NA	NA	1	0	0.0114	0.0228	159	Arg	Cys	NA	0
0	0.0121	0.0431	160	Arg	Trp	1	0	0	0.0114	0.0304	160	Ser	Leu	NA	0
0.023	0.0143	0.023	161	Ile	NA	NA	1	0.019	0.0083	0.019	161	Tyr	NA	NA	1
0	0.0171	0.0221	162	Asp	Asn	1	0	0	0.0144	0.0133	162	Met	Arg	1	0
0	0.0165	0.0278	163	Val	GlyfsTer	1	0	0.0228	0.0144	0.0228	163	Pro	NA	NA	1
0.0297	0.0151	0.0297	164	Lys	NA	NA	1	0.0304	0.014	0.0304	164	Ser	NA	NA	1
0.0268	0.0151	0.0268	165	Ala	NA	NA	1	0	0.0186	0.0342	165	Thr	Ala	NA	0
0.0221	0.0128	0.0221	166	Asp	NA	NA	1	0.0171	0.0209	0.0171	166	His	NA	NA	1
0.0278	0.0156	0.0278	167	Val	NA	NA	1	0.0228	0.0148	0.0228	167	Arg	NA	NA	1
0.0221	0.0176	0.0221	168	Asp	NA	NA	1	0.0342	0.0197	0.0342	168	Thr	NA	NA	1
0	0.0147	0.0412	169	Ser	Trp	1	0	0	0.0197	0.0342	169	Thr	Lys	NA	0
0	0.0133	0.023	170	Ile	Thr	NA	0	0.0247	0.0152	0.0247	170	Glu	NA	NA	1
0	0.0111	0.0259	171	His	Tyr	NA	0	0.0171	0.0083	0.0171	171	Lys	NA	NA	1
0.0278	0.0109	0.0278	172	Pro	NA	NA	1	0	0.0083	0.0038	172	Trp	GlufsTer	1	0
0.0201	0.01	0.0201	173	Thr	NA	NA	1	0	0.0095	0.0342	173	Asn	Ser	NA	0
0	0.01	0.0316	174	Glu	SerfsTer	1	0	0	0.0095	0.0304	174	Ser	Asn	NA	0
0.0134	0.01	0.0134	175	Phe	NA	NA	1	0.0304	0.0095	0.0304	175	*Val*	NA	NA	1
0	0.01	0.0125	176	Cys	Phe	NA	0	0.0171	0.0118	0.0171	176	Ala	NA	NA	1
0.0259	0.0095	0.0259	177	His	NA	NA	1	0	0.0175	0.0285	177	Asp	Asn	NA	0
0.0125	0.0075	0.0125	178	Asn	NA	NA	1	0.0114	0.0114	0.0114	178	Cys	NA	NA	1
0	0.0101	0.0125	179	Cys	Ser	1	0	0.0285	0.0102	0.0285	179	Leu	NA	NA	1
0	0.013	0.0048	180	Trp	GlyfsTer	NA	0	0	0.0163	0.0228	180	Pro	Thr	NA	0
0	0.013	0.0412	181	Ser	Gly	1	0	0.0114	0.014	0.0114	181	Cys	NA	NA	1
0.023	0.0105	0.023	182	Ile	NA	NA	1	0.0304	0.0083	0.0304	182	Val	NA	NA	1
0	0.0092	0.0192	183	Met	Ile	NA	0	0	0.0144	0.0171	183	Phe	Leu	NA	0
0.0259	0.0092	0.0259	184	His	NA	NA	1	0	0.0178	0.0285	184	Leu	Pro	NA	0
0.0431	0.0119	0.0431	185	Arg	NA	NA	1	0.0304	0.0152	0.0304	185	Val	NA	NA	1
0	0.0147	0.0297	186	Lys	Thr	NA	0	0.0285	0.0201	0.0285	186	Asp	NA	NA	1
0	0.0124	0.0134	187	Phe	Ile	NA	0	0.0171	0.0235	0.0171	187	Phe	NA	NA	1
0	0.0124	0.0412	188	Ser	Ile	NA	0	0.0247	0.025	0.0247	188	Glu	NA	NA	1
0	0.0098	0.0412	189	Ser	Ile	NA	0	0.0171	0.0216	0.0171	189	Phe	NA	NA	1

0.0268	0.0062	0.0268	190	Ala	NA	NA	1	0.038	0.0182	0.038	190	Gly	NA	NA	1
0.0278	0.0076	0.0278	191	Pro	NA	NA	1	0.0114	0.0201	0.0114	191	Cys	NA	NA	1
0	0.0076	0.0125	192	Cys	Tyr	NA	0	0	0.0167	0.0171	192	Ala	Thr	NA	0
0	0.0119	0.0316	193	Glu	Lys	NA	0	0.0342	0.0091	0.0342	193	Thr	NA	NA	1
0	0.0119	0.0278	194	Val	Phe	NA	0	0	0.0068	0.0304	194	Ser	Thr	NA	0
0.0077	0.0092	0.0077	195	Tyr	NA	NA	1	0	0.0125	0.019	195	Tyr	His	NA	0
0.0134	0.0084	0.0134	196	Phe	NA	NA	1	0	0.0102	0.0266	196	Ile	Phe	NA	0
0	0.0084	0.0278	197	Pro	Gln	1	0	0.0285	0.0102	0.0285	197	Leu	NA	NA	1
0.0431	0.0084	0.0431	198	Arg	NA	NA	1	0.0228	0.0159	0.0228	198	Pro	NA	NA	1
0	0.0084	0.0125	199	Asn	Asp	NA	0	0	0.0171	0.0247	199	Glu	Ter	NA	0
0	0.0077	0.0278	200	Val	Met	NA	0	0.0285	0.0171	0.0285	200	Leu	NA	NA	1
0.0201	0.0063	0.0201	201	Thr	NA	NA	1	0.0057	0.0201	0.0057	201	Gln	NA	NA	1
0	0.0063	0.0192	202	Met	Thr	NA	0	0.0285	0.0258	0.0285	202	Asp	NA	NA	1
0	0.004	0.0316	203	Glu	Gln	NA	0	0.038	0.0201	0.038	203	Gly	NA	NA	1
0	0.004	0.0048	204	Trp	Ter	NA	0	0.0285	0.0224	0.0285	204	Leu	NA	NA	1
0	0.0081	0.0259	205	His	Gln	NA	0	0	0.0167	0.0304	205	Ser	Ala	NA	0
0	0.0074	0.0278	206	Pro	Ser	NA	0	0.0171	0.0152	0.0171	206	Phe	NA	NA	1
0	0.0074	0.0259	207	His	Asn	1	0	0	0.0156	0.0171	207	His	Leu	NA	0
0.0201	0.0097	0.0201	208	Thr	NA	NA	1	0.0304	0.0156	0.0304	208	Val	NA	NA	1
0	0.0109	0.0278	209	Pro	Leu	NA	0	0.0304	0.0121	0.0304	209	Ser	NA	NA	1
0.0412	0.0109	0.0412	210	Ser	NA	NA	1	0	0.0156	0.0266	210	*Ile*	Val	NA	0
0.0125	0.0129	0.0125	211	Cys	NA	NA	1	0	0.0095	0.0171	211	Ala	Thr	NA	0
0	0.0129	0.0221	212	Asp	Asn	NA	0	0.0171	0.0091	0.0171	212	Lys	NA	NA	1
0.023	0.0109	0.023	213	Ile	NA	NA	1	0	0.0091	0.0342	213	Asn	Thr	NA	0
0.0125	0.0109	0.0125	214	Cys	NA	NA	1	0.0285	0.0144	0.0285	214	Asp	NA	NA	1
0	0.0082	0.0125	215	Asn	Asp	NA	0	0	0.0148	0.0342	215	Thr	Asn	1	0
0.0201	0.0095	0.0201	216	Thr	NA	NA	1	0.0266	0.0201	0.0266	216	Ile	NA	NA	1
0	0.0125	0.0268	217	Ala	Asp	1	0	0.019	0.0201	0.019	217	Tyr	NA	NA	1
0	0.0145	0.0431	218	Arg	Cys	1	0	0.0266	0.0277	0.0266	218	Ile	NA	NA	1
0	0.0162	0.0431	219	Arg	Trp	1	0	0.0285	0.03	0.0285	219	Leu	NA	NA	1
0.0144	0.0162	0.0144	220	Gly	NA	NA	1	0.038	0.0296	0.038	220	Gly	NA	NA	1
0.0249	0.0167	0.0249	221	Leu	NA	NA	1	0.038	0.0304	0.038	221	Gly	NA	NA	1
0.0297	0.0167	0.0297	222	Lys	NA	NA	1	0.0171	0.0304	0.0171	222	His	NA	NA	1
0.0431	0.0195	0.0431	223	Arg	NA	NA	1	0.0304	0.0262	0.0304	223	Ser	NA	NA	1
0.0297	0.0207	0.0297	224	Lys	NA	NA	1	0.0285	0.0186	0.0285	224	Leu	NA	NA	1
0	0.0218	0.0412	225	Ser	Ile	NA	0	0.0171	0.022	0.0171	225	Ala	NA	NA	1
0.0249	0.0204	0.0249	226	Leu	NA	NA	1	0	0.0159	0.0342	226	Asn	Thr	1	0
0	0.0174	0.0115	227	Gln	Leu	1	0	0.0342	0.0102	0.0342	227	Asn	NA	NA	1
0.0278	0.0131	0.0278	228	Pro	NA	NA	1	0	0.0114	0.0266	228	Ile	Val	NA	0
0.0125	0.0131	0.0125	229	Asn	NA	NA	1	0	0.0148	0.0228	229	*Arg*	Gln	1	0
0.0249	0.0131	0.0249	230	Leu	NA	NA	1	0.0228	0.008	0.0228	230	Pro	NA	NA	1
0.0115	0.0106	0.0115	231	Gln	NA	NA	1	0.0171	0.008	0.0171	231	Ala	NA	NA	1
0	0.0136	0.0249	232	Leu	Val	NA	0	0	0.0118	0.0342	232	Asn	Ser	NA	0
0	0.0108	0.0412	233	Ser	Asn	NA	0	0	0.0072	0.0285	233	Leu	Val	NA	0
0.0297	0.0124	0.0297	234	Lys	NA	NA	1	0.019	0.0091	0.019	234	Tyr	NA	NA	1
0	0.0099	0.0297	235	Lys	Glu	NA	0	0	0.0091	0.0228	235	Arg	Thr	NA	0
0	0.0109	0.0249	236	Leu	ThrfsTer	NA	0	0.0266	0.0091	0.0266	236	Ile	NA	NA	1
0.0297	0.0109	0.0297	237	Lys	NA	NA	1	0	0.011	0.0228	237	Arg	Ser	NA	0
0	0.0109	0.0201	238	Thr	Asn	NA	0	0	0.011	0.0304	238	Val	Ile	NA	0
0.0278	0.0123	0.0278	239	Val	NA	NA	1	0.0285	0.0057	0.0285	239	Asp	NA	NA	1
0	0.0134	0.0249	240	Leu	Phe	NA	0	0	0.0114	0.0285	240	Leu	Ile	1	0
0.0221	0.0161	0.0221	241	Asp	NA	NA	1	0	0.0114	0.0228	241	Pro	Leu	NA	0
0	0.0131	0.0115	242	Gln	Arg	NA	0	0.0285	0.0118	0.0285	242	Leu	NA	NA	1
0	0.0131	0.0268	243	Ala	Thr	NA	0	0	0.0163	0.038	243	Gly	Ser	NA	0
0.0431	0.0104	0.0431	244	Arg	NA	NA	1	0.0304	0.0197	0.0304	244	Ser	NA	NA	1
0.0115	0.0133	0.0115	245	Gln	NA	NA	1	0.0228	0.0201	0.0228	245	Pro	NA	NA	1
0.0268	0.0111	0.0268	246	Ala	NA	NA	1	0.0171	0.0269	0.0171	246	Ala	NA	NA	1
0	0.0111	0.0431	247	Arg	Gly	1	0	0.0304	0.0231	0.0304	247	Val	NA	NA	1
0	0.0138	0.0115	248	Gln	Glu	NA	0	0.0342	0.0186	0.0342	248	Asn	NA	NA	1
0	0.0095	0.0259	249	His	Arg	NA	0	0.0114	0.0152	0.0114	249	Cys	NA	NA	1

