T-cell repertoire analysis in Paroxysmal Nocturnal Haemoglobinuria

Bethany Laura Kuszlewicz

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy in Medicine

The University of Leeds
School of Medicine
Leeds Institute of Medical Research
Division of Haematology and Immunology

September, 2020

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to take this opportunity to acknowledge and thank all of the people who have made the work in this PhD thesis possible. Firstly, I would like to thank the University of Leeds for allowing me to carry out this project work and for funding my studies through the "Leeds Anniversary Research Scholarship". I'd like to thank my main supervisor Dr Darren Newton for his advice and guidance over the duration of this project. In addition, his expertise and patience during the method development stages of this project when no science wanted to work. I would also like to thank my co-supervisors, Professor Peter Hillmen and Dr Gina Doody for their support during this project. In particular, thank you to Pete for his insight and expertise into the world of PNH and AA. Thank you to Dr Stephen Richards for helping me improve my knowledge of PNH and AA in a diagnostic context and providing invaluable insight into patient selection for this study. It was a real privilege to be able to research and learn about PNH and AA alongside experts in the field. The PNH Research Tissue Bank played an important role in this project providing both patient and control samples and I'd like to thank the team for their contributions. Significant amounts of sequencing data were produced in this project thanks to the Leeds Next Generation Sequencing Facility MiSeq® services.

Most importantly, I would like to take this opportunity to thank all of the people who kindly donated samples to this study. In particular, a big thank you to the PNH and AA patients. Without these donations, the work would not have been possible and I appreciate your contributions no end. I hope that this work can help contribute further to the understanding of the mechanisms of PNH to work towards therapeutics and cures.

Finally, I would like to thank all of my friends and family who have supported me over the years to get me to this point, for their kindness, encouragement and belief in my work. To Amy, Helen and Meg, I want to thank you for being some of my biggest supporters from Day 1 and always taking the time to understand what my research is all about! To my family, in particular my Mum, Dad and Brother, I want to thank you all for your continuous support in all that I do, for your encouragement, guidance and belief in my abilities. Without you, none of this would have been possible.

Abstract

Paroxysmal Nocturnal Haemoglobinuria (PNH) is a rare, acquired, haematological disorder caused by the clonal expansions of haematopoietic stem cells in the bone marrow that have acquired a somatic *PIG-A* mutation. This mutation leads to glycosylphosphatidylinositol (GPI) anchors not being formed resulting in cells lacking over 25 GPI linked proteins on their surfaces. These proteins are important for many cellular functions including controlled inhibition of the complement pathway in the immune system.

However, the mutation alone does not result in PNH pathogenesis. Disruption of the bone marrow is required. Recent research on PNH long term bone marrow cultures, combined with links with the T-cell mediated disease Aplastic Anaemia (AA) has led to a hypothesis that T-cells are involved in the pathogenesis of PNH.

The main project aim was to assess whether there were specific T-cell receptor beta (TCRB) clones present exclusively in PNH patient repertoires. Firstly, 454 sequencing data from 18 PNH patients and 10 normals was analysed to identify a case for T-cells in PNH (**Chapter 3**). Subsequently, a high throughput TCRB sequencing method, along with a bioinformatics workflow, was designed, developed, and tested to analyse TCRB repertoires (**Chapter 4**). "Normal" (**Chapter 4**) and AA TCRB repertoires were analysed alongside PNH patients (**Chapter 5 and 6**) serving as comparisons. Over 150 million TCRB sequencing reads from 31 normals, 43 PNH patients, 26 AA patients with PNH and 6 AA patients with no PNH repertoires were analysed.

This project is one of the largest TCRB repertoire studies in the context of PNH, identifying 26 novel TCRB clones exclusive to PNH/AA patients and evidence to suggest links between TCRB clonal populations and PNH clinical status. This will aid further research into the role of T-cells in the pathogenesis and/or progression of PNH (**Chapter 7**).

Table of Contents

BLANK PAGE

List of Tables

List of Figures

List of abbreviations

AA	Aplastic Anaemia
aa	amino acid
APC	Antigen presenting cell
BC	Buffy coat
BM	Bone marrow
cDNA	Complementary DNA
cDNA	Complementary DNA
CDR1	Complementarity determining region 1
CDR2	Complementarity determining region 2
CDR3	Complementarity determining region 3
CMV	Cytomegalovirus
cTEC	Cortical thymic epithelial cells
DC	Dendritic cell
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
EBV	Epstein Barr virus
gDNA	Genomic DNA
GPI	Glycosylphosphatidylinositol
GRAVY	Grand average of hydropathy
HLA	Human leukocyte antigen
HSC	Haematopoietic stem cell
HTS	High throughput sequencing
HTS	High throughput sequencing
IMGT®	The international ImMunoGeneTics information system
iNKT	Invariant natural killer T-cells
ITAM	Immunoreceptor tyrosine-based activation motif
LGL Leukaemia	Large Granular Lymphocytic Leukaemia
LOH	Loss of heterozygosity
LTBMC	Long term bone marrow culture
MHC	Major histocompatibility complex
MNCs	Mononuclear cells
mRNA	Messenger RNA
mTEC	Medullary thymic epithelial cells
NGS	Next-generation sequencing
NK T-cells	Natural Killer T-cells
nt	nucleotide
PB	Peripheral blood
PCR	Polymerase chain reaction
PIG-A	Phosphatidylinositol N-acetylglucosaminyltransferase subunit A
pMHC	Peptide MHC interaction
PNH	Paroxysmal Nocturnal Haemoglobinuria
PNH	Paroxysmal Nocturnal Haemoglobinuria
RAG	Recombinase activating gene
RNA	Ribonucleic acid

RSS	Recombination signal sequences
RTB	Research Tissue Bank
T-cell	Thymocytes
TCR	T-cell receptor
TCRA	T-cell receptor alpha chain
TCRB	T-cell receptor beta chain
TCRBJ	T-cell receptor beta J region
TCRBV	T-cell receptor beta V region
TdT	Terminal deoxynucleotidyl transferase
Tfh	Follicular T helper
Th	T-helper
Tregs	Regulatory T-cells

Chapter 1 - Introduction

1.1 . A brief introduction to T-cells

T-cells, also known as thymocytes, are important mediators of cellular immunity capable of providing both short and long-term protection against a diverse range of pathogens [**1**]. They begin their lives as haematopoietic stem cells in the bone marrow [**2**] from which, in the form of lymphocytic progenitor cells, they migrate to the thymus where they undergo a series of developmental processes including V(D)J genetic recombination events detailed in **Section 1.3.1.** [**3,4**].

V(D)J gene recombination events lead to the generation of a unique range of T-cell receptors (TCRs) being expressed on the surfaces of T-cells [**5**]. Research currently suggests that only one type of unique TCR is expressed on the surface of a single T-cell [**6**]. The cell bearing the receptor then undergoes a series of selection events to ensure it is i) capable of binding major histocompatibility class molecules/complexes (MHC) to initiate immune responses ii) releasing MHC complexes so that immune responses can be dampened down once the antigen has been cleared and iii) to ensure T-cells do not attack uninfected or healthy host cells [**7,8,9**]. This process is part of the theory of clonal selection. T-cells are thought to express different receptors, which are antigen specific, as a result of genetic recombination and mutations. Once the TCR is presented with the antigen that it recognises via the process of antigen presentation, it undergoes proliferation and clonal expansion. The term clonal selection was coined to explain the theory that the lymphocytes are antigen specific before coming into contact with the antigen and go on to proliferate as a result of selection by an antigen [**276**].

MHC molecules are glycoproteins whose genes are located on chromosome 6. The gene complex is often referred to as human leukocyte antigen (HLA) and is made up of more than 200 genes. The MHC locus is located in the short arm of chromosome 6p21.31. MHC molecules are a critical part of the immune system. They allow T-cells to differentiate between antigens from self (an individual's own tissues and cells) or non-self, which can include other individual's tissues, for instance through transplants or foreign antigens such as environmental pathogens and viruses [**10**].

Firstly, positive selection is carried out on the TCR expressing T-cells. T-cells that do not express a TCR, or the TCR does not recognise self MHC class molecules and, therefore, cannot recognise antigens to mount an immune response, die. MHCs expressed by an individual's immune system are referred to as self (discussed in more detail in **Section 1.2**.) [**11**]. In this process, cortical thymic epithelial cells (cTECs) in the thymus act as antigen presenting cells (APCs) and are able to present private peptides (those created within the human body of an individual) to the T-cells in order to test the TCRs recognition capabilities [**12**].

The second selection event is known as negative selection, whereby successful self-MHC specific T-cells migrate to the medulla portion of the thymus [**13**]. In the medulla, self-peptides are presented to the T-cells by dendritic cells or medullary thymic epithelial cells (mTECs) using MHC molecular machinery [**14**]. T-cells that express receptors that have too high a binding affinity for MHC complex/self-peptides will die at this stage [**15**]. This helps reduce the chances of an individual developing autoimmune diseases caused by auto reactive T-cells whose receptors readily bind host tissues [**16**] as well as removing T-cells that bind too strongly to MHC complex/self-peptides, which could cause heightened and prolonged immune responses [**17**].

Successful T-cells that have passed all stages of the selection process leave the thymus as naïve T-cells expressing a functioning alpha-beta or gamma-delta T-cell receptor [**18**]. Naïve T-cells then circulate around the body via the peripheral blood and lymphatic system [**19**]. They pass through secondary (or peripheral) lymphoid organs such as the spleen and lymph nodes, which house mature naïve T-cells along with mature APCs such as dendritic cells [**20**]. Naïve T-cells become primed when they interact with an APC that is presenting an antigen in the context of an MHC molecule to which its receptor can bind. This is providing that there are also other necessary co-stimulatory signals. These costimulatory molecules such as surface molecules are necessary to form immunological synapses and secrete chemokines for processes such as migration [**21**].

T-cell signalling occurs through the CD3 complex of the T-cell receptor. Signalling is essential for the activation and subsequent proliferation of immune cells [**22**]. Within the CD3, in the cytoplasmic tail, there are conserved regions of four amino acids that are repeated twice. These are immunoreceptor tyrosine-based activation motifs (ITAMS)[**23**].

When the receptor binds to the specific antigen, the tyrosine kinase *Lck* is recruited by CD4 or CD8. *Lck* phosphorylates the tyrosine residues in the ITAMs allowing other proteins to dock. This induces a signalling cascade [**24**]. *Lck* has a higher affinity for CD4 than CD8 and it has been proposed that the amount of *Lck* signal could determine which lineage the T-cell commits to [**25**]. The T-cells subsequently become activated and undergo clonal expansion of these specific populations expressing the antigen specific TCR [**26**]. These clonal expansions then differentiate into effector T-cells and memory T-cells such as central memory and tissue-resident memory T-cells (**Section 1.2.1.**) [**27**]. Effector T-cells will carry out functions such as killing infected cells. Memory T-cells allow the immune system to recognise the antigen should it re-infect the human body and can then mount a quicker immune response the second time round, a process known as immunological memory [**28,29**]. Dependent on the surface markers expressed by the T-cells after differentiation, cytokine and chemokine signalling, T-cells can migrate to inflamed tissues and other organs, enter the lymph nodes or will remain in the peripheral blood or tissues [**30**].

1.2 . T-cell subtypes and their functions

All T-cells express the co-factor cluster differentiation 3, CD3 on their cell surface. The CD3 is made up of six individual polypeptide chains. The TCR complex is formed when these chains, a CD3γ chain, a CD3δ chain, and two CD3ε chains, come together with the CD3 and TCR. This is important in intracellular signalling required for T-cell activation [**31**]. However, in addition, during developmental stages in the thymus, T-cells can fall into several different types of developmentally and functionally distinct lineages dependent on other co-factors expressed on the T-cell surface [**32**]. T-cell interactions with antigens presented by the immune system via molecular complexes (as mentioned above) can determine the T-cell subtype [**33**] some of which are discussed in this section.

1.2.1. Conventional T-cell subtypes

The first lineage is the conventional T-cell subset which consists of CD4+ and CD8+ T-cells [**34**]. These are T-cells that either express CD4 or CD8 respectively and bind peptides via MHC class complexes (**Figure. 1**) . The single positive thymocytes originated from common progenitor double positive cells that expressed both CD4 and CD8 during positive selection [**34**] **(Section 1.1**). The mechanisms for this CD4/CD8 T-cell lineage choice are not fully understood.

CD4+ T-cells bind MHC class II/ antigen complexes that are expressed on the surface of professional antigen presenting cells, such as dendritic cells. These antigens are extracellular proteins that have been digested within the cell and then presented to the T-cell via MHC class II molecules [**35**].

CD4+ T-cells are often described as helper T-cells (Th) and are involved in immune functions such as releasing cytokines that regulate downstream immune responses [**36**]. Subsets include Th1, Th2, Th9, Th17, Th22 and Tfh and are characterised by their corresponding cytokine profiles. Some CD4+ subsets have regulatory roles and are known as regulatory T-cells (Tregs) [**37**]. Tregs are involved in maintaining immune tolerance [**38**]. After an infection is cleared, the immune response needs dampening down, a process Tregs contribute to [**39**]. They, therefore, help prevent autoimmunity by suppressing auto reactive T-cells that have managed to escape the negative selection processes occurring in the thymus [**40**]. They also prevent autoimmunity by supressing innate cells and autoreactive B-cells that are products of autoantibodies and can present auto-antigens to T-cells [**41**].

In contrast, CD8+ T cells express the CD8 cofactors, CD8 alpha and beta, on their surface which bind to peptide/MHC class I molecules expressed on the surface of all nucleated cells. MHC class I molecules present endogenous proteins from within cells, such as virus infected cells and cancerous cells, allowing cytotoxic immune responses to be initiated. CD8+ T cells, therefore, have immune functions that kill cells that are infected, cancerous or damaged, by inducing apoptosis and are also known as cytotoxic T-cells [**42**].

CD4+/CD8+ T-cells can be separated further into categories according to their immunological state [**43**]. When a naïve T-cell becomes activated by presentation of a recognisable antigen by an APC, the T-cell proliferates and differentiates into effector T-cells. These effector T-cells can migrate to the site of the infection. These cells are short-lived however, these T-cells can differentiate into another subtype called memory T-cells. Memory T-cells are long-lived and can also be either CD4+ or CD8+ and respond to a specific antigen that has previously been encountered by the immune system, for example from a recurring infection [**44**]. These T-cells then differentiate into effector T-cells to mount immune responses against this antigen, usually faster than the first time the antigen was encountered [**45**] as first mentioned in **Section 1.1**. Subtypes of memory T-cell include central memory and tissue resident memory [**46**], the latter of which was only discovered in the past decade [**47**] and is unusual in regards to T-cell behaviour as this type does not circulate in the periphery [**48**]. Tissue resident T-cells present in organs such as the liver and are thought to contribute to an immune system's long-lived T-cell memory pool [**49**].

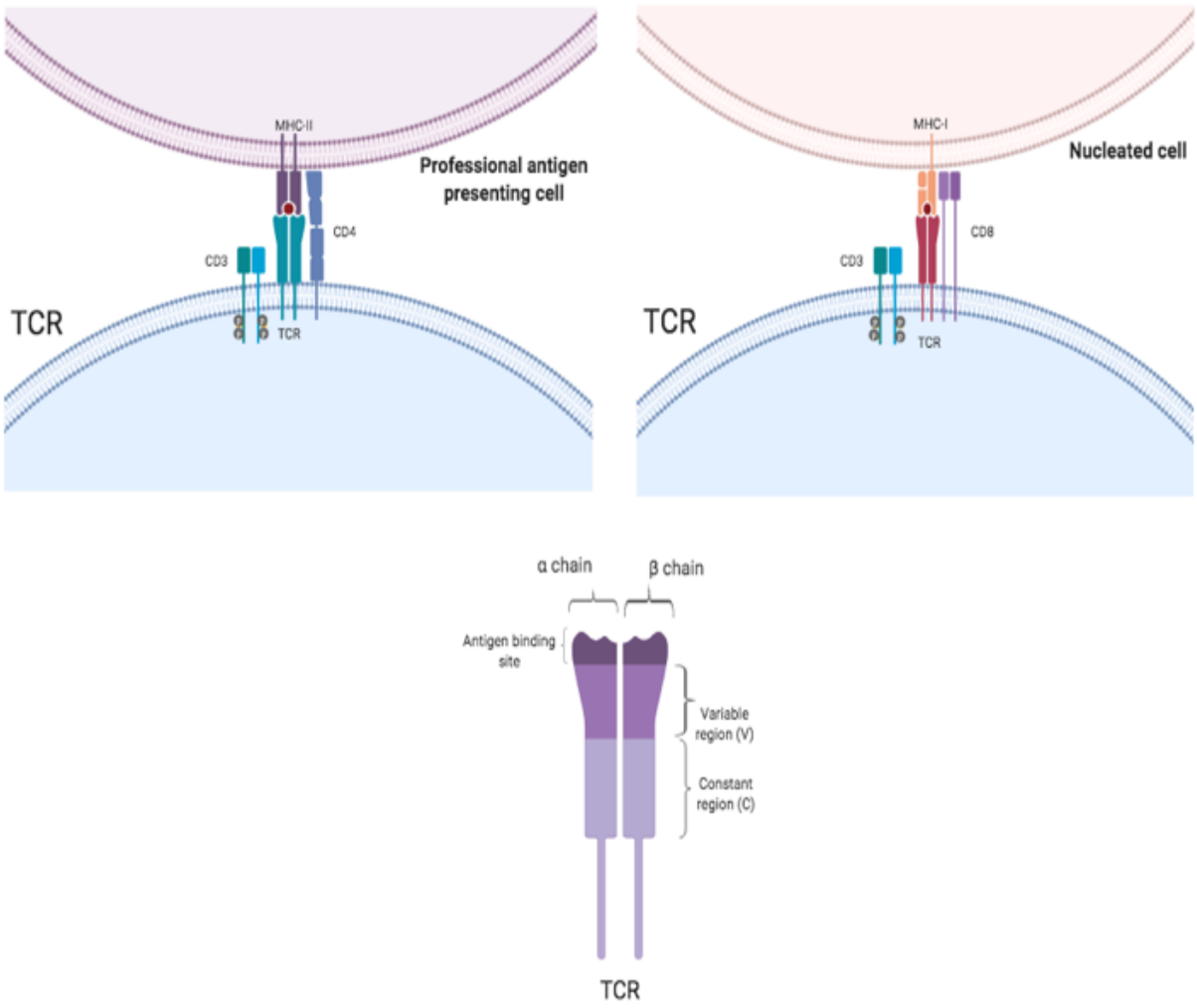


Figure 1. Simplified T-cell receptor/pMHC complex structures in CD4+ and CD8+ αβ T-cells.

Top left and top right represent T-cell receptor/pMHC bound complexes on the surfaces of CD4+ and CD8+ T-cells respectively. Both types of T-cell express co-factor CD3 which is important in intracellular signalling to activate T-cells. When T-cell receptors bind antigens, CD4+ and CD8+ T-cells bind to MHC class molecules presented on the surface of cells that present antigens. In CD4+ T-cells, T-cells express a CD4 co-factor. This binds to MHC class II molecules expressed on the surface of professional antigen presenting cells such as B cells. MHC class II molecules present antigens that are extracellular proteins. CD8+ T-cells express CD8 co-factors. These bind to MHC class I molecules expressed on the surface of nucleated cells and present endogenous proteins from virally infected cells and cancerous cells. The bottom image depicts a close-up version of the alpha beta TCR in the TCR/pMHC complex structures. The TCR is made up of an alpha and beta chain with a variable and constant region. The top of the variable region contains the antigen binding site, the portion of the TCR that interacts most closely with the antigen and contains the CDR3 region. To note, TCR complex signalling is not included in this diagram and the CD3 complex is not displayed in full for simplicity.

1.2.2. Non-conventional T-cell subtypes

The second lineage, unconventional T-cells, includes gamma delta T-cells and a T-cell population known as natural killer T-cells that express surface markers and display functions of both conventional T-cells and natural killer cells [**50**]. They, therefore, share characteristics observed in both innate and adaptive immune responses [**51**]. NKT cells are activated during inflammatory conditions and many infections [**52**].

Many NKT cells recognise antigens via the cluster of differentiation 1 complex (CD1) which is a glycoprotein expressed on the surface of some antigen presenting cells [**53**]. CD1 presents lipid antigens to T-cells [**54**] and as a result, NK-T cells can recognise self-lipids and lipids derived from foreign sources. Studies have shown in cord blood that NK T-cells display an activated memory phenotype indicative of an endogenous response during gestation [**55**]. There are four isoforms of CD1 that can make up the complex [**56**]. CD1a/b/c make up group 1 CD1 molecules and CD1d forms group 2 [**57**]. Group 2 CD1 T-cells generally expand after antigen recognition in the periphery and are considered polyclonal [**58**]. This group of T-cells are alternatively named CD1d restricted T-cells. Amongst the most common self-reactive T-cells found circulating in peripheral blood are the CD1a restricted T-cell subset which express a diverse range of T-cell receptors [**59,60**].

One group of NKT cells that have been extensively studied are known as invariant NK T-cells (iNKT) [**61**]. This population expresses TCRs belonging to one TCR alpha family and a limited number of TCR betas. In humans this is TRAV10-TRAJ18 and most commonly VB25 [**62**]. iNKT have been found to be long term residents in some tissues. Their adaptation to responding to lipids allows the immune system to survey and respond to a wider range of pathogens. However, iNKT-cells constitute 0.01-0.1% of T-cells in human blood and are therefore challenging to isolate and sequence in TCRB repertoire studies, therefore, are not the focus of this project work [**63**].

1.3. Types of T-cell receptors and their structures

As described in **Section 1.2**, the differences between conventional and unconventional T-cells are observed when looking at the complexes that they bind with (MHC class or CD1). These complexes are presented by APCs. MHC class complexes and CD1 complexes present antigens found in the human body to T-cells in order to generate downstream immune responses [**64**]. T-cells can then be split further into populations according to the type and structure of the TCR that they express.

The TCR is the portion of the T-cell that interacts with these complexes and there are two classes, alpha-beta and gamma-delta [**65**]. An individual's TCR repertoire is the combination of unique TCRs that are expressed on the surface of T-cell populations in a human being at a given time. Gamma delta TCRs only contribute between 1-10% of the entire TCR repertoire and tend to be attributed to mucosal gut immunology [**66**] and, therefore, only alpha beta TCRs are analysed in this thesis.

1.3.1. Structure of the alpha-beta T-cell receptor

Alpha-beta TCRs are heterodimers consisting of an alpha and beta chain [**67**] (**Figure 1.**). Each chain consists of a constant region (c), a joining region (j) and a variable region (v) [**68**] (**Figures 2 and 3**). The variable region is the portion of the receptor that binds to antigens presented by MHC class and CD1 complexes [**69**]. The human immune system is capable of recognising a wide range of pathogens that enter the body, the majority of which bind to different TCRs [**70**]. In order to understand how the immune system has evolved to recognise such an extensive range of pathogens, it is important to understand how TCRs are formed at a genetic level.

1.3.1.1. Combinatorial and junctional diversity

Before detailing the structures of the alpha and beta chains, it is important to understand the processes by which the chains rearrange to generate diversity. These processes are known as combinatorial and junctional diversity events. The VDJ Recombinase complex is essential in these processes.

Combinatorial diversity arises when different V, (D) and J gene segments randomly recombine to form the V-region exon (**Figures 2. and 3.**) [**71**]. Two of the proteins required are located at the ends of the VDJ genes and are essential for activation of VDJ recombination. These are the 'recombination activating genes' *RAG1* and *RAG2* and are only expressed in the developing T-cells in order to allow for TCR gene re-arrangement [**72**]. In combination with additional proteins, they form a complex enabling the separation and subsequent re-arranging and then re-joining of the VDJ genes. The *RAG* genes encode enzymes that cleave the dsDNA between the antigen receptor coding segment and a region known as the flanking recombination signal sequence (RSS) creating junctions [**73**]. The RAG enzymes remain until other proteins are recruited to the site which repair the junctions.

The next steps lead to junctional diversity which occurs in the junctional regions encoded by the V, D and J gene segments. One of the proteins that is recruited to the site where the *RAG* genes are present is a lymphoid specific enzyme called terminal deoxynucleotidyl transferase (TdT). This enzyme randomly adds P and N nucleotides at these gene segment junctions to repair the nicks in the DNA [**74**]. This generates the many combinations of VDJ genes that form a TCR receptor resulting in the diversity. Between TCRA and TCRB chains, the estimated number of V gene pairs at $5.8x10^6$ and junctional diversity at approximately 2×10^{11} estimates a total diversity of the TCRAB repertoire of 10^{18} [**75**].

1.3.1.2. TCR alpha locus

The TCR alpha locus consists of 70-80 V alpha gene segments and then a cluster of 61 J alpha segments a considerable distance from the V genes (**Figure 2.)** [**76**]. The TCR alpha locus is interrupted between the V and J gene segments by the TCR delta locus [**77**]. The J genes are followed by one C gene, known as the constant gene segment, because this is identical for all TCR alpha chains [**78**]. During T-cell development, the germline genes rearrange and undergo a number of recombination events and nucleotide insertion, deletion and substitution events to form a unique TCR alpha chain [**79**]. This consists of just one V gene, one J gene and the constant gene segment. Due to the large number of TCRAJ genes in the TCR alpha locus, variability in CDR3s is even greater.

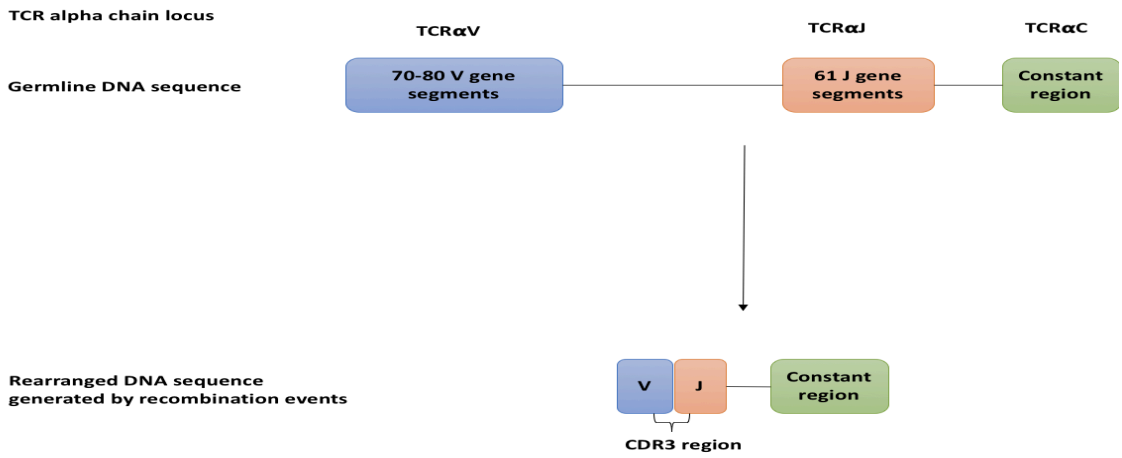


Figure 2. TCR alpha VJ recombination events that occur in the thymus.

The germline TCR alpha chain DNA sequence contains 70-80 V gene segments, 61J gene segments and a constant region. Recombination events occur to generate a TCR alpha chain containing one V gene segment, one J and one constant region. The CDR3 straddles the VJ junction. The diagram does not show the TCR delta locus that spans the TCR alpha and the distances are not in proportion, the diagram is for illustrative purposes.

1.3.1.3. TCR beta locus

The fundamental difference between alpha and beta chains are that the TCR beta locus has an additional diversity gene segment [**80**]. The TCR beta locus has 52 functional V beta gene segments located away from two separate clusters that each contain a single diversity, D, gene segment, with six or seven J segments and then a constant region (**Figure. 3**). Each TCR beta C gene has separate exons encoding the constant domain [**81**]. The diversity gene segment present in the TCR beta locus allows for greater diversity in TCR beta chains, because of the potential for a greater number of different gene combinations produced during recombination, than in alpha chains that lack the additional D gene segments [**82**].

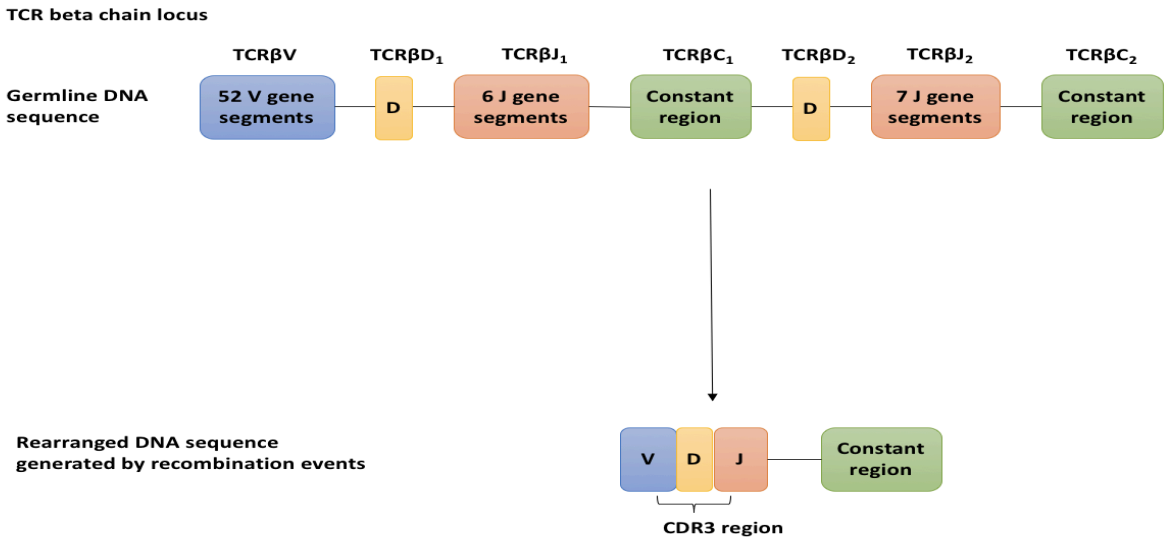


Figure 3. TCR beta VDJ recombination events that occur in the thymus.

The germline TCR beta chain DNA sequence contains 52 V gene segments, followed by two clusters each containing a diversity gene segment, 6 or 7 J gene segments and a constant region. Recombination events occur to generate a TCR beta chain containing one V gene segment, one J and one constant region. The CDR3 region straddles the rearranged VDJ section of the locus and accounts for the majority of the variation seen in an individual's T-cell receptor repertoire. The distances between gene segments are not in proportion, the diagram is for illustrative purposes.

1.3.1.4. CDR3 region and its importance in T-cell receptor repertoire sequencing

A TCR repertoire is the characterisation of a population of T-cells, generally circulating in the peripheral blood, by the TCR that they express. A TCR contains three complementarity determining regions, CDR1, CDR2 and CDR3. They constitute part of the variable region in the TCR and are important in the binding of the MHC class or CD1 presented peptides to the TCR, which is needed to recognise the antigen and to promote a subsequent immune response if necessary [**83**]. CDR1 and CDR2 only bind to the MHC class molecules/CD1 complexes and are located in the germline encoded V domains so are less commonly used in repertoire sequencing studies [**86**]. CDR3 is most commonly used when defining and assessing TCR clones in TCR repertoire research because it straddles the TCR V(D)J junction, is unique for every T-cell clonotype (defined in **2.9.2.2.**) and encodes the receptor portion that interacts most closely with the antigenic peptide (**Figure 4.**) [**84**]. According to IMGT© (the international ImMunoGeneTics information system®) the gold standard knowledge base for TCR repertoire analysis, structurally speaking, the CDR3 is the 'codon positions 105 to end of the V-REGION in germline gDNA or cDNA, codon positions 105 to 117 in V-DOMAIN' [**85**]. As CDR3s also contain D-J junctions, TCRBJ contributes to significant diversity within the repertoire, especially in TCRA with 61 TCRBJ gene segments [**87**].

The diversity of the CDR3 amino acid sequence, in combination with the alpha and beta chain genes, provides insight into T-cell diversity and shows if the TCR repertoire is skewed towards a particular antigen (antigen driven) due to infection or disease, for example. The amino acid properties of the CDR3 can also shed light on possible immune recognition mechanisms of T-cell clones. For example, CDR3s that harbour hydrophobic amino acids at specific points in the CDR3 sequence are thought to promote the development of self-reactive T-cells that are involved in autoimmune disease in mice [**88**].

CDR3 lengths are increased from germline sizes by the addition of nucleotides. This process generates a greater diversity of TCRs with both beneficial and negative effects. Longer CDR3s have a higher chance of sequence variation and they can potentially reach into narrower antigenic pockets [**89,90**]. Shorter CDR3s would struggle to interact with these pockets. These CDR3s however, have been found to be highly enriched during thymic selection. Antigen driven TCRs also tend to have shorter CDR3s than their naïve counterparts and public T-cell responses also tend to involve shorter CDR3s [**91**].

The CDR3 region of the beta chain accounts for the majority of the variation seen in a person's TCR repertoire [**92**], and, therefore, will be the starting point for analysis in this project. From here on out, a TCR clonotype, in the context of this research project, will refer to a T-cell population that expresses a receptor with an identical TCR beta V gene family, J gene family and CDR3 amino acid sequence. The frequency of a specific CDR3 sequence indicates the abundance of its TCR clone and, combined with the TCR V and J gene family usage data, is an important aspect of analysis for deciphering immune repertoire properties that may be linked to illnesses such as autoimmune diseases or viral responses [**93**].

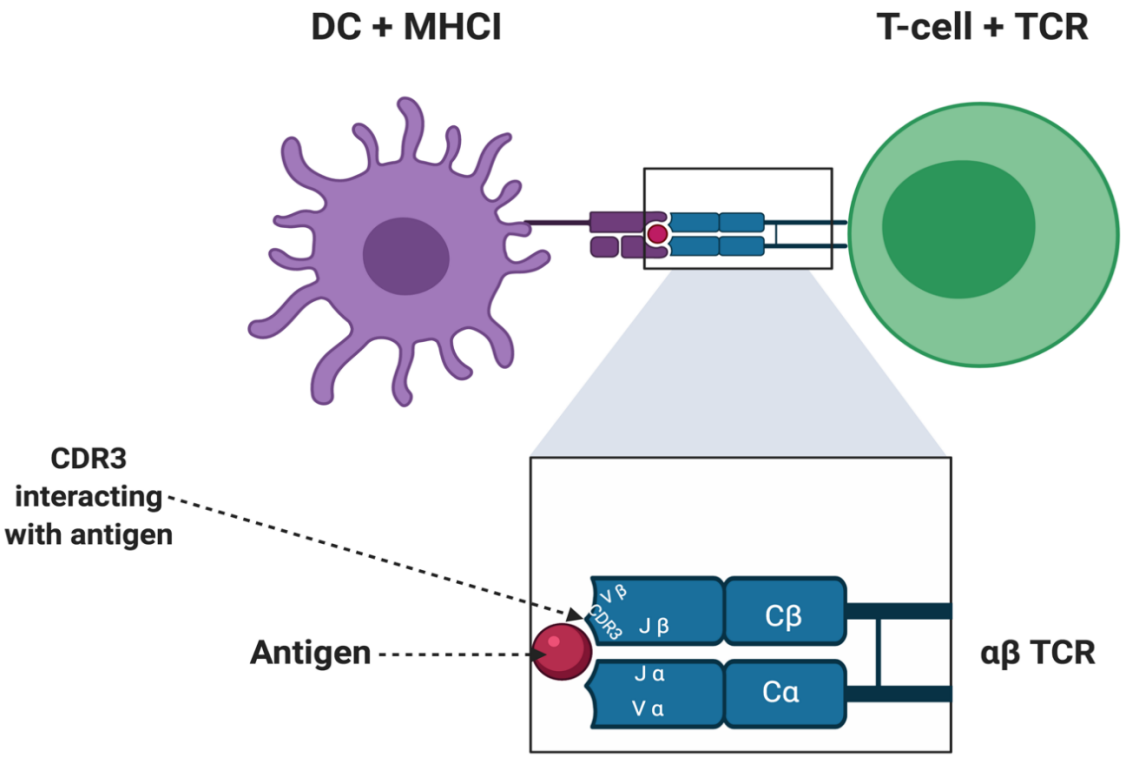


Figure 4. Interaction of CDR3β of an alpha beta TCR with an antigen presented via an MHC I.

The dendritic cell, acting as an antigen presenting cell, is presenting the antigen via an MHC I to the alpha beta TCR. Each alpha beta TCR is made up of an alpha and beta chain. The alpha chain contains a constant region (C), a joining region (J) and variable region (V). The beta chain contains a constant region, variable region, joining region and diversity region. The region that straddles the V-D-J junction is known as the CDR3. The CDR3 is the portion of the alpha beta TCR that interacts most closely with an antigen and CDR3 beta is thought to account for the majority of the variation observed in a TCR repertoire.

1.4. Factors affecting the diversity of T-cell receptor repertoires

TCR repertoires are dynamic, they are constantly changing [**94**]. Many factors can influence an individual's TCR variation over time and variation between individuals, including stress, diet, illness, sex and age (with regards to increasing TCRB clonality) as well as factors such as genetic biases that can influence overall TCR repertoire composition [**95-101**]. TCR repertoire changes with age are described in **Figure 5**.

1.4.1. Genetic variation and pre-selection of the T-cell receptor repertoire

Genetic variation accounts for 20-40% of immunological variation observed. Copy number variants of genes and rare variants in the genome can cause structural variations which attribute to this [**102**]. Only a small set of genes encode TCRs but, as a result of random deletion, insertion and substitution events, as well as genetic recombination processes, it is estimated that around 10^{15} to 10^{20} T-cell clonotypes are capable of being produced in a human being [**103**]. However, this is not realistic because there are only 10^{13} human cells in the entire human body, and T-cells only contribute to 70-85% of lymphocyte populations [**104**]. It does, however, illustrate the magnitude of diversity that a repertoire can attain. The majority of TCR clonotypes do not even make it to the periphery after undergoing the strict thymic selection events discussed in **Section 1.1.** to eliminate autoreactive T-cells. It is thought that in reality only one in a 100 thymocytes reach the periphery [**105**] and that there are only approximately 25 million clonotypes circulating in the peripheral blood [**106**]. The majority of T-cells are thought to be present in lymphoid tissues with only around 2-3% found in peripheral blood [**107**].

One known genetic cause for the restricted and scaled down diversity of TCR receptors observed in peripheral blood is the composition of the pre-selection TCR repertoire (**Figure 6.**) [**108**]. The pre-selection TCR repertoire is the repertoire that exists before positive and negative selection of TCRs occurs in the thymus [**109**]. The diversity of alpha-beta TCRs is heavily restricted and structured before thymic selection, with genetic factors influencing the composition of this pre-selection repertoire and, therefore, influencing the TCR clonotypes that successfully enter the peripheral blood [**110**].

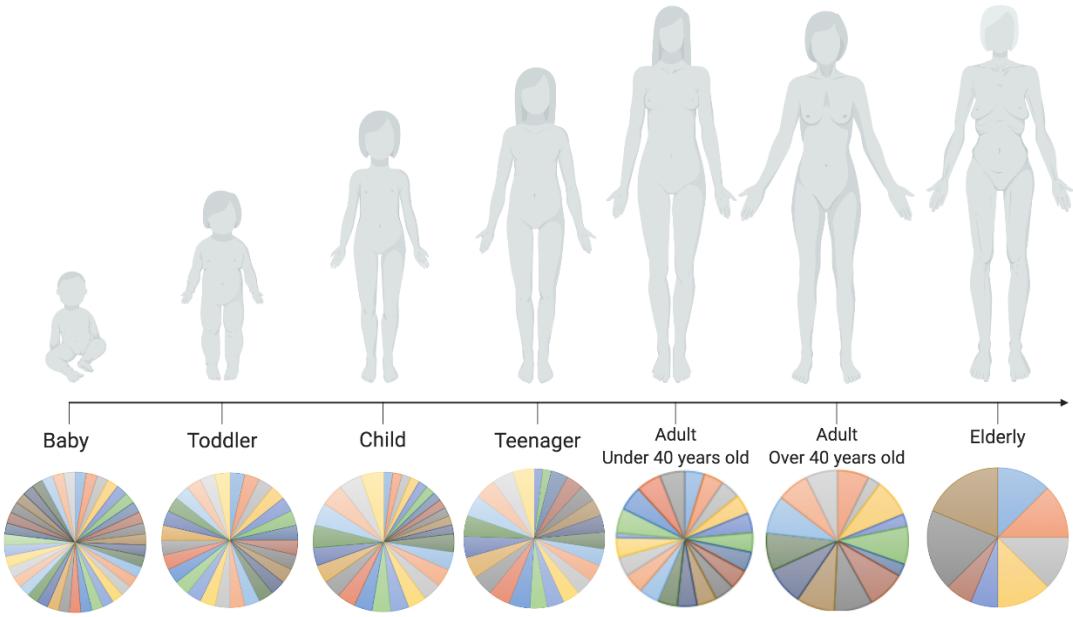


Figure 5. Diversity of TCRB repertoires decrease with the ageing process.

The seven stages of human ageing are shown in relation to what an average TCRB repertoire would expect to look like at each stage. The circles depict the TCRB repertoire as a whole and the segments are simplistic indicating TCRB clones. The greater the number of segments in the sphere, the greater the TCRB repertoire diversity. The more segments of similar size in the repertoire, the more stable it is. Larger segments indicate expanded TCRB clones that are proliferating. This can be in response to an antigen to drive an immune response, or proliferation to sustain homeostasis of naïve T-cells and mature subtypes. As a baby, the TCRB has the capacity to be extremely diverse and is influenced by the maternal TCRB repertoire [**391**]. As the baby ages into a toddler, environmental factors such as diet and exposure to antigens are involved. TCRB diversity remains high and some TCRB clones may expand slightly in response to infection or vaccinations. As a child, TCRB diversity remains high, but the thymic involution process gradually starts, therefore, diversity is slightly lower than at younger ages. Even healthy children have been found to have persistent monoclonal TCRB clones to antigens [**336**]. Moving into the teenage years and adulthood below the age of 40 years old, TCRB diversity decreases over time and larger clonal expansions occurring to infection and other factors such as stress will affect the TCRB repertoire stability [**95,355**]. Once an adult passes the age of 40 years old, studies have suggested that the thymus drastically decreases the output of naïve T-cells, TCRB diversity decreases and the immune system is less capable of mounting immune responses to new antigens. In old age, immunosenescence is in full progress, TCRB diversity is at its lowest and large sustained TCRB clones will appear in response to recurrent infections such as EBV. The elderly often see a reverse in their CD4:CD8 T-cell ratios, a marker of immunosenescence [**135-136**].

1.4.2. Effects of environment on TCR diversity

One study estimated that environmental factors can influence over 50% of immunological variation. It suggested that convergence of immune status could be linked to cohabiting. Individuals who shared living quarters had a higher chance of sharing unique TCRs than those who did not. This could be attributed to multiple factors that are associated with cohabitation, such as sharing diets, stress and exposure to similar antigens, allergens and same courses of vaccines for example for overseas travel [**111**].

1.4.3. Generation and prevalence of "public" T-cell receptor clonotypes

Another layer of diversity that can be added to the TCR repertoire is evaluating how common a unique TCR is within a given population size consisting of many individuals' TCR repertoires. For example, the TCRs that are shared amongst the cohabiting individuals from the study cited above, can be referred to as "public TCR clonotypes" as they occur in more than one individual. "Public" clonotypes can also be categorised further into "semi-public" and "public" dependent on whether they are observed in just a couple of individuals, particularly relatives, or amongst many individuals in a population [**112**]. Both genetic and environmental factors are thought to be linked to "public" clonotypes evolving and some examples are outlined below [**113**].

1.4.3.1. V(D)J recombination events, convergent recombination and recombinatorial biases

The V(D)J recombination events in T-cell development that lead to TCR clonotype diversity were originally thought to be entirely random. However, recent research has suggested that some rearrangements are more favourable and occur more frequently than others, suggesting a reason for 'public' clonotypes arising [**114-115**]. Unequal frequency distribution of TCR clonotypes is thought to be attributed to convergent recombination leading to some TCRs having structural advantages over others by having more than one way of being produced. For example, particular nucleotide sequences can be produced using a variety of recombination events, so are more likely to occur, some amino acids can be made by a greater number of nucleotide triplets and certain TCRs require fewer insertion, deletion and substitution events, creating potential skewing of clonotype frequency distributions before the receptor has even met any antigens [**116**].

1.4.3.2. Common antigens and shared MHC class complexes

Individuals will share TCR clonotypes that are responsive to antigens that are abundant in a population and commonly infect humans, for example antigens to Cytomegalovirus, CMV [**117**]. Another example would be vaccination, where a substantial number of people are infected with the same live or attenuated antigen. This is because individuals will show some level of similarity in the MHC class complexes that they express at a genetic level, particularly related individuals.

As the MHC class complexes present the antigens to TCRs, it is likely structurally similar T-cells have evolved to bind to these alike MHC class molecules. This can lead to public TCR clonotypes that are identified in many individuals [**118**].

1.4.4. Generation and prevalence of "private" T-cell receptor clonotypes

Each person's TCR repertoire is generally highly diverse if they are "healthy" and is unique to that individual. Despite the presence of some public clonotypes, the majority of the TCRs found in a person's TCR repertoire are unique to that individual and are known as "private" clonotypes. Many factors can affect TCR repertoire diversity and can account for why the majority of clonotypes are "private" and not shared between individuals [**119**]. It would be a sensible assumption that a TCR repertoire's diversity is attributed to genetics and in some part this is true. An example is in the pre-selection TCR repertoire as detailed above.

However, a recent study sequencing the TCR repertoires of a set of twins showed that despite the twins sharing the same genes, their TCR repertoires shared no greater number of clonotypes than two randomly selected, non-related individuals [**120**]. Another study found that only 1.1% of TCR clonotypes were shared between two donors when defining a clonotype using a CDR3 nucleotide sequence, increasing to 14.2% when using CDR3 amino acid sequences [**121**]. This highlights that the majority of TCR clonotypes observed in a repertoire will be "private", however, the percentages of shared TCRBs may differ according to environmental, genetic and technical factors such as blood sampling.

Factors, for example the age of the individual (**Figure 5.**) and the environment that an individual person is exposed to, especially from an early age like infections, vaccines and allergens, greatly affect the composition and profile of their TCR repertoire, leading to a group of clonotypes only observed in the individual [**122**].

Research has also suggested that private clonotypes have evolved in response to preventing the spread of viruses and infections in contrast to herd immunity [**123**]. If every individual mounted exactly the same immune response to exactly the same antigens, it would be very easy for pathogens, such as viruses, to evolve to evade the immune responses and therefore, be able to infect and spread amongst the whole population. Having variations in responses means that the spread of a virus would be much slower within a population [**124**]. The unique composition of an individual's TCR repertoires have lead some researchers to believe it could be used as an "immunological fingerprint" that one day can be used to discover the immunological history of a person's immune system and how they are responding to a current function or illness and is why TCR repertoire sequencing is becoming more and more prevalent in research [**125-126**].

1.4.5. Homeostatic regulation of T-cells and thymic involution

The most effective immune systems are those that have highly diverse TCR repertoires [**127**]. Even though TCR repertoires are dynamic, when focussing on the most abundant TCRB clonotypes in an individual's repertoire, it remains fairly stable and constant throughout life, subject to events such as infections or illness. This was highlighted in a study where an individual's TCR repertoire was studied over the course of a year. When analysing the most abundant TCRs, 63% were shared across the two time points and the two samples clustered together when compared to other individuals' TCRs, showing stability [**128**]. This is attributed to homeostatic regulation which controls T-cell numbers and T-cell distribution using regulatory mechanisms including selection events that drive the specialisation of T-cells into subtypes that are best suited to occupy specific niches [**129-130**]. Homeostatic regulation is also maintained by balancing the death of naïve T-cells with the export of naïve T-cells from the thymus and the proliferation of naïve lymphocytes in the periphery in the absence of any known antigenic stimulation (homeostatic proliferation) [**131**]. In reality, it is thought that there may be some degree of antigenic stimulation but with low affinity self-antigens. This is because TCR interaction with MHC:peptide, along with signals generated from cytokine receptors are thought to be needed to elicit T-cell proliferation of any kind. CD4+ T-cells undergo homeostatic proliferation, dividing every 3-4 days. CD8+ T-cells perform this process much faster [**132**].

However, exceptions to this are infancy and old age. In infancy the TCR repertoire will undergo many changes in response to the environment [**133**]. During early years, infants are also given vaccinations, and this will also alter the TCRB repertoire as the immune system responds to the vaccine.

These T-cells will then shrink, differentiate into memory cells and circulate at low levels in the periphery, attributing to the increasing diversity of the repertoire.

Programming of the thymus in early life is essential for immune stability. Children who had undergone thymectomies were shown to have accelerated T-cell ageing usually attributed to the natural ageing process where the thymus begins the involution process (shrinks in size reducing export of naïve T-cells) from childhood [**134**]. Once an individual is approximately 40 years old [**99**], the immune response is no longer capable of educating naïve T-cells to new antigens and the involution process can be attributed to immunosenescence [**135-136**]. This is why the elderly tend to have larger clonal populations, for example T-cells responding to recurrent infections such as Epstein Barr Virus, and therefore, have a less diverse repertoire [**137**]. Involution of the thymus may allow the maintenance of an optimal peripheral TCR repertoire attained through years of antigen exposure in adulthood [**138**].

1.4.6. Antigen skewed T-cell receptor repertoires

TCR diversity is at its highest in naïve T-cell populations before T-cells have interacted with specific antigens [**139**]. It is important that the naïve populations are diverse to ensure that they can recognise a range of antigens. The genes that code the MHC molecules are amongst the most polymorphic genes in humans. MHC polymorphisms can have an effect on a person's TCR repertoire by determining which peptides are presented to T-cells and how efficiently a T-cell clonotype can interact with APCs that express the varying MHCs. Antigen experienced repertoires are seen to be skewed towards certain antigen specificities and pathogens that commonly infect human beings [**140**]. It is thought that public TCR clonotypes in these antigen experienced repertoires could help identify markers of particular diseases. Analysing an individual's TCR repertoire could act as an indicator of current immune status and the identification of antigen specific T-cells that are involved in specific disease development and autoimmunity, for instance Autoimmune Encephalomyelitis [**141**].

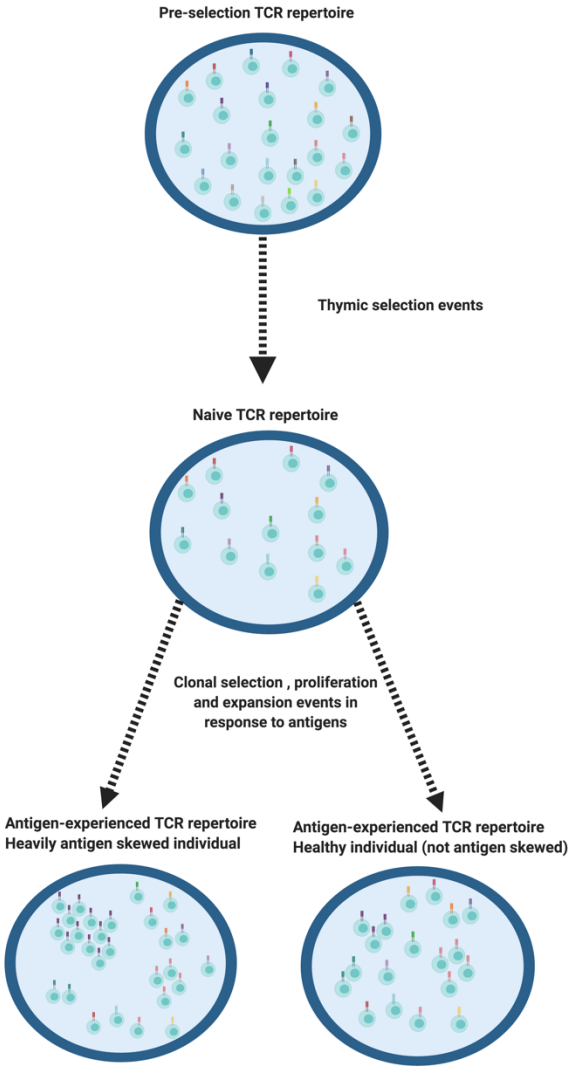


Figure 6. Developmental processes generating TCR repertoire diversity.

Each circle depicts a TCR repertoire at each stage. Within these circles are T-cells expressing a TCR. Each colour depicts a different type of TCR. The pre-selection TCR repertoire is biased by genetics and forms the basis for any subsequent thymic selection of TCRs. In the thymus, the pre-selection repertoire undergoes a number of positive and negative selection events which lead to the death of auto-reactive TCRs. The subsequent naïve TCR repertoire is less diverse than the pre-selection repertoire. These naïve T-cells enter the periphery where they undergo clonal selection, proliferation and expansion events in response to antigens. A healthy, non-antigen skewed TCR repertoire that is not responding to current infection may have a small number of clonal expansions (depicted by multiple TCRs of the same colour) but is generally diverse. The antigen skewed TCR repertoire on the bottom left shows some TCRs that are not clonally expanded along with groups of highly expanded TCRs. This symbolises how antigen driven TCR repertoires, such as those responding to infection, are less diverse and more clonally expanded.

1.5. T-cell receptor repertoire studies and their importance

The immune repertoire can be investigated by sequencing rearranged T-cell receptors, particularly those expressed by circulating T-cells in peripheral blood. It therefore allows for the profiling of the adaptive immune system [**142**] and can be applied to a range of research areas from clonal evolution, cancers and infectious diseases, to interpreting the immune status of a patient, for instance after a haematopoietic stem cell transplantation [**143-146**].

1.5.1. Historical T-cell receptor repertoire sequencing methods

With the advent of high throughput sequencing technologies, the TCR sequencing field has rapidly changed allowing researchers to identify key aspects of the TCR repertoire. However, prior to these technologies a number of alternative approaches were employed and provided research findings that were the foundation of the field, some of which are used in parallel with current high throughput sequencing techniques.

One method uses monoclonal antibodies that target the variable region of the T-cell receptor chain to analyse heterogeneous populations of T-cells using flow cytometry [**147**]. This method is good for studying selected T-cell populations and changes over time, for instance, variations affected by immune system ageing. However, currently the number of antibodies available does not cover the full range of TCRBV gene families and, therefore, cannot capture absolute diversity of T-cells. Flow cytometry is also unable to address more complex and detailed T-cell receptor analysis approaches such as V-D-J gene junctional studies and CDR3 analysis.

Spectratyping is another technique that has been employed to investigate TCR repertoires using the electrophoresis capillary method to perform CDR3 analysis [**148**]. CDR3 length distributions are analysed in the context of different TCRBV genes which is good for clonality studies such as those researching into T-cell Leukaemia where T-cells are known to be pathological.

TCR methodologies have advanced to incorporate sequencing technologies, from less high throughput methods such as Sanger sequencing to higher throughput methodologies such as Illumina® sequencing [**149**]. Sequencing TCRs allows the analysis of both CDR3 distributions and TCRBV gene usage as with the more historical TCR methods, but also allows for diversity of TCR sequences in a repertoire to be evaluated in the context of both TCRBV and J gene usage and analysis such as CDR3 amino acid properties and nucleotide insertions.

Different sequencing methods contribute advantages to the TCR sequencing field as well as caveats. For instance, 454 sequencing allows for longer sequencing reads which can capture the full length of TCR genomic DNA compared to sequencing technologies such as Illumina® [**150**]. However, the way in which 454 sequencing works (**Section 3.1.**) means that it is prone to homopolymer errors which can make sequence identification inaccurate. Illumina® sequencing works in a way that does not incur these homopolymer errors and provides a high throughput and greater sequencing depth for a sequencing library which is important when trying to assess TCR diversity [**151**].

When employing sequencing methods for TCR studies there are three main methods for amplifying TCR genes as part of the sequencing library preparation. The method type is dependent on the genetic material that the genes are being amplified from, RNA or gDNA. When using RNA only, a nested PCR method can be employed using 5'RACE which incorporates an adaptor at the 5'end of cDNA during its synthesis. The two other approaches can be used on both gDNA and cDNA (from mRNA). Target enrichment involves the fragmentation of gDNA/cDNA and then end repairing which incorporates multiple adenosine monophosphates to form a poly A tail using standard sequencing library preparations. RNA baits complementary to the sequence of interest hybridise to the targets, performing the target enrichment step. These target enriched sequences are then selected using magnetic beads and can undergo further amplification PCR steps and are subsequently sequenced. The final method, multiplex PCR, involves the PCR of the genetic material with multiple primer sets to amplify TCR genes, followed by subsequent PCR reactions to carry out adaptor ligation before sequencing and is the preferred method for the work in this project [**152-154**].

1.5.2. TCRB clonality versus TCRB diversity studies

TCRB repertoire studies are important for being able to identify T-cell dynamics in the context of human beings rather than having to rely on mouse models. Mouse models have been used in numerous studies to investigate areas such as naïve T-cell dynamics. However, mouse immune systems differ in some aspects from human beings [**155-156**]. Therefore, some conclusions drawn from the findings that are then assumed to be similar in human beings may be inaccurate. TCRB repertoire sequencing allows for a non-invasive analysis of T-cell dynamics in human beings.

TCR repertoire sequencing methods can be used to study both clonality and diversity of TCRB repertoires [**157-158**]. In this thesis, TCR diversity refers to the number of unique TCRB receptor clonotypes, (same TCRBV/J gene combination, and CDR3 amino acid sequence) which can sometimes be phrased as species richness, a term originally coined for ecology research [**159**].

TCR clonality in this thesis is defined as the number of unique TCRs within the repertoire that have undergone clonal expansion. The measurements of TCR clonality are defined in **Section 4.5.4.** using experimental methods.

Depending on the question the data is required to answer, the type of study used changes. Autoimmune disease and cancer studies tend to be more interested in clonal expansions of T-cells in response to disease and, therefore, focus more on clonal TCRB analysis [**160**]. In TCR repertoire studies, clonality can arise from two biological scenarios (**Figure 7.**). The first is homeostatic proliferation of naïve T-cells, which is required to maintain a healthy balance of T-cells circulating in the periphery (**Section 1.4.5.).** These clonal expansions tend to be lower in frequency and are characteristic of a healthy, diverse, stable TCRB repertoire. The second is clonal proliferation of TCRB clones in response to an antigen, triggering downstream immune responses **(Section 1.4.6.**). These clonal expansions are characteristic of antigen skewed TCR repertoires in response to infections and disease, tend to be less diverse and are associated with the elderly in particular [**132**].

Immunological genetic diseases that do not have an autoimmune basis will be more concerned with the overall diversity and stability of a TCRB repertoire and variations of nucleotide substitution, deletion and insertion events that occur in the TCR development stages in the thymus [**161**]. Many diversity measures such as the Chao1 estimator [**162**] are used in repertoire studies and have been adapted from uses in population ecology which look at species variation [**163**].

Diversity studies in both B- and T-cell repertoires are being used to infer immunological fitness of an individual, for instance, to predict the success of immunotherapies in treatments of some cancers. Research is attempting to define biomarkers according to repertoire characteristics and link these to predictions of patient frailty [**164-165**]. Immune repertoire studies combined with medical and treatment records would be needed (**Chapter 7.**) which is a challenge for the field.