0.0297	0.011	0.0297	250	Lys	NA	NA	1	0	0.0148	0.0342	250	Thr	SerfsTer	NA	0
0	0.0127	0.0431	251	Arg	SerfsTer	1	0	0	0.0125	0.0304	251	Val	SerfsTer	1	0
0	0.015	0.0431	252	Arg	dup	NA	0	0.0285	0.0178	0.0285	252	Leu	NA	NA	1
0.0268	0.0191	0.0268	253	Ala	NA	NA	1	0.0228	0.0178	0.0228	253	Pro	NA	NA	1
0	0.0191	0.0115	254	Gln	Ter	NA	0	0.038	0.0178	0.038	254	Gly	NA	NA	1
0.0268	0.0161	0.0268	255	Ala	NA	NA	1	0	0.0182	0.038	255	Gly	Glu	NA	0
0.0431	0.0161	0.0431	256	Arg	NA	NA	1	0	0.0197	0.0266	256	Ile	Met	NA	0
0.023	0.0189	0.023	257	Ile	NA	NA	1	0.0304	0.0121	0.0304	257	Ser	NA	NA	1
0.0412	0.0162	0.0412	258	Ser	NA	NA	1	0.0304	0.0121	0.0304	258	Val	NA	NA	1
0	0.0192	0.0412	259	Ser	Asn	1	0	0	0.0156	0.0304	259	Ser	Phe	NA	0
0	0.0165	0.0297	260	Lys	Asn	NA	0	0	0.0095	0.0304	260	Ser	Arg	NA	0
0	0.0145	0.0221	261	Asp	SerfsTer	1	0	0.0171	0.0091	0.0171	261	Ala	NA	NA	1
0.0278	0.0122	0.0278	262	Val	NA	NA	1	0	0.0091	0.0266	262	Ile	Thr	NA	0
0	0.0081	0.0192	263	Met	Arg	NA	0	0.0285	0.0102	0.0285	263	Leu	NA	NA	1
0.0297	0.0093	0.0297	264	Lys	NA	NA	1	0	0.0137	0.0342	264	Thr	Ile	NA	0
0	0.0093	0.0297	265	Lys	del	NA	0	0.0057	0.0205	0.0057	265	Gln	NA	NA	1
0.023	0.0123	0.023	266	Ile	NA	NA	1	0.0342	0.0148	0.0342	266	Thr	NA	NA	1
0	0.0095	0.0268	267	Ala	Thr	1	0	0.0342	0.0205	0.0342	267	Asn	NA	NA	1
0	0.0095	0.0125	268	Asn	Ser	NA	0	0	0.0194	0.0342	268	Asn	Ser	1	0
0.0125	0.009	0.0125	269	Cys	NA	NA	1	0.0285	0.0159	0.0285	269	Asp	NA	NA	1
0	0.009	0.0412	270	Ser	Thr	NA	0	0	0.0091	0.0247	270	Glu	Ala	1	0
0.0297	0.0087	0.0297	271	Lys	NA	NA	1	0.0171	0.0144	0.0171	271	Phe	NA	NA	1
0	0.0087	0.023	272	Ile	Val	1	0	0	0.0087	0.0304	272	Val	Ile	1	0
0	0.0087	0.0259	273	His	Arg	NA	0	0.0266	0.0163	0.0266	273	Ile	NA	NA	1
0.0249	0.01	0.0249	274	Leu	NA	NA	1	0	0.0205	0.0304	274	Val	Leu	1	0
0	0.0127	0.0412	275	Ser	Gly	1	0	0.038	0.0205	0.038	275	Gly	NA	NA	1
0.0201	0.0125	0.0201	276	Thr	NA	NA	1	0.038	0.0152	0.038	276	Gly	NA	NA	1
0	0.0125	0.0297	277	Lys	Arg	NA	0	0	0.0209	0.019	277	Tyr	Cys	NA	0
0	0.0138	0.0249	278	Leu	Phe	NA	0	0	0.0182	0.0057	278	Gln	His	NA	0
0.0249	0.0113	0.0249	279	Leu	NA	NA	1	0.0285	0.0175	0.0285	279	Leu	NA	NA	1
0.0268	0.0113	0.0268	280	Ala	NA	NA	1	0.0247	0.0186	0.0247	280	Glu	NA	NA	1
0.0278	0.0093	0.0278	281	Val	NA	NA	1	0.0342	0.022	0.0342	281	Asn	NA	NA	1
0	0.0106	0.0221	282	Asp	His	NA	0	0.0057	0.0209	0.0057	282	Gln	NA	NA	1
0.0134	0.0134	0.0134	283	Phe	NA	NA	1	0.0171	0.0186	0.0171	283	Lys	NA	NA	1
0	0.0139	0.0278	284	Pro	Ser	NA	0	0.0228	0.0118	0.0228	284	Arg	NA	NA	1
0	0.0112	0.0316	285	Glu	Gln	NA	0	0.0133	0.0106	0.0133	285	Met	NA	NA	1
0	0.0084	0.0259	286	His	Tyr	NA	0	0	0.0072	0.0266	286	Ile	Phe	NA	0
0.0134	0.0084	0.0134	287	Phe	NA	NA	1	0	0.0027	0.0114	287	Cys	AlafsTer	NA	0
0.0278	0.0083	0.0278	288	Val	NA	NA	1	0	0.0053	0.0342	288	Asn	ThrfsTer	1	0
0.0297	0.0083	0.0297	289	Lys	NA	NA	1	0	0.0053	0.0266	289	Ile	Asn	NA	0
0	0.0106	0.0412	290	Ser	Phe	NA	0	0.0266	0.011	0.0266	290	Ile	NA	NA	1
0	0.0119	0.023	291	Ile	Val	NA	0	0	0.011	0.0304	291	Ser	Tyr	NA	0
0	0.0137	0.0412	292	Ser	Phe	NA	0	0.0285	0.011	0.0285	292	Leu	NA	NA	1
0.0125	0.0109	0.0125	293	Cys	NA	NA	1	0	0.0125	0.0247	293	Glu	Gly	1	0
0	0.0103	0.0115	294	Gln	Arg	1	0	0	0.0159	0.0285	294	Asp	Glu	NA	0
0.023	0.0128	0.023	295	Ile	NA	NA	1	0.0342	0.0156	0.0342	295	Asn	NA	NA	1
0.0125	0.0154	0.0125	296	Cys	NA	NA	1	0.0171	0.0156	0.0171	296	Lys	NA	NA	1
0.0316	0.0154	0.0316	297	Glu	NA	NA	1	0.0266	0.0209	0.0266	297	Ile	NA	NA	1
0	0.0142	0.0259	298	His	Tyr	1	0	0	0.014	0.0247	298	Glu	Asp	NA	0
0.023	0.017	0.023	299	Ile	NA	NA	1	0.0266	0.0156	0.0266	299	Ile	NA	NA	1
0.0249	0.0147	0.0249	300	Leu	NA	NA	1	0	0.0129	0.0228	300	Arg	Cys	1	0
0.0268	0.0134	0.0268	301	Ala	NA	NA	1	0.0247	0.0129	0.0247	301	Glu	NA	NA	1
0	0.0115	0.0221	302	Asp	Glu	NA	0	0.0133	0.0144	0.0133	302	Met	NA	NA	1
0	0.0128	0.0278	303	Pro	His	NA	0	0	0.0144	0.0247	303	Glu	Gly	1	0
0.0278	0.0134	0.0278	304	Val	NA	NA	1	0.0342	0.0152	0.0342	304	Thr	NA	NA	1
0	0.0109	0.0316	305	Glu	Asp	NA	0	0	0.0133	0.0228	305	Pro	Ala	1	0
0	0.011	0.0201	306	Thr	Ile	NA	0	0.0285	0.0201	0.0285	306	Asp	NA	NA	1
0.0125	0.0124	0.0125	307	Asn	NA	NA	1	0.0038	0.0178	0.0038	307	Trp	NA	NA	1
0.0125	0.0136	0.0125	308	Cys	NA	NA	1	0.0342	0.0235	0.0342	308	Thr	NA	NA	1
0.0297	0.0151	0.0297	309	Lys	NA	NA	1	0.0228	0.0231	0.0228	309	Pro	NA	NA	1