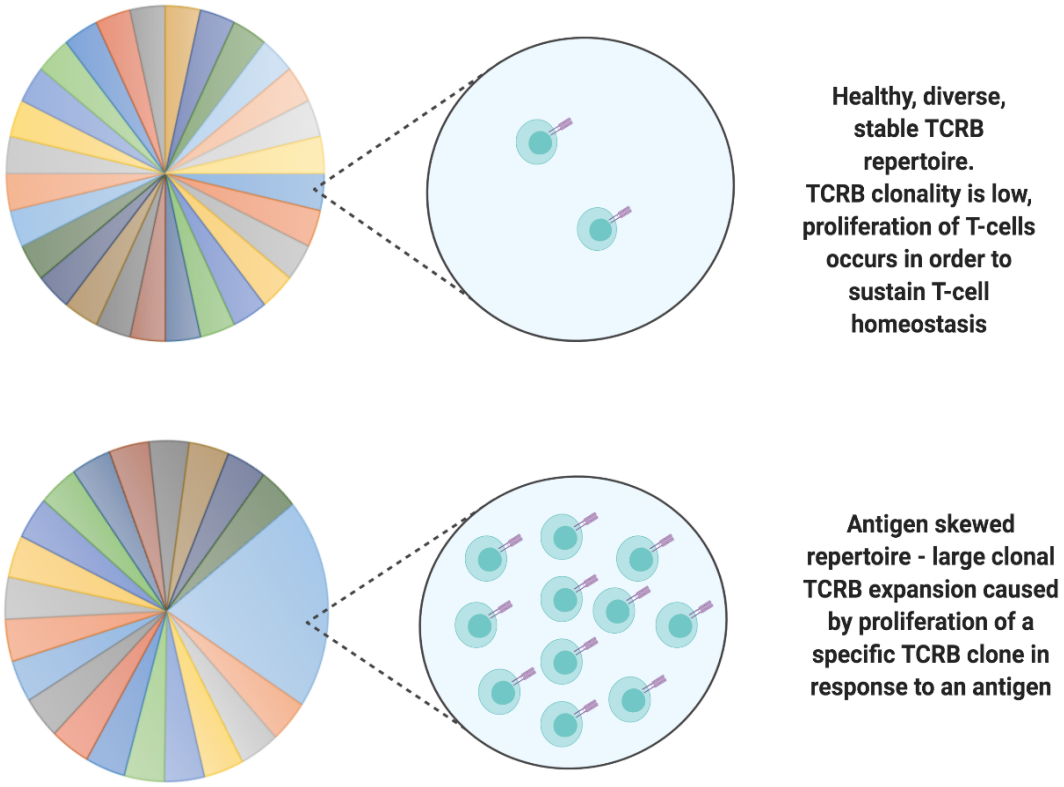


Figure 7. Types of clonality caused by proliferation in the TCRB repertoire.

T-cell proliferation is essential to sustain a healthy, diverse TCRB repertoire. Two types of proliferation can lead to varying levels of TCRB clonality in the repertoire. Within a healthy, stable, diverse repertoire, indicated by the circle with equal segments in the top diagram, slight TCRB clonal expansion will be observed. These are naïve T-cell proliferations that are essential in maintaining immune cell. The other type of proliferation is TCRBs that are proliferating in response to an antigen in order to promote a downstream immune response. These TCRBs can range in size and can take up a considerable percentage of the TCRB repertoire. This leads to skewed repertoires, indicated by the circle in the bottom left hand corner. Homeostatic proliferation will be much lower in percentage in the context of the entire TCRB repertoire, indicated by two TCRs in the top right. Antigen skewed responses will tend to be larger, represented by multiple TCRs of the same type in the bottom right.

1.6. A brief overview of bioinformatic analysis in TCRB repertoire studies

As the TCRB sequencing field expands and moves towards high throughput sequencing technology, generating millions upon millions of sequencing reads, the bioinformatic analysis supporting TCRB sequencing methods is ever changing. Bioinformatic analysis is required for both the pre-processing of sequencing reads as they come off a sequencer and subsequent downstream analysis, taking TCRB sequencing reads and then accurately re-creating TCRB repertoires. Analysis steps and tools will be discussed fully in the **Material and Methods Section**, however, below, is a brief summary of the steps to be expected in a standard TCRB repertoire study.

Many software packages exist to cater for different steps in the analysis, which are summarised in **Table 1**. Samples are de-multiplexed according to the sequencing adapters, prior to being taken off the sequencer. If unique molecular identifiers have been used in the sequencing methods, an additional de-multiplexing step will be carried out, collapsing PCR duplicates into biological TCR repertoires according to UMI consensus [**166**] (discussed in **Section 2.8.9. and Section 6.7.**).

Once the sequencing data has come off a sequencer the data needs to undergo strict quality control steps to ensure that the sequencing data is of good quality, especially if it is being used to detect genetic nucleotide variations in genetic studies. Quality is usually assessed base by base, with expected quality dropping towards the end of a sequencing read in the case of Illumina® MiSeq sequencing. Usually this is carried out using the base Phred scores and filtering accordingly. Phred scores [**167**] are discussed in **Section 2.9.1.2.**

Depending on whether the sequencing method used is single read or paired-end reads, the paired sequencing reads may need aligning. When sequencing paired-end reads, there is usually a region of overlap between the forward and reverse reads which is used by an algorithm to generate a consensus sequence leading to the overlap of the reads into one accurate sequencing read.

This sequencing read then needs to be annotated with TCRB genes. Usually the standard database that the TCRB data is aligned to is IMGT® [**85**] which gives a comprehensive overview of possible TCRB genes.

Once the reads are annotated into TCRB reads, in order to create a TCR repertoire, they must be assembled into TCRB clones using pre-determined definitions. Additional error corrective steps can then be undertaken to ensure accurate TCRB repertoires are assembled from the data.

Once the TCRB clones are assembled, there are a great deal of interpretations that can be made from the data. Again, depending on the biological question being asked, the analysis can include anything from antigen specificity, TCRBV/J gene usage, public and private clonotypes, clonal abundance and 3D protein modelling. Diversity can be calculated using measures such as Gini-Simpson [**168**] and rarefaction plots [**169**].

Table 1. Analysis steps and complementary tools and softwares for TCR repertoire analysis.

Analysis step	**Examples of tools and softwares available**
De-multiplexing and analysing UMIs	Presto [**170**], MIGEC [**171**], clipUMI [**172**], UMI-reducer [**173**], Decombinator [**174**]
Quality control	FastQC [**175**]
Quality filtering according to Phred scores and trimming reads	Trimmomatic [**176**], TrimGalore [**177**], Cutadapt [**178**], FastX toolkit [**179**]
Overlapping paired-end reads to build a consensus read	Pandaseq [**180**], pear [**181**], FLASH [**182**]
Annotating TCR reads	MiXCR [**183**] aligning to IMGT® IMGT/high V-quest [**184**]
Assembling TCR clones	MiXCR [**183**]
Downstream TCR analysis	Immunarch(formally tCR) [**185**], MiXCR [**183**], VDJtools [**186**], scTCRseq (single cell TCRseq) [**187**], MiTCR [**188**]

1.7. Paroxysmal Nocturnal Haemoglobinuria as a model for TCR repertoire sequencing

In this project, experimental and bioinformatics analysis of TCRB repertoires was used to investigate T-cells in the context of the disease Paroxysmal Nocturnal Haemoglobinuria (PNH). Patients' peripheral blood and bone marrow samples from the PNH Research Tissue Bank were used to evaluate whether changes in TCR repertoires occur in patients with PNH and whether there are TCR clonotypes specific to the disease. The following section will describe PNH, its pathology and the involvement of T-cells in PNH to show how TCR repertoire sequencing benefits research into this disease.

1.7.1. A brief introduction to PNH

PNH, is a rare, life threatening and chronic disease. It is caused by the clonal expansion of haematopoietic stem cells that harbour a somatic mutation in the *PIG-A* gene [**189-190**]. It is classed as a non-malignant, clonal, haematological disease and affects 1-5 people per million per year

[**191-192**]. People of any age can be diagnosed with PNH; however, the median age of diagnosis is 30 years old and there is not a skew towards a particular sex or prevalence in particular countries [**193**]. Recent research has suggested that haemolytic PNH reaches a peak between the ages of 30-49 years of age [**194**]. The disease is characterised by complement-mediated chronic intravascular haemolysis resulting in haemolytic anaemia and haemosiderinuria (brown urine caused by chronic intravascular haemolysis). Unpredictable worsening of the condition and recurrent haemoglobinuria is seen with some patients also exhibiting variable degrees of bone marrow failure and susceptibility to thromboembolism. Thrombosis, in fact is the most frequent complication of PNH and can be fatal. Haematopoiesis is impaired in almost all patients and it is unclear as to why exacerbations in the disease occur at night [**195-196**].

1.7.2. Pathophysiology of PNH

1.7.2.1. *PIG-A* mutation in PNH

PNH is an acquired genetic disorder and it is caused by a somatic mutation in haematopoietic stem cells (HSCs) in the phosphatidylinositol glycan class A gene (*PIG-A*). *PIG-A* is a housekeeping gene and encodes an enzyme that is essential (in combination with other proteins) for the transfer of N-acetyl glucosamine to phosphatidyl inositol [**197**]. This is the first step of glycosylphosphatidylinositol (GPI) anchor synthesis and GPI anchored proteins are involved in many important cellular processes, such as membrane trafficking and signal transduction, for example at the immunological synapse [**198**]. The *PIG-A* gene is located on the short arm of the X chromosome and is, therefore, X linked.

PIG-A is the only X linked gene involved in GPI anchor synthesis. Therefore, only a single mutation is needed to impair this gene, even in females, because of X-chromosome functional inactivation [**199**]. One individual can have multiple somatic *PIG-A* mutations. Around 20% of PNH patients are believed to have 2 or more *PIG-A* loss of function mutations, which leads to multiple PNH clones, that can co-exist [**200**].

1.7.2.2 Complement system in PNH

The complement system is part of the innate immune response that aids antibodies' and phagocytic cells' abilities to target and break down microbes and damaged cells, removing them from the immune system. Its functions include induction of phagocytosis, inflammation and damaging the membranes of bacteria [**201**]. The complement response is essential for immune responses. However, if unregulated at some level due to lack of regulatory proteins, like in PNH, it can cause damage to host tissues [**202**].

PNH cells lack GPI associated proteins on their surface and over 25 different proteins are thought to be absent from the surface of red blood cells of PNH patients [**203**]. Two essential proteins absent from these cells are CD59 and CD55, both complement regulatory proteins [**204**]. CD55 controls early complement activation by the inhibition of C5 and C3 convertases, whereas, CD59 interferes with the terminal effector complement which blocks the incorporation of C9 and the C5b-C8 complex and is thought to play an important role in the progression of PNH [**205**].

The absence of the proteins on the surface of PNH red blood cells leaves the cells exposed to complement mediated lysis resulting in intravascular haemolysis [**206**]. These proteins are also deficient on white blood cells; however, nucleated cells have an additional complement inhibitor, CD46, which is not GPI linked. One suggestion as to why PNH patients are more susceptible to thromboembolism is that it is thought to be attributed to the uncontrolled complement activation acting on platelets [**207**].

1.7.2.3. Hypothesis for the cause of PNH clonal expansions

In order for PNH to progress, the mutated HSC clone must expand allowing for differentiation into different blood cell types such as red blood cells and platelets [**208**]. The *PIG-A* mutation does not intrinsically cause this clonal expansion alone. Healthy individuals have been found to have around 0.003% of peripheral blood granulocytes deficient in GPI anchored proteins with somatic *PIG-A* mutations without progression to PNH [**209-211**]. In addition to this, currently no effective PNH model mouse exists as *PIG-A* knockout mice do not show clonal expansion of PNH clones in bone marrows [**212**].

Although the cause of PNH clonal expansion is unknown, the PNH community have two main hypotheses. The first is known as the 'two hit hypothesis', whereby HSCs acquire additional mutations after the *PIG-A* mutation event. These mutations provide survival and growth advantages over the non-PNH stem cells, allowing for the PNH clones to expand [**213**]. Secondary mutation events that are thought to lead to clonal expansion have included studies linking dysregulation of *HMGA2* and separately the *JAK2* gain-of-function mutation in PNH patients [**214,215**]. Another deep sequencing study found that other genes such as *TET2* were mutated in some PNH patients [**216,217**]. Whether these are crucial to PNH pathogenesis or coincidental is another area of current interest. Studies have shown that when multiple PNH clones exist, with different loss of function *PIG-A* mutations, along with secondary mutation events, some of the mutations can initially convey a proliferative growth advantage over others. For instance, mutations in *JAK2* and *PIG-A* had been found to have proliferative advantages initially in PNH patients [**218**].

The second hypothesis involves clonal selection by extrinsic factors, referring to immune system mechanisms [**219**]. Aplastic Anaemia is caused by CD8+ T-cell mediated destruction of the bone marrow and is considered an autoimmune disease [**220**]. It has many associations with PNH [**221**]. One hypothesis for the cause of AA pathogenesis is that the immune system reacts to the virus EBV to try and kill the infection, but the immune system becomes out of balance and can lead to detrimental autoimmune responses attacking the bone marrow attributing to AA pathogenesis [**222**]. Forty to sixty percent of AA patients have a PNH clone which is generally small [**223**]. This has contributed to the thinking that the disrupted bone marrow environment in PNH that leads to the clonal expansion of the PNH clones could be immune mediated . One study suggested that PNH clonal expansions could be caused by autoimmunity towards normal HSCs [**203**].

Another suggestion is that autoreactive cytotoxic lymphocytes target GPI anchored proteins or the GPI anchor itself expressed on normal HSCs. Therefore, PNH clones that lack these proteins avoid the autoimmune attack, 'immune escape', allowing them to homeostatically proliferate and can fill available niches, for instance in the bone marrow or in peripheral blood, out competing normal HSCs [**224**].

1.7.2.4. Aplastic Anaemia and PNH

Research has provided evidence that there is T-cell mediated suppression of haematopoiesis in both bone marrow failure syndromes PNH and AA [**225**]. PNH develops in over 10% of patients with AA, generally as patients are recovering from the disease, supporting the hypothesis that blood cells from PNH patients are more resistant to an autoimmune environment [**226**]. Some studies have suggested as many as 40-70% of AA patients have detectable levels of PNH clones [**227**]. Research by Platanias [**228**] showed that Th17 cells (T helper) were increased in the peripheral blood and bone marrow of AA patients. Kordasti *et al.* (2009) discovered that Th1 and Th2 cells were significantly higher in AA patients and severe AA patients saw significantly lower numbers of Tregs than the normal controls [**229**]. As PNH clones are able to survive and grow in a hypoplastic bone marrow environment, such as in AA, some researchers have suggested that the evolution of the PNH clone may have been nature's way of restoring some form of bone marrow function even in the most adverse autoimmune environments [**230**].

1.7.2.5. Treatment for PNH

The current standard for the treatment for PNH is the monoclonal antibody, Eculizumab. Eculizumab is a terminal complement inhibitor administered to patients every two weeks. It has been considered revolutionary in the treatment of PNH. Treatment with Eculizumab prevents intravascular haemolysis and downstream events, but does not affect the bone marrow failure aspect of PNH [**231**].

1.8. Project Rationale

Reiterating **Section 1.7.**, the majority of patients suffering from PNH have mutations affecting the *PIG-A* gene. However, as mentioned above, research has shown that normal controls also have small numbers of PNH clones present, but these clones do not expand.

Therefore, there must be a selection pressure to expand the mutated clones in PNH. When using a mouse model with PNH HSCs, the clones do not persist and PNH does not progress in the mouse. This led to the idea that it is unlikely that the mutation alone causes the disease and that a combination of events cause the progression of PNH, including a disrupted bone marrow environment. The work outlined in this thesis, therefore, looked to identify a potential cause for the disrupted bone marrow environment, which was thought to possibly be T-cell mediated as in AA.

1.8.1. Role of T-cells in PNH

The links between the lack of GPI associated proteins on the PNH HSCs and the bone marrow failure aspect of PNH are still unclear. A major question in PNH is what factors are allowing for the expansion of GPI-deficient clones within the disrupted bone marrow environment. PNH clones do not seem to have a proliferative advantage to overcome normal haematopoiesis [**232**]. However, their expansion is strongly linked to histocompatibility antigen HLA-DR2 [**233**]. Studies have suggested that there is antigen driven clonal T-cell proliferations that act on haematopoietic stem cells in PNH selectively suppressing normal stem cells when compared with PNH stem cells [**234**].

However, whether the T-cells kill the normal HSCs directly allowing for PNH clonal expansion is only hypothesised. Karadimitris *et al.* found that the frequency of individuals with skewed repertoires was higher in PNH than in the control groups. They found that the TCRBV families 4,7,23,24, had T-cell clonal expansions occurring at a higher frequency in PNH patients than normal controls, with TCRBV7 being the most frequently skewed [**235**]. Gargiulo *et al.* identified a novel, auto reactive population of T-cells that were more abundant in PNH patients [**236**]. However, on further inspection of the data, the invariant TCRVα21Jα31-1, was only found in 1 out of 3 PNH patients' mRNA, highlighting that not all TCRB responses are the same in all PNH patients. These CD8+ T-cells were CD1d restricted and GPI specific, reactive against antigen presenting cells that were loaded with GPI, suggesting the possibility that T-cells target GPI in normal HSCs in PNH. Studies mentioned in this section suggest long term persistence of T-cell clonal expansions could lead to T-cell mediated mechanisms that underlie the pathogenesis of PNH. However, at present no specific T-cell clone has been identified in the pathogenesis of PNH. The majority of T-cell repertoire studies in PNH used older techniques, such as spectra-typing, 454-sequencing and flow cytometry which provide valuable insight but have limitations. For instance, flow cytometry covers around half of the TCRBV families and does not assess TCRBJ gene families. Another limitation of the majority of PNH TCR studies is PNH patient numbers due to the rarity of the disease.

1.9. Project aims and objectives

1.9.1. Project aims

The hypothesis was that there was a single or series of exclusive TCRBs shared in the repertoires of PNH patients . The overall aim of the work in this thesis was to identify these TCRBs, if present, and to provide more evidence to help clarify the role of T-cells in PNH progression by analysing TCRB dynamics in PNH patients. This was achieved by developing experimental TCRB repertoire sequencing techniques and designing a bioinformatics workflow for analysing and interpreting over 200 million sequencing reads to establish accurate TCRB repertoires.

These techniques were applied to PNH, AA and healthy control samples of peripheral blood and bone marrow to analyse the TCRB repertoires of 77PNH and AA patients. Thirty-one normal TCRB repertoires were sequenced alongside this and analysed to gather background information on clonotypes that appear in '"public" repertoires, to help identify truly PNH specific T-cells if present. This allowed an investigation into whether there are TCRB clones that occur as clonal populations specifically in patients with PNH. Alternatively, whether there are TCRB clones in both healthy controls and PNH that are more prevalent or less prevalent in PNH patients was investigated. Equally, by studying TCRB repertoires in AA patients both with and without PNH, whether there are differences in AA patient repertoires depending on or attributing to PNH development could be assessed.

1.9.2. Objectives of the work carried out in this thesis

The work outlined in this thesis is split across four results chapters. Much of the work in this thesis involved the development of new experimental and bioinformatic pipelines. Rationale for the use of methods and types of analysis along with extensive optimisation steps, will be discussed in **Chapters 2-6**. This thesis can be broken down into six main objectives outlined below:

- Analyse 454 sequencing data from 18 PNH patients and 10 controls as pilot data to assess whether TCRB repertoires differ in PNH patients. The objective was to establish a reason for developing high throughput sequencing technologies for analysing PNH TCRB repertoires (**Chapter 3**).

- Identify a cohort of patient samples from the PNH Research Tissue Bank for TCRB repertoire sequencing and analysis, including time course studies, spontaneous remission patients, paired bone marrow and peripheral blood samples, and AA patients (**Chapter 2**).

- Develop, optimise (**Chapter 2**) and implement TCRB repertoire experimental techniques, including library preparation, for sequencing TCRB from peripheral blood and bone marrow samples (**Chapter 4**).

- Establish a normal background variation of TCRB repertoires from healthy controls to act as a comparison for PNH TCRB repertoires for assessing whether or not observed differences in TCRBs are PNH specific (**Chapter 4**).

- Design, optimise (**Chapter 2**) and implement a bioinformatics workflow to efficiently process the large amounts of data produced by Illumina® MiSeq sequencing and implement insightful analysis of TCRB repertoire data (**Chapters 4, 5,6**).

- Implement methods to analyse and compare a large number of PNH and AA samples from peripheral blood and bone marrow, as well as analysis of clonotype tracking in longitudinal samples to investigate changes in TCR repertoires according to PNH status over time (**Chapter 5 and 6**).

Chapter 2 - Materials and Methods

A large proportion of this project involved the development, optimisation and implementation of experimental and bioinformatics methods and analysis. Along with the finalised pipelines used to generate the data, rationale and thought processes behind choices of methods and analysis whilst developing the pipelines will be discussed in this section.

2.1. Patient and healthy control sample categorisation and selection

All healthy controls and patient samples were selected from the Leeds PNH Research Tissue Bank (RTB) in accordance with their ethical guidelines. The healthy controls' metadata included their age and sex. Patient samples included age, sex and their Paroxysmal Nocturnal Haemoglobinuria (PNH) or Aplastic Anaemia (AA) status. AA patients were selected as T-cells are known to be involved in the pathogenesis of the disease and there is a link between PNH and AA.

2.1.1. Healthy control selection

Thirty-one healthy controls, defined as not being diagnosed with PNH or AA at the time of sampling, were selected from the Leeds PNH RTB. The ages ranged from 22 to 57 years old with an average age of 37 and median of 35 years of age. The ratio of females:males was 2.4:1. Where possible healthy controls were selected as matches for age groups of the patient samples. Healthy controls were broken down into the age brackets depicted in **Table 2**.

2.1.2. Patient categorisation

Patients were categorised by Dr S. Richards in the Division of Haematology and Immunology, who had extensive diagnostic experience as part of the Leeds PNH team, according to the following parameters: PNH status, AA status, size and status of PNH clone (**Table 3.**). All patients were either diagnosed with PNH or AA. Clone size refers to the size of the PNH peripheral blood granulocyte clone and was identified using clinical flow cytometry methods [**194,237**]. Seventy-seven patients from the RTB had at least one TCRB repertoire sample successfully sequenced and analysed. Three patients had bone marrow and peripheral blood matched samples sequenced. Two patients had undergone spontaneous remission from PNH prior to the sample being taken for sequencing. Ages of the patients ranged from 19 to 85 years old.

The average age was 46 and the median was 43 years of age. The ratio of females to males was 1:1.3. The aim of the patient selection was to have good representation of all the major categories, for example, PNH patients with a large stable clone, but also as many as possible of the more obscure groups such as PNH patients who were recovering.

Table 2. Age-range of 31 healthy controls and 77 PNH and AA patient samples from the Leeds PNH RTB. TCRB repertoires were successfully sequenced and used in the TCRB repertoire analysis.

Age-range	Number of normals	Number of PNH or AA patient samples
18-24	1	9
25-29	6	8
30-34	8	4
35-39	4	13
40-44	4	9
45-49	4	4
50-54	2	3
55-59	2	5
60-64	-	4
65-69	-	8
70-74	-	5
75-80	-	3
80-84	-	1
85-89	-	1

2.1.2.1. Longitudinal TCRB repertoire studies

In order to assess variability of TCRB repertoires over short periods of time, 4 patients had samples sequenced and analysed from multiple time points taken over the short-term (up to two years apart). In order to assess changes in TCRB repertoires over longer periods of time, 9 patients were selected from a 2013 trial for TCRB sequencing and these were compared with samples taken over 4 years later.

Table 3. Categorisation of 77 PNH and AA patients selected from the Leeds PNH RTB. Patient samples were categorised according to PNH status, AA status and size and status of the PNH clone. Samples were selected from the Leeds PNH Research Tissue Bank.

Patient category	Number of patients samples sequenced
PNH large clone (>25%), stable disease	17
PNH decreasing clone	6
PNH new clone/increasing clone	17
Aplastic anaemia; no PNH clone	6
Aplastic anaemia; small PNH clone (< 25%)	19
Aplastic anaemia; increasing clone	7
Aplastic anaemia; large clone (>25%)	1
Other (interesting cases to be discussed on an individual basis – **Chapter 6**)	4

2.2. Sample preparation

2.2.1. Mononuclear cell extraction from peripheral blood

Peripheral blood was extracted from patients and healthy controls according to the guidelines and ethics outlined by the Leeds PNH RTB. Mononuclear cells (MNCs) were extracted from the peripheral blood using Lymphoprep™(Axis-Shield) (according to the manufacturer's guidelines) by creating a density gradient. The MNCs were counted using a microscope and haemocytometer and then aliquoted into 1 x 10^6 cells per aliquot.

2.2.2. Flow cytometry

Flow cytometry was performed on an aliquot of the fresh 1 x 10^6 mononuclear cells using a Cytoflex (Beckman Coulter) flow cytometer. To prepare the cells for flow cytometry, 200µl of the cell suspension followed by 2ml of MACS buffer (Miltenyi Biotec) was put into each flow tube (5 in total) and the samples were centrifuged at 1200rpm for 3 minutes. The supernatant was removed, discarded and the cells were re-suspended in the residual volume. One tube was unstained (no antibody) and the remaining tubes were stained with 10µl of the antibodies CD3-FITC, CD15-FITC, CD11b-FITC, and CD19-PE respectively.

The samples were kept at 4°C for 20 minutes and were then re-suspended in 2ml MACS buffer, centrifuged at 1200rpm for 3 minutes and the supernatant was removed and discarded.

0.5ml of MACS buffer was added to each sample and these were stored in the dark at 4°C before being analysed on the Cytoflex within two hours. Flow cytometry data analysis was performed using CytExpert(1.2.11.0). CD3 marker percentages were used to assess the percentage of T-cells present in each sample for use in **Section 2.9.2.2**. In general, **Section 2.2.** was the standardised method used for samples supplied by the Leeds PNH RTB, supplying samples of MNCs ready for preparation of cDNA or gDNA. However, on occasion, a buffy coat method was used by the RTB. As a result, these samples contained fewer T-cells than the standardised methods. This has been taken into account for the calculations in **Section 2.9.2.1**.

2.3. Preparation of genetic material

Optimisation of library sequencing methods was performed using both cDNA and gDNA as genetic material input. This was before gDNA was chosen as the preferred genetic material for repertoire sequencing methods of PNH patients and the healthy controls (**Section 4.3.2**) .

2.3.1. RNA extraction

Aliquots of the fresh MNCs were centrifuged for five minutes at 700g. The supernatant was removed and 1ml of TRIzol® reagent was added to the pellet containing the MNCs. RNA was then extracted from the samples using the Direct-zol ™ RNA Kit (Zymo Research) according to the manufacturer's protocol using spin column technology.

2.3.2. DNase treatment

To remove any gDNA contamination in the RNA samples before cDNA synthesis, the Ambion® DNA-free™ DNase Treatment and Removal Reagents (Life Technologies) was used to treat the RNA samples according to the manufacturer's protocol. Firstly, a DNase enzyme was added to the RNA sample to remove any contaminating DNA or reduce it to levels that are not detectable by PCR. The "DNase inactivation reagent" was then used to remove the DNase enzyme along with any divalent cations including magnesium and calcium which could accelerate RNA degradation. The quality and yield of the RNA samples were then measured using the NanoDrop (ThermoFisher Scientific) which calculated absorbance levels. "Pure" RNA had a 260/280 ratio of 2.0 and a 260/230 ratio between 2.0-2.2.

2.3.3. cDNA synthesis

The methods for cDNA synthesis were used in optimisation methods to assess whether a targeted TCRβ chain cDNA synthesis approach produced better TCRB repertoire data than a general cDNA synthesising methodology. Reverse transcription was performed using the ImProm-II™ Reverse Transcription System (Promega) as per the manufacturer's protocol. In brief, the reverse transcription primer (oligo(dT)((0.5µg/reaction) or gene-specific primer (10–20pmol/reaction)) was added to the RNA sample (up to 1 µg) on ice and made up to 5µl using nuclease free water. The mix was then heated to 70^{O}C for at least five minutes and then spun down in a microcentrifuge for 10 seconds to ensure any condensation had returned to the mixture. It was then kept on ice. A master mix kept on ice containing: ImProm-II™ 5X Reaction Buffer, $MgCl_2$ (final concentration 1.5–8.0mM), dNTP Mix (final concentration 0.5mM each dNTP), Recombinant RNasin® Ribonuclease Inhibitor (optional) and 20u ImProm-II™ Reverse Transcriptase was made up to a final volume of 15µl per reaction with nuclease free water. This was vortexed gently and kept on ice before dispensing into PCR tubes on ice (15µl master mix per cDNA reaction). 5µl of the RNA primer mix was added to each 15µl of master mix for a final volume of 20µl. The reaction mix then underwent PCR to synthesise DNA. Firstly, the PCR cycled through the annealing stage, whereby the mixture was incubated at 25°C for five minutes. Then the mixture underwent the extension stage where it was incubated at 42°C for up to one hour. As the cDNA would be used in future PCR reactions an additional step was needed to thermally inactivate the reverse transcriptase. The mixture was incubated at 70°C for 15 minutes and then the cDNA could be used in PCR reactions or stored at -80°C until needed. The only modification to the protocol was the addition of a TCRβ chain C region primer (CTCAGCTCCAGTG) [**396**] for half the samples instead of a standard oligo T primer (Sigma) to carry out a TCRB region specific cDNA synthesis.

2.3.4. Genomic DNA extraction

The QIAamp DNA Mini Kit (Quiagen) was used according to the manufacturer's instructions to extract gDNA from an aliquot of fresh MNCs. Genomic DNA (gDNA) was extracted using spin-column procedures. In brief, the cell sample was incubated with a protease and a buffer containing surfactants or detergents, vortexed for 15 seconds and then incubated for 10 minutes at 56 ^{O}C. This was the length of time needed for cell lysis to be at its maximum. Physical destruction of the cell wall occurred when vortexing the mixture which allowed the nucleic DNA to be free in solution. The buffers then removed membrane lipids and the proteases broke down enzymes and other protein contaminants into amino acids which meant that they could no longer interact with or destroy the nucleic DNA.

The subsequent buffer washes and spin column methodology was used to purify and precipitate the DNA for use in reactions such as PCR. The sample's DNA quality and yield were then measured using the NanoDrop (ThermoFisher Scientific) which calculated absorbance levels. "Pure" DNA had a 260/280 ratio of 1.8 and a 260/230 ratio between 2.0-2.2.

2.4. Sequencing Library preparation

In order to generate TCRB repertoire sequencing data from gDNA or cDNA samples, TCRB library preparation methods were developed. Two TCRB sequencing library preparation methods were developed as part of this project to allow accurate comparison and validation of results, the *Robins et al. primer method* [**238**] and the *BIOMED-2 primer method* [**154**].

Development of a sequencing library method can be broken down into 8 stages. These are outlined below in bold alongside important developmental considerations, such as optimisation steps.

Genetic material selection	**The right genetic material must be selected dependent on the biological question the experiments need to answer.** Selection of genetic input into PCR reactions is needed. This could be gDNA or cDNA. The quality and concentration of the genetic material needs to be evaluated to make the subsequent TCRB data both accurate and the experiments viable.
TCRB amplification PCR	**The first PCR reaction uses primers that specifically amplify TCRB genes from genetic material.** Primer selection and optimisation experiments are needed to amplify the most TCRB product from the genetic material. This includes PCR cycle number and PCR condition optimisation, such as Mg^{2+} concentration and annealing temperatures.
PCR clean up steps	**Removal of PCR contaminants including reaction buffers, primer dimers and non-specific binding PCR products.** Selection and optimisation of clean-up methods including gel electrophoresis, bead and column size selection is needed. Removal of lower molecular weight products than the desired TCRB PCR product is essential to ensure non-biased sequencing of the TCRB library.

Indexing PCR	**The index PCR reaction adds sequencing adapters and identification index codes to each end of the TCRB product. This allows the TCRB product to bind to flow cells during sequencing and the sequenced pooled library (multiple samples) to be de-multiplexed back into individual samples.** Optimising PCR conditions to yield the most TCRB-adapter product are essential. This includes optimising adapter concentration and PCR cycle number.
Product selection	**After the indexing reaction, the TCRB-adapter PCR product is selected out of the mix. This can be using similar methods to the PCR clean up.** Select a method that is effective and accurate at selecting the product but does not lose too much product in the process.
DNA quality check	**Quality of the DNA is checked to ensure that no lower molecular weight contaminants such as primer dimers are present in the final library samples.** Assess which products present in each sample are the desired product and what are contaminants. Check library samples only have the TCRB-adapter product present.
DNA quantification	**When pooling TCRB products from multiple samples into one library, equimolar amounts of each sample must be added.** Decide what concentration of samples are needed for a good quality library and to prevent sample skewing or variation in read coverage and depth.
Sequencing	**Sequencing of the TCRB sequencing library is carried out to generate data for bioinformatic analysis.** It is important to decide before running experiments which sequencing library method is to be used and in the case of Illumina®, what cycle length is necessary to successfully sequence the product size. A decision needs to be made whether to use paired or single - end sequencing technologies.

2.5. *Robins et al. primer method's* optimisation steps and workflow

The first sequencing library preparation method was based around the use of 45 forward TCRBV gene primers and 13 reverse TCRBV J gene primers from a paper published by Robins *et al.* (2009) [**238**] and will be referred to as the *Robins et al. primer method*. Outlined below are the optimisation steps and methods used to develop this methodology.

2.5.1. Computationally validating TCR beta chain primers

The 45 forward primers for TCR beta V gene segments and the 13 reverse primers for the TCR beta J gene segments were validated using NCBI Gene [**239**] and UCSC *In Silico* PCR [**240**].

2.5.2. Adapting TCR beta chain primers for MiSeq sequencing

In order to make sure that the primers were compatible with the Illumina® MiSeq sequencing platform and Illumina® Nextera Index kit, the following sequences were added to the beginning of the V gene primers, and the J gene primers:

V **TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**

J **GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**

This allowed the PCR products to be recognised in the second-round indexing reaction. The Nextera codes are identical in sequence and position in the adapted primer sequence to those in the *BIOMED-2 primer method* (**Tables 6 and 7**).

2.5.3. Experimentally validating and optimising TCR beta chain primers

It was not feasible to test all combinations of the 45V and 13J primers on genetic material individually, therefore the two most common combinations [**241**] of V/J genes, TRBV20-1/TRBJ2-1 and TRBV5-1/TRBJ1-1, were used to test whether the primers worked on gDNA and cDNA. They were also used for the optimisation of the Mg ion concentration and annealing temperatures in the multiplex PCR to ensure the primers effectively amplified the TCRB CDR3 region. The primer annealing temperature was tested at 61^{o}C, 63^{o}C, 65^{o}C and 67^{o}C. The Mg ion concentration was varied between 1.5 mM, 2.5mM and 3.5mM.

0.2µl of both the V primer and J primer were added to a 20µl PCR reaction, with 10µl of the Phusion Flash High-Fidelity Master Mix (ThermoFisher Scientific), up to 2µl of cDNA or gDNA (equivalent to 100ng per PCR reaction) and nuclease free water. The final concentration of the primers in the PCR were 22nM each.

The Phusion Flash Master Mix contained a Pfu based proofreading polymerase, Phusion Flash II DNA polymerase, along with all necessary components for a PCR minus the primers and DNA template. The master mix had a Mg ion concentration of 1.5mM. To make samples up to a Mg ion concentration of 2.5mM and 3.5mM, magnesium chloride was added to the PCR reactions. The thermal cycling conditions used are detailed in **Table 4.** A positive control using the housekeeping gene *GAPDH* was used each time as it amplifies readily and was a good indicator if there was an issue with the PCR reaction. 1µl each of forward and reverse *GAPDH* amplifying primers (10µm stock) amplified gDNA and cDNA in a parallel PCR reaction. A negative control was also used each time using nuclease free water in place of any genetic material to show that the primers were not contaminated. PCR products were then mixed with a 6x loading buffer and run on a 2% agarose gel (1xTBE) containing GelGreen® dye at 70mA and imaged using a GelDoc.

Table 4. PCR conditions used in optimisation PCR reactions for the *Robins et al. primer method*.

The below PCR conditions were used for optimising the TCRB amplifying PCR reaction in the *Robins et al. primer method*. Annealing temperatures were optimised between 61 -67°C and the best annealing temperature was found to be 65°C and used for subsequent library preparation.

1 cycle	**30-35 cycles**			**1 cycle**
Denaturation		*Annealing*	*Extension*	
98°C	98°C	65°C (61 -67°C)	72°C	72°C
10 seconds	1 seconds	5 seconds	10 seconds	60 seconds

2.6. Library preparation workflow for the *Robins et al. primer method*

The *Robins et al. primer method* was used to generate a number of TCRB libraries for optimisation (**Table 5.**) and analysis of healthy controls.

2.6.1. Amplification of TCR beta chains using multiplex PCR

1µm stock solutions of the pooled 45 TCRB V gene primers and the 13 TCRB J gene primers were kept separately to prevent cross reaction. Each PCR reaction was 20µl in volume. 10µl of the Phusion Flash High-Fidelity Master Mix (ThermoFisher Scientific) and 0.2µl each of the V and J primer pools were added to each reaction. Each primer would be present at 22nm making the final concentrations of V and J primer in each PCR reaction 1.0µM. The volume of genetic material was dependent on the concentration required and the reaction was made up to 20µl using nuclease free water. A PCR reaction using 1µl each of forward and reverse *GAPDH* amplifying primers (10µm stock) was used as a control for the amplification reactions and a control containing no genetic material was used to ensure the primers were not contaminated. The PCR samples were run as stated in **Table 4**. using 1.5mM as the optimum Mg ion concentration and 65^{o}C for the annealing temperature.

2.6.2. Cleaning up library components

The samples were cleaned up using AMPure XP purification beads according to the Illumina® Nextera "clean up libraries" protocol which was sufficient to remove any lower base contaminants that could affect sequencing efficacy.

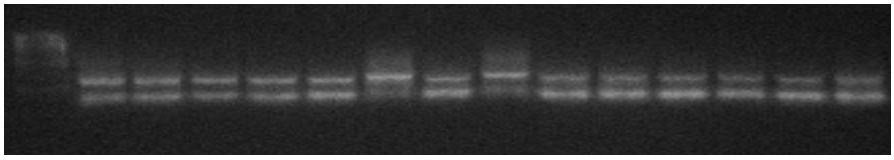


Figure 8. *Robins et al. primer method* amplified a TCRB product of approximately 230bp.

The first round TCRB amplification PCR reaction using the *Robins et al. primers* amplified a TCRB product of approximately 230bp. Fourteen cDNA and gDNA samples were amplified alongside the Hyperladder™ 100bp (Bioline) ladder for comparison. The lower band was caused by non-specific binding events producing products that partly amplified or were primer dimers. PCR clean-up was necessary to remove these bands before the indexing PCR reaction.

2.6.3. DNA extraction and second round PCR for Illumina® Nextera adapter tagging

PCR products were then run on a 2% agarose gel (1xTBE) at 70mA using GelGreen® dye and imaged using a GelDoc. Gel bands at around 230 bp were cut from the gel and the DNA was purified from the agarose using the Zymoclean™ Gel DNA Recovery Kit (as per the manufacturer's protocol, discussed in greater detail in **Section 2.8.5.**). 6µl of the purified DNA and 10µl of Phusion Flash High-Fidelity Master Mix were added to each PCR reaction along with 2µl of Nextera N index, and 2µl of Nextera S index from the Nextera XT Indexing kit. A different combination of N and S indexes were used for each sample. The two step PCR conditions below were used to eliminate lower sequencing bands from the mixture. The PCR cycling conditions, for each cycle, was 98°C for 1 second followed by 72°C for 10 seconds, for a total of 10 cycles. PCR products were then run on a 2% agarose gel (1xTBE) at 70mA using GelGreen® dye and imaged using a GelDoc. Gel bands at around 250 bp were cut from the gel and the DNA was purified from the agarose using the Zymoclean™ Gel DNA Recovery Kit (as per the manufacturer's protocol, discussed in greater detail in **Section 2.8.5.**) finally eluting the DNA product into 15 µl of nuclease free water.

2.6.4. Determining library component concentrations

Samples were analysed using the AgiLent Tapestation to assess the molecular weight of peaks that appeared in the samples. A peak would be expected at around 250bp (**Figure 9.**). The "Invitrogen Quant-IT broad range dsDNA assay kit" was used to determine the library component concentrations using PicoGreen and the data from the fluorimeter was analysed with the software *Fluostar*. From the data, the volume of each sample in the library that needed to be added to the final sequencing library was identified and combined to form a final library mix.

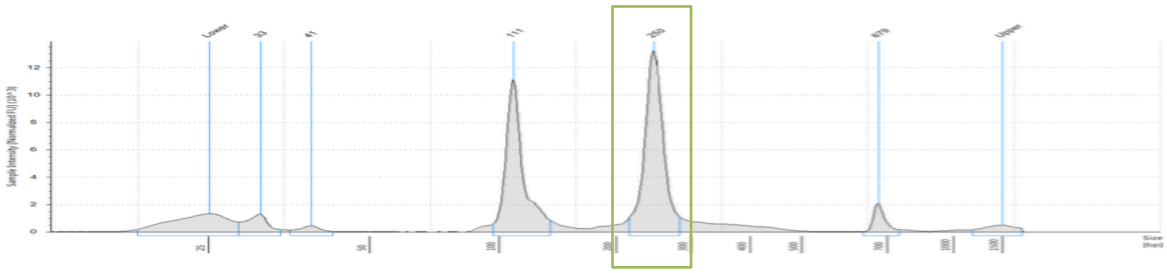


Figure 9. Molecular weight profile of a gDNA sample using the *Robins et al. primer method*. Following the indexing step, TCRB amplicon reaction mixtures that were generated using the *Robins et al. primer method* showed molecular profiles as above. Running the samples on an agarose gel allowed the separation of the two larger peaks. The peak at 250bp (green box) was the TCRB amplicon final library product and was cut from the gel, and purified ready for sequencing. This process removed the 111bp peak attributed to primer dimerization from the final product.

2.6.5. Sequencing

The library was sent to the University of Leeds Next Generation Sequencing Facility to be sequenced on the Illumina® MiSeq platform. Each sequencing was 300bp in length and paired-end. The sequencing data was then de-multiplexed according to the Nextera codes ready for analysis.

Table 5. The first sequencing library performed using the *Robins et al. primer method*.

Design of the first MiSeq sequencing run using the *Robins et al. primer method*. Samples were taken from three normals and were cDNA and gDNA. Two types of cDNA, one using the standard oligo T primer in reverse transcription and one using a TCRB constant chain specific primer were used to compare a targeted approach versus a general approach. The amount of genetic material refers to the amount of genetic material inputted into the first multiplex PCR reaction to amplify the TCRB CDR3 region which was varied to assess TCRB diversity observed with varying amounts of genetic material. Sample replicates were important when designing libraries to allow the effects of sequencing and PCR errors to be assessed on the TCRB repertoire accuracy and diversity observed. An additional sample from an inflammatory condition in which T-cells are thought to have a role, but were few in number, was used as a control to see how well the primers amplified T-cells at lower concentrations.

Sample number	Type of genetic material as input	Amount of genetic material ng	Number of replicates to be included in the library	Additional Notes
0001	gDNA	100	4	0001 was a sample from a normal donor. One sample was used for sequencing without a second gel purification to assess whether this is an essential step
0001	gDNA	100	1	A repeat of one of the samples already used was to test gDNA at 100ng as a check for sequencing errors
	gDNA	200	3	
	gDNA	400	3	
0001	cDNA using oligo T primer in reverse transcription	100	3	
	cDNA using oligo T primer in reverse transcription	200	3	
	cDNA using oligo T primer in reverse transcription	400	3	
0001	cDNA using C region primer in reverse transcription	100	3	C region primer used to assess whether performing reverse transcription specifically around the TCR beta chain, before amplifying using the TCR beta specific primers had an effect on T cell receptor repertoire sequencing

	cDNA using C region primer in reverse transcription	200	3	
	cDNA using C region primer in reverse transcription	400	3	
0002	gDNA	100	1	A sample prepared from a different normal donor to assess repertoire diversity between two individuals
0003	gDNA	400	1	An inflammatory disorder sample thought to have few T-cells acted as a good test for amplifying T-cells at low levels

2.7. *BIOMED-2 primer method's* optimisation steps and workflow

In order to evaluate the TCRB sequencing results generated using the *Robins et al. primer method* it was important to develop a method using alternative primer sets. For this method, primers from the BIOMED-2 paper [**154**] were used. This method will be referred to as the *BIOMED-2 primer method*.

2.7.1. Evaluating the optimal concentration of genetic material

TCRB diversity observed in a sample was expected to vary with concentration of genetic material inputted in the first TCRB amplifying PCR reaction. The higher the concentration of the genetic material, the more TCRB clones detected in a sample and thus the diversity increases. Too low a concentration could result in minimal detection of rarer clones. However, when dealing with clinical samples, many may not have a high DNA concentration. It is important to keep the concentrations of the genetic material as similar as possible between samples, to allow for accurate comparisons of TCRB repertoires. To assess the optimal and realistic concentration of genetic material to be used, 100ng, 200ng and 400ng of DNA per 20 µl were trialled as input for the first PCR reaction using the same sample. The workflow used was the standard BIOMED-2 optimised primer method (**Section 2.8.**). These tests were replicated at least three times to account for natural variation. The optimal concentration was found to be 200ng, the results are discussed in **Chapter 4**.

2.7.2. PCR optimisation for the *BIOMED-2 primer method*

Although the *BIOMED-2 primer* sets themselves had been heavily tested by the Euro Clonality Consortium [**379**] and are routinely used in NHS hospitals for diagnostics, a number of optimisation steps were needed to successfully create TCRB sequencing libraries using these primers.

Three factors were important when optimising PCR conditions for TCRB sequencing. The first, Mg^{2+} concentration was important for ensuring specific binding of TCRB primers to the gDNA to amplify the TCRB product. Mg ions act as catalysts for the DNA polymerases allowing efficient DNA replication, however, too high a concentration can lead to a drop in the polymerase's specificity and more non-specific binding. DNA strands may also not denature at high temperatures if the Mg^{2+} concentration is too high. Too low a concentration could lead to the primer failing to anneal to template DNA and lower efficiency of the polymerase enzyme. Secondly, annealing temperatures are the stage in a PCR where primers bind to opposite strands of denatured dsDNA to allow replication. The annealing temperature is primer specific. Too low a temperature will see non-specific products and too high will result in primers being unable to bind. The final factor was the number of PCR cycles. Increasing the number of cycles in a PCR will increase the PCR product yield to a point where it reaches saturation. Each PCR cycle will allow for PCR errors and bias to be incorporated into the PCR product each time which could cause inaccuracies. It is, therefore, important to find the right balance between PCR product yield and minimal biases.

In order to optimise these factors, test PCRs were set up using the PCR cycling conditions and reagents as stated in **Table 9 and 10**. Cycle numbers were trialled at the standard 35 cycles and then 50 cycles, which is the maximum at which saturation occurs in product yield. The optimised conditions were 3.5mM and an annealing temperature of 61^{o}C for the primers with UMIs added (**Figure 10.**) and 63^{o}C for non-UMI primers and a first PCR of 35 cycles.

For the index adapter reaction, an additional factor was optimised, the sequencing adapter concentration. Too high a concentration would lead to non-specific binding events and too low would mean less efficient addition of the adapters to the PCR product and then less of the product would bind successfully to the flow cell for sequencing. In order that the TCRB amplicons can be sequenced by the MiSeq, the ends of the sequence must be complementary to bind to the sequencing flow cell. In order for this to happen, adapter sequences must be added to the TCRB amplicons. The adapters contain N and S Nextera index sequences which bind to the start and end of the sequence respectively. Each sample going into a library has a different N and S index combination. The reads can then be demultiplexed after sequencing back into the original sample sets using the index codes. The adapter concentration was optimised to 2µl of each N and S index per index PCR reaction (detailed in full in **Section 2.8.4**).

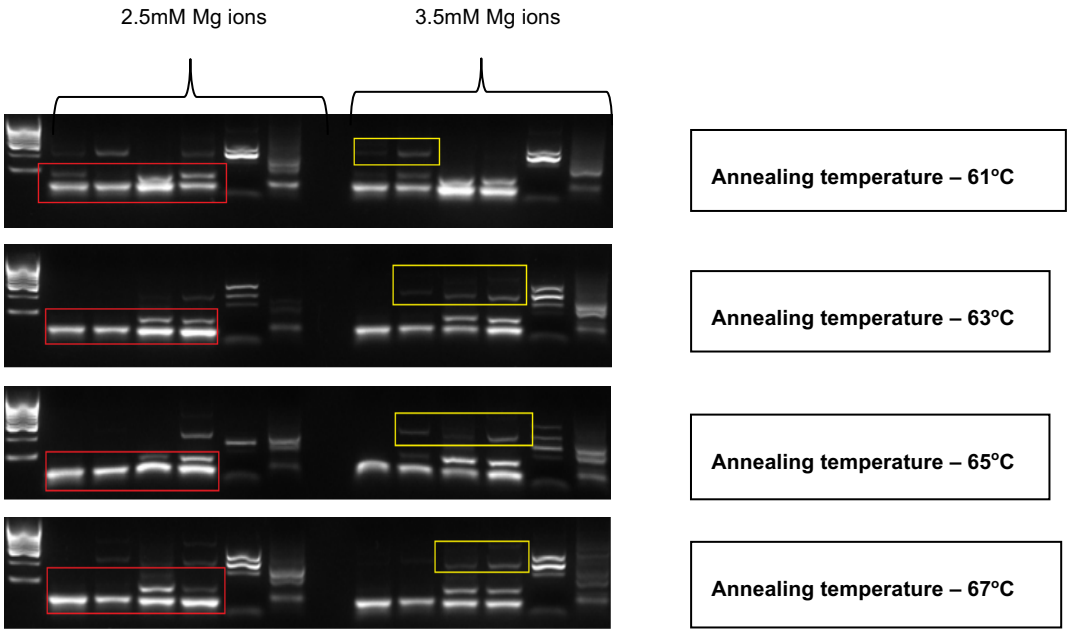


Figure 10. Optimisation of first round PCR reactions for the *BIOMED-2 primer method*.

Mg^{2+} ion concentration in the PCR was optimised at 2.5mM (left side) and 3.5mM (right hand side) in the PCR reactions. Each set of reactions were optimised for an annealing temperature between 61°C and 67°C. PCR reactions were run on a 2% agarose gel, 14 wells per row. The first well was the HyperLadder™ 100bp (Bioline), followed by UMI-JA, UMI-JB, Non-UMI-JA, Non-UMI-JB, GAPDH, the same sample amplified using the "Robins et al." primer method and then a blank. The same order was used for the next 7 wells using the higher 3.5mM Mg ion concentration. The same gDNA sample was used for all the PCRs at 200ng per PCR. The red boxes indicate regions where non-specific binding and primer dimers were expected to be present. The yellow box indicates the region where the desired TCRB product band should be present. The brighter the band the higher the yield. There will always be non-specific products and primer dimer present in the mixture, the aim of the optimisation was to get the brightest TCRB band possible with lower intensity bands for the non-specific products.

2.7.3. PCR clean up optimisation workflow for the *BIOMED-2 primer method*

During library preparation it is important to make sure that the sample is pure and only contains the TCRB amplicons for sequencing. No lower molecular weight contaminants can be present in the TCRB library preparation. This is because NGS (next-generation sequencing) efficiency is size determined, small products will exponentially fill all of the sequencing space. This is why the size selection processes in TCRB library preparation are of high importance.

As stated in **Section 2.5.** the *Robins et al. primer method* used gel selection and AxyPrep beads after every PCR reaction to remove and prevent any products formed from non-specific primer binding or primer dimerisation being included in the final TCRB amplicon library. This method was initially used for the *BIOMED-2 primer method*, where the JA and JB primer reactions (explained in detail in **Section 2.8.**) would be run in separate lanes, and then the bands cut out and purified together for DNA extraction. However, as the samples were run separately this could introduce initial experimental bias. Therefore, a new selection method using beads was optimised and used in its place.

After the first PCR reaction, the JA and JB 20µl reactions were vortexed (to return any mixture back to the PCR that had condensed with heating) and combined into one 40µl reaction mixture. Promega Pronex®beads were used to perform the selection process according to the manufacturer's protocol. The first bead selection used "protocol 6A" from the manufacturer, using 1.5x the volume of beads to the volume of sample (40µl). This step removed most contaminants below 250bp such as primer dimers.

The second round of bead selection used a dual size selection method optimised to select a product peak of approximately 325bp, which is complementary to a standard curve for the TCRB amplicon size at this stage. A 1.1.x bead selection volume was used to remove smaller base pair products such as primer dimers and contaminants such as buffers that might have still been present after the first selection. Smaller contaminants than the TCRB amplicon were bound to the beads in the mixture. The tubes of mixture were then placed onto a magnet causing these bead bound contaminants to bind to the magnet. This allowed the supernatant containing the TCRB amplicons to be moved into a new tube. The second bead selection used a lower bead concentration of 0.35x this time allowing the TCRB amplicons to bind to the beads. The supernatant containing remaining contaminants was removed, and the purified TCRB mixture underwent washing steps and was then eluted in a volume of 20µl of elution buffer for downstream processes.

Initially, the dual selection process was trialled without the need for the first 1.5x step, however the selection around the peak of TCRB amplicons was not accurate enough, with much of the lower base pair contaminants still present in the sample (**Figure 11.**). Therefore, the first 1.5x step was necessary to remove the lower base pair contaminants to a level to allow successful dual bead size selection to occur in the subsequent step.

Without this, the dual selection was ineffective in selecting for the 325bp peak (**Figure 11.**). This meant that some TCRB product was lost, but overall, this method was the most accurate for TCRB library preparation.

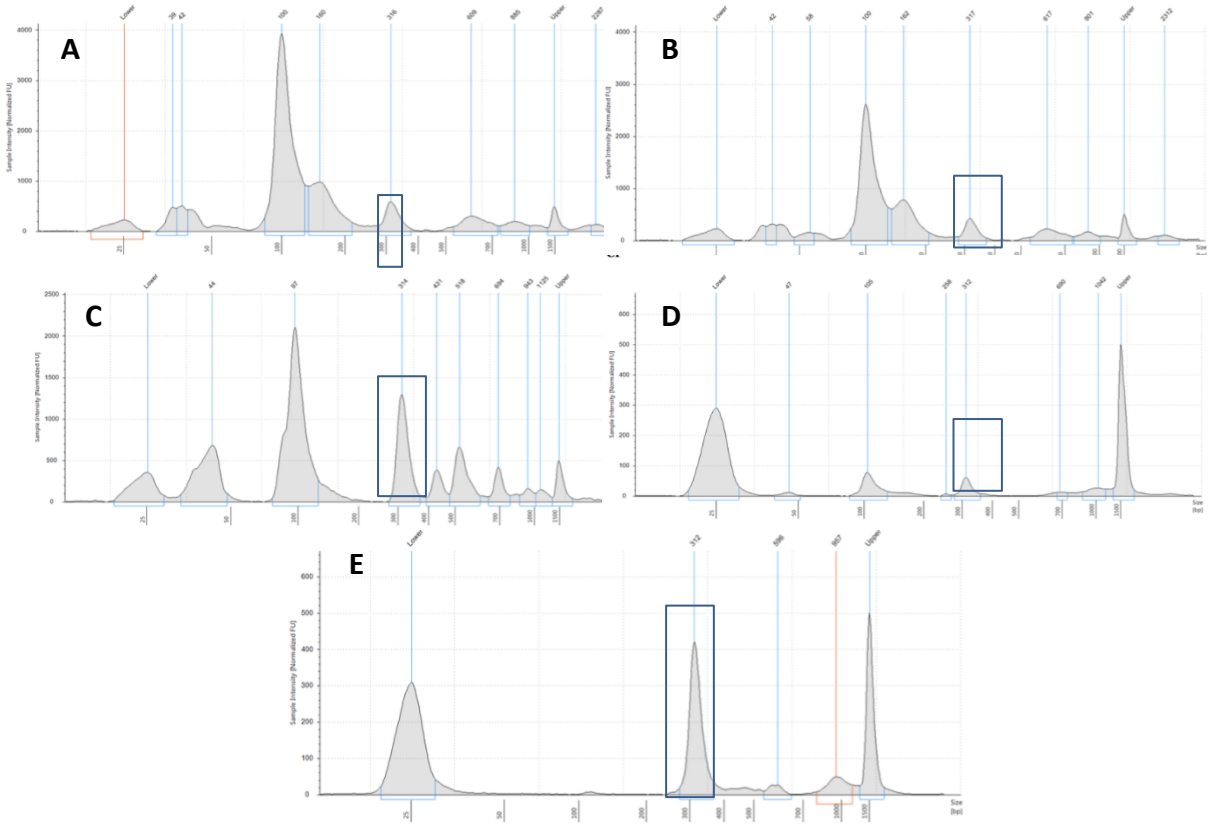


Figure 11. Molecular weight profiles of TCRB amplicons after the first PCR reaction.

The *BIOMED-2 primer method* produced a TCR product peak at 316/317 bp after the first TCRB amplifying PCR reaction when using gDNA from a specific sample (**A**). **A** and **C** show the profile before a size selection clean up peak, showing large amounts of contamination indicated by peaks above and below 316-317bp. After using a 1.5x Promega Pronex bead size selection, the primer dimer contamination peak at 100bp had reduced in size by about a quarter (**B**). However, there was still a considerable amount of contamination that needed to be removed. The same gDNA underwent a separate first round PCR TCRB amplifying reaction producing a peak at 314bp (**C**). This sample had more TCRB product and less primer contamination but still significant levels. The sample was cleaned up using Promega Pronex beads and only using the 1.1x/0.35x dual size selection discussed in **Section 2.7.3**. Although the sample appeared purer after clean up (**D**), primer was still present and there was significant product loss. The optimised PCR clean-up method using both the 1.5x step and then the 1.1x/0/35x dual selection step produced a pure TCRB product peak at 312bp with minimal primer contamination (**E**). Dark blue boxes indicate the desired TCRB amplicon peak.

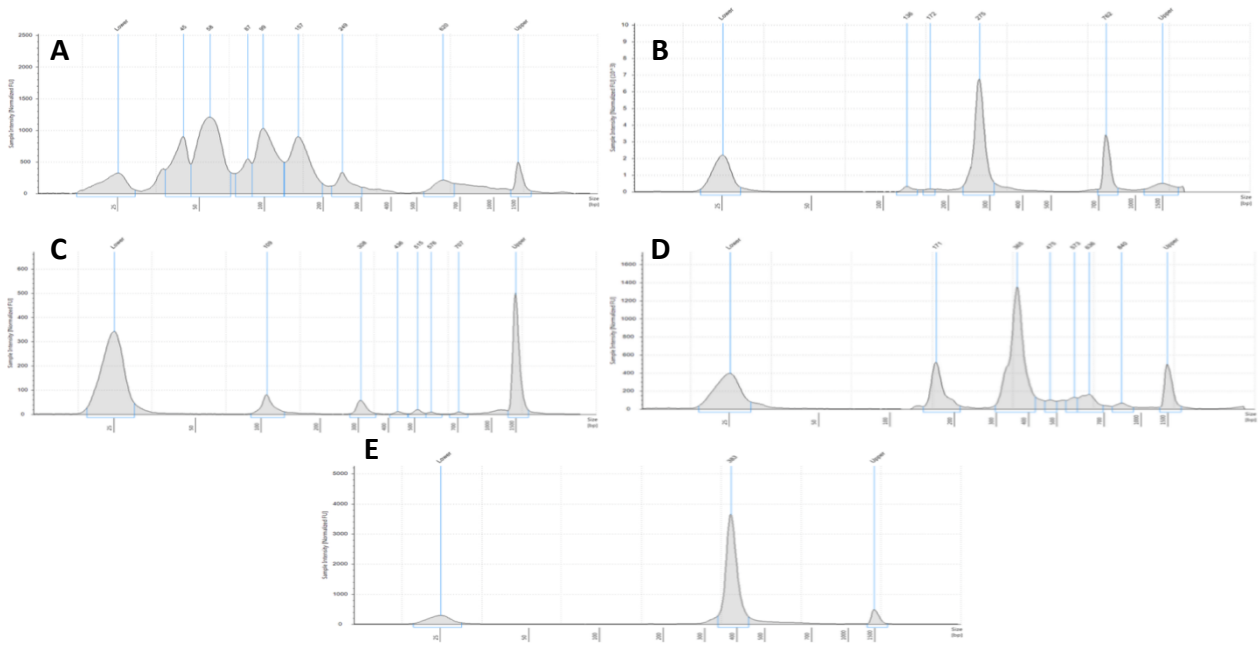


Figure 12. Molecular profiles for the optimisation of PCR clean-ups using the BIOMED-2 primers.