0	0.0151	0.0259	310	His	Gln	NA	0	0.0285	0.0258	0.0285	310	Asp	NA	NA	1
0.0278	0.0164	0.0278	311	Val	NA	NA	1	0.0266	0.019	0.0266	311	Ile	NA	NA	1
0.0134	0.0174	0.0134	312	Phe	NA	NA	1	0.0171	0.0144	0.0171	312	Lys	NA	NA	1
0.0125	0.0187	0.0125	313	Cys	NA	NA	1	0	0.0121	0.0171	313	His	Arg	NA	0
0.0431	0.02	0.0431	314	Arg	NA	NA	1	0	0.0068	0.0304	314	Ser	Ile	1	0
0	0.02	0.0278	315	Val	Ile	1	0	0.0171	0.0042	0.0171	315	Lys	NA	NA	1
0.0125	0.0198	0.0125	316	Cys	NA	NA	1	0	0.0076	0.0266	316	Ile	Met	NA	0
0.023	0.0214	0.023	317	Ile	NA	NA	1	0.0038	0.0152	0.0038	317	Trp	NA	NA	1
0.0249	0.0229	0.0249	318	Leu	NA	NA	1	0.0171	0.0178	0.0171	318	Phe	NA	NA	1
0.0431	0.0205	0.0431	319	Arg	NA	NA	1	0.038	0.0178	0.038	319	Gly	NA	NA	1
0	0.0205	0.0125	320	Cys	Tyr	NA	0	0.0304	0.0197	0.0304	320	Ser	NA	NA	1
0.0249	0.0193	0.0249	321	Leu	NA	NA	1	0	0.0239	0.0342	321	Asn	Ser	NA	0
0.0297	0.0177	0.0297	322	Lys	NA	NA	1	0.0133	0.0163	0.0133	322	Met	NA	NA	1
0.0278	0.0152	0.0278	323	Val	NA	NA	1	0.038	0.0178	0.038	323	Gly	NA	NA	1
0.0192	0.0137	0.0192	324	Met	NA	NA	1	0	0.0178	0.0342	324	Asn	Ser	NA	0
0	0.0178	0.0144	325	Gly	Asp	NA	0	0.038	0.0152	0.038	325	Gly	NA	NA	1
0	0.0153	0.0412	326	Ser	Asn	NA	0	0	0.011	0.0342	326	Thr	Asn	1	0
0.0077	0.0124	0.0077	327	Tyr	NA	NA	1	0	0.0167	0.0304	327	Val	Ile	NA	0
0	0.0104	0.0125	328	*Cys*	Tyr	NA	0	0.0171	0.0167	0.0171	328	Phe	NA	NA	1
0.0278	0.0112	0.0278	329	Pro	NA	NA	1	0.0285	0.022	0.0285	329	Leu	NA	NA	1
0.0412	0.0112	0.0412	330	Ser	NA	NA	1	0.038	0.022	0.038	330	Gly	NA	NA	1
0	0.0126	0.0125	331	Cys	Gly	NA	0	0.0266	0.0186	0.0266	331	Ile	NA	NA	1
0	0.0146	0.0431	332	Arg	Gly	1	0	0	0.0129	0.0228	332	Pro	Arg	NA	0
0.0077	0.0146	0.0077	333	Tyr	NA	NA	1	0	0.0121	0.038	333	Gly	Glu	NA	0
0.0278	0.014	0.0278	334	Pro	NA	NA	1	0	0.0068	0.0285	334	Asp	Asn	NA	0
0	0.0124	0.0125	335	Cys	Phe	NA	0	0.0342	0.008	0.0342	335	Asn	NA	NA	1
0.0134	0.0155	0.0134	336	Phe	NA	NA	1	0	0.008	0.0171	336	Lys	Glu	NA	0
0.0278	0.0155	0.0278	337	Pro	NA	NA	1	0.0057	0.014	0.0057	337	Gln	NA	NA	1
0	0.0175	0.0201	338	Thr	Pro	NA	0	0	0.0133	0.0304	338	Val	LeufsTer	NA	0
0.0221	0.0148	0.0221	339	Asp	NA	NA	1	0.0304	0.0133	0.0304	339	Val	NA	NA	1
0.0249	0.0177	0.0249	340	Leu	NA	NA	1	0.0304	0.0121	0.0304	340	Ser	NA	NA	1
0.0316	0.0164	0.0316	341	Glu	NA	NA	1	0	0.0121	0.0247	341	Glu	Gln	NA	0
0	0.0136	0.0412	342	Ser	Gly	NA	0	0	0.0061	0.038	342	Gly	Glu	1	0
0.0278	0.0161	0.0278	343	Pro	NA	NA	1	0	0.0034	0.0171	343	Phe	Ser	1	0
0	0.018	0.0278	344	Val	Leu	NA	0	0	0.0034	0.019	344	Tyr	His	1	0
0.0297	0.0155	0.0297	345	Lys	NA	NA	1	0.0171	0.0034	0.0171	345	Phe	NA	NA	1
0	0.0149	0.0412	346	Ser	Cys	NA	0	0	0.0091	0.019	346	Tyr	Cys	NA	0
0	0.0161	0.0134	347	Phe	Leu	NA	0	0	0.0125	0.0133	347	Met	Thr	1	0
0.0249	0.0174	0.0249	348	Leu	NA	NA	1	0.0285	0.0114	0.0285	348	Leu	NA	NA	1
0.0412	0.0174	0.0412	349	Ser	NA	NA	1	0.0171	0.0114	0.0171	349	Lys	NA	NA	1
0	0.0145	0.0278	350	Val	Ile	NA	0	0.0114	0.0163	0.0114	350	Cys	NA	NA	1
0.0249	0.0145	0.0249	351	Leu	NA	NA	1	0	0.0163	0.0171	351	Ala	Thr	NA	0
0.0125	0.0174	0.0125	352	Asn	NA	NA	1	0.0247	0.0129	0.0247	352	Glu	NA	NA	1
0.0412	0.015	0.0412	353	Ser	NA	NA	1	0.0285	0.0106	0.0285	353	Asp	NA	NA	1
0	0.0108	0.0249	354	Leu	Met	NA	0	0	0.0175	0.0285	354	Asp	Asn	NA	0
0	0.0108	0.0192	355	Met	Thr	NA	0	0	0.0175	0.0342	355	Thr	Ile	NA	0
0	0.0113	0.0278	356	Val	Met	NA	0	0.0342	0.0167	0.0342	356	Asn	NA	NA	1
0.0297	0.0132	0.0297	357	Lys	NA	NA	1	0.0247	0.0167	0.0247	357	Glu	NA	NA	1
0	0.0091	0.0125	358	*Cys*	Phe	NA	0	0.0247	0.0167	0.0247	358	Glu	NA	NA	1
0	0.0091	0.0278	359	Pro	Leu	NA	0	0	0.0167	0.0057	359	Gln	Arg	NA	0
0	0.0091	0.0268	360	Ala	Thr	1	0	0	0.0152	0.0342	360	Thr	Ala	NA	0
0.0297	0.0091	0.0297	361	Lys	NA	NA	1	0.0342	0.0102	0.0342	361	Thr	NA	NA	1
0.0316	0.0089	0.0316	362	Glu	NA	NA	1	0.0171	0.0102	0.0171	362	Phe	NA	NA	1
0	0.0089	0.0125	363	Cys	Tyr	1	0	0	0.0163	0.0342	363	Thr	Ile	NA	0
0	0.0089	0.0125	364	Asn	Asp	1	0	0	0.0106	0.0342	364	Asn	Lys	1	0
0	0.0089	0.0316	365	Glu	Lys	NA	0	0.0304	0.014	0.0304	365	Ser	NA	NA	1
0	0.0089	0.0316	366	Glu	Gly	1	0	0.0057	0.0201	0.0057	366	Gln	NA	NA	1
0.0278	0.0058	0.0278	367	Val	NA	NA	1	0.0342	0.0269	0.0342	367	Thr	NA	NA	1
0	0.007	0.0412	368	Ser	Asn	NA	0	0.0304	0.0258	0.0304	368	Ser	NA	NA	1
0	0.007	0.0249	369	Leu	Phe	NA	0	0.0342	0.0304	0.0342	369	Thr	NA	NA	1