The PCR product constituents were measured according to their molecular weight using the Agilent TapeStation to assess what profiles are associated with successful TCRB reactions and whether size selection methods were successful in removing contaminants. When no DNA was inputted in any of the PCR reactions, the final "library" (**A**) contained contaminants below 150bp which were primer dimers and contaminants above the desired size that were most likely attributed to non-specific binding events. The TCRB final product size of between 376 and 433 bp was notably absent acting as a successful negative control. Successful size selection was often dependent on the quality and concentration of the DNA sample which made finding a method suitable for lower quality and concentrations challenging. When the sample DNA was of low concentration, the TCRB product was more likely to be lost completely through size selection due to less effective PCR amplification (**B**). When using the bead selection process (**C**-**D**), depending on the amount of primer dimer formed during the PCR reaction, the effectiveness of the selection process differed. When similar peak levels of primer dimer and TCRB product were observed (**C**), or high levels of primer dimer (sometimes more than the TCRB product) were present (**D**), several size selection clean ups were implemented to get rid of the primer peak, but each time a similar amount of TCRB product was lost. Consequently, a choice had to be made as to whether to go through another clean-up step and risk more TCRB product loss, or fewer clean ups leaving some primer behind, which could cause inaccurate sequencing. After extensive optimisation of methods, the two step method was implemented in **Section 2.7.3.** providing the best results of a smooth TCRB product peak with minimal primer contaminants and as little product loss as possible (**E**).

2.7.4. Experimentally evaluating the effect and subsequently compensating for PCR and sequencing errors associated with TCRB library preparation

One of the biggest caveats with TCRB sequencing is how to know when differences in the TCRB repertoire are truly biological or caused by factors such as PCR bias, sequencing errors or technical errors. This is why it was important to test for these caveats when developing the *BIOMED-2 primer method* further. Experimental methods were employed to evaluate the effect and then compensate for variations in sequencing attributed to sequencing and PCR when generating the TCRB sequencing reads. All samples designed to assess the effect of PCR and sequencing error were replicated at least three times in a library to account for natural variation observed in TCRB data.

2.7.4.1. Methods to identify and compensate for sequencing errors

Technical rather than biological variation can occur in TCRB repertoires when they are run as part of different sequencing libraries. This can be attributed to variations in sequencing machinery and sequencing tile quality variants. For example, one flow cell in the sequencer may have an associated sequencing error bias. To ensure any drastic changes in library sequencing attributed to these factors were identified and consequently compensated for, identical samples were run on each sequencing run from a mixed gDNA sample (Promega). This commercially sourced gDNA was a mixture of male and females from the ages of 18-60 years old. The concentration of gDNA inputted into the first TCRB amplifying reaction was 200ng. This sample was pooled with each library and run alongside patient and control samples. The most abundant TCRB clonotypes were compared between libraries (**Section 4.3.5.**). Natural variation was expected to occur. However, if there were any drastic differences in TCBV/J gene usage, for instance, it could be assumed that these were due to sequencing and this information would be taken into account when comparing samples between library preparations.

Each sequencing run will generate undetermined reads which cannot be sorted according to Nextera index, generally because of sequencing errors that occurred during the sequencing of the adapter codes. For every sequencing run, these files were analysed to ensure particular TCRB clonotypes were not all being discarded due to inherent and potentially biased adapter sequencing errors, thereby potentially inaccurately skewing a TCRB repertoire.

To test for sequencing error variation between samples on the same sequencing run, a number of samples were split into two reactions prior to the second round adapter PCR. These were then sequenced as separate samples and assessed to see whether a significant difference occurs in TCRB repertoire analysis because of errors or biases such as sequencing lane bias, incorporated during the sequencing reactions. Top TCRB clones and TCRBV/J gene usage were analysed as well as repertoire overlap to take into account natural variation in rarer clones (**Section 4.3.6.**).

2.7.4.2. Identifying and compensating for PCR errors

During the library preparation process, original TCRB receptor sequences are amplified in PCR steps; the first when initially amplifying the TCRB region with TCRBV and TCRBJ gene primers and the second when adding Nextera® adapters ready for pooling samples to make a library and then for sequencing. The number of PCR cycles may affect the diversity results of the TCRB repertoire by incorporating PCR biases and errors. To evaluate the potential effect of this, additional samples were run alongside the standard *BIOMED-2* method samples. These samples had undergone either 50 cycles compared to the usual 35 cycles at the first stage PCR (PCR cycles tend to reach a maximum yield by the 50th cycle) or 20 instead of the usual 10 at the adapter indexing PCR step. The TCRB repertoire data was then compared to the same gDNA sample that had undergone the normal PCR cycles to evaluate the effect.

2.7.5. Determining the sensitivity of the *BIOMED-2 primer method*

Pilot data from a simultaneously running study into viral based treatment of tumours was used to investigate the sensitivity of the *BIOMED-2 primer method*. Samples from a number of patients were taken at different time points and TCRB sequencing data was produced using the optimised *BIOMED-2 method* detailed in full in **Section 2.8.** The time points from the study, before, at the time of viral treatment and after the viral treatment, were expected to differ in time enough that variations in TCRB clonotypes could be observed. To assess this, patient samples at the three time points were compared and any novel TCRB clonotypes that emerged in the final time point could be assumed to be linked to the viral treatment. This indicated sensitivity of the *BIOMED-2 primer method* as it was able to decipher and identify new emerging TCRB clones with time (**Section 4.4.1.**).

2.7.6. Validating the *BIOMED-2 primer method*

As the *BIOMED-2 primer method* was developed as part of this project, it was important to validate the results using data from alternative methods that identify TCRB clonotypes. In order to evaluate the accuracy of this high throughput sequencing method, the TCRB sequencing data of a RAG deficient patient whose repertoire was analysed using a flow cytometry TCRBV identifying method, was compared. This method used 23 TCRB antibodies capable of identifying 29 TCRB family and subfamily genes. TCRBV gene family usage was used as the comparison.

The BIOMED-2 primers had been used to generate the 454 sequencing data in **Chapter 3**. These primers had been adapted for use with 454 sequencing, whereas in this research they had been adapted for use with Illumina® sequencing. Therefore, comparing TCRBV gene usage between the two methods would be a good indicator of effective adaptation of the primers to a high throughput method. Two samples from the 454 sequencing cohort were identified and run on a library using the optimised *BIOMED-2 primer method*.

2.8. *BIOMED-2 primer method* for TCRB library preparation

As results in **Chapter 4** will show, the *BIOMED-2 primer method* was considered to be the most accurate and was, therefore, used to generate TCRB sequencing libraries for the PNH patients and healthy controls. Outlined below is the final optimised pipeline used to generate the TCRB repertoires for the patient samples and healthy controls with a schematic shown in **Figure 13**.

2.8.1. Genetic material

Immune repertoire research appears to be split on whether it is best to use genomic DNA or complementary DNA (cDNA) derived from mRNA as the genetic material for the initial PCR TCRB amplifying reactions. Choice of genetic material very much depends on the biological question that the sequencing data needs to answer. Sequencing cDNA does not provide absolute quantification of the T-cell receptor repertoire because a single T-cell can express multiple copies of mRNA (cDNA is synthesised from mRNA). These levels will vary depending on T-cell activation status and therefore cannot be used to estimate diversity within the repertoire. In this project, the hypothesis is that in PNH, T-cell clonal expansions should be observed. Therefore, cDNA may not be the best choice as proportions of TCRB could be biased by expression levels.

However, cDNA removes the large intron that spans the TCRBJ and TCRBC regions, allowing for more immune specific genetic input into a PCR reaction.

This is advantageous to maximise the number of TCRB clonotypes that can be identified at a certain sequencing depth. Methods such as 5'RACE can be used on cDNA using primers that span the length of the constant region. Therefore, only a select number are needed to amplify different TCRB V and J genes, reducing potential for multiplex primer bias at the TCRB amplification stage unlike with gDNA.

Genomic DNA is beneficial as it provides absolute quantification of diversity as only one productive copy is produced by each T-cell, so it will not be biased by factors such as disease progression. The majority of PNH patient samples are genomic DNA which is why genomic DNA was used for 454-sequencing in **Chapter 3**.

To assess whether the use of cDNA had specific benefits over using gDNA for TCRB sequencing in this project, both types were used in the initial stages of **Chapter 4** results. However, gDNA was selected as the input of choice for the main *BIOMED-2 primer* method. Genomic DNA was selected for use in this method as the majority of the clinical samples from the RTB were gDNA. Based on results discussed in **Section 4.3.1.,** 200ng per 20µl was selected as the optimal concentration of gDNA for the first round PCR reaction. This was selected as the concentration was large enough to allow TCRB diversity to be observed, but also not too high a concentration that there would be too much sample to fit into the 20µl PCR reaction. Lower concentrations were accepted from lower yield gDNA samples, but where possible 200ng was used for consistency. cDNA can also be used for this method, however, this was not optimised further as part of this project.

2.8.2. TCRB amplification PCR reaction

The primers used in the first PCR reaction are detailed in **Table 6 and 7**. These were adapted from the original to make them compatible for Illumina® sequencing by the addition of Nextera adapters (**Section 2.5.2**). For the first round PCR, two reactions were used for each sample, splitting the TCRBJ primers to stop unspecific interactions. These TCRBJ primer aliquots will now be referred to as JA and JB (**Table 8.**). Each PCR reaction was 20µl in total to limit PCR associated biases and set up as in **Table 9**. The only variants in volume from **Table 9** were nuclease free water and sample volume, which were dependent on sample concentration and made up to 20µl using the nuclease free water. Each sample had an equimolar amount of sample put into each JA and JB PCR and the PCRs were run at the same time in the same PCR machine to reduce PCR bias attributable to the PCR machine.

The PCR conditions and cycle numbers used were found to be optimal (**Section 2.7.2**) to produce a good yield of TCRB product with minimal non-specific binding (**Table 10.**).

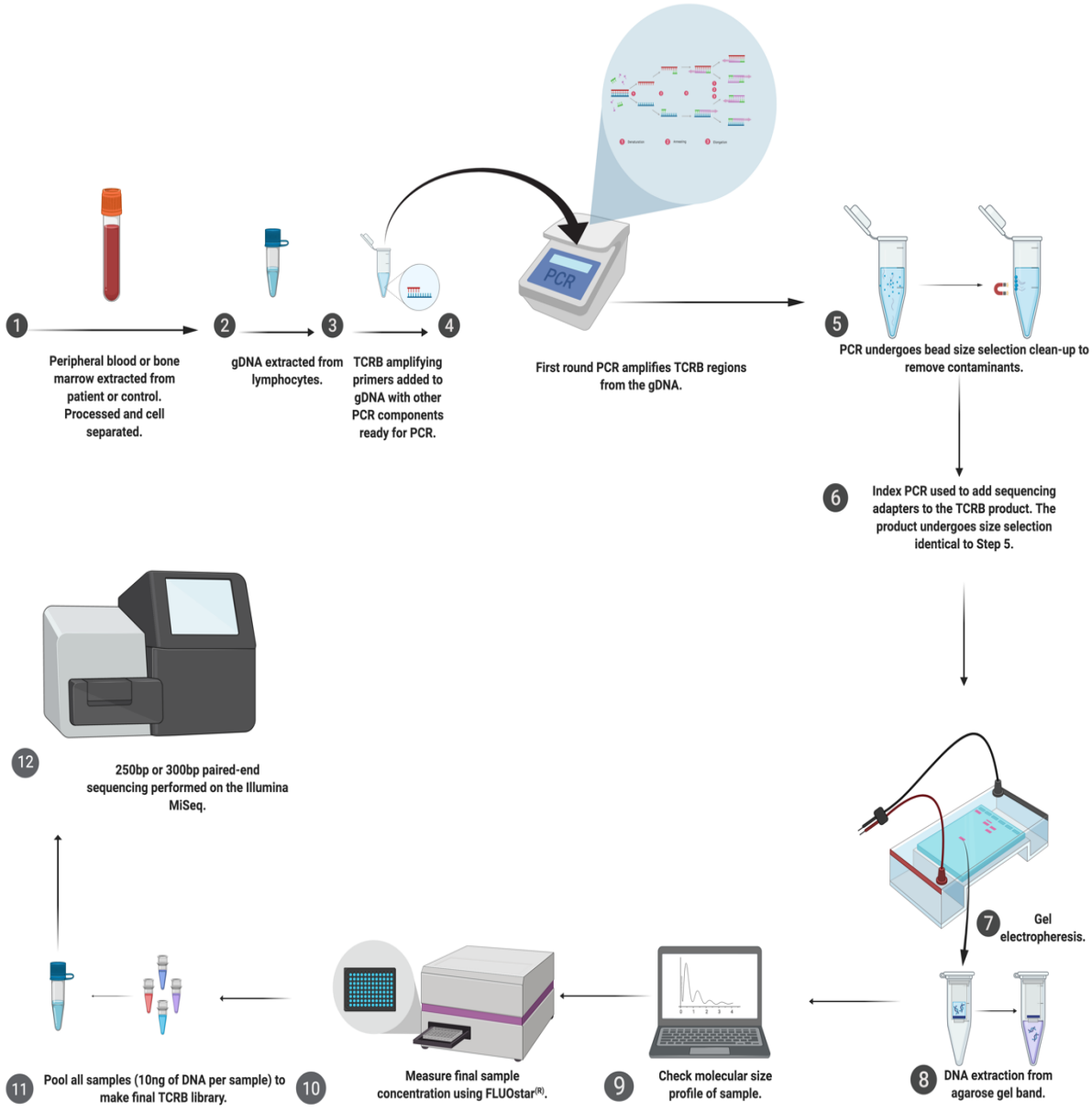


Figure 13. Optimised pipeline for TCRB sequencing library preparation using the BIOMED-2 primers.

A pipeline was developed and optimised in order to generate TCRB sequencing libraries from genetic material (**Section 2.8.**). This method is referred to as the *BIOMED-2 primer method* and is compatible with Illumina® MiSeq sequencing. This pipeline was used to evaluate the TCRB repertoires of 31 healthy controls and 77 patients.

Table 6. *BIOMED-2* TCRBV primers adapted for MiSeq sequencing. Twenty-three *BIOMED-2* TCRBV primers (Sigma) were adapted by the addition of an Illumina® Nextera adapter code to the forward primers allowing Nextera Indexes to be added to the TCRB product during the second PCR reaction. This meant that the TCRB library product could bind to the flow cells for sequencing. The Nextera code is indicated by the nucleotides in capital letters and the *BIOMED-2* primers are in lower case. The same Nextera sequence was added to all TCRBV primers.

Nextera code + TCRBV family (forward primer)	Primer nucleotide sequence
VB2nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGaactatgttttggtatcgtca
VB4nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcacgatgttctggtaccgtcagca
VB5/1nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcagtgtgtcctggtaccaacag
VB6a/11nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGaaccctttattggtaccgaca
VB6b/25nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGatccctttttttggtaccaacag
VB6cnex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGaaccctttattggtatcaacag
VB7nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcactatgtattggtacaagca
VB8anex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGctcccgttttctggtacagacagac
VB9nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcgctatgtattggtataaacag
VB10nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGttatgtttactggtatcgtaagaagc
VB11nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcaaaatgtactggtatcaacaa
VB12a/3/13a/15nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGatacatctactggtatcgacaagac
VB13bnex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGggccatgtactggtatagacaag
VB13c/12b/14nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGgtatatgtcctggtatcgacaaga
VB16nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtaacctttattggtatcgacgtgt
VB17nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGggccatgtactggtaccgaca
VB18nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtcatgtttactggtatcggcag
VB19nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGttatgtttattggtatcaacagaatca
VB20nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcaaccatatctggtaccgaca
VB21nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGtaccctttactggtaccggcag
VB22nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGatacttctattggtacagacaaatct
VB23/8bnex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcacggtctactggtaccagca
VB24nex	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcgtcatgtactggtaccagca

Table 7. *BIOMED-2* TCRBJ primers adapted for MiSeq sequencing. Thirteen *BIOMED-2* TCRBJ primers(Sigma) were adapted by the addition of an Illumina® Nextera adapter code to the reverse primers allowing Nextera Indexes to be added to the TCRB product during the second PCR reaction. This meant that the TCRB library product could bind to the flow cells for sequencing. The Nextera code is indicated by the nucleotides in capital letters and the BIOMED-2 primers are in lower case. The same Nextera sequence was added to all TCRBJ primers. The TCRBJ primers were split into two aliquots for separate amplification reactions when combined with the TCRBV primers in the first PCR. This was to prevent cross-annealing of primers. The aliquot JA was made up of nine TCRBJ primers and JB contained four.

Nextera code + TCRBJ family (reverse primer)	Primer nucleotide sequence
JA primer mix	
JB1.1nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacctacaactgtgaatctggtg
JB1.2nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacctacaacggttaacctggtc
JB1.3nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacctacaacagtgagccaactt
JB1.4nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcatacccaagacagagagctgggttc
JB1.5nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacctaggatggagagtcgagtc
JB1.6nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcatacctgtcacagtgagcctg
JB2.2nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacccagtacggtcagcct
JB2.6nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGctcgcccagcacggtcagcct
JB2.7nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcttacctgtaaccgtgagcctg
JB primer mix	
JB2.1nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGccttcttacctagcacggtga
JB2.3nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcccgcttaccgagcactgtca
JB2.4nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGccagcttacccagcactgaga
JB2.5nex	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcgcgcacaccgagcac

Table 8. *BIOMED-2 primers* split according to TCRBJ region primers. In the *BIOMED-2 method*, the TCRBJ primers were split into two for use in the TCRB amplification reaction with all TCRB V region primers. The J primers were split into groups to avoid cross annealing and unspecific interactions.

JA Primer mix	JB Primer mix
J1-1	J2-1
J1-2	J2-3
J1-3	J2-4
J1-4	J2-5
J1-5	
J1-6	
J2-2	
J2-6	
J2-7	

Table 9. Reagents for the first round TCRB amplification PCR using the *BIOMED-2 method*.

For the first round PCR reaction in the *BIOMED-2 primer method*, TCRB amplification occurs. To achieve this, PCR reactions were set up as detailed below. The volumes below remained constant for all samples. An x, indicates variables that occurred from sample to sample. Depending on the gDNA sample concentration, the volume of DNA needed to achieve 200ng would vary and this would be made up to a final volume of 20µl with nuclease free water to ensure all PCR reactions were equal in volume to reduce biases.

PCR reagent	Volume / µl	PCR reagent	Volume / µl
Nuclease free water	x	Nuclease free water	x
$MgCl_2$ (25mM)	1.6	$MgCl_2$ (25mM)	1.6
Phusion Flash master mix	10	Phusion Flash master mix	10
V primer aliquot	0.96	V primer aliquot	0.96
JA primer aliquot	0.36	JB primer aliquot	0.16
DNA at 200ng	x	DNA at 200ng	x
Total volume	20ul	Total volume	20ul

Table 10. PCR cycle conditions for the first TCRB amplification PCR for the *BIOMED-2 method*.

During the first round TCRB amplification, the following temperatures and time-lengths for the denaturation, annealing and extension stages as well as the PCR cycle number were found to be optimal to generate a good TCRB product yield.

1 cycle	**35 cycles**			**1 cycle**
Denaturation		*Annealing*	*Extension*	
98°C	98°C	63°C	72°C	72°C
10 seconds	1 seconds	5 seconds	10 seconds	60 seconds

2.8.3. PCR clean-up steps

Promega Pronex® beads were used for the PCR clean up steps. The clean-up step needed to successfully remove PCR contaminants whilst preserving the TCRB product band in order to improve the subsequent index PCR. The method that provided optimal TCRB product recovery with minimal loss, was a 1.5x bead step followed by a dual size bead selection. The 20µl JA and JB PCR paired mixes for each sample were combined to form one 40µl mix for one sample.

The first step used a 1.5x bead volume (60µ**l**) and was used to remove smaller base pair products, such as primer dimers, and contaminants, such as buffers. The second step involved dual size selection using a 1.1x bead volume followed by a 0.35x volume (as outlined in more detail in **Section 2.7.3**.). The second step selected for product with a peak of 325bp which effectively selected for the TCRB product of base size (307-364bp). The final product was eluted into 20µ**l** of elution buffer. Samples could be stored in the fridge overnight, however, it was found that carrying out the entire library preparation in one day provided the best quality sequencing.

2.8.4. Second round index tagging PCR conditions

The Illumina® 96 Nextera Indexing kit was used to perform the indexing reaction adding sequencing adapters to the TCRB amplicon. Each sample was given a different N/S Nextera Index combination, to allow it to be identified when all samples were pooled together for sequencing. The PCR reaction was carried out in a 20µl volume using 10µl of Phusion Flash master mix, 2µ**l** each of the specified N and S Nextera index chosen for that sample and 6 µ**l** of the DNA from the clean-up PCR (**Table 11.**). The reaction was briefly vortexed to mix. The PCR was run according to the cycle conditions in **Table 11.** after which the PCR tubes were briefly vortexed and could be stored overnight in the fridge. The index PCR added a unique combination of codes to each sample's TCRB product. This enabled the samples to be de-multiplexed once all samples were pooled together to form a sequencing library. The PCR reactions were carried out in a volume of 20µl using 6µl of the TCRB product. During the indexing PCR the temperatures and time-lengths for the denaturation, annealing and extension stages as well as the PCR cycle number in **Table 11.** were found to be optimal.

Table 11. PCR reagent set up and conditions for the index PCR for the *BIOMED-2 method*.

Index PCR reagent	**Volume of reagent (µl) / 20 µl**
Phusion flash master mix	10
N Nextera index	2
S Nextera index	2
DNA from clean up PCR (TCRB product)	6

1 cycle	**10 cycles**		**1 cycle**
Denaturation		*Annealing*	*Extension*
98 °C	98°C	72°C	72°C
30 seconds	1 seconds	10 seconds	60 seconds

2.8.5. TCRB product Selection

After the index reaction, the addition of the index codes increased the TCRB product size to between 376 and 433 bp. The majority of samples were size selected using gel electrophoresis. The 20µl samples were mixed with 4µl of TriTrack DNA loading dye (6X) (Thermo Fisher Scientific) and run on a 2% GTG agarose gel using a 1xTBE buffer alongside a Hyperladder IV (BIoline) until a good separation of bands was observed. The gel was imaged using a light box with UV and a blue filter guard. Two bands were observed in the sample, one at the size of the TCRB product and one much lower (around 100bp) which was excess adapter and dimerised adapter which had not been used in the PCR. The TCRB product band was cut out of the gel. The DNA was extracted from the gel using the standard protocol from the Zymoclean ™ Gel DNA recovery kit, avoiding the heating step which could lead to GC bias in the TCRB product when sequenced. The method used ADB buffer to digest the agarose and the sample was then added to a spin column where it underwent a number of centrifuge steps, finally being eluted in 15µl of nuclease free water rather than the elution buffer provided in the kit. Some samples which were of poorer quality may have produced some non-specific binding products in the index PCR which were hard to separate using gel electrophoresis. For these few samples, the PCR clean up in **Section 2.7.3.** was repeated and then gel electrophoresis was performed as above.

2.8.6. DNA quality check

To ensure that the TCRB product was present in the final sample and that there was no contamination, the molecular weight profile for each sample was analysed using the AgiLent Tapestation. One microlitre of sample was mixed with 3µl of D1000 sample buffer which had previously reached room temperature. The samples were then run on the 2200 TapeSation using a D1000 screen tape which had previously been brought to room temperature. 2200 TapeStation Controller Software was used to generate reports for each sample to assess whether the samples were pure for sequencing. Good reports showed strong peaks of TCRB product at a size between 376 and 433 bp, with no contamination below this. Some samples had small amount of product above this but, as long as the peaks were broad and low, the sample was acceptable for sequencing (**Figure 14**). If contamination was found below the TCRB product in the occasional sample, an additional 1.5x ProNex bead selection was carried out and the sample re-tested on the TapeStation to ensure purity.

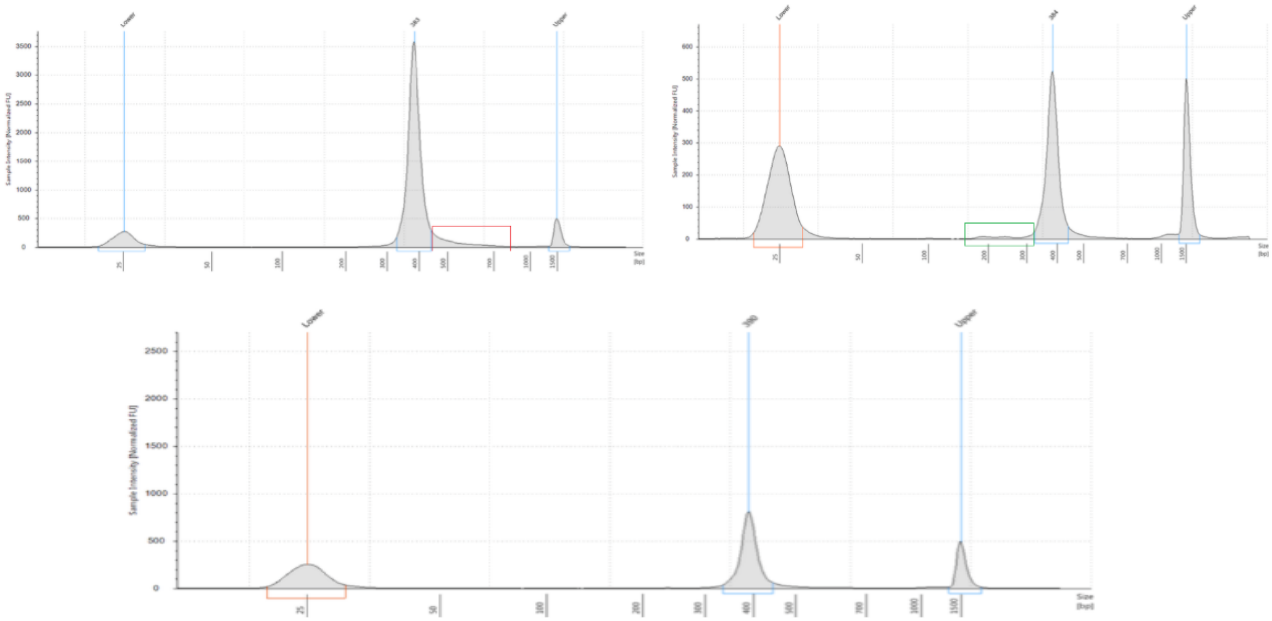


Figure 14. Molecular weight profiles of TCRB library product using the *BIOMED-2 method*.

The 2200 Agilent Tapestation was used to generate molecular weight profiles for TCRB samples during the library preparation process. At the final stage, where the samples were ready for sequencing, if there was some non-specific binding present in the sample, above the TCRB product size, as long as it was a low, wide peak (**Top left,** indicated by the red rectangle), it was accepted for sequencing. If there were even small amounts of non-specific products below the TCRB amplicon size (**Top right,** indicated by the green rectangle), the sample underwent an additional 1.5.x bead selection step to ensure it was removed. The majority of good quality samples produced pure TCRB peaks between 376 and 433bp (**Bottom**).

2.8.7. DNA quantification

In order to assess the concentration of TCRB product in a sample, DNA quantification was carried out using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (InvitrogenTM, ThermoFIsher Scientific) according to the manufacturer's instructions. The protocol first set up a DNA concentration gradient of standards in the first column of wells in a black fluorimetry 96 well plate. This served as a comparison for the samples to assess their individual concentration. One microlitre of each sample was placed in a new well. A buffer was then made using Quant-iT™ PicoGreen® dsDNA reagent and TE buffer. This was then used to dilute the DNA gradient and the samples. PicoGreen® is a nucleic acid stain that specifically quantifies double stranded DNA and so will not bind to primers. The reagent is ultrasensitive and fluorescent which allows its signal to be measured using a machine.

The plate was read using the FluoStar Software which generated a standard curve for the DNA concentration gradient and then subsequent reports for the concentration of DNA in the samples. These samples were then normalised to calculate what volume of sample was needed for 10ng. The aim was to then add 10ng of each sample into the pooled TCRB sequencing library for sequencing. Equimolar samples would reduce biases in sequencing depth. If 10ng could not be achieved by poorer quality samples, a maximum volume of sample allowed in one library was 7µl, again to reduce bias. Once all samples were added to the TCRB pooled library, sequencing could commence.

2.8.8. Sequencing of TCRB libraries

Sequencing was performed by the Leeds Next Generation Sequencing Facility. The pooled TCRB library underwent quality control checks prior to being sequenced. The Illumina® MiSeq was used to sequence the TCRB libraries and cycles of either 250bp or 300bp were used. Paired-end sequencing was carried out on all samples.