0	0.007	0.0316	370	Glu	Lys	NA	0	0.0247	0.0281	0.0247	370	Glu	NA	NA	1
0.0297	0.0093	0.0297	371	Lys	NA	NA	1	0.0285	0.022	0.0285	371	Asp	NA	NA	1
0	0.0106	0.0077	372	Tyr	Cys	NA	0	0.0228	0.0152	0.0228	372	Pro	NA	NA	1
0.0125	0.0148	0.0125	373	Asn	NA	NA	1	0	0.0102	0.038	373	Gly	Arg	NA	0
0	0.0148	0.0259	374	His	Leu	NA	0	0	0.0046	0.0285	374	Asp	His	NA	0
0	0.0148	0.0259	375	*His*	del	1	0	0	0	0.0304	375	Ser	Tyr	NA	0
0.023	0.015	0.023	376	Ile	NA	NA	1	0	0.0034	0.0342	376	Thr	Asn	1	0
0.0412	0.0191	0.0412	377	Ser	NA	NA	1	0	0.0083	0.0228	377	Pro	Leu	NA	0
0.0412	0.0208	0.0412	378	Ser	NA	NA	1	0.0171	0.0083	0.0171	378	Phe	NA	NA	1
0	0.024	0.0259	379	His	GlnfsTer	NA	0	0.0247	0.0083	0.0247	379	Glu	NA	NA	1
0	0.024	0.0297	380	Lys	ArgfsTer	1	0	0	0.0083	0.0285	380	Asp	Tyr	NA	0
0.0316	0.023	0.0316	381	Glu	NA	NA	1	0	0.0049	0.0304	381	*Ser*	Ter	NA	0
0.0412	0.0189	0.0412	382	Ser	NA	NA	1	0	0	0.0247	382	Glu	Gly	NA	0
0.0297	0.0148	0.0297	383	Lys	NA	NA	1	0	0	0.0247	383	Glu	Lys	NA	0
0.0316	0.0171	0.0316	384	Glu	NA	NA	1	0	0	0.0171	384	Phe	Cys	NA	0
0	0.0183	0.023	385	Ile	Thr	NA	0	0	0.0061	0.0114	385	Cys	SerfsTer	NA	0
0.0134	0.0181	0.0134	386	Phe	NA	NA	1	0	0.0061	0.0171	386	Phe	Cys	1	0
0	0.014	0.0278	387	Val	Leu	1	0	0.0304	0.011	0.0304	387	Ser	NA	NA	1
0	0.0125	0.0259	388	His	Tyr	NA	0	0	0.011	0.0171	388	Ala	Glu	NA	0
0.023	0.0136	0.023	389	Ile	NA	NA	1	0.0247	0.0178	0.0247	389	Glu	NA	NA	1
0.0125	0.0164	0.0125	390	Asn	NA	NA	1	0	0.0118	0.0171	390	Ala	Glu	1	0
0.0297	0.0151	0.0297	391	Lys	NA	NA	1	0.0342	0.0118	0.0342	391	Asn	NA	NA	1
0	0.0162	0.0144	392	Gly	Arg	NA	0	0	0.0125	0.0304	392	Ser	Gly	NA	0
0.0144	0.0162	0.0144	393	Gly	NA	NA	1	0	0.0125	0.0171	393	Phe	LeufsTer	NA	0
0.0431	0.0164	0.0431	394	Arg	NA	NA	1	0.0285	0.0057	0.0285	394	Asp	NA	NA	1
0.0278	0.0176	0.0278	395	Pro	NA	NA	1	0	0.0057	0.038	395	Gly	Asp	1	0
0	0.0147	0.0431	396	*Arg*	Cys	1	0	0	0.0106	0.0285	396	Asp	Gly	NA	0
0.0115	0.0172	0.0115	397	Gln	NA	NA	1	0	0.0049	0.0285	397	Asp	Gly	1	0
0	0.0157	0.0259	398	His	Arg	NA	0	0.0247	0.0049	0.0247	398	Glu	NA	NA	1
0.0249	0.0114	0.0249	399	Leu	NA	NA	1	0	0.0049	0.0171	399	Phe	Leu	1	0
0.0249	0.0129	0.0249	400	Leu	NA	NA	1	0	0.0049	0.0285	400	*Asp*	His	NA	0
0	0.0156	0.0412	401	*Ser*	Leu	NA	0	0	0	0.0342	401	Thr	Ser	1	0
0.0249	0.0156	0.0249	402	Leu	NA	NA	1	0	0.0049	0.019	402	Tyr	Cys	NA	0
0	0.0186	0.0201	403	*Thr*	Ser	NA	0	0	0.0049	0.0342	403	Asn	MetfsTer	NA	0
0	0.0187	0.0431	404	*Arg*	Trp	1	0	0.0247	0.0106	0.0247	404	Glu	NA	NA	1
0.0431	0.0162	0.0431	405	Arg	NA	NA	1	0	0.0156	0.0285	405	Asp	Glu	NA	0
0.0268	0.0187	0.0268	406	Ala	NA	NA	1	0.0285	0.0156	0.0285	406	Asp	NA	NA	1
0.0115	0.0162	0.0115	407	Gln	NA	NA	1	0.0247	0.0163	0.0247	407	Glu	NA	NA	1
0.0297	0.0194	0.0297	408	Lys	NA	NA	1	0	0.0213	0.0247	408	Glu	Lys	1	0
0.0259	0.0219	0.0259	409	His	NA	NA	1	0.0285	0.0216	0.0285	409	Asp	NA	NA	1
0	0.0205	0.0431	410	*Arg*	Trp	1	0	0.0247	0.0167	0.0247	410	Glu	NA	NA	1
0.0249	0.0203	0.0249	411	Leu	NA	NA	1	0.0304	0.0235	0.0304	411	Ser	NA	NA	1
0	0.0203	0.0431	412	Arg	Leu	1	0	0	0.0254	0.0247	412	Glu	Gln	NA	0
0.0316	0.0174	0.0316	413	Glu	NA	NA	1	0.0342	0.0243	0.0342	413	Thr	NA	NA	1
0.0249	0.0148	0.0249	414	Leu	NA	NA	1	0.038	0.0182	0.038	414	Gly	NA	NA	1
0.0297	0.0174	0.0297	415	Lys	NA	NA	1	0.019	0.0235	0.019	415	Tyr	NA	NA	1
0.0249	0.0163	0.0249	416	Leu	NA	NA	1	0	0.0167	0.0038	416	Trp	Leu	NA	0
0.0115	0.0163	0.0115	417	Gln	NA	NA	1	0.0266	0.0114	0.0266	417	Ile	NA	NA	1
0	0.0153	0.0278	418	Val	Ile	NA	0	0	0.0099	0.0342	418	Thr	Ile	NA	0
0	0.0158	0.0297	419	Lys	Arg	NA	0	0.0114	0.0099	0.0114	419	Cys	NA	NA	1
0.0268	0.016	0.0268	420	Ala	NA	NA	1	0.0114	0.0114	0.0114	420	Cys	NA	NA	1
0.0134	0.0167	0.0134	421	Phe	NA	NA	1	0	0.0114	0.0228	421	Pro	Ser	NA	0
0	0.017	0.0268	422	Ala	Ser	NA	0	0.0342	0.0148	0.0342	422	Thr	NA	NA	1
0.0221	0.0184	0.0221	423	Asp	NA	NA	1	0	0.0186	0.0114	423	Cys	Tyr	NA	0
0.0297	0.0206	0.0297	424	Lys	NA	NA	1	0.0285	0.0186	0.0285	424	Asp	NA	NA	1
0.0316	0.0179	0.0316	425	Glu	NA	NA	1	0.0304	0.0171	0.0304	425	Val	NA	NA	1
0.0316	0.0196	0.0316	426	Glu	NA	NA	1	0	0.0171	0.0285	426	Asp	Asn	1	0
0.0144	0.0237	0.0144	427	Gly	NA	NA	1	0.0266	0.0182	0.0266	427	Ile	NA	NA	1
0.0144	0.0215	0.0144	428	Gly	NA	NA	1	0	0.0129	0.0342	428	Asn	His	1	0
0.0221	0.0198	0.0221	429	Asp	NA	NA	1	0.0342	0.019	0.0342	429	Thr	NA	NA	1

0	0.0166	0.0278	430	Val	Met	NA	0	0.0038	0.0137	0.0038	430	Trp	NA	NA	1
0.0297	0.0134	0.0297	431	Lys	NA	NA	1	0.0304	0.0171	0.0304	431	Val	NA	NA	1
0.0412	0.0145	0.0412	432	Ser	NA	NA	1	0	0.0102	0.0228	432	Pro	Arg	NA	0
0	0.0144	0.0278	433	*Val*	Ala	NA	0	0.0171	0.0156	0.0171	433	Phe	NA	NA	1
0.0125	0.0147	0.0125	434	Cys	NA	NA	1	0	0.0095	0.019	434	Tyr	Cys	NA	0
0	0.0172	0.0192	435	*Met*	Val	NA	0	0.0304	0.0095	0.0304	435	Ser	NA	NA	1
0	0.0169	0.0201	436	Thr	Ile	NA	0	0	0.0061	0.0342	436	Thr	Asn	NA	0
0.0249	0.0152	0.0249	437	Leu	NA	NA	1	0	0.0061	0.0247	437	*Glu*	Lys	NA	0
0.0134	0.0152	0.0134	438	Phe	NA	NA	1	0	0	0.0285	438	Leu	Phe	1	0
0.0249	0.014	0.0249	439	Leu	NA	NA	1	0	0	0.0342	439	Asn	Ser	NA	0
0.0249	0.0183	0.0249	440	Leu	NA	NA	1	0	0	0.0171	440	Lys	Arg	1	0
0.0268	0.0183	0.0268	441	Ala	NA	NA	1	0	0	0.0228	441	Pro	His	NA	0
0.0249	0.019	0.0249	442	Leu	NA	NA	1	0	0.0053	0.0171	442	Ala	Thr	NA	0
0	0.0202	0.0431	443	Arg	Lys	NA	0	0	0.0091	0.0133	443	Met	Thr	1	0
0	0.0177	0.0268	444	*Ala*	Val	NA	0	0.0266	0.0114	0.0266	444	Ile	NA	NA	1
0.0431	0.0164	0.0431	445	Arg	NA	NA	1	0.019	0.0114	0.019	445	Tyr	NA	NA	1
0	0.0164	0.0125	446	Asn	Tyr	1	0	0.0114	0.0114	0.0114	446	Cys	NA	NA	1
0.0316	0.0161	0.0316	447	Glu	NA	NA	1	0	0.0137	0.0304	447	Ser	Cys	NA	0
0.0259	0.0161	0.0259	448	His	NA	NA	1	0	0.0156	0.0171	448	His	Arg	NA	0
0	0.0186	0.0431	449	Arg	Lys	NA	0	0.038	0.0133	0.038	449	Gly	NA	NA	1
0.0115	0.0174	0.0115	450	Gln	NA	NA	1	0.0285	0.0167	0.0285	450	Asp	NA	NA	1
0.0268	0.0174	0.0268	451	Ala	NA	NA	1	0	0.0175	0.038	451	*Gly*	Ala	NA	0
0.0221	0.0143	0.0221	452	Asp	NA	NA	1	0.0171	0.0159	0.0171	452	His	NA	NA	1
0	0.0136	0.0316	453	Glu	Asp	NA	0	0.0038	0.0102	0.0038	453	Trp	NA	NA	1
0.0249	0.0148	0.0249	454	Leu	NA	NA	1	0.0304	0.0137	0.0304	454	Val	NA	NA	1
0.0316	0.0151	0.0316	455	Glu	NA	NA	1	0	0.0114	0.0171	455	His	Arg	NA	0
0	0.0153	0.0268	456	Ala	ProfsTer	1	0	0.0171	0.0106	0.0171	456	Ala	NA	NA	1
0	0.0146	0.023	457	Ile	Thr	1	0	0.0057	0.0046	0.0057	457	Gln	NA	NA	1
0.0192	0.0146	0.0192	458	Met	NA	NA	1	0	0.0102	0.0114	458	Cys	Gly	1	0
0.0115	0.0135	0.0115	459	Gln	NA	NA	1	0	0.0125	0.0133	459	Met	Val	1	0
0.0144	0.0128	0.0144	460	Gly	NA	NA	1	0.0285	0.0114	0.0285	460	Asp	NA	NA	1
0.0297	0.014	0.0297	461	Lys	NA	NA	1	0.0285	0.0163	0.0285	461	Leu	NA	NA	1
0.0144	0.014	0.0144	462	Gly	NA	NA	1	0	0.0163	0.0171	462	Ala	Thr	1	0
0	0.0121	0.0412	463	Ser	Cys	NA	0	0.0247	0.0175	0.0247	463	Glu	NA	NA	1
0.0144	0.0137	0.0144	464	Gly	NA	NA	1	0	0.0175	0.0228	464	Arg	Cys	1	0
0.0249	0.0123	0.0249	465	Leu	NA	NA	1	0.0342	0.0228	0.0342	465	Thr	NA	NA	1
0.0115	0.0118	0.0115	466	Gln	NA	NA	1	0.0285	0.0178	0.0285	466	Leu	NA	NA	1
0	0.013	0.0278	467	Pro	Leu	NA	0	0.0266	0.0235	0.0266	467	Ile	NA	NA	1
0	0.0153	0.0268	468	Ala	Val	NA	0	0	0.0228	0.0171	468	His	Arg	NA	0
0.0278	0.0139	0.0278	469	Val	NA	NA	1	0.0285	0.0205	0.0285	469	Leu	NA	NA	1
0	0.0142	0.0125	470	Cys	TrpfsTer	NA	0	0.0304	0.0152	0.0304	470	Ser	NA	NA	1
0.0249	0.0143	0.0249	471	Leu	NA	NA	1	0.0171	0.0213	0.0171	471	Ala	NA	NA	1
0.0268	0.0163	0.0268	472	Ala	NA	NA	1	0	0.0156	0.038	472	Gly	Arg	1	0
0.023	0.0163	0.023	473	Ile	NA	NA	1	0.0304	0.0129	0.0304	473	Ser	NA	NA	1
0	0.016	0.0431	474	*Arg*	Cys	1	0	0	0.0133	0.0342	474	Asn	Ser	NA	0
0.0278	0.016	0.0278	475	Val	NA	NA	1	0.0171	0.0133	0.0171	475	Lys	NA	NA	1
0.0125	0.0148	0.0125	476	Asn	NA	NA	1	0.019	0.0072	0.019	476	Tyr	NA	NA	1
0.0201	0.0162	0.0201	477	Thr	NA	NA	1	0	0.014	0.019	477	Tyr	His	NA	0
0	0.0151	0.0134	478	*Phe*	Leu	NA	0	0	0.0156	0.0114	478	Cys	Tyr	NA	0
0.0249	0.0158	0.0249	479	Leu	NA	NA	1	0.0342	0.0118	0.0342	479	Asn	NA	NA	1
0	0.013	0.0412	480	Ser	Gly	NA	0	0.0247	0.0118	0.0247	480	Glu	NA	NA	1
0.0125	0.0148	0.0125	481	Cys	NA	NA	1	0	0.0118	0.0171	481	His	Asp	NA	0
0.0412	0.0147	0.0412	482	Ser	NA	NA	1	0	0.0102	0.0304	482	Val	Met	NA	0
0.0115	0.0147	0.0115	483	Gln	NA	NA	1	0	0.0053	0.0247	483	Glu	Lys	NA	0
0.0077	0.0122	0.0077	484	Tyr	NA	NA	1	0.0266	0.0099	0.0266	484	Ile	NA	NA	1
0	0.0122	0.0259	485	His	Tyr	NA	0	0	0.0099	0.0171	485	Ala	Thr	NA	0
0.0297	0.0109	0.0297	486	Lys	NA	NA	1	0.0228	0.0099	0.0228	486	Arg	NA	NA	1
0.0192	0.0098	0.0192	487	Met	NA	NA	1	0	0.008	0.0171	487	Ala	Gly	NA	0
0	0.0086	0.0077	488	Tyr	Phe	NA	0	0	0.008	0.0285	488	Leu	Val	1	0
0	0.0079	0.0431	489	Arg	Ser	NA	0	0.0171	0.0034	0.0171	489	His	NA	NA	1

0	0.0099	0.0201	490	Thr	Ile	NA	0	0	0.0046	0.0342	490	Thr	Asn	NA	0
0	0.0083	0.0278	491	Val	Glu	NA	0	0	0.0046	0.0228	491	Pro	Ser	1	0
0.0297	0.0107	0.0297	492	Lys	NA	NA	1	0.0057	0.0072	0.0057	492	Gln	NA	NA	1
0	0.0119	0.0268	493	Ala	Thr	NA	0	0	0.0072	0.0228	493	Arg	Gly	1	0
0	0.0119	0.023	494	Ile	Val	NA	0	0.0304	0.0118	0.0304	494	Val	NA	NA	1
0.0201	0.0132	0.0201	495	Thr	NA	NA	1	0	0.0163	0.0285	495	Leu	Pro	NA	0
0.0144	0.0132	0.0144	496	Gly	NA	NA	1	0.0228	0.0163	0.0228	496	Pro	NA	NA	1
0.0431	0.013	0.0431	497	Arg	NA	NA	1	0.0285	0.0137	0.0285	497	Leu	NA	NA	1
0.0115	0.0155	0.0115	498	Gln	NA	NA	1	0	0.0182	0.0171	498	Lys	Gln	1	0
0	0.0181	0.023	499	Ile	Thr	NA	0	0.0171	0.0137	0.0171	499	Lys	NA	NA	1
0.0134	0.0188	0.0134	500	Phe	NA	NA	1	0.0228	0.008	0.0228	500	Pro	NA	NA	1
0	0.0198	0.0115	501	Gln	Glu	NA	0	0	0.0114	0.0228	501	Pro	Thr	NA	0
0.0278	0.0155	0.0278	502	Pro	NA	NA	1	0	0.014	0.0133	502	Met	Val	NA	0
0.0249	0.0156	0.0249	503	Leu	NA	NA	1	0.0171	0.0095	0.0171	503	Lys	NA	NA	1
0.0259	0.0183	0.0259	504	His	NA	NA	1	0.0304	0.0095	0.0304	504	Ser	NA	NA	1
0.0268	0.017	0.0268	505	Ala	NA	NA	1	0	0.0129	0.0285	505	Leu	Phe	NA	0
0.0249	0.017	0.0249	506	Leu	NA	NA	1	0	0.0129	0.0228	506	Arg	Cys	1	0
0	0.0142	0.0431	507	*Arg*	Gln	NA	0	0.0171	0.0068	0.0171	507	Lys	NA	NA	1
0.0125	0.0117	0.0125	508	Asn	NA	NA	1	0.0171	0.0129	0.0171	508	Lys	NA	NA	1
0.0268	0.0116	0.0268	509	Ala	NA	NA	1	0	0.0129	0.038	509	Gly	ArgfsTer	1	0
0	0.0089	0.0316	510	Glu	Lys	NA	0	0.0304	0.0129	0.0304	510	Ser	NA	NA	1
0	0.0079	0.0297	511	Lys	Asn	NA	0	0	0.0148	0.038	511	Gly	Arg	NA	0
0	0.0086	0.0278	512	Val	Ile	1	0	0.0171	0.0148	0.0171	512	Lys	NA	NA	1
0	0.01	0.0249	513	Leu	Phe	NA	0	0.0266	0.0156	0.0266	513	Ile	NA	NA	1
0.0249	0.0099	0.0249	514	Leu	NA	NA	1	0	0.0201	0.0285	514	Leu	Phe	NA	0
0	0.0099	0.0278	515	Pro	Leu	NA	0	0.0342	0.0201	0.0342	515	Thr	NA	NA	1
0.0144	0.013	0.0144	516	Gly	NA	NA	1	0.0228	0.0182	0.0228	516	Pro	NA	NA	1
0.0077	0.013	0.0077	517	Tyr	NA	NA	1	0.0171	0.0216	0.0171	517	Ala	NA	NA	1
0.0259	0.013	0.0259	518	His	NA	NA	1	0.0171	0.0209	0.0171	518	Lys	NA	NA	1
0.0259	0.0105	0.0259	519	His	NA	NA	1	0.0171	0.0197	0.0171	519	Lys	NA	NA	1
0	0.0105	0.0134	520	Phe	Leu	1	0	0.0304	0.0163	0.0304	520	Ser	NA	NA	1
0.0316	0.0116	0.0316	521	Glu	NA	NA	1	0.0171	0.0175	0.0171	521	Phe	NA	NA	1
0	0.0138	0.0048	522	*Trp*	Cys	NA	0	0	0.014	0.0285	522	Leu	Arg	NA	0
0	0.0125	0.0115	523	Gln	Lys	NA	0	0.0228	0.0137	0.0228	523	Arg	NA	NA	1
0	0.0127	0.0278	524	Pro	Ala	1	0	0	0.0137	0.0228	524	Arg	del	1	0
0	0.0168	0.0278	525	Pro	Ser	NA	0	0.0285	0.0194	0.0285	525	Leu	NA	NA	1
0.0249	0.0177	0.0249	526	Leu	NA	NA	1	0.0171	0.0185	0.0171	526	Phe	NA	NA	1
0.0297	0.0177	0.0297	527	Lys	NA	NA	1	0.0285	0.0261	0.0285	527	Asp	NA	NA	1

RAG2 end

RAG1 continued

RAG1 continued

0.0125	0.0177	0.0125	528	Asn	NA	NA	1	0.0278	0.0153	0.0278	786	Pro	NA	NA	1
0.0278	0.0199	0.0278	529	Val	NA	NA	1	0.0134	0.0153	0.0134	787	Phe	NA	NA	1
0.0412	0.0227	0.0412	530	Ser	NA	NA	1	0	0.0195	0.023	788	Ile	Val	NA	0
0.0412	0.0217	0.0412	531	Ser	NA	NA	1	0	0.0191	0.0316	789	Glu	Lys	NA	0
0	0.021	0.0412	532	Ser	Asn	NA	0	0	0.0183	0.0201	790	Thr	Ile	NA	0
0	0.0198	0.0201	533	Thr	GlnfsTer	NA	0	0.0278	0.0182	0.0278	791	Val	NA	NA	1
0.0221	0.0192	0.0221	534	Asp	NA	NA	1	0.0278	0.0194	0.0278	792	Pro	NA	NA	1
0.0278	0.0151	0.0278	535	Val	NA	NA	1	0.0412	0.022	0.0412	793	Ser	NA	NA	1
0.0144	0.0134	0.0144	536	Gly	NA	NA	1	0.023	0.0232	0.023	794	Ile	NA	NA	1
0.023	0.0175	0.023	537	Ile	NA	NA	1	0.0221	0.0254	0.0221	795	Asp	NA	NA	1
0	0.019	0.023	538	Ile	LeufsTer	NA	0	0.0268	0.0226	0.0268	796	Ala	NA	NA	1
0.0221	0.0168	0.0221	539	Asp	NA	NA	1	0.0249	0.0213	0.0249	797	Leu	NA	NA	1
0	0.0181	0.0144	540	Gly	Glu	NA	0	0.0259	0.0184	0.0259	798	His	NA	NA	1
0.0249	0.0167	0.0249	541	Leu	NA	NA	1	0.0125	0.0161	0.0125	799	Cys	NA	NA	1
0.0412	0.0185	0.0412	542	Ser	NA	NA	1	0.0221	0.0166	0.0221	800	Asp	NA	NA	1
0.0144	0.0185	0.0144	543	Gly	NA	NA	1	0	0.0171	0.023	801	Ile	Val	NA	0