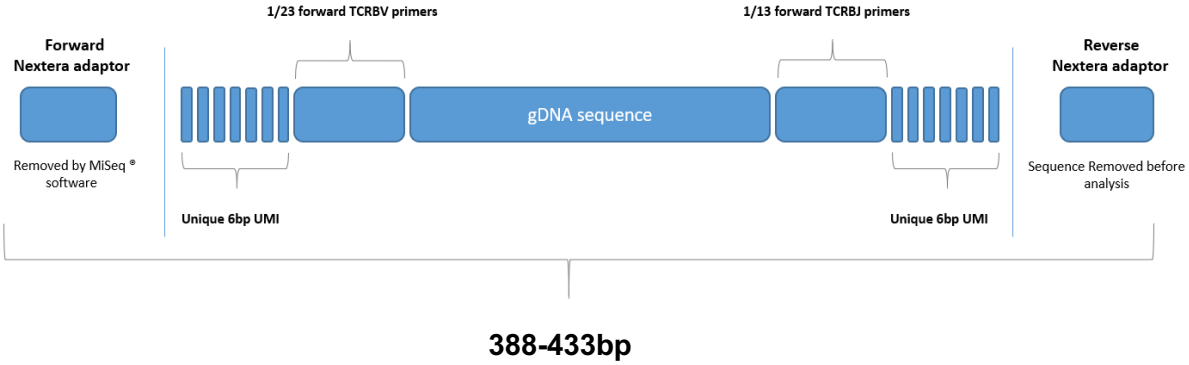


Figure 15. UMI adapted *BIOMED-2 primers* to reduce amplification bias in TCRB reads. BIOMED-2 primers were first adapted for compatibility with the Illumina® MiSeq sequencer. This was done by the implementation of a Nextera adaptor to the start of the forward TCRBV primer and reverse TCRBJ primer. The forward and reverse adapter had a different sequence. The adapters bind to Nextera codes in the index reaction allowing each sample to be given a unique Nextera code combination. This allows the sequencing data of each sample to be de-multiplexed from the pooled library. For UMI-adapted BIOMED-2 primers an additional, randomly allocated, six base pair code was added to the primer in between the adapter and the TCRB amplifying gene region. This gave a final TCRB-UMI library product range of between 388-433bp. In non-UMI BIOMED-2 primers, the only difference is the absence of these six base pair sequences.

2.8.9. Unique molecular identifier adapted primers reduce TCRB amplification bias

During the library preparation method, the PCR reactions amplify and replicate the original TCRB sequences so that they are at detectable levels for subsequent sequencing. However, this can lead to associated PCR errors. An experimental method, known as unique molecular identifiers (UMIs) can be implemented to help reduce inherent PCR bias. This method is more common for cDNA and RNA methods where it is essential, because T-cell receptor mRNA can have multiple copies dependent on expression levels which can skew TCRB repertoires, rather than gDNA which is only present in one functional form per T-cell.

Unique molecular identifiers are incorporated in TCRB sequences during the first TCRB amplifying PCR to decipher biological replicates (TCRB clones that are identical and a result of clonal expansion) from technical replicates caused by the PCR reactions (**Figure 16.**). This method was developed and trialled on a number of samples.

Six random base pair UMIs were incorporated into the TCRB primers(sigma) (**Figure 15**) for both the reverse and forward primers between the Nextera-Adapter sequence and TCRB region primer. Within each TCRB V or J family primer aliquot, 16,777,216 unique combinations of primers were present because of the UMIs, each ready to bind to, and amplify alongside a different TCRB biological sequence.

In order for the UMIs to be incorporated into the sequences, PCR is needed. In theory, only one PCR cycle should be used to incorporate these UMIs to reduce PCR bias. More PCR cycles would risk incorporating more bias and defeat the objective of UMIs. However, in practice, very little TCRB product is produced in one PCR cycle and is likely to be lost during the rest of the library preparation process. For this reason, cycle lengths of 1,2,3 and 10 for the first PCR were trialled in the experiments. PCRs were set up in exactly the same way as described in the TCRB library preparation sections, with the same primer concentrations.

However, the annealing temperature was optimised to slightly lower than the normal primers at 61^{o}C (**Figure 10.**) with the rest of the cycle conditions being identical to the usual non-UMI *BIOMED-2 primer methods*.

Clean up methods were identical to the library preparation pipeline as it was still important to ensure primers were removed from the reaction as they would out replicate the TCRB product.

For the indexing PCR, the reaction conditions and temperatures were kept the same. However, cycles of 10, 20 and 35 were trialled, trying to keep the cycle numbers as low as possible to prevent PCR bias, but high enough to generate enough TCRB product to be sequenced. Product selection, final DNA quality and quantification checks as well as sequencing were identical to the library preparation pipeline in **Section 2.8.1-2.8.8.** UMI samples were not pooled with non-UMI samples. Considerably lower amounts of TCRB product were present at the end of the library preparation for sequencing than when using the non-UMI *BIOMED-2 primers* (**Figure 17.**) but this was to be expected as fewer replication cycles took place.

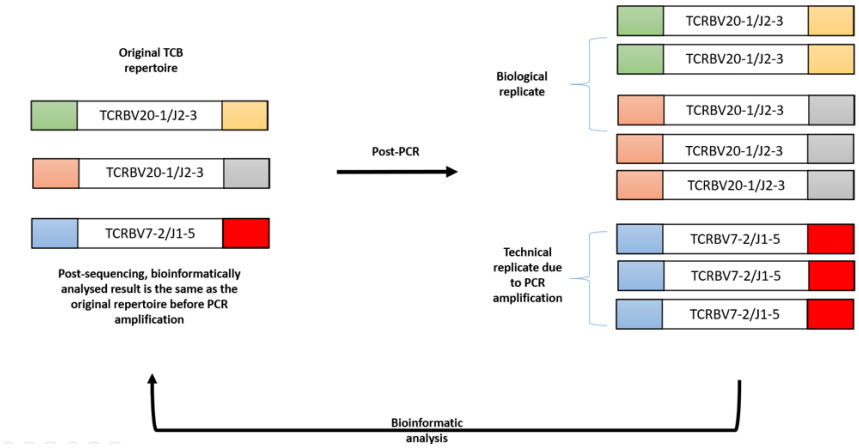


Figure 16. Using unique molecular identifiers in TCRB sequencing helps reduce PCR bias.

Unique molecular identifiers that are incorporated into the TCRB amplifying primers are incorporated into each separate biological TCRB sequence at the point of TCRB amplification. Once this has occurred any subsequent amplification PCR will not inflate bias for the reads. After sequencing, the UMI reads can be collapsed according to groups with the same UMI which are technical replicates rather than biological replicates. This means that the original TCRB repertoire is shown by the data, rather than PCR inflated expansions of reads.

After sequencing, an additional step was implemented into the bioinformatic pipeline discussed in the next section to process the UMIs. The software clipUMI [**172**] was used to generate a list of all the 12 base pair UMI combinations (forward and reverse primers). This list of UMIs was then parsed through to the program pRESTO [**170**]. Here, the software generated consensus UMI sequences using the UMI combinations that matched with a maximum of one base different over the 12 bases and had the same TCRB sequences.

After consensus building and quality control steps, the TCRB sequences in each sample were able to be collapsed (removing technical replicates) into a TCRB repertoire similar to that observed in the original sample before PCR took place (**Figure 16.**). After these processing steps, the bioinformatics pipeline was identical to non-UMI sequences.

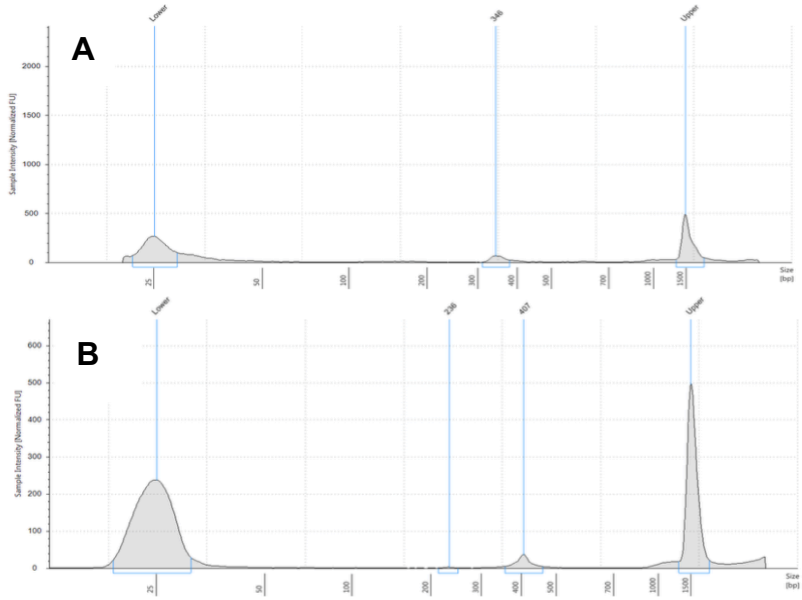


Figure 17. Molecular profiles of UMI incorporated TCRB product using the BIOMED-2 primers. When using the UMI primers for the BIOMED-2 primers, the TCRB amplification PCR was cycled at either 1,2,3 or 10 at an annealing temperature of 61°C and then underwent the same clean up steps as the non-UMI primers to produce a peak of TCRB product at 346bp for the sample shown (**A**). The indexing PCR (**B**) was cycled for either 10, 20 or 35 cycles before undergoing clean up and final library preparation steps. For this sample a clean peak was observed at 407bp. Less product was generated using the UMI method due to restrictions in PCR cycle numbers.

2.9. Bioinformatic analysis methods

Many different types of bioinformatics analysis can be used for TCRB repertoire studies. Over the course of the project many bioinformatics tools and methods were assessed as this is a rapidly developing field with new methods constantly being added. However, for the simplicity of this thesis, only the final pipelines are discussed in the **Material and Methods** section.

2.9.1. 454-sequencing analysis

Chapter 3 discusses the analysis of previously generated 454-sequencing data and the beginnings of building a bioinformatics pipeline to analyse TCRB repertoires.

The 454-sequencing data had previously undergone stringent and standardised quality control methods for 454-sequencing data before being downloaded from the sequencer. Therefore, trimming and quality control parameters were not necessary in this specific pipeline. 454-sequencing is lower throughput than newer sequencing technologies, such as Illumina® . Therefore, sequencing files could quickly and efficiently be processed using the online platform Galaxy [**262**], which contains suites of online tools for biomedical data analysis.

2.9.1.1. Pre-processing reads and overlap alignment of paired-end reads

Firstly, the program *Trim* (Galaxy version 0.0.1) was used to trim off the primer sequences from the ends of the sequencing reads. The two trimmed paired end sequencing files for each sample were then overlapped using the software *FLASH v 1.2.11* with its default parameters.

2.9.1.2. Phred scoring and quality control filtering

Filter by quality (Galaxy version 1.0.0) was used to filter the sequencing reads according to their Phred scores. The parameters were cut off value = 20, and percentage of bases in sequence that must have a cut off value of equal to or higher than 20 = 98. Sequencing reads that had a Phred score base quality of 20 or above were kept, other reads were discarded. In sequencing, base calling is when a nucleotide base is assigned on comparing it to a fluorescence peak generated through the sequencing cycle process. There is a small error associated with this process for modern high throughput sequencing technologies. A Phred score is logarithmically related to the error probabilities associated with this base calling process in sequencing. It is calculated as follows:

$$\mathbf{Q = -10log10P}$$

Where **Q** equals Phred score and **P** is the error probability of base-calling [**397**]. A Phred score of 20 means that each sequencing base has a probability of 1 in 100 of being incorrectly called. The Phred score could have been increased to higher than 20, keeping only the greater quality sequencing reads. However, the number of sequences and data lost because they did not meet the parameters was assessed and as a result the Q score kept at 20. *FastQC* (v 0.11.5) [**175**] was then used to analyse the quality of the remaining sequencing reads, taking into account factors such as per base sequence quality and per base sequence length.

2.9.1.3. Annotating T-cell receptor beta sequencing data

For the following workflow, all work was carried out using a *Virtual Box* (v5.1) virtual machine with an Ubuntu(16.04) operating system and the Linux kernel. When R (v3.3.2) was used in the downstream analysis, R studio (v1.0.44) was used as the accompanying software platform.

To identify the TCRB clones present in the sequencing files, the software *MiXCR* v 2.1 [**183**] was implemented in the Linux command line. IMGT®, the international ImMunoGeneTics information system®, is an online database storing information such as T-cell receptor genes and is used as the gold standard in immune repertoire sequencing analysis [**85**]. The IMGT® database files were downloaded locally on to the virtual machine as a .json file. Each pre-processed sequence file was then aligned to the locally downloaded IMGT® database, annotated with TCRB gene annotations and assembled into clones. A clone was a T-cell receptor that shared the same TCRBV gene family, TCRBJ gene family and amino acid CDR3 sequence. Then only the productive TCRB sequences were exported. These were sequences that were not out of frame or did not contain stop codons and were therefore functional in the TCRB repertoire.

2.9.1.4. Downstream analysis of T-cell receptor beta clones

VDJtools (v1.1.4) [**186**] was implemented in the Linux command line and used to produce graphical data output from the sequencing data, such as Circos plots identifying TCRBV and J gene family usage and spectratypes showing abundant CDR3 clones and rarefaction plots. VDJtools was also used to pool the sequencing reads to analyse overall CDR3 amino acid sequence commonality between controls and between PNH groups. The sequence files previously generated by *MiXCR* and *VDJtools* were parsed into R. Then the R package, *tcR* (v2.2.1) [**185**] was used to generate TCRB V and J bar plots, basic clonotype statistics, and clonal homeostasis plots. Biostatistics methods used include: Chao1 estimator [**162**], Inverse Simpson [**249**] and the Shannon Diversity Function [**393**].

2.9.2. High throughput sequencing analysis

In order to analyse the larger volumes of data produced by Illumina® sequencing, the methods used **in Section 2.9.1.** were modified and expanded upon. In this portion of the project, over 150 million sequences from the Illumina® MiSeq were analysed. It was therefore essential that the new pipeline was time efficient and accurate at analysing the data, but also robust to withstand the high throughput nature of the data (**Figure 18.**).

Instead of continuing to use the Galaxy platform for pre-processing, the analysis was carried out locally using the virtual machine outlined in **Section 2.9.1**. A combination of bash scripting, Linux commands, R scripting and python scripting were used to create the new pipeline. All sequencing reads processed were paired-end and either 250bp or 300bp cycles.

2.9.2.1. Quality control and pre-processing TCRB sequencing reads

Quality control analysis was carried out on the separate forward and reverse sequencing reads using *FastQC* (v 0.11.5 or above) [**175**]. Forward reads (**Figure 19A.**) in general were better quality than reverse reads (**Figure 19C.**). FastQC was used to ensure that the sequencing had performed correctly looking at factors such as read length, Nextera® adapter content and duplicated and overrepresented sequences. Tile quality was assessed to ensure there was no positional sequencing bias from the sequencing machine itself. Duplication error warnings were common due to the nature of TCRB repertoire sequencing where if clonal expansions are present, highly represented reads will be common.

The next step involved the trimming of any Nextera® adapters still present and primers from each end of the sequence reads (**Figures 19 and 20**). The softwares, TrimGalore, which uses Cutadapt, and Trimmomatic, (**Table 1.**) were used in combination to achieve the best quality control trimming. In general, the end bases in a sequence were of poorer quality and needed to be trimmed to ensure successful overlap in subsequent steps. Bases were trimmed starting from the end of the sequence until all sequences achieved a Phred score of 30 (q30) (99.9% base call accuracy). FastQC analysis was then repeated on the sequencing reads and the quality of the bases were assessed.

If q30 had not been achieved an additional trimming was implemented. This was essential for runs that used 300bp cycles. These saw a dip in quality at around 265bp and then an improvement in quality as the sequence reached the end, which was not necessarily corrected with one trimming step (**Figure 20.**). Forward reads generally saw better quality reads than reverse reads. If an additional trimming step was required, an additional FastQC step was used to check that q30 had successfully been achieved. Very few reads were discarded at this step. They were only discarded if their length fell below 20bp once trimmed.

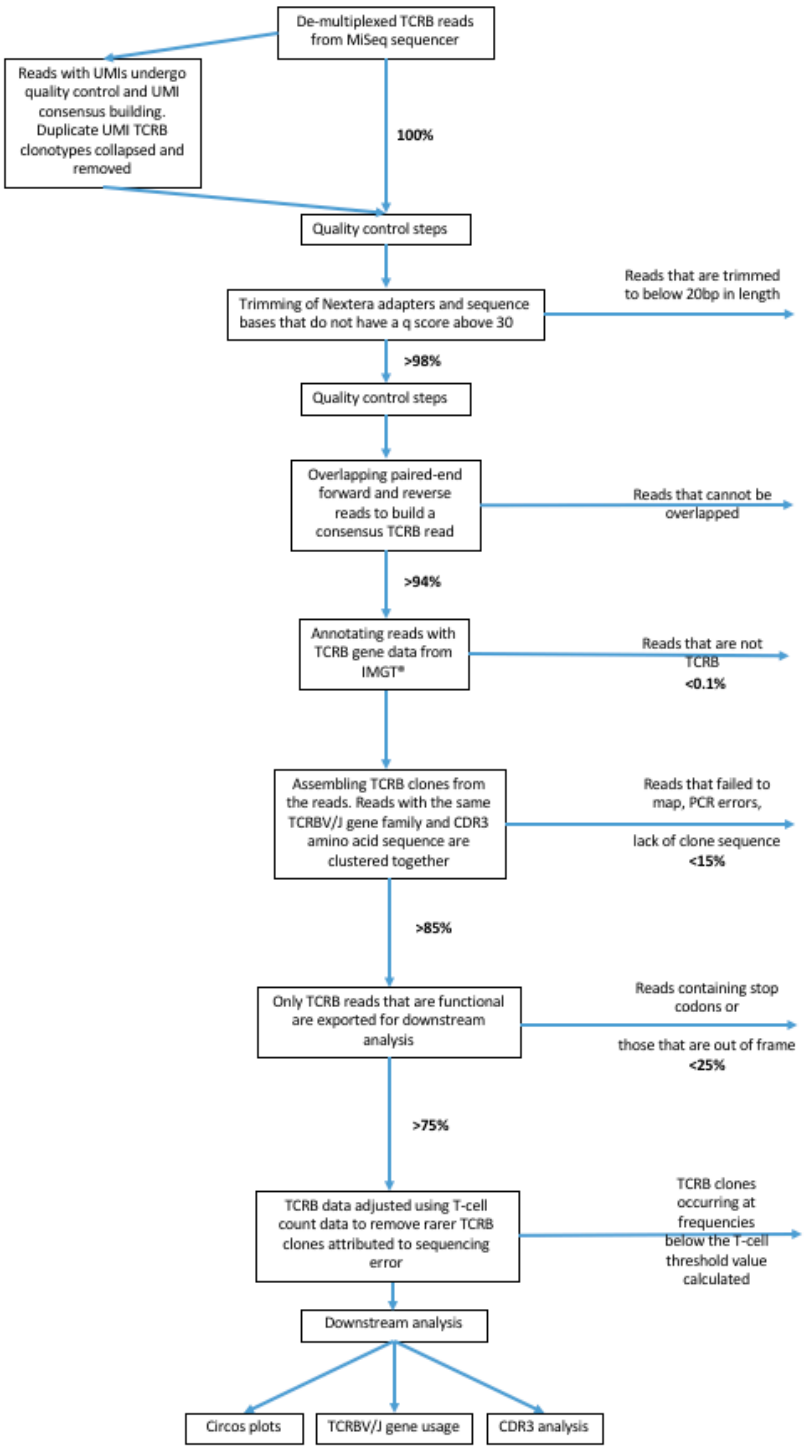


Figure 18. Bioinformatics pipeline for the analysis of TCRB sequencing data generated using the *BIOMED-2 primer method*.

The bioinformatics pipeline above processed over 150 million TCRB sequencing reads including 31 healthy controls and 77 PNH or AA patients. Percentages in the central section indicate the number of expected TCRB reads to successfully filter through to the next step. Percentages and arrows to the right indicate reads discarded at each step as they did not meet the requisite criteria.

2.9.2.2. Alignment processes

Overlapping paired-end sequencing reads

The first alignment process involved the overlapping of the paired-end forward and reverse reads. As the TCRB products ranged from 376bp to 433bp, 250bp and 300bp paired-end cycle sequencing achieved sufficient overlap for this step. This overlap was used to build a consensus sequence using the algorithm from the software FLASH [**182**]. Standard parameters were used with the exception of an optional orientation step that meant all reads were aligned in both "outie" and "innie" orientations to ensure as high a percentage of overlap as possible was achieved. In general, an overlap value over 94% should be achieved on good quality data. FLASH merged the paired R1 and R2 files so that just one consensus TCRB read was taken forward in the analysis.

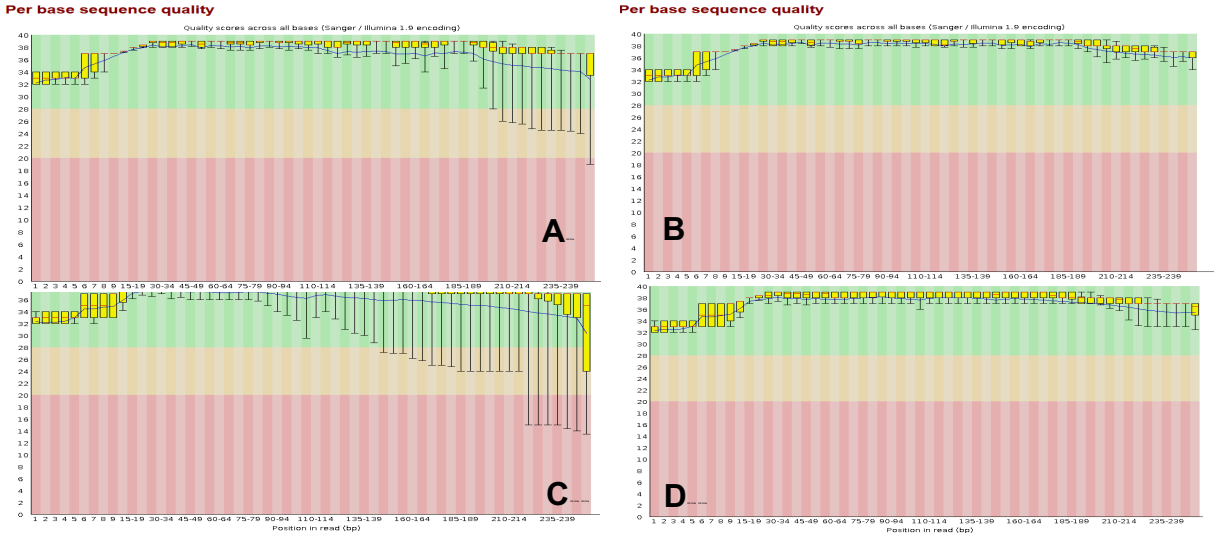


Figure 19. Sequence base qualities of TCRB repertoires using the *BIOMED-2 primer method* (250bp).

After the TCRB methods had undergone TCRB library preparation using the *BIOMED-2 primer method* they were sequenced using the Illumina® MiSeq. Each sample generated forward and reverse reads. Forward reads (**A**) generally saw higher quality base scores than the reverse reads (**C**). It is usual for sequence base quality to lower over the course of a sequencing cycle. The bioinformatics pipeline trimmed reads until a base score of q=30 was achieved using TrimGalore and Trimmomatic softwares (**B** and **D**). FastQC was used to generate the base quality plots. The green areas indicate base quality Phred scores of 28 and above. Each yellow box plot indicates the interquartile ranges of the base quality at each base of the sequencing read across the sample. The black lines indicate the 10% and 90% range. The red line represents the median quality and the blue line the mean. The plots above were from the same TCRB sample sequenced using a paired-end 250bp cycle.

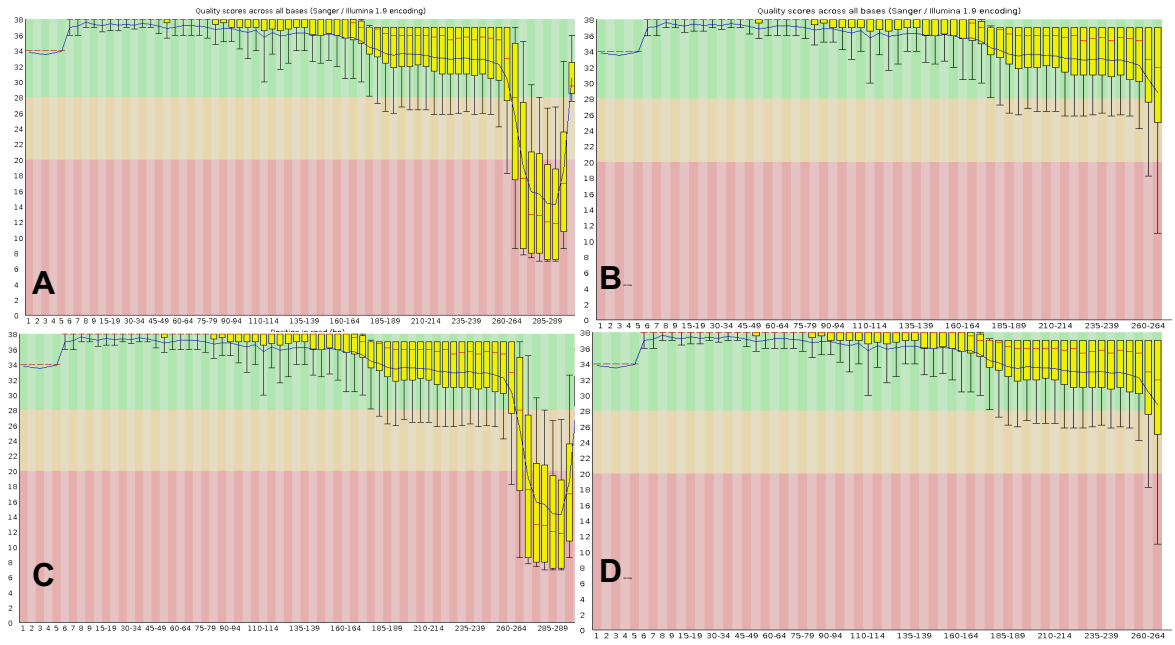


Figure 20. Sequence base qualities of TCRB repertoires using the *BIOMED-2 primer method* (300bp).

After the TCRB methods had undergone TCRB library preparation using the *BIOMED-2 primer method* they were sequenced using the Illumina® MiSeq. The bioinformatics pipeline trimmed reads until a base score of q=30 was achieved using TrimGalore and Trimmomatic softwares (**A** and **C**). FastQC was used to generate the base quality plots. The green areas indicate base quality Phred scores of 28 and above. Each yellow box plot indicates the interquartile ranges of the base quality at each base of the sequencing read across the sample. The black lines indicate the 10% and 90% range. The red line represents the median quality and the blue line the mean. The plots above were from the same TCRB sample sequenced using a paired-end 300bp cycle. Samples generally sequenced using 300bp cycles, did not achieve a Q score of 30 after the first trimming step (**A** and **C**). At around 260-270bp a dip in quality was observed after trimming. A second trimming step was repeated on the new data set which achieved the desired Q score of 30 for the mean and median values (**B** and **D**).

Alignment to IMGT® for T-cell annotation of sequences

As mentioned, the International ImMunoGeneTics information system® was used as the database of choice for the alignment of the sequencing reads, annotating them with T-cell receptor beta gene data. The software MiXCR was used to align the reads to IMGT® again as described in **Section 2.9.1**. At this stage, any sequencing reads that were not TCRB or were unable to be assigned a TCRB gene annotation, for instance if there was a lack of TCRBJ gene, were discarded. A negligible number of reads were discarded for not being TCRB.

Defining, assembling and filtering TCRB clonotypes

In this study a TCRB clonotype is defined as any sequence containing the same TCRB V, TCRB J genes and CDR3 amino acid sequence. CDR3 amino acid sequence was chosen over nucleotide sequence to avoid biases in analysis caused by the biological biases associated with convergent recombination (**Section 1.4.3.1.**). In order to group the T-cell receptor beta gene annotated sequences according to this definition of a clonotype, the software MiXCR was used.

From good quality data, it would be expected that over 85% of the reads from the alignment stage pass the clonotype clustering process. Those reads that do not pass, may be discarded due to PCR error correction or failed mapping or, the majority, due to lack of clone sequence (attributed to sequencing errors). PCR and sequencing associated errors were compensated for bioinformatically. During the assembly stages of TCRB clones using MiXCR, some reads were dropped by the analysis due to PCR errors. The algorithm clustered TCRB sequencing reads believed to be clones, and compared the non-hypermutation origin regions (T-cells unlike B-cells do not undergo hypermutation). These are portions of the TCRB gene, such as the non CDR3 portion of the TCRB V and J genes that should be the same sequence for all clones of the same family. The probability of a single nucleotide mutation in the clonal sequences was calculated and then used to create a threshold. Any number of nucleotide mutations that surpassed this threshold when compared to the IMGT® sequence data were attributed to sequencing or PCR errors and subsequently discarded and filtered out from further analysis by MiXCR.

The TCRB clonotypes that passed this stage then went on to be exported using MiXCR. When using gDNA as the sequencing material, a maximum of two TCRB receptors gene sequences can be detected per T-cell. If the first receptor does not pass the selective processes in the thymus (**Section 1.1.**) the TCR has one more chance to re-arrange to produce a viable TCR. The first receptor will not be functional so will, therefore, either contain a stop codon in the genetic sequence, or the sequence will be out of frame. If the TCR passed the selective processes first time, only one gene sequence will be present in the data and it will be productive. Therefore, as only the functional, active TCRB needs to be analysed in the context of the TCRB repertoire, MiXCR was used to export only the productive/functional clonotypes. This can range in frequency from sample to sample, but usually 75% and above of clonotype reads are expected to be productive. The clones output file was then converted to VDJtools format using the VDJtools software for downstream analysis.