0	0.0185	0.0249	544	Leu	Ile	NA	0	0.0144	0.0159	0.0144	802	Gly	NA	NA	1
0.0412	0.0207	0.0412	545	Ser	NA	NA	1	0.0125	0.0141	0.0125	803	Asn	NA	NA	1
0	0.019	0.0412	546	Ser	Phe	NA	0	0	0.0128	0.0268	804	Ala	Thr	NA	0
0.0412	0.0149	0.0412	547	Ser	NA	NA	1	0.0268	0.0106	0.0268	805	Ala	NA	NA	1
0	0.0162	0.0278	548	Val	Leu	NA	0	0.0316	0.012	0.0316	806	Glu	NA	NA	1
0.0221	0.0184	0.0221	549	Asp	NA	NA	1	0.0134	0.0117	0.0134	807	Phe	NA	NA	1
0.0221	0.0143	0.0221	550	Asp	NA	NA	1	0.0077	0.0129	0.0077	808	Tyr	NA	NA	1
0.0077	0.0143	0.0077	551	Tyr	NA	NA	1	0	0.0129	0.0297	809	Lys	Arg	NA	0
0	0.0128	0.0278	552	Pro	Ala	NA	0	0	0.0103	0.023	810	Ile	Val	NA	0
0.0278	0.0128	0.0278	553	Val	NA	NA	1	0.0134	0.0071	0.0134	811	Phe	NA	NA	1
0.0221	0.0106	0.0221	554	Asp	NA	NA	1	0.0115	0.0089	0.0115	812	Gln	NA	NA	1
0	0.0098	0.0201	555	Thr	Ile	NA	0	0.0249	0.0081	0.0249	813	Leu	NA	NA	1
0	0.009	0.023	556	Ile	Thr	NA	0	0	0.0081	0.0316	814	Glu	Asp	NA	0
0.0268	0.0098	0.0268	557	Ala	NA	NA	1	0	0.0081	0.023	815	Ile	Leu	1	0
0	0.0092	0.0297	558	Lys	Asn	NA	0	0	0.0081	0.0144	816	Gly	Glu	NA	0
0	0.0111	0.0431	559	Arg	Ser	NA	0	0.0316	0.0069	0.0316	817	Glu	NA	NA	1
0.0134	0.0138	0.0134	560	Phe	NA	NA	1	0	0.0044	0.0278	818	Val	Leu	1	0
0	0.0138	0.0431	561	Arg	Cys	1	0	0	0.0071	0.0077	819	Tyr	Cys	NA	0
0.0077	0.0111	0.0077	562	Tyr	NA	NA	1	0	0.0112	0.0297	820	Lys	Glu	1	0
0.0221	0.0152	0.0221	563	Asp	NA	NA	1	0.0125	0.0112	0.0125	821	Asn	NA	NA	1
0.0412	0.0179	0.0412	564	Ser	NA	NA	1	0	0.0081	0.0278	822	Pro	Arg	NA	0
0.0268	0.0191	0.0268	565	Ala	NA	NA	1	0	0.0112	0.0125	823	Asn	Thr	1	0
0	0.0191	0.0249	566	Leu	Phe	NA	0	0.0268	0.0112	0.0268	824	Ala	NA	NA	1
0	0.0205	0.0278	567	Val	Leu	1	0	0.0412	0.0112	0.0412	825	Ser	NA	NA	1
0.0412	0.0183	0.0412	568	Ser	NA	NA	1	0	0.01	0.0297	826	Lys	Asn	NA	0
0.0268	0.0174	0.0268	569	Ala	NA	NA	1	0	0.0105	0.0316	827	Glu	Asp	NA	0
0.0249	0.0178	0.0249	570	Leu	NA	NA	1	0.0316	0.0116	0.0316	828	Glu	NA	NA	1
0	0.0178	0.0192	571	Met	Ile	NA	0	0	0.0089	0.0431	829	Arg	Ser	NA	0
0.0221	0.0201	0.0221	572	Asp	NA	NA	1	0	0.0048	0.0297	830	Lys	Ter	1	0
0	0.0185	0.0192	573	Met	Ile	NA	0	0	0.0048	0.0431	831	Arg	Met	NA	0
0.0316	0.019	0.0316	574	Glu	NA	NA	1	0.0048	0.007	0.0048	832	Trp	NA	NA	1
0.0316	0.0165	0.0316	575	Glu	NA	NA	1	0.0115	0.0068	0.0115	833	Gln	NA	NA	1
0	0.0165	0.0221	576	Asp	Val	NA	0	0	0.0068	0.0268	834	Ala	Ser	NA	0
0.023	0.0143	0.023	577	Ile	NA	NA	1	0	0.0093	0.0201	835	Thr	Ala	NA	0
0.0249	0.0143	0.0249	578	Leu	NA	NA	1	0	0.0093	0.0249	836	Leu	Met	NA	0
0.0316	0.0123	0.0316	579	Glu	NA	NA	1	0.0221	0.0088	0.0221	837	Asp	NA	NA	1
0	0.0091	0.0144	580	Gly	Asp	1	0	0.0297	0.0077	0.0297	838	Lys	NA	NA	1
0	0.0091	0.0192	581	Met	Thr	NA	0	0	0.0077	0.0259	839	His	Tyr	1	0
0	0.0068	0.0431	582	Arg	Lys	NA	0	0.0249	0.0089	0.0249	840	Leu	NA	NA	1
0	0.0065	0.0412	583	Ser	Phe	NA	0	0	0.0114	0.0431	841	Arg	Trp	1	0
0.0115	0.0034	0.0115	584	Gln	NA	NA	1	0	0.0122	0.0297	842	Lys	Glu	NA	0
0	0.0058	0.0221	585	Asp	Val	NA	0	0	0.0092	0.0297	843	Lys	Glu	1	0
0	0.0071	0.0249	586	Leu	Phe	NA	0	0	0.0092	0.0192	844	Met	Thr	1	0
0	0.0085	0.0221	587	Asp	Asn	NA	0	0.0125	0.0067	0.0125	845	Asn	NA	NA	1
0.0221	0.0085	0.0221	588	Asp	NA	NA	1	0.0249	0.0067	0.0249	846	Leu	NA	NA	1
0	0.0087	0.0077	589	Tyr	Ter	NA	0	0.0297	0.0067	0.0297	847	Lys	NA	NA	1
0.0249	0.0107	0.0249	590	Leu	NA	NA	1	0	0.008	0.0278	848	Pro	Ala	NA	0
0.0125	0.0135	0.0125	591	Asn	NA	NA	1	0	0.008	0.023	849	Ile	Thr	NA	0
0.0144	0.0163	0.0144	592	Gly	NA	NA	1	0	0.008	0.0192	850	Met	Arg	NA	0
0	0.0141	0.0278	593	Pro	Ser	NA	0	0	0.0068	0.0431	851	Arg	Lys	NA	0
0.0134	0.0171	0.0134	594	Phe	NA	NA	1	0	0.0038	0.0192	852	Met	Leu	NA	0
0.0201	0.0177	0.0201	595	Thr	NA	NA	1	0.0125	0.0081	0.0125	853	Asn	NA	NA	1
0.0278	0.0206	0.0278	596	Val	NA	NA	1	0	0.0111	0.0144	854	Gly	Asp	NA	0
0.0278	0.0204	0.0278	597	Val	NA	NA	1	0.0125	0.0136	0.0125	855	Asn	NA	NA	1
0	0.0226	0.0278	598	Val	Met	1	0	0.0134	0.0136	0.0134	856	Phe	NA	NA	1
0.0297	0.0227	0.0297	599	Lys	NA	NA	1	0	0.0156	0.0268	857	Ala	Ser	1	0
0.0316	0.0226	0.0316	600	Glu	NA	NA	1	0.0431	0.0144	0.0431	858	Arg	NA	NA	1
0.0412	0.0213	0.0412	601	Ser	NA	NA	1	0.0297	0.0144	0.0297	859	Lys	NA	NA	1
0.0125	0.0185	0.0125	602	Cys	NA	NA	1	0.0249	0.0151	0.0249	860	Leu	NA	NA	1
0.0221	0.0185	0.0221	603	Asp	NA	NA	1	0	0.0138	0.0192	861	Met	Ile	1	0

0.0144	0.0155	0.0144	604	Gly	NA	NA	1	0.0201	0.016	0.0201	862	Thr	NA	NA	1
0.0192	0.0124	0.0192	605	Met	NA	NA	1	0	0.0117	0.0297	863	Lys	Thr	1	0
0.0144	0.0082	0.0144	606	Gly	NA	NA	1	0	0.0087	0.0316	864	Glu	Ter	NA	0
0	0.0096	0.0221	607	Asp	Glu	NA	0	0.0201	0.0075	0.0201	865	Thr	NA	NA	1
0	0.0088	0.0278	608	Val	Met	NA	0	0	0.0075	0.0278	866	Val	GlyfsTer	NA	0
0	0.0074	0.0412	609	Ser	Gly	NA	0	0.0221	0.0055	0.0221	867	Asp	NA	NA	1
0	0.0069	0.0316	610	Glu	Gln	NA	0	0	0.0078	0.0268	868	Ala	Ser	1	0
0	0.0082	0.0297	611	Lys	Thr	NA	0	0	0.0105	0.0278	869	Val	Ile	NA	0
0.0259	0.0082	0.0259	612	His	NA	NA	1	0.0125	0.0127	0.0125	870	Cys	NA	NA	1
0.0144	0.0082	0.0144	613	Gly	NA	NA	1	0	0.0127	0.0316	871	Glu	Lys	NA	0
0	0.0082	0.0412	614	Ser	Thr	NA	0	0	0.0136	0.0249	872	Leu	Ter	1	0
0.0144	0.0114	0.0144	615	Gly	NA	NA	1	0.023	0.0179	0.023	873	Ile	NA	NA	1
0.0278	0.0144	0.0278	616	Pro	NA	NA	1	0.0278	0.0205	0.0278	874	Pro	NA	NA	1
0	0.0118	0.0278	617	Val	Leu	1	0	0.0412	0.0193	0.0412	875	Ser	NA	NA	1
0	0.0131	0.0278	618	Val	Ala	NA	0	0	0.022	0.0316	876	Glu	Lys	NA	0
0	0.0131	0.0278	619	Pro	GlnfsTer	1	0	0.0316	0.0244	0.0316	877	Glu	NA	NA	1
0.0316	0.0117	0.0316	620	Glu	NA	NA	1	0.0431	0.0221	0.0431	878	Arg	NA	NA	1
0.0297	0.013	0.0297	621	Lys	NA	NA	1	0.0259	0.0225	0.0259	879	His	NA	NA	1
0	0.0144	0.0268	622	Ala	Thr	NA	0	0	0.0184	0.0316	880	Glu	Lys	1	0
0.0278	0.0164	0.0278	623	Val	NA	NA	1	0.0268	0.0184	0.0268	881	Ala	NA	NA	1
0	0.0164	0.0431	624	Arg	Cys	1	0	0.0249	0.0152	0.0249	882	Leu	NA	NA	1
0	0.0151	0.0134	625	Phe	Ile	NA	0	0	0.0109	0.0431	883	Arg	Lys	NA	0
0.0412	0.0151	0.0412	626	Ser	NA	NA	1	0.0316	0.0091	0.0316	884	Glu	NA	NA	1
0.0134	0.0174	0.0134	627	Phe	NA	NA	1	0	0.0116	0.0249	885	Leu	HisfsTer	1	0
0.0201	0.0147	0.0201	628	Thr	NA	NA	1	0	0.0119	0.0192	886	Met	Thr	NA	0
0	0.0147	0.023	629	Ile	Val	NA	0	0	0.0113	0.0221	887	Asp	Asn	NA	0
0.0192	0.0147	0.0192	630	Met	NA	NA	1	0	0.0143	0.0249	888	Leu	Phe	1	0
0.0297	0.0105	0.0297	631	Lys	NA	NA	1	0.0077	0.0111	0.0077	889	Tyr	NA	NA	1
0.023	0.0092	0.023	632	Ile	NA	NA	1	0.0249	0.0139	0.0249	890	Leu	NA	NA	1
0	0.0113	0.0201	633	Thr	Ile	NA	0	0.0297	0.0144	0.0297	891	Lys	NA	NA	1
0	0.0125	0.023	634	Ile	Val	NA	0	0.0192	0.0144	0.0192	892	Met	NA	NA	1
0	0.0118	0.0268	635	Ala	Thr	NA	0	0.0297	0.0185	0.0297	893	Lys	NA	NA	1
0	0.0116	0.0259	636	His	Arg	NA	0	0	0.0219	0.0278	894	Pro	Gln	NA	0
0	0.0093	0.0412	637	Ser	Gly	NA	0	0.0278	0.0206	0.0278	895	Val	NA	NA	1
0.0412	0.0121	0.0412	638	Ser	NA	NA	1	0.0048	0.0204	0.0048	896	Trp	NA	NA	1
0.0115	0.0134	0.0115	639	Gln	NA	NA	1	0	0.0185	0.0431	897	Arg	Ter	1	0
0.0125	0.0166	0.0125	640	Asn	NA	NA	1	0.0412	0.0155	0.0412	898	Ser	NA	NA	1
0.0278	0.0198	0.0278	641	Val	NA	NA	1	0.0412	0.0187	0.0412	899	Ser	NA	NA	1
0	0.0224	0.0297	642	Lys	Gln	NA	0	0.0125	0.0159	0.0125	900	Cys	NA	NA	1
0.0278	0.0213	0.0278	643	Val	NA	NA	1	0.0278	0.0182	0.0278	901	Pro	NA	NA	1
0.0134	0.0229	0.0134	644	Phe	NA	NA	1	0	0.0182	0.0268	902	Ala	Asp	NA	0
0.0316	0.0229	0.0316	645	Glu	NA	NA	1	0	0.0182	0.0297	903	Lys	ArgfsTer	NA	0
0.0316	0.0201	0.0316	646	Glu	NA	NA	1	0.0316	0.0141	0.0316	904	Glu	NA	NA	1
0.0268	0.0233	0.0268	647	Ala	NA	NA	1	0	0.0141	0.0125	905	Cys	Ser	NA	0
0.0297	0.023	0.0297	648	Lys	NA	NA	1	0.0278	0.0125	0.0278	906	Pro	NA	NA	1
0.0278	0.0229	0.0278	649	Pro	NA	NA	1	0	0.0132	0.0316	907	Glu	Asp	NA	0
0.0125	0.021	0.0125	650	Asn	NA	NA	1	0.0412	0.0174	0.0412	908	Ser	NA	NA	1
0	0.0208	0.0412	651	Ser	Pro	NA	0	0	0.0142	0.0249	909	Leu	Pro	NA	0
0.0316	0.0209	0.0316	652	Glu	NA	NA	1	0.0125	0.0154	0.0125	910	Cys	NA	NA	1
0.0249	0.0204	0.0249	653	Leu	NA	NA	1	0.0115	0.0168	0.0115	911	Gln	NA	NA	1
0.0125	0.0176	0.0125	654	Cys	NA	NA	1	0.0077	0.0179	0.0077	912	Tyr	NA	NA	1
0.0125	0.0189	0.0125	655	Cys	NA	NA	1	0.0412	0.0138	0.0412	913	Ser	NA	NA	1
0.0297	0.0189	0.0297	656	Lys	NA	NA	1	0	0.0151	0.0134	914	Phe	Ser	1	0
0.0278	0.0182	0.0278	657	Pro	NA	NA	1	0.0125	0.0166	0.0125	915	Asn	NA	NA	1
0.0249	0.0184	0.0249	658	Leu	NA	NA	1	0.0412	0.0154	0.0412	916	Ser	NA	NA	1
0	0.0172	0.0125	659	Cys	Tyr	NA	0	0.0115	0.0172	0.0115	917	Gln	NA	NA	1
0.0249	0.0191	0.0249	660	Leu	NA	NA	1	0	0.0155	0.0431	918	Arg	Cys	1	0
0	0.0161	0.0192	661	Met	del	NA	0	0.0134	0.0197	0.0134	919	Phe	NA	NA	1
0.0249	0.0155	0.0249	662	Leu	NA	NA	1	0.0268	0.0184	0.0268	920	Ala	NA	NA	1
0.0268	0.013	0.0268	663	Ala	NA	NA	1	0	0.0143	0.0316	921	Glu	Lys	1	0

0	0.0162	0.0221	664	Asp	Glu	NA	0	0.0249	0.0131	0.0249	922	Leu	NA	NA	1
0.0316	0.0137	0.0316	665	Glu	NA	NA	1	0.0249	0.0161	0.0249	923	Leu	NA	NA	1
0	0.0162	0.0412	666	Ser	Thr	NA	0	0.0412	0.0148	0.0412	924	Ser	NA	NA	1
0.0221	0.0157	0.0221	667	Asp	NA	NA	1	0	0.0164	0.0201	925	Thr	Met	NA	0
0	0.0157	0.0259	668	His	Tyr	NA	0	0	0.0172	0.0297	926	Lys	Thr	NA	0
0.0316	0.0157	0.0316	669	Glu	NA	NA	1	0	0.0178	0.0134	927	Phe	Ser	NA	0
0	0.0151	0.0201	670	Thr	Lys	1	0	0.0297	0.0153	0.0297	928	Lys	NA	NA	1
0.0249	0.0192	0.0249	671	Leu	NA	NA	1	0	0.0112	0.0077	929	Tyr	His	NA	0
0.0201	0.017	0.0201	672	Thr	NA	NA	1	0.0431	0.0135	0.0431	930	Arg	NA	NA	1
0.0268	0.0195	0.0268	673	Ala	NA	NA	1	0.0077	0.0135	0.0077	931	Tyr	NA	NA	1
0	0.0163	0.023	674	Ile	TyrfsTer	1	0	0.0316	0.0135	0.0316	932	Glu	NA	NA	1
0.0249	0.019	0.0249	675	Leu	NA	NA	1	0	0.0113	0.0144	933	Gly	Val	NA	0
0.0412	0.0197	0.0412	676	Ser	NA	NA	1	0	0.0113	0.0297	934	Lys	Glu	NA	0
0	0.0176	0.0278	677	Pro	Thr	1	0	0.023	0.0096	0.023	935	Ile	NA	NA	1
0.0249	0.015	0.0249	678	Leu	NA	NA	1	0	0.0118	0.0201	936	Thr	Ala	NA	0
0	0.015	0.023	679	Ile	Val	NA	0	0	0.0086	0.0125	937	Asn	His	NA	0
0.0268	0.0144	0.0268	680	Ala	NA	NA	1	0.0077	0.0111	0.0077	938	Tyr	NA	NA	1
0.0316	0.0132	0.0316	681	Glu	NA	NA	1	0	0.0138	0.0134	939	Phe	Leu	NA	0
0	0.0132	0.0431	682	Arg	Lys	NA	0	0.0259	0.0141	0.0259	940	His	NA	NA	1
0	0.0107	0.0316	683	Glu	Val	NA	0	0.0297	0.0141	0.0297	941	Lys	NA	NA	1
0	0.0139	0.0268	684	Ala	ProfsTer	1	0	0	0.0169	0.0201	942	Thr	Ile	NA	0
0.0192	0.0112	0.0192	685	Met	NA	NA	1	0.0249	0.0161	0.0249	943	Leu	NA	NA	1
0.0297	0.0081	0.0297	686	Lys	NA	NA	1	0.0268	0.0184	0.0268	944	Ala	NA	NA	1
0	0.0105	0.0412	687	Ser	Asn	NA	0	0.0259	0.0181	0.0259	945	His	NA	NA	1
0	0.0137	0.0412	688	Ser	Gly	NA	0	0	0.0183	0.0278	946	Val	Leu	NA	0
0.0316	0.0162	0.0316	689	Glu	NA	NA	1	0.0278	0.0226	0.0278	947	Pro	NA	NA	1
0	0.0157	0.0249	690	Leu	Ser	NA	0	0	0.0201	0.0316	948	Glu	Asp	NA	0
0	0.0142	0.0192	691	Met	Val	NA	0	0.023	0.0189	0.023	949	Ile	NA	NA	1
0.0249	0.0165	0.0249	692	Leu	NA	NA	1	0.023	0.0204	0.023	950	Ile	NA	NA	1
0.0316	0.0165	0.0316	693	Glu	NA	NA	1	0.0316	0.0204	0.0316	951	Glu	NA	NA	1
0.0249	0.0133	0.0249	694	Leu	NA	NA	1	0.0431	0.0176	0.0431	952	Arg	NA	NA	1
0.0144	0.0153	0.0144	695	Gly	NA	NA	1	0	0.0203	0.0221	953	Asp	Tyr	NA	0
0.0144	0.0167	0.0144	696	Gly	NA	NA	1	0.0144	0.018	0.0144	954	Gly	NA	NA	1
0.023	0.0172	0.023	697	Ile	NA	NA	1	0.0412	0.0157	0.0412	955	Ser	NA	NA	1
0	0.0153	0.0249	698	Leu	Phe	NA	0	0	0.0126	0.023	956	Ile	Thr	NA	0
0	0.0151	0.0431	699	Arg	Trp	1	0	0	0.0082	0.0144	957	Gly	Trp	NA	0
0.0201	0.0151	0.0201	700	Thr	NA	NA	1	0.0268	0.0082	0.0268	958	Ala	NA	NA	1
0.0134	0.0179	0.0134	701	Phe	NA	NA	1	0	0.0081	0.0048	959	Trp	Ter	NA	0
0.0297	0.0171	0.0297	702	Lys	NA	NA	1	0	0.0071	0.0268	960	Ala	Thr	1	0
0.0134	0.0191	0.0134	703	Phe	NA	NA	1	0	0.0112	0.0412	961	Ser	MetfsTer	1	0
0.023	0.0205	0.023	704	Ile	NA	NA	1	0	0.0127	0.0316	962	Glu	Asp	NA	0
0.0134	0.0193	0.0134	705	Phe	NA	NA	1	0	0.01	0.0144	963	Gly	Glu	NA	0
0.0431	0.0201	0.0431	706	Arg	NA	NA	1	0.0125	0.0129	0.0125	964	Asn	NA	NA	1
0.0144	0.0203	0.0144	707	Gly	NA	NA	1	0.0316	0.0154	0.0316	965	Glu	NA	NA	1
0.0201	0.022	0.0201	708	Thr	NA	NA	1	0.0412	0.0168	0.0412	966	Ser	NA	NA	1
0.0144	0.0221	0.0144	709	Gly	NA	NA	1	0.0144	0.0211	0.0144	967	Gly	NA	NA	1
0.0077	0.0208	0.0077	710	Tyr	NA	NA	1	0	0.0211	0.0125	968	Asn	Lys	NA	0
0.0221	0.0165	0.0221	711	Asp	NA	NA	1	0.0297	0.0198	0.0297	969	Lys	NA	NA	1
0.0316	0.0182	0.0316	712	Glu	NA	NA	1	0.0249	0.0167	0.0249	970	Leu	NA	NA	1
0.0297	0.019	0.0297	713	Lys	NA	NA	1	0.0134	0.0155	0.0134	971	Phe	NA	NA	1
0.0249	0.0207	0.0249	714	Leu	NA	NA	1	0.0431	0.016	0.0431	972	Arg	NA	NA	1
0	0.0199	0.0278	715	Val	CysfsTer	1	0	0	0.0173	0.0431	973	Arg	Cys	1	0
0	0.0202	0.0431	716	Arg	Trp	1	0	0	0.017	0.0134	974	Phe	Leu	NA	0
0.0316	0.0171	0.0316	717	Glu	NA	NA	1	0	0.0188	0.0431	975	Arg	Trp	1	0
0.0278	0.0168	0.0278	718	Val	NA	NA	1	0.0297	0.0186	0.0297	976	Lys	NA	NA	1
0.0316	0.0184	0.0316	719	Glu	NA	NA	1	0.0192	0.0143	0.0192	977	Met	NA	NA	1
0	0.0198	0.0144	720	Gly	Asp	NA	0	0.0125	0.0173	0.0125	978	Asn	NA	NA	1
0.0249	0.024	0.0249	721	Leu	NA	NA	1	0.0268	0.0185	0.0268	979	Ala	NA	NA	1
0	0.0236	0.0316	722	Glu	Lys	NA	0	0.0431	0.0185	0.0431	980	Arg	NA	NA	1
0.0268	0.0216	0.0268	723	Ala	NA	NA	1	0.0115	0.0187	0.0115	981	Gln	NA	NA	1