Adjusting TCRB clonotypes according to T-cell sample numbers

Within a natural TCRB repertoire there can be more abundant clones (higher percentage) present along with rarer clones that will be present at very low levels, with some TCRB receptors only appearing once in a repertoire. However, this will not be true for the sequencing data representation. This is because Illumina® sequencing uses a method known as bridge amplification [**260**]. This allows the TCRB library product to be amplified and cluster in flow cells, allowing the product to be present at detectable levels for the sequencer. This however, means that the TCRB reads are artificially inflated pre-sequencing. Therefore, legitimate TCRB sequences, even those present at very small levels in the original TCRB repertoire, would be expected to be present more than once in the sequenced sample. The rarer sequences that appear only once in the VDJtools files may be attributed to sequencing errors of other legitimate clones. In order to more accurately decipher biologically low frequency clones present in the repertoire from sequencing errors, where applicable, T-cell percentage data was used for filtering. This data was collected using flow cytometry (**Section 2.2.2**). The following calculations were carried out to achieve a cut off threshold for real TCR clonotypes in each sample individualised by T-cell count data. Where T-cell data was not available, for PNH samples the T-cell count percentage was assumed to be 80%, and healthy controls and aplastic anaemia samples at 60%. If samples were prepared from buffy coat, the percentages were halved as buffy coat contains around half as many T-cells.

T-cell concentration was calculated by dividing the concentration of the genetic material inputted into the first PCR (most samples were 200ng per PCR) over the total amount of genetic material (6 picograms per cell). Total number of T-cells was calculated by dividing the T-cell percentage by 100 and multiplying by the T-cell concentration. The number of total TCRB clonotype reads was then divided by the T-cells in the sample to calculate the reads per cell. Each value for each TCRB clonotype in the VDJtools file was then divided by this number. Due to variation in sequencing, some cells may be sequenced more times than the threshold value and some less. To make sure the lesser values were not discarded, the threshold was defined as the reads per cell divided by 2. Any TCRB clonotypes that had newly calculated clonotype reads below this value were assumed to be sequencing errors rather than rare clones and discarded from the downstream analysis. An example is detailed in **Table 12**.

2.9.2.3. Downstream analysis - Creating a background for normal variation observed in TCRB repertoires

Due to the dynamic nature of TCRB repertoires outlined in **Section 1.4.** it was important in any downstream analysis to assess what "normal" TCRB repertoires in the general population consist of, allowing for accurate comparison with patient TCRB repertoires linked with AA and PNH. TCRB repertoires from 31 normals described in **Section 2.1.1.** were sequenced**.** A number of downstream analysis methods were performed on the T-cell receptor beta data in order to investigate repertoire dynamics. Concepts and rationale of methodologies are discussed in more detail in the context of their results when first introduced in the result sections. The following will give a brief overview of how the analysis was carried out and the primary aim.

TCRB clonotype variation analysis

In order to assess the effect of age and sex on the number of unique TCRB clonotypes, the total number of TCRB receptors in each sample and their abundance in a person's repertoire for both healthy controls and patients was investigated using the R package *immunarch*. Both CDR3 nucleotide sequence and amino acid sequence were investigated in each analysis. Values were plotted as bar plots, with statistical p values added from Wilcoxon rank sum tests [**242**]. P values were adjusted using the Holm method [**243**] when multiple comparisons were made, for instance, between age ranges.

TCRB V and J gene family usage analysis

In order to investigate changes in TCRBV and J genes within a TCRB repertoire, the R packages *TCR* and *immunarch* were utilised. TCRBV and J gene family usages were plotted separately as bar charts according to the gene identified in the TCRB clonotype. Plots were analysed both on an individual basis and when patients and controls were grouped according to age and sex, assessing the effect these factors may have on outcomes. The software *VDJtools* was used to analyse TCRB clone size irrespective of CDR3 in an individual's repertoire by generating a Circos plot [**244**]. This assessed whether particular TCRBV/J gene combinations were commonly observed. Circos plots are circle plots that represent an individual's TCRB repertoire. The top hemisphere depicts TCRB V genes and the bottom TCRBJ genes. Each TCRBV/J gene combination in a repertoire joins up to make a segment in the Circos plot. The larger the size of the segment, the larger the abundance of the unique TCRB in the repertoire. Therefore, these plots are excellent for easy visualisation of clonality in a repertoire.

TCRB CDR3 analysis

CDR3 analysis was carried out using several methods. Length distribution of CDR3 (both amino acid and nucleotide sequence) was assessed using *Immunarch*. Differences in CDR3 distributions between sex and ages were statistically tested using the Wilcoxon test and Holm adjusted P values. Spectratyping graphs were created using *VDJtools* to assess CDR3 length in the context of abundance of clonal TCRB CDR3 amino acid sequences to assess whether clonal sequences shared a particular CDR3 length.

CDR3 amino acid physiochemical property analysis

The amino acid properties that make up a TCRB CDR3 sequence are thought to provide insight into the properties of the T-cell receptor. Amino acid properties of CDR3 sequences were assessed using the R package *Alakazam* [**245**]. The amino acid sequences were analysed in the context of clonal response levels to assess whether specific amino acids were more common in clonally expanded TCRB sequences or in TCRB that were not clonally expanded. Clonal response levels used in this project are defined in **Section 4.5.4**. Significant difference in these factors between response levels was statistically tested (Kruskal Wallis) [**246**]. Additionally, each response level property result was compared to one another pairwise (Wilcoxon). To investigate positional effects of amino acids in CDR3 sequences, kmers were created from the CDR3 sequences of length 4, 8 and 15 which was the maximum kmer length of the CDR3. This enabled analysis into more central amino acid variation within the CDR3. Position probability matrices were produced to investigate this.

Defining and investigating variances in clonal TCRB response levels in the TCRB repertoire

In a TCRB repertoire, as it is a snapshot in time of the immune system, a number of combinations of TCRB events could be occurring. Clonal expansions occur through T-cell proliferation to i) maintain homeostasis and ii) when T-cells are activated in response to an antigen to generate an immune response. Repertoires are dynamic but the blood sample taken is static, which means that it can be difficult to assess the ongoing T-cell dynamics.

For example, a mid-level clonal expansion could be in a mid-activation state, gearing up to fight infection or it could be contracting after the infection has passed and the immune response has been dampened down. Although these events cannot be deciphered, an attempt was made to measure clonal responses. In order to associate a numerical value with clonal response, the percentage of the individual's repertoire accounting for the top TCRB clone across all the 31 normals was plotted and a linear regression smooth curve of best fit was plotted using the 'loess' method [**247**]. A y intercept of the median value for the top clone proportion across the normals was added to this graph. The plots showed groupings in clonal responses across the healthy controls. This would be expected as a number of controls could have expanded TCRB clones promoted by the process of fighting infections, such as those caused by the common cold or suffering from allergies or other diseases not detailed by the PNH RTB data. However, the majority of the repertoires were expected to show no large T-cell clonal expansions. Therefore, their levels depicted would show natural percentages of sequences that occur due to PCR amplification processes, creating a baseline value. Naturally, the normal data (**Figure 37.**) grouped into clonal expansion levels according to natural variation in the population. These figures were used subsequently to define clonal expansions (**Table 13.**).

Assessment of monoclonality versus polyclonality in the TCRB repertoire

When investigating TCRB clonal responses these can be monoclonal, one TCRB clone, or polyclonal, comprising of multiple simultaneously clonally expanded responses. The TCRB clones across the 31 healthy controls making up more than 2.42 % of the total TCRB repertoire were investigated for monoclonal and polyclonal responses.

Investigating effects of thymic involution

As introduced in **Section 1.4.5.**, the thymus involutes as a person gets older and the diversity of the TCRB repertoire would be expected to decrease with age. Diversity can further be decreased because of chronic exposure to infections such as CMV and EBV. To investigate this effect on the TCRB repertoire in healthy controls, the number of unique TCRB clonotypes, number of TCRB clonotypes, CDR3 length distributions, TCRBV/J gene usage and diversity measures, including d50 [**248**], inverse Simpson [**249**] and Gini-Simpson [**168**], were investigated in the context of age. It is suggested after the age of 40 the TCR repertoire is no longer able to generate more unique TCRB attributed to the thymic involution process [**355**]. Therefore, for this analysis, healthy controls were split into below 40 years old and above 40 years old, with a ratio of samples of 19:12 respectively.

TCRB repertoire overlap studies

In order to assess the extent of overlap expected between two healthy controls in a population, overlap studies were performed using *immunarch*. Two methods for similarity were used, "public" and "overlap" [**250, 251**]. The "public" method calculated the number of shared TCRB clonotypes between two individuals and summed the counts of these clonotypes. The "overlap" method measured the overlap or intersection between two finite TCRB repertoires, normalising the data according to TCRB repertoire size.

TCRB clone identification using database cross validation and scientific literature

To assess whether TCRB clonal responses in the 31 healthy controls were observed in other diseases or infections, such as autoimmune disease, cross validation of these clones across multiple TCRB data platforms was carried out. This was stratified into clonal response levels. Internet searches, scientific literature searches using Pubmed [**252**] and Google Scholar [**253**], alongside cross validation with pre-existing TCRB knowledge bases, including the McPAS-TCR and VDJ-db databases [**254-255**] were carried out on TCRB clones that were considered to be clonally expanded.

Generating a TCRB public repertoire

Public TCRB clonotypes, as mentioned in **Section 1.4.3.** are defined as TCRB clonotypes with the same TCRBV/J gene combination and amino acid CDR3 sequence that appear in two or more individuals. These TCRB clonotypes were assessed using *immunarch* with parameters set so that any TCRB clonotype in one of the 31 healthy controls that appeared in one or more other controls was added to a matrix. This generated a database of public TCRB clonotypes from the experiment which were cross-referenced with databases using the methods detailed above.

Diversity measures to evaluate TCRB repertoire diversity in healthy controls

There is an extensive range of statistical tests and diversity measures that can be used in TCRB repertoire studies. Seven diversity measures were used in this project as they provided different aspects of information on TCRB repertoire data (**Table 14.)**. The data was pre-processed in R followed by statistical analysis carried out using *immunarch*. The diversity measurements were used to provide an insight into the types of TCRB diversity expected to be observed when no known pathological immune responses are taking place.

For this analysis one of the controls was removed as it contained a TCRB clone accounting for 53% of the TCRB repertoire which was abnormal and would skew diversity statistics. Therefore, only 30 healthy controls were included in this method.

Table 12. Adjusting TCRB sequencing reads for technical amplification during sequencing.

All TCRB processed sequencing reads were adjusted for artificially inflated TCRB clone numbers attributed to amplification during the sequencing process. In this example, 200ng of gDNA was used in the first TCRB amplifying PCR. In the sample 90% of the cells were identified as T-cells using flow cytometry. The number of TCRB sequencing reads that passed pre-processing was 300,000. This generated the value of 10 for the number of sequencing reads expected per cell. All TCRB clonotype numbers were divided by 10 to reduce bias attributed to amplification. A cut off threshold of 5 was then used to filter out any TCRB clonotypes that potentially arose from sequencing errors. TCRB clonotypes with fewer than five reads in this particular repertoire were discarded from downstream analysis.

Concentration of genetic material inputted into first PCR (TCRB amplification)	**200ng**		
Concentration of genetic material in each cell	**6 picograms**		
T-cell percentage of sample	**90%**		
TCRB clonotype reads	**300000**		
T-cell concentration in PCR sample	Concentration of genetic material inputted into first PCR / concentration of genetic material in each cell	200000/6	**33333**
T-cells in sample	T-cell concentration in PCR sample x(T-cell percentage in sample /100)	33333 x (90/100)	**30000**
TCRB sequencing reads per cell	TCRB clonotype reads/T-cells in sample	300000/30000	**10**
Number of reads for each TCRB clonotype	Each individual clonotype number /TCRB sequencing reads per cell	Reads/10	
Cut-off threshold for adjusted TCRB clonotypes **>x refers biological TCRB clones** **<x TCRB clones generated due to sequencing error and subsequently discarded from further analysis**	TCRB sequencing reads per cell / 2	Reads/10/2	**Reads/5**

Table 13. TCRB clonal response thresholds calculated from TCRB sequencing data of 31 normals.

In order to assess natural variation in TCRB clonotype response levels, 31 healthy controls had their TCRB repertoires sequenced using the *BIOMED-2 primer method* and analysed. The percentage of each control's top clonotype was plotted and a curve of best fit applied. Natural variation of these values occurred across the healthy controls allowing threshold values to be assigned for non-expanded TCRB clonotypes, low level expansions, moderately expanded and hyper-expanded clonotypes.

TCRB clonotype response level	**TCRB clonotype percentage in the entire TCRB repertoire/ %**	**Potential factors causing response level**
Non-expanded clonotypes	$0 \leq x < 2.42$	TCRBs identified by level may be higher than biologically present. Artificially amplified expansion levels as a result of experimental/technical expansion of clonotypes, attributed to PCR and sequence amplification of reads rather than biological expansion.
Low level expansion	$2.42 \leq x < 4.85$	Beginning of a T-cell activation response to an antigen, recent T-cell activation, or towards the end of an infection/immune response where the T-cell response is dampening down, contracting and almost returned to normal levels.
Moderately expanded	$4.85 \leq x < 20$	In the midst of a clonal T-cell response, could be long or short-term.
Hyper-expanded	$20 \leq x \leq 100$	In the midst of a clonal T-cell response, could be long or short-term, could be a chronic infection, or T-cell mediated disease. Unusual for one clonotype to take up such a large portion of the repertoire.

Table 14. Diversity measures used to investigate TCRB variability within a TCRB repertoire.

Diversity measure	Rationale for choice in TCRB repertoire analysis
Chao1 estimator [**162**]	Investigates the number of rarer clones in a repertoire that potentially were not discovered by the sequencing sample.
d50 [**248**]	Calculates the number of TCRB clonotypes needed to represent 50% of the entire repertoire using the most abundant clonotypes first. The smaller the d50, fewer TCRB clonotypes are needed to make up 50% and therefore the repertoire is more clonally expanded.
Gini index [**257**]	The closer the index to zero, the more even the TCRB clones in proportion in the repertoire. Combined with d50 values provides insight into stability of the TCRB repertoire.
Gini-Simpson index [**168**]	Indicates diversity by calculating the probability two selected TCRB receptors are different clonotypes.
Hill values [**256**]	Family of diversity indices. Q =0 measured repertoire evenness, q =1 represented the Shannon index and q=2 was the Simpson diversity index.
Inverse Simpson [**249**]	Calculates species richness and evenness of TCRB clonotype populations. Indicator of repertoire stability.
Rarefaction [**169**]	Assesses species richness using an extrapolation method. Can indicate TCRB richness and species diversity saturation attributed to sequencing depth.

2.9.2.4. Patient data downstream analysis

Each of the analysis methods outlined above in **Section 2.9.2.3.** were repeated using the PNH and AA patient sequencing data. The information gathered for generating background values for TCRB repertoires using healthy controls was compared with each analysis for patients. This allowed any findings in the PNH or AA patients for factors such as TCRBV/J gene usage to be assessed in the context of natural variation, and allowed accurate assessment as to whether specific TCRB repertoire findings were AA or PNH specific, or natural population variance. Analysis methods were repeated splitting patients into the categories in **Table 3.** to compare responses according to PNH status and progression. TCRB clonality and clonal expansion analysis were of particular interest for the PNH and AA patients. This is because, if there is some form of T-cell involvement in PNH, it would expect to be observed using the variety of methods outlined above.

Chapter 3 - T-cell receptor beta repertoire analysis in Paroxysmal Nocturnal Haemoglobinuria using 454-sequencing technologies

3.1. Introduction

Originally, **Chapter 3** displayed the complete data analysis carried out on previously generated 454 sequencing data from 18 PNH patients and 10 healthy controls. Due to the page limit of this thesis, the full analysis could not be displayed. The introduction to the project work, objectives of this chapter, along with a brief summary of the important results and how they shaped the direction the project took in **Chapters 4** and **5** will be outlined below.

As discussed in **Chapter 1**, there are many methods that can be used to generate T-cell receptor beta (TCRB) repertoire data. One more common sequencing method is 454 pyrosequencing (Roche 454 sequencing) which was launched in 2005 [**150**]. The method (outlined in more detail in **Figures 21. and 22.**) generates a sequencing library bound to a bead. This bead is populated by PCR within a droplet. Sequencing then takes place and each read is from a single bead in a single well. A nucleotide is assigned on the basis of light emissions caused by an enzymatic reaction involving the enzyme pyrophosphate. This occurs when, in each round of sequencing, only one type of nucleotide is added to the sequence at one time. If there are multiple adenine nucleotides in the sequence, for example, these will all be added in one cycle producing a large bright flash of light. The recorded peaks from the light emitted produce a pyrogram shown in **Figure 21**. In each round of sequencing only one type of nucleotide can be added to the sequence at one time.

There are a few advantages to using 454 sequencing over earlier sequencing methods such as the Sanger method, one being that many cells can be sequenced at a lower cost. Another being that the method does not rely on cloning template DNA. It has therefore been suggested that it will not skip sections of the DNA that cannot be cloned, for instance regions such as heterochromatin [**258**], that are compact, enabling the sequence to be more accurate.

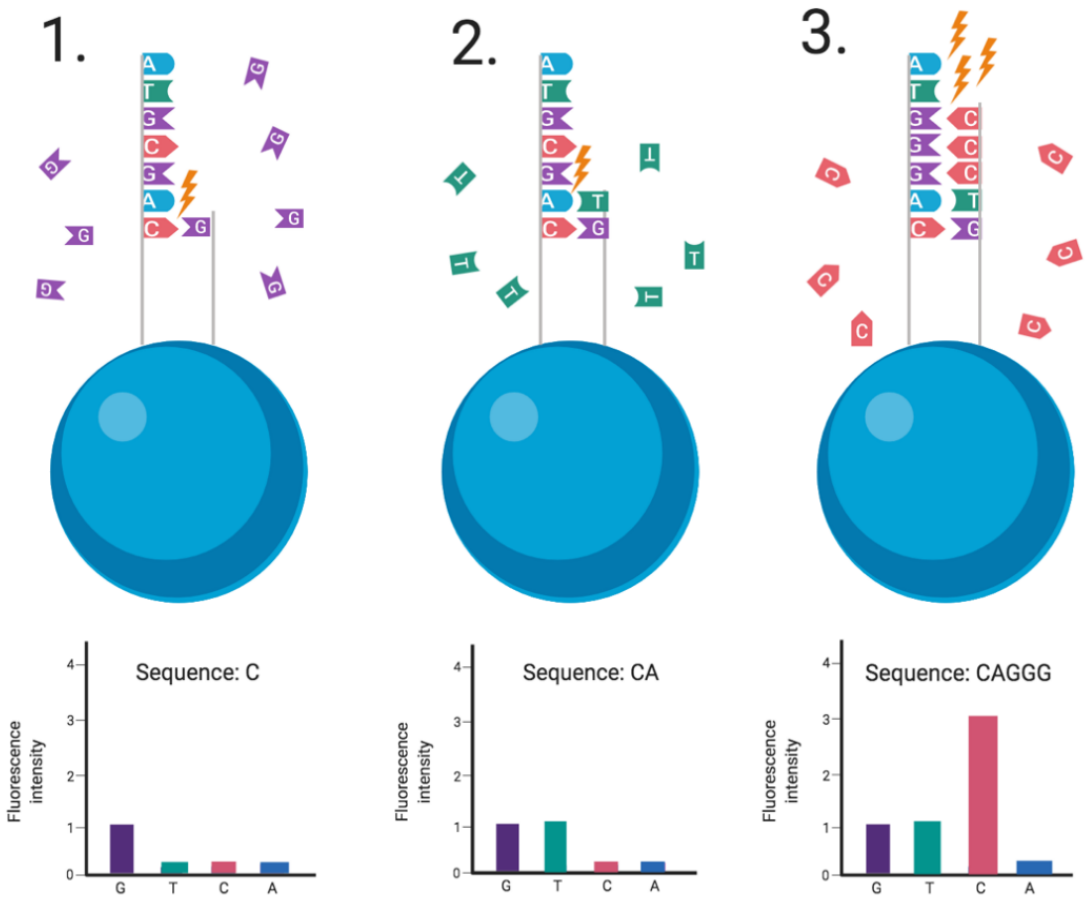


Figure 21. Sequencing cycle method in 454 sequencing. The blue spheres represent the tiny resin beads mentioned in **Figure 22** with the complementary sequence attached on the left-hand side of the sequencing strand. **1.** Represents the cycle where guanine nucleotides are added to the wells that house the specific sequence bead. Here, a guanine is incorporated into the sequence and light is emitted via the enzymatic reaction signified by the lightning bolt. The pyrogram below indicates that one guanine was added, which equates to one cytosine in the sequence (complementary base). **2.** Represents a second cycle where thymine nucleotides are added to each well and one T base is incorporated emitting light. It appears on the pyrogram below and the sequence is now CA. **3.** In the final sequence, three cytosine nucleotides are added to the sequence, emitting three times the amount of light, which is registered on the pyrogram, showing a final sequence of CAGGG.

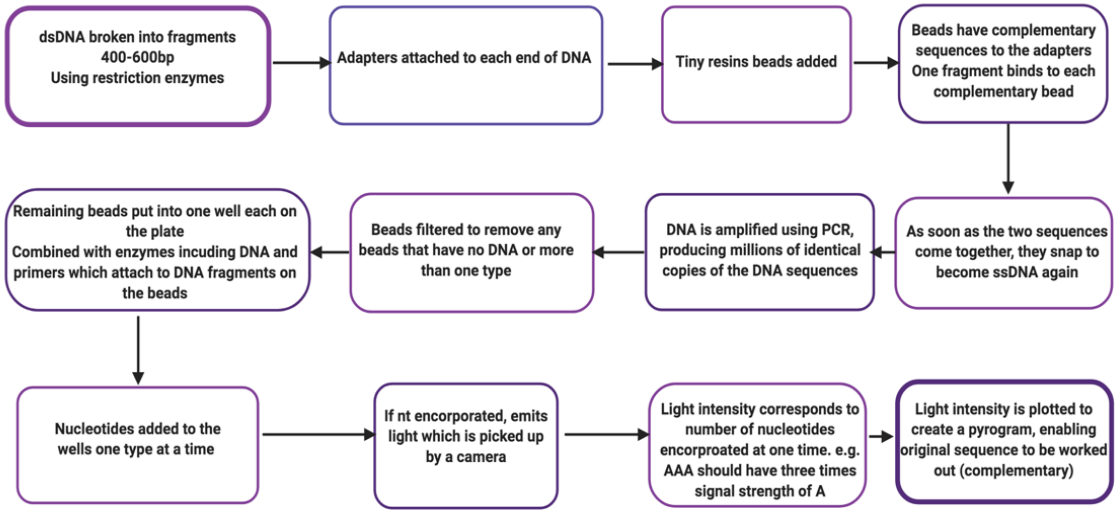


Figure 22. Overview of the 454-sequencing method. The 454-sequencing method involves a number of steps that are outlined in the workflow above. The first step starts with dsDNA being broken into smaller fragments. *BIOMED-2 primers* are used to amplify the TCRB region with a linker based on the "m13" sequence. A second round PCR binds with the linker and adds the sequencing primer sequence and MID unique identifier. The final step consists of plotting a pyrogram based on light intensity from nucleotide additions, also shown in **Figure 21**.

The main aim of the work carried out in this chapter was to assess as to whether the TCRB repertoire data produced by 454- sequencing provided evidence for and assessed the feasibility of TCRB repertoire studies in PNH. The secondary aim was to identify potential pitfalls in the data which could make biological inferences challenging and would need improving in the HTS method development (**Section 3.3**). Firstly, a bioinformatics workflow was developed to allow manageable amounts of sequencing data to be analysed and recreated into TCRB repertoires (**Section 2.9.1.**). This would provide insight into whether there were differences in TCRB repertoires between PNH patients and normals. The 454-sequencing data used *BIOMED-2 primers* [**154**]. As will become clear in **Chapter 4**, this was the preferred primer method for the subsequent high throughput sequencing analysis. The results in **Chapter 3**, therefore, also serve as a comparison for an adapted two step PCR method between two sequencing methodologies, 454 and Illumina®.

3.2. Results

Sequencing data from gDNA isolated from the peripheral blood of 10 healthy controls and 18 PNH patients was analysed in this chapter. TCRB was amplified using *BIOMED-2 primer sets* [**154**]. The 454-sequencing data had undergone quality control tests standardised by the sequencer software beforehand and only the sequences that passed these checks were analysed. The pipeline developed involved five main steps: pre-processing, quality control, alignment of the reads to a TCRB database, assembling TCRB clones and finally downstream analysis (**Section 2.9.1**. and **Figure 23.**). Pre-processing involved trimming sequencing adapters and aligning paired-end reads. Quality control steps trimmed sequences where the bases fell below a Phred score of 20. In order to annotate the reads with TCRB information such as TCRBV and J families, all reads were aligned to the IMGT® database [**85**]. TCRB clones then needed to be grouped from the data. These were defined as any reads annotated with the same TCRBV/J and CDR3 amino acid sequence. Once clones were assembled, any productive reads were then ready for downstream analysis. Analysis of TCRB repertoire data is vast and the methods selected provided different insights into the data to help aid further work.

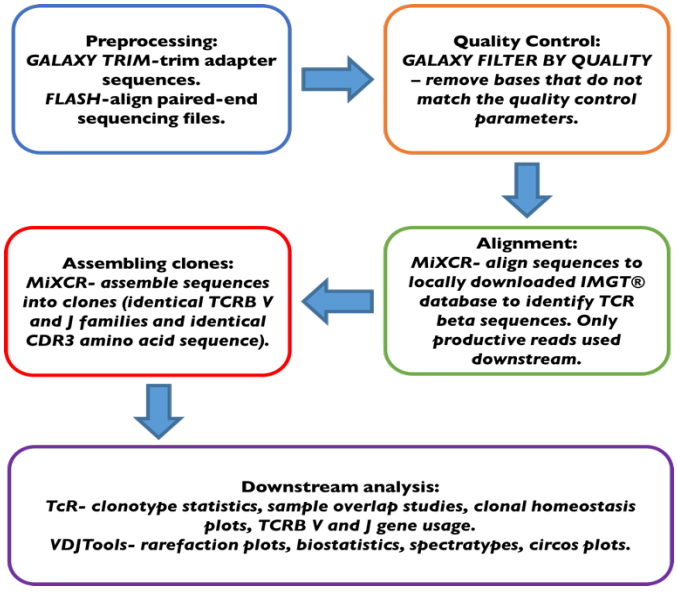


Figure 23. Pipeline developed to construct T-cell receptor beta repertoires from 454-sequencing data.

In total, 232,177 TCRB sequencing reads were analysed downstream for the control data sets and 12,884,563 TCRB sequencing reads in the PNH cohort. A number of analysis measures were carried out including TRCBV/J gene usage, CDR3 spectratyping plots, diversity measures including Chao1 estimators, inverse Simpson and d50 (**Table 14.**) along with clonal homeostasis analysis measuring the number of clonal TCRBs in a repertoire (data not shown).

Circos plots showing the diversity in TCRBV and J gene combinations in 4 PNH patients (**Figure 24.**) and graphs indicating TCRB V and J gene usage plots across the 18 PNH and 10 healthy controls (**Figure 25**. **and 26.**) have been included as examples of the data analysis performed.

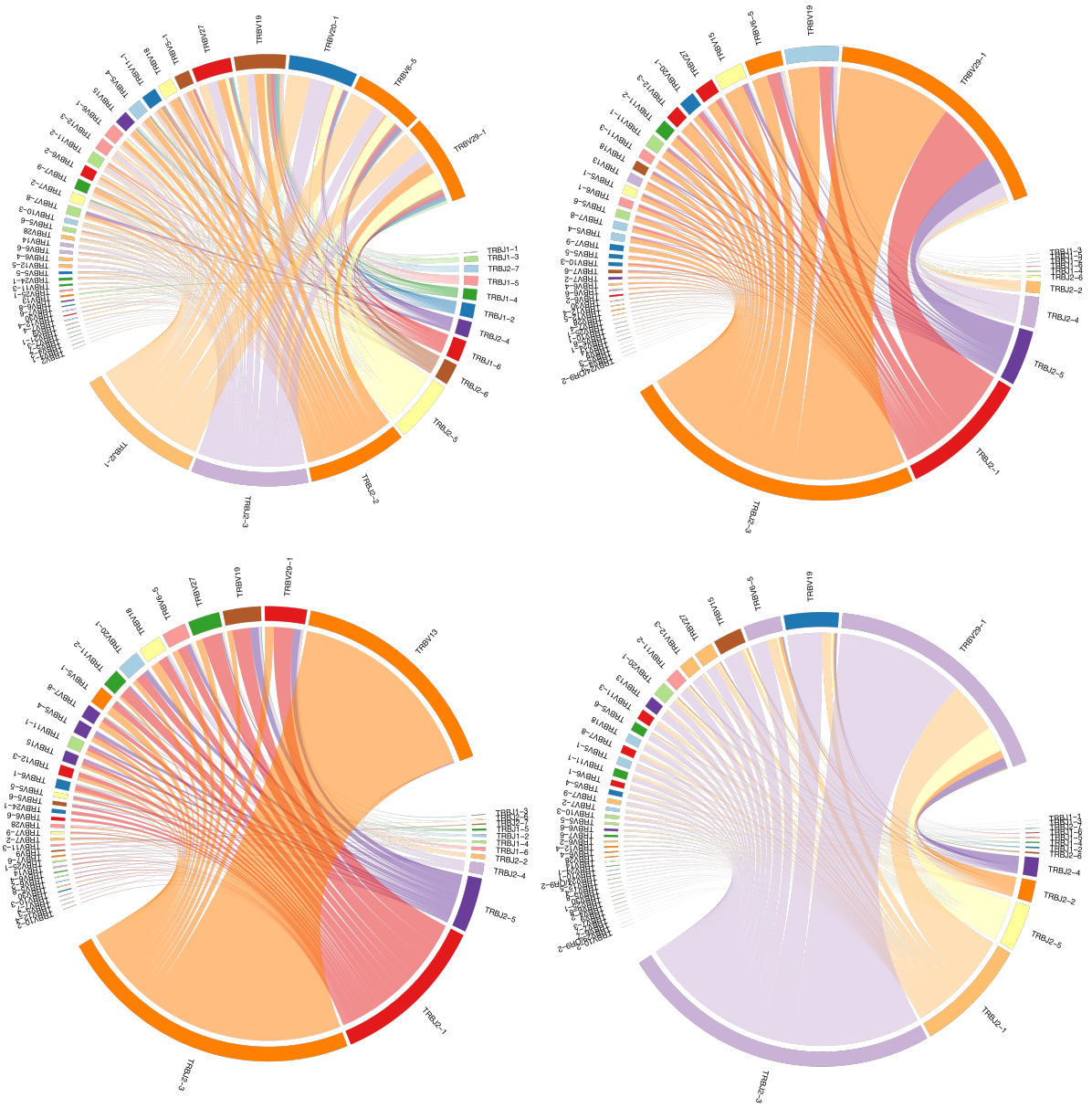


Figure 24. Circos plots of TCRB clonality in 4 PNH patients.