0.0412	0.0184	0.0412	724	Ser	NA	NA	1	0	0.0168	0.0412	982	Ser	Tyr	NA	0
0.0144	0.0184	0.0144	725	Gly	NA	NA	1	0.0297	0.0187	0.0297	983	Lys	NA	NA	1
0.0412	0.0159	0.0412	726	Ser	NA	NA	1	0.0125	0.0182	0.0125	984	Cys	NA	NA	1
0.0278	0.0184	0.0278	727	Val	NA	NA	1	0	0.0167	0.0077	985	Tyr	His	1	0
0.0077	0.017	0.0077	728	Tyr	NA	NA	1	0.0316	0.018	0.0316	986	Glu	NA	NA	1
0	0.0151	0.023	729	Ile	Leu	1	0	0	0.018	0.0192	987	Met	Thr	NA	0
0	0.0136	0.0125	730	Cys	Arg	1	0	0.0316	0.0151	0.0316	988	Glu	NA	NA	1
0	0.0095	0.0201	731	Thr	Ser	NA	0	0.0221	0.0138	0.0221	989	Asp	NA	NA	1
0.0249	0.0067	0.0249	732	Leu	NA	NA	1	0.0278	0.0138	0.0278	990	Val	NA	NA	1
0.0125	0.0084	0.0125	733	Cys	NA	NA	1	0.0249	0.0131	0.0249	991	Leu	NA	NA	1
0.0221	0.0084	0.0221	734	Asp	NA	NA	1	0	0.0131	0.0297	992	Lys	Glu	NA	0
0	0.0084	0.0268	735	Ala	Val	NA	0	0	0.01	0.0259	993	His	Arg	1	0
0	0.0084	0.0201	736	Thr	Asn	NA	0	0	0.0119	0.0259	994	His	Arg	NA	0
0	0.0071	0.0431	737	Arg	Cys	1	0	0	0.0121	0.0048	995	Trp	Gly	1	0
0.0249	0.0058	0.0249	738	Leu	NA	NA	1	0.0249	0.0096	0.0249	996	Leu	NA	NA	1
0	0.0061	0.0316	739	Glu	Gln	NA	0	0	0.0096	0.0077	997	Tyr	HisfsTer	NA	0
0	0.0089	0.0268	740	Ala	Gly	NA	0	0	0.0096	0.0201	998	Thr	LeufsTer	NA	0
0	0.0089	0.0412	741	Ser	Phe	NA	0	0.0412	0.0126	0.0412	999	Ser	NA	NA	1
0.0115	0.0089	0.0115	742	Gln	NA	NA	1	0.0297	0.0139	0.0297	1000	Lys	NA	NA	1
0	0.0064	0.0125	743	Asn	IlefsTer	NA	0	0	0.0114	0.0077	1001	Tyr	Cys	NA	0
0.0249	0.0064	0.0249	744	Leu	NA	NA	1	0	0.0114	0.0249	1002	Leu	Phe	NA	0
0.0278	0.0084	0.0278	745	Val	NA	NA	1	0	0.0141	0.0115	1003	Gln	Ter	NA	0
0	0.0128	0.0134	746	Phe	Cys	NA	0	0.0297	0.01	0.0297	1004	Lys	NA	NA	1
0	0.0116	0.0259	747	His	Gln	NA	0	0.0134	0.0082	0.0134	1005	Phe	NA	NA	1
0	0.0116	0.0412	748	Ser	Thr	1	0	0	0.0082	0.0192	1006	Met	Val	1	0
0	0.0118	0.023	749	Ile	Val	NA	0	0	0.0107	0.0125	1007	Asn	Ser	1	0
0.0201	0.009	0.0201	750	Thr	NA	NA	1	0.0268	0.0137	0.0268	1008	Ala	NA	NA	1
0.0431	0.0103	0.0431	751	Arg	NA	NA	1	0	0.0128	0.0259	1009	His	Tyr	NA	0
0	0.0103	0.0412	752	Ser	Arg	NA	0	0.0125	0.0155	0.0125	1010	Asn	NA	NA	1
0	0.0134	0.0259	753	His	Arg	1	0	0	0.017	0.0268	1011	Ala	Gly	NA	0
0.0268	0.0134	0.0268	754	Ala	NA	NA	1	0.0249	0.0183	0.0249	1012	Leu	NA	NA	1
0	0.0122	0.0316	755	Glu	Asp	NA	0	0.0297	0.0156	0.0297	1013	Lys	NA	NA	1
0.0125	0.011	0.0125	756	Asn	NA	NA	1	0.0201	0.0156	0.0201	1014	Thr	NA	NA	1
0	0.011	0.0249	757	Leu	Val	NA	0	0.0412	0.0156	0.0412	1015	Ser	NA	NA	1
0.0316	0.0115	0.0316	758	Glu	NA	NA	1	0.0144	0.0156	0.0144	1016	Gly	NA	NA	1
0	0.0088	0.0431	759	Arg	Cys	1	0	0.0134	0.0143	0.0134	1017	Phe	NA	NA	1
0.0077	0.0129	0.0077	760	Tyr	NA	NA	1	0	0.014	0.0201	1018	Thr	Ile	NA	0
0.0316	0.0129	0.0316	761	Glu	NA	NA	1	0	0.012	0.0192	1019	Met	Val	1	0
0	0.0129	0.0278	762	Val	Ile	NA	0	0.0125	0.0104	0.0125	1020	Asn	NA	NA	1
0.0048	0.0105	0.0048	763	Trp	NA	NA	1	0	0.0104	0.0278	1021	Pro	Ser	NA	0
0	0.0131	0.0431	764	Arg	Cys	1	0	0.0115	0.009	0.0115	1022	Gln	NA	NA	1
0.0412	0.0155	0.0412	765	Ser	NA	NA	1	0.0268	0.009	0.0268	1023	Ala	NA	NA	1
0.0125	0.0124	0.0125	766	Asn	NA	NA	1	0	0.0115	0.0412	1024	Ser	Asn	NA	0
0	0.0151	0.0278	767	Pro	Arg	NA	0	0.0249	0.0103	0.0249	1025	Leu	NA	NA	1
0.0077	0.0147	0.0077	768	Tyr	NA	NA	1	0.0144	0.0103	0.0144	1026	Gly	NA	NA	1
0.0259	0.0147	0.0259	769	His	NA	NA	1	0	0.0123	0.0221	1027	Asp	GlyfsTer	1	0
0.0316	0.0105	0.0316	770	Glu	NA	NA	1	0	0.0096	0.0278	1028	Pro	Gln	1	0
0	0.0093	0.0412	771	Ser	Thr	1	0	0.0249	0.0137	0.0249	1029	Leu	NA	NA	1
0.0278	0.0093	0.0278	772	Val	NA	NA	1	0	0.0112	0.0144	1030	Gly	Ala	1	0
0	0.0085	0.0316	773	Glu	Gly	NA	0	0	0.0129	0.023	1031	Ile	Val	1	0
0	0.0087	0.0316	774	Glu	Lys	NA	0	0.0316	0.0129	0.0316	1032	Glu	NA	NA	1
0	0.0085	0.0249	775	Leu	Arg	NA	0	0	0.0141	0.0221	1033	Asp	Asn	1	0
0	0.01	0.0431	776	Arg	Trp	1	0	0.0412	0.0116	0.0412	1034	Ser	NA	NA	1
0	0.01	0.0221	777	Asp	Asn	NA	0	0	0.0157	0.0249	1035	Leu	Pro	NA	0
0	0.01	0.0431	778	Arg	Gly	1	0	0.0316	0.0157	0.0316	1036	Glu	NA	NA	1
0.0278	0.0127	0.0278	779	Val	NA	NA	1	0	0.0157	0.0412	1037	Ser	Arg	1	0
0.0297	0.0156	0.0297	780	Lys	NA	NA	1	0.0115	0.0171	0.0115	1038	Gln	NA	NA	1
0.0144	0.0184	0.0144	781	Gly	NA	NA	1	0	0.0144	0.0221	1039	Asp	Asn	1	0
0.0278	0.0198	0.0278	782	Val	NA	NA	1	0.0412	0.0179	0.0412	1040	Ser	NA	NA	1
0	0.0198	0.0412	783	Ser	Ter	NA	0	0	0.0162	0.0192	1041	Met	Val	NA	0

0.0268	0.017	0.0268	784	Ala	NA	NA	1	0.0316	0.0185	0.0316	1042	Glu	NA	NA	1
0.0297	0.014	0.0297	785	Lys	NA	NA	1	0.0134	0.0197	0.0134	1043	Phe	NA	NA	1