Showcasing the natural diversity and similarities in TCRB V and J families in PNH patients. Circos plots display clonality irrespective of CDR3. Each clone is mapped between a TCRB V gene segment and TCRB J gene segment. The width of the segment signifies the percentage of the overall TCRB repertoire represented by this clone. The wider the segment the larger the clone in the repertoire. PNH repertoires varied from patient to patient when assessing TCRBV/J combinations. The patient in the bottom left exhibited unusual V/J gene usage of V13/J2-3 rarely observed in other samples' TCRB repertoires at a clonal level in the data sequenced in this project.

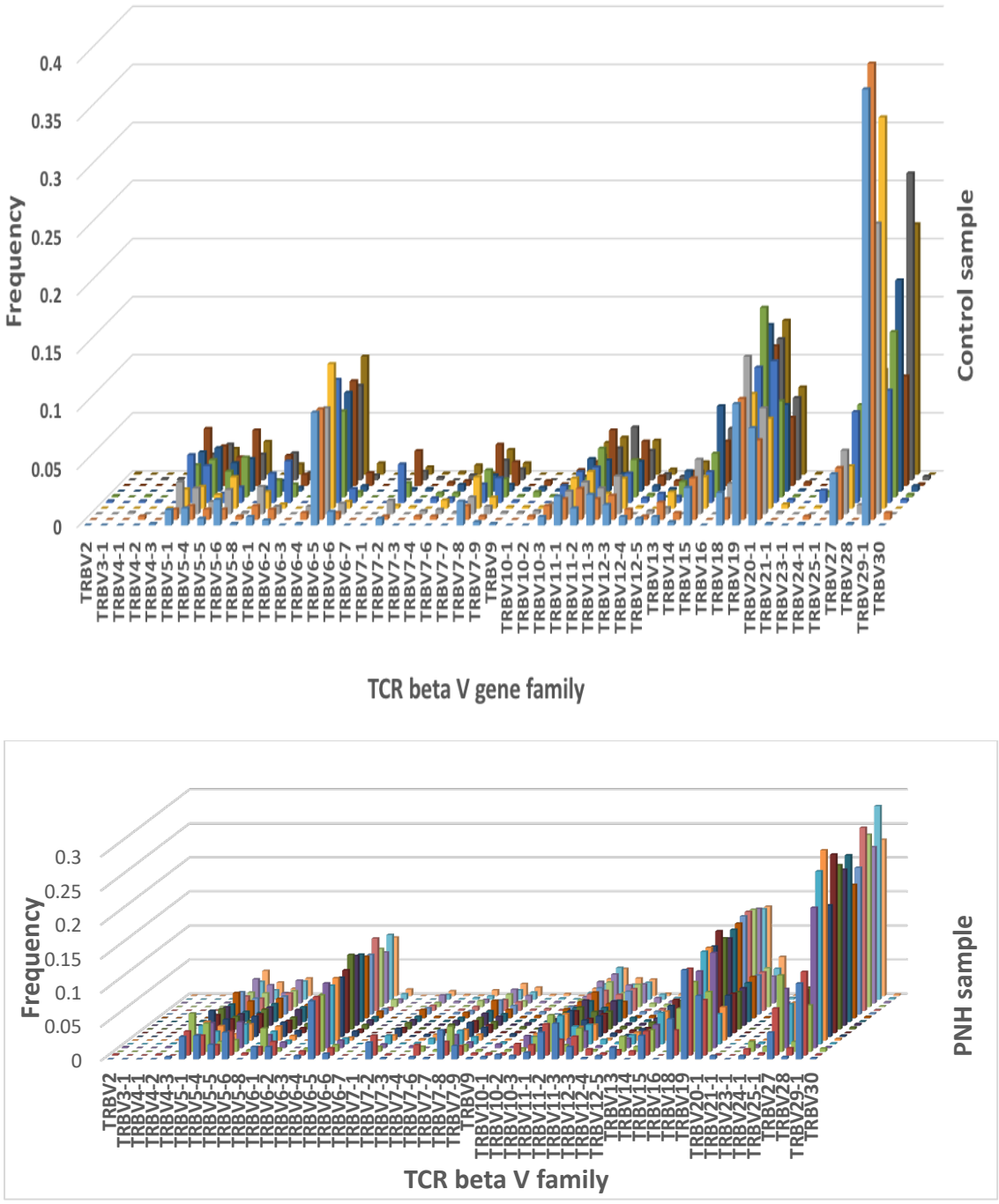


Figure 25. TCR beta V gene family usage in 10 healthy controls (top) and 18 PNH patients (bottom). Analysed from 454 TCR beta sequencing data. TCR beta V gene families were identified using the IMGT® database. The Z axis represents each control sample.

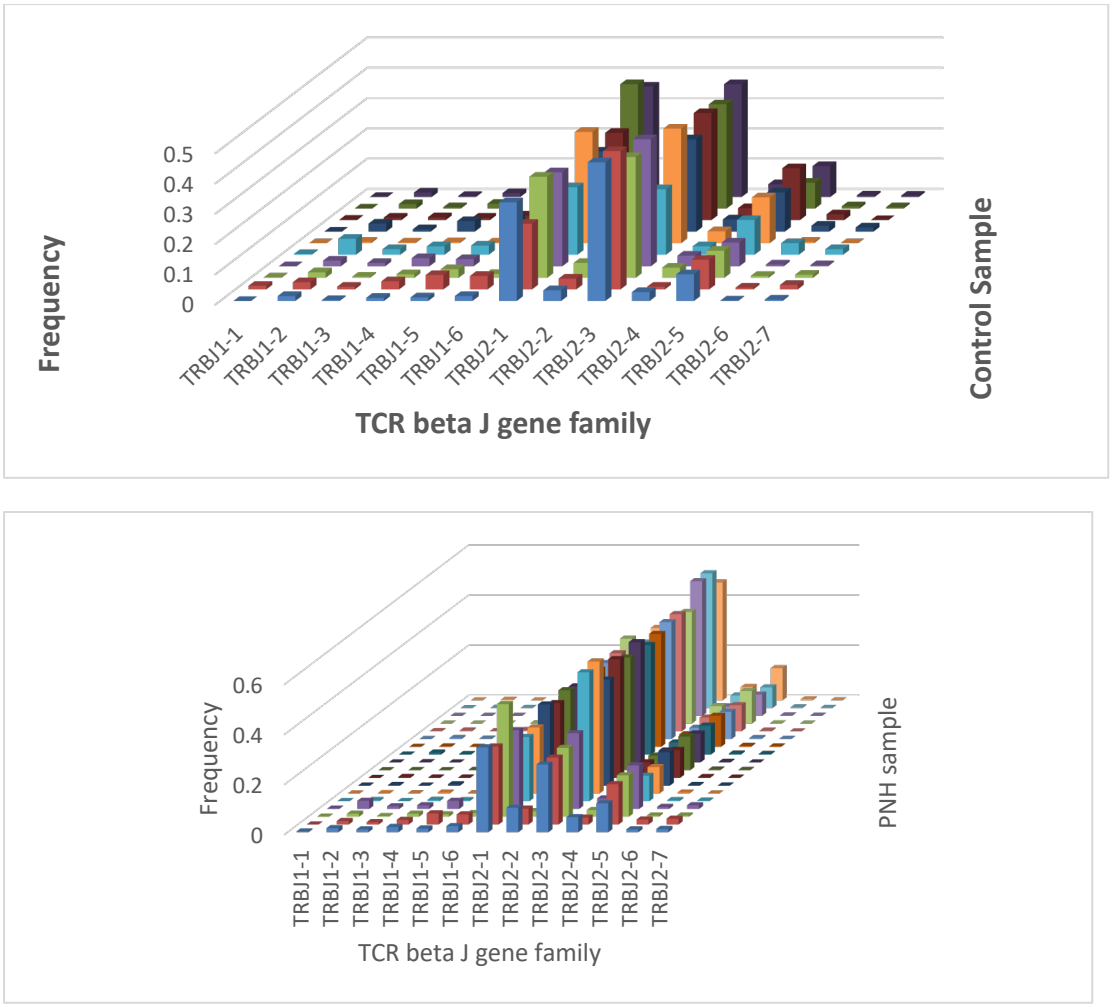


Figure 26. TCR beta J gene family usage I0 healthy controls (top) and 18 PNH patients (bottom). Analysed from 454 TCR beta sequencing data. TCR beta J gene families were identified using the IMGT® database. The Z axis represents each control sample.

3.3. Discussion

3.3.1. Summary of results, key findings and future project considerations

The aim of the work in this chapter was to provide evidence for pursuing a HTS TCRB repertoire method in PNH and identify potential caveats in the methods to make improvements in the HTS method. From the analysis of the 454-sequencing data it was evident how diverse TCRB repertoires can be and how they vary considerably between individuals regardless of disease status. Studies investigating what a "normal" TCRB repertoire looks like are few in number [**241**], highlighting the importance of having a control cohort sequenced alongside PNH samples. Having healthy controls sequenced alongside the PNH group using the same method also allowed for an accurate comparison between groups rather than using publicly available healthy control data generated using 5' RACE for example. In order to improve the TCRB repertoire analysis based on the findings in this chapter, the project focussed on expanding and improving the background for TCRB repertoire data from healthy controls to compare with PNH patients. The methodologies and findings will be detailed in **Chapter 4**. Generating a background of normals will also allow for a better definition of clonal expansions to be calculated in this project, as some normals will have clonal expansions as a result of current infections and other factors but the majority should not have clonal populations.

Both similarities and differences were observed between the PNH and control cohorts. Similar trends in TCRBV and J gene usage were observed with V19, V29-1 and J2-1, J2-3 being the more commonly used V/J gene combinations. On an individual level, different CDR3 sequences were identified as most common amongst the PNH patients when compared with the control group. Average CDR3 length was 44bp. It could suggest that a particular CDR3 is involved in PNH, but more research was needed to prove this and in a larger PNH cohort, with patients stratified according to clinical status (**Chapters 5 and 6**). In the PNH cohort 14 patients shared V29-1/J2-3 as their most abundant clonotype in line with healthy controls which makes deciphering pathogenic responses challenging. On average PNH repertoires appeared more clonal than the control group with some CDR3 distributions being skewed and not as normally distributed as in the control data-set, however if the normal group had equal numbers of samples this trend may have been observed. Non-normally distributed sequence lengths have been linked to problems, such as autoimmunity [**91**] which will be investigated further in the context of these PNH patients in this thesis.

Combining all TCRB reads in each cohort together was done to see whether this was a better technique for analysing TCRB repertoires than on an individual basis.

From a computational and analytical point of view it would also be less complex than comparing many different repertoires from many different samples. Using this method, TCRBV29-1 and TCRBJ2-5 with a CDR3 amino acid sequence 'CSALRRYSQETQYF' appeared in all 18 PNH and in 9 out of 10 controls as the top clone. This was a good example of a "public" clone. The problem with combining the reads, rather than looking at repertoires on an individual basis, means that any clonotype that is specifically skewed in a repertoire can be overrepresented in the combined reads. This will make a clonotype appear more common when in reality it is only present in one sample. Combining reads makes some analysis approaches more difficult, for instance looking at monoclonality in an individual's repertoire. Therefore, this approach was used with care in subsequent analysis.

Variation was observed in TCRB repertoires within both the PNH and the control cohorts. Variety across a human population is to be expected as TCRB repertoires are dynamic and affected by factors such as infection, age and HLA-type (**Section 1.4.**). It is important to be able to distinguish between, where possible, variation that is considered "natural" and variation that is considered disease specific. An example of the variation observed was one CDR3 sequence in the PNH group (also observed in the control group). Six PNH shared the CDR3 'CSVERGLSSYNEQFF' as their most abundant CDR3 but the majority were not hugely clonal compared with the controls. However, one of the PNH patients had a very highly expanded TCRB clone with that CDR3 sequencing accounting for 40% of the entire repertoire.

Seven PNH patients shared the CDR3 amino acid sequence 'CATSRVADTDTQYF' as their most abundant clonotype but this was only shared with 3 out of the 10 controls. Metadata was not available for these samples therefore it could not be established if perhaps larger clonal CDR3s were found in patients with progressive PNH compared to those with stable PNH. The sample sizes were large for PNH as it is a rare disease but on the small size to enable trends in cohorts to be identified. In **Chapter 5** and **6**, more patient samples were able to be sequenced, allowing the cohort to be split into different stages of PNH for example new, progressive PNH or large PNH clones that are stable. Metadata was available about the individual's age and sex in the subsequent results chapters which improved the analysis from this chapter.

When applying biostatistics to the data, it would be expected that the controls would have higher diversity of TCRB clonotypes in their repertoires, and based on previous research in PNH, PNH patients would have less diverse, more clonal, antigen skewed TCRB repertoires (**Section 1.7.**). This was not the case with the 454-sequencing data.

However, a possible cause of this is variance in the sequencing depth. PNH samples on average had five times the sequencing reads than the normal controls (data not shown). This allowed for the detection of more TCRB clonotypes and, therefore, the PNH repertoires appeared more diverse when using the biostatistics, which used measures such as TCRB clonotype. Therefore, PNH patients may have biologically more diverse TCRB repertoires than normals, but to assess this, the technical variance in sequencing depth would have to be improved, so that similar reads were achieved for most samples. Although, this is not an exact science, one way to achieve this would be to ensure equimolar TCRB gDNA from each sample is going into the final TCRB sequencing library. This was incorporated into the HTS technique (**Section 2.8.7.**). Another reason for the difference in sequencing depths or differences observed in the number of TCRB clonotypes could be attributed to the concentration of genetic input in the initial TCRB amplifying PCR reactions. The concentrations in the 454-sequencing library were not known. However, in the future HTS methods in the next chapters these factors were assessed and taken into account to achieve better sequencing depth comparisons between samples (**Section 4.3**).

Interestingly, despite increased sequence depth in the PNH cohort which allowed for the detection of rarer clonotypes, Chao1 estimators counting the number of TCRBs missed by sequencing, was higher. This may be down to the way the estimator is calculated. Possibly, this finding combined with the fact that more and more clonotypes were discovered as the sequencing reads in PNH increased, may suggest that in PNH samples there are a greater number of rarer clonotypes. The trend observed in normals, however, could be suggestive of the need for exhaustive sequencing (increasing sequencing depth no longer increases the number of TCRB clonotypes) [**121**] although the HTS have increased sequence depth which may solve this. In repertoire studies underrepresentation of the entire individual's TCRB repertoire will always occur as it is not possible to capture the diversity from one blood sample [**106**]. However, underrepresentation in the sample can be reduced, with measures such as sequencing the samples multiple times. This can involve taking multiple samples from the gDNA and amplifying and sequencing separately, or taking an amplified sample, splitting before sequencing, then sequencing multiple samples to collate TCRB clonotypes. These methods will be carried out for the HTS sequencing method (**Section 4.3**).

3.3.2. Chapter conclusions

A bioinformatics pipeline was successfully developed to analyse 454-sequencing reads to create TCRB repertoires. These repertoires allowed for the comparison of TCRB features within and between PNH and healthy control cohorts. The data served as a useful comparison with the PNH and normals in **Chapters 4**,**5** and **6**. Although 454-sequencing allows for longer read lengths than Illumina®, which is ideal for sequencing the length of a TCR without the need for paired-end reads, 454 has a problem with homopolymer repeats [**151**]. For example, if multiple adenines are incorporated in one cycle, it is hard to discriminate between 3 or 4 of them, as the bright light will not necessarily become more intense. Therefore, many errors occur in these regions, this effect is lessened with Illumina®. The 454-sequencing method can sequence theoretically, up to 20 million bases per run. In reality about 200,000 sequences is the maximum with sequence lengths significantly lower than 1000bp [**150**]. High throughput sequencing technologies such as the Illumina® MiSeq are capable of generating around 20 million reads per sequencing run [**259**]. This generates a greater sequencing depth per sample than the 454-sequencer, potentially allowing more TCRB clonotypes to be identified in a given sample. This is why the project will now focus on high throughput sequencing methods for future TCRB repertoire analysis.

The analysis in this chapter highlighted both similarities and differences between and within the two cohorts. To investigate this further, the project needed to progress in the direction of using high throughput sequencing technologies such as Illumina® sequencing, analyse more PNH patients' TCRB repertoires and create a "normal" background for TCRB data to allow accurate comparisons and conclusions to be made about TCRB repertoires in PNH. The addition of metadata for patients and normals was essential for project progression. The results of which will be discussed in the next chapters.

Chapter 4 - Optimisation of TCRB high throughput sequencing methodologies and defining a 'healthy' TCRB repertoire

4.1. Introduction

The 454-sequencing results discussed in **Chapter 3** showed the potential for the use of genomic sequencing in deciphering further the role of T-cells in Paroxysmal Nocturnal Haemoglobinuria. It allowed the development of an analysis pipeline and highlighted important considerations to be made when designing and implementing the high throughput sequencing method. This included factors such as patient and control sample sizes, the importance of clinical data, read depth, cell number calculations and consistency between sequencing library runs. As discussed, there were, however, some aspects of 454 sequencing, such as the homopolymer errors and low throughput of sequencing reads, that made the method less efficient than some of the next-generation sequencing techniques now accessible in research.

In order to generate high throughput sequencing for this project, methods were developed using Illumina® sequencing. In Illumina® sequencing technologies DNA that has undergone library preparation (**see methods**), undergoes cluster generation on sequencing flow cells using the process called "bridge amplification". The DNA is then sequenced by the method "sequencing by synthesis". This involves the incorporation of a fluorescently labelled nucleotide into a nucleic chain during a sequencing cycle. This acts as a terminator stopping any more nucleotides being added to the chain. After each nucleotide is incorporated, the fluorescence produced is imaged to identify the base [**160**].

Although Illumina® sequencing produces short reads compared to 454-sequencing, there are many advantages to using this technology. In this project the MiSeq sequencer was utilised which sequences DNA to produce 20 million reads per sequencing run [**259**] providing high throughput results. Illumina® routinely uses paired-end sequencing technologies where the same DNA sequence is sequenced in both the forward and reverse directions, allowing these reads to be overlapped, providing a consensus sequence with accuracy. Sequencing by synthesis rather than pyrosequencing has fewer errors in homopolymer regions [**151**].

Before comparing T-cell receptor repertoires of PNH patients, it was important to test the methodologies on healthy controls. Firstly, to assess whether the methods were working and producing accurate results, with similar TCRBV/J usage profiles to normals in **Chapter 3**. Secondly, to account for natural variations that can be observed in populations and trends that may also appear in PNH patients unspecific to the disease.

As discussed in the main introduction, T-cell receptor repertoires are dynamic. They are easily affected by factors such as disease, age, sex and stress (**Section 1.4.**). Even though PNH does not seem to have an age bias in current published literature [**194**], it is still important to assess age when observing the normal cohorts. Sex can also play a role in the differences observed, for instance females seem to have a bias towards CD8+ T cells which links with their increased susceptibility to autoimmune diseases [**261**]. It is therefore important, where possible, to investigate these findings in the healthy control cohort of this project.

The bioinformatics analysis pipeline in **Chapter 3** was significantly expanded in this chapter for a number of reasons. In the 454-sequencing data-set there were approximately 13 million reads. Using Illumina® sequencing, one sequencing run produced over 20 million reads. The high throughput nature of the data led to coding and scripting being introduced into the pipeline rather than using online servers such as Galaxy [**262**]. This allowed for the analysis to be more efficient, reproducible and involve a greater range of possible analyses of the data, giving detailed TCRB repertoire results. A greater number of analysis types were also introduced into the new pipeline, as the increased sequencing depth allowed these to be successfully implemented. As the project aims for this chapter involved the development of the bioinformatics pipeline detailed and summarised in **Section 2.9.2**, for completeness, rationale for subsequent analysis and how important parameters in the pipeline were obtained, will be detailed in this chapter, along with the results from the samples themselves.

Carrying on from the work in **Chapter 3**, important concepts needed to be defined including how to define a TCRB clonotype experimentally. Using the optimised sequencing methods to accurately define hyperexpanded TCRB clones would also be important when investigating a suspected T-cell clonally expanded disease response in PNH.

The main aim of the work in this chapter was to adapt the T-cell receptor repertoire sequencing methods in **Chapter 3** to high-throughput next generation sequencing technologies and to use the techniques to create a base line of values established from healthy controls.

Generating larger volumes of sequencing data, firstly on healthy controls, would provide more accurate information on T-cell receptor repertoires in normals and natural variations in populations. This is with the aim of providing benchmarks for comparisons when Paroxysmal Nocturnal Haemoglobinuria samples are analysed, to decipher possible PNH specific changes in T-cells, which will be discussed in **Chapter 5 and 6**.

The primary objective of this chapter was to develop an accurate and efficient way to experimentally amplify and sequence TCRBs in DNA samples. The secondary objective was to then adapt and improve the bioinformatics pipeline from **Chapter 3** to successfully analyse the sequencing reads in order to create TCRB repertoires. The pipeline would need to be robust and streamlined to process the high throughput sequencing data quickly and efficiently. The final objective was to compare the TCRB repertoires of 31 healthy controls to understand better natural changes in the TCRB repertoire with factors such as age and assess potential variations and caveats that may be attributed to the experimental methods themselves.

4.2. Results - Comparison of two methods for the amplification of TCRB

In this project two methods for amplifying the regions of the T-cell receptor beta chain were optimised and performed on healthy controls. This would help identify any discrepancies in the data caused by the specific methods rather than natural variation in the repertoires.

The first method was based around the use of 45 forward TCRBV gene primers and 13 reverse TCRBJ gene primers from a paper published by Robins et al. (2009) and will be referred to as the *Robins et al. primer method* [**238**]. This paper was one of the most highly cited TCRB sequencing method publications and the method is now part of one of the commercial market leads for TCRB sequencing [**238**]. The second method was based on the primer sets from the EuroClonality/BIOMED-2 consortium using 23 TCRBV primers and 13 TCRBJ gene primers [**154**] which will be referred to as the *BIOMED-2 primer* method from here onwards. The primers amplify all functional TCRBV/J genes. This method was developed and validated by a large consortium of researchers and has been routinely used in diagnostic laboratories for many years.

It was important to test two methods for the TCRB PCR amplification as this process can produce many biases which can artificially skew the identification of TCRB clones in a repertoire sample. Any biases caused, for example one TCRBV primer having a higher binding affinity to gDNA than others, would be amplified in the repertoire sequencing reads due to the exponential nature of PCR amplification. By adapting two well established methods and comparing the results, it would help identify any potential discrepancies or inaccuracies in TCRB repertoire sequencing.

4.2.1. TCRB V and J gene usage differed depending on TCRB gene primer method

When observing TCRB V and J gene usage across control samples when using the *Robins et al.* and *BIOMED-2 primer* methods, it would be expected to observe similar trends in high and low gene usage, as well as high diversity in normals. However, large differences in the repertoires were observed dependent on the method used (**Figure 27** and **28**.).

Despite being adapted from 454 sequencing to Illumina®, the *BIOMED-2 primer method* produced similar V family trend results to that of the 454 sequencing in **Chapter 3**. Both methods showed V29-1/J2-1 and J2-3 as the most common combinations across PNH and healthy control cohorts, highlighting that the HTS adaptation of *the BIOMED-2 primers* for Illumina® did not affect or skew their performance. The *BIOMED-2 method* (**Figure 27, bottom**) indicated a variety of TCBV gene usage observed across 30 healthy controls. V29-1 (blue box) and V19 (yellow box) sharing the greatest usages across the repertoires with some samples showing over 25% usage in the repertoire. The trends were also similar in usage across the normals.

When using the *Robins et al. primer method*, there was a skewing towards TRCRBV 7 families (**Figure 27, top**). which was considerably different from the normal repertoires observed with the 454 sequencing and the *BIOMED-2 primer method*. V7-2 had the highest usage using this method, in one sample accounting for over 30% of the entire V beta gene usage. This was followed by V7-6. Apart from some gene usage across the V6 family primers, the majority of the other 23 families saw little to no usage across the repertoire. V19 (yellow box), and V29-1(blue box), families with the highest family usage using the *BIOMED-2 method*, again showed little and no usage respectively in the *Robins et al. primer method*. The origins of the V7-2 skewing, biological or experimental, were further investigated and the results detailed below.

In agreement with the 454 sequencing in **Chapter 3**, TCRBJ 2-1 and 2-3 had the highest usage in an individual TCRB repertoire across 30 heathy controls (**Figure 28, bottom**). Only 11 of the 13 genes were shown as J1-1 and 1-3 were not present or rare in clonotypes across the healthy controls. Using the *Robins et al. primer method*, J2-1 and J2-3 were also amongst the highest genes used in an individual's repertoire. J1-1 and J1-3, were present but at varying frequencies and generally the lower of the TCRBJ gene usage observed. Interestingly, the J2-5, 2-6 and 2-7 were highly used in repertoire using the *Robins et al. primer method*. **Figures 27 and 28.** show 3 of the samples analysed using the *Robins et al. primer method* for simplicity and were selected as they show the general trends observed across all the samples. In conclusion, there were some similarities shared across the two methods for TCRBJ gene usage but also significant differences.

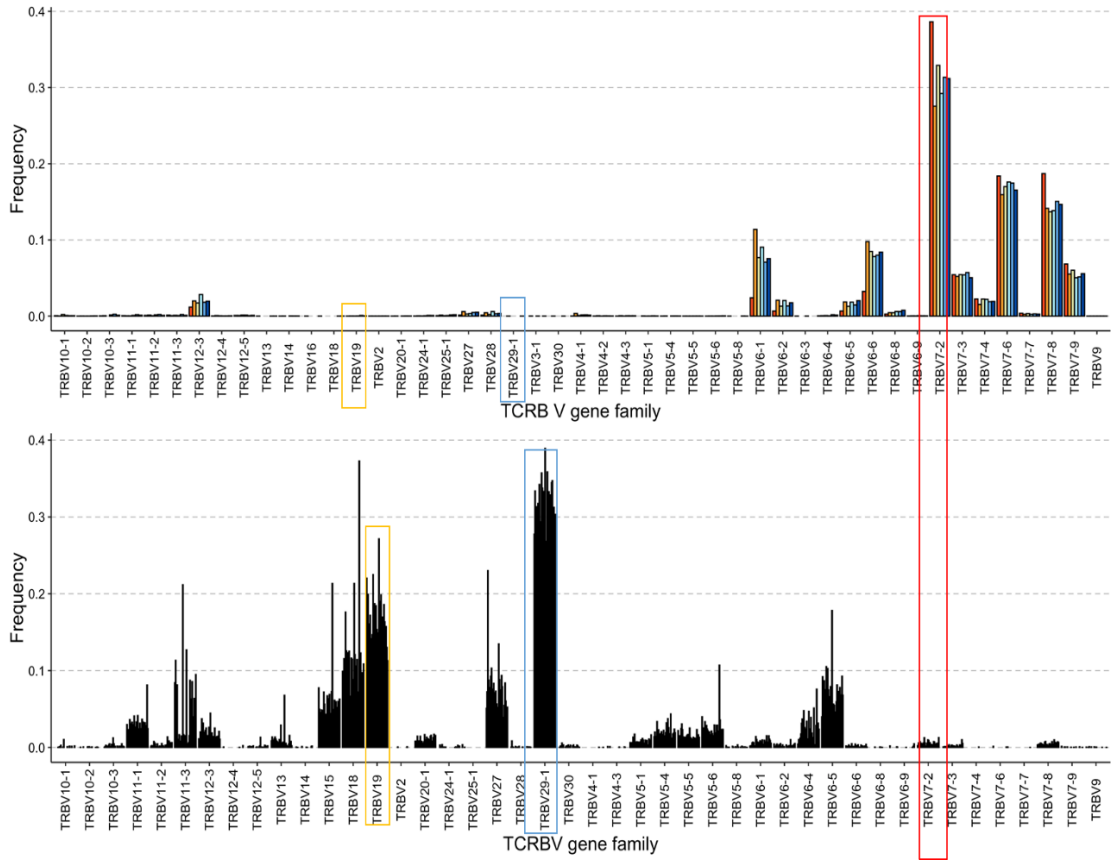


Figure 27. TCRB V gene usage in healthy controls. TCRBV gene usage values from six randomly selected control samples using the *Robins et al. primer method* (top) and in 30 healthy controls using the *BIOMED-2 method* (bottom). Only six samples are shown for the *Robins et al.* method for simplicity showing the general trend observed across all samples. The V gene family 7-2 is highlighted with a red box in both plots, V29-1 blue, and V19 in yellow to highlight differences in the primer methods. The number of TCRBV gene families on the x axis differs as each method uses a different number of TCRBV gene family primers. The y axis indicates frequency of the TCRBV gene in the sample's TCRB repertoire.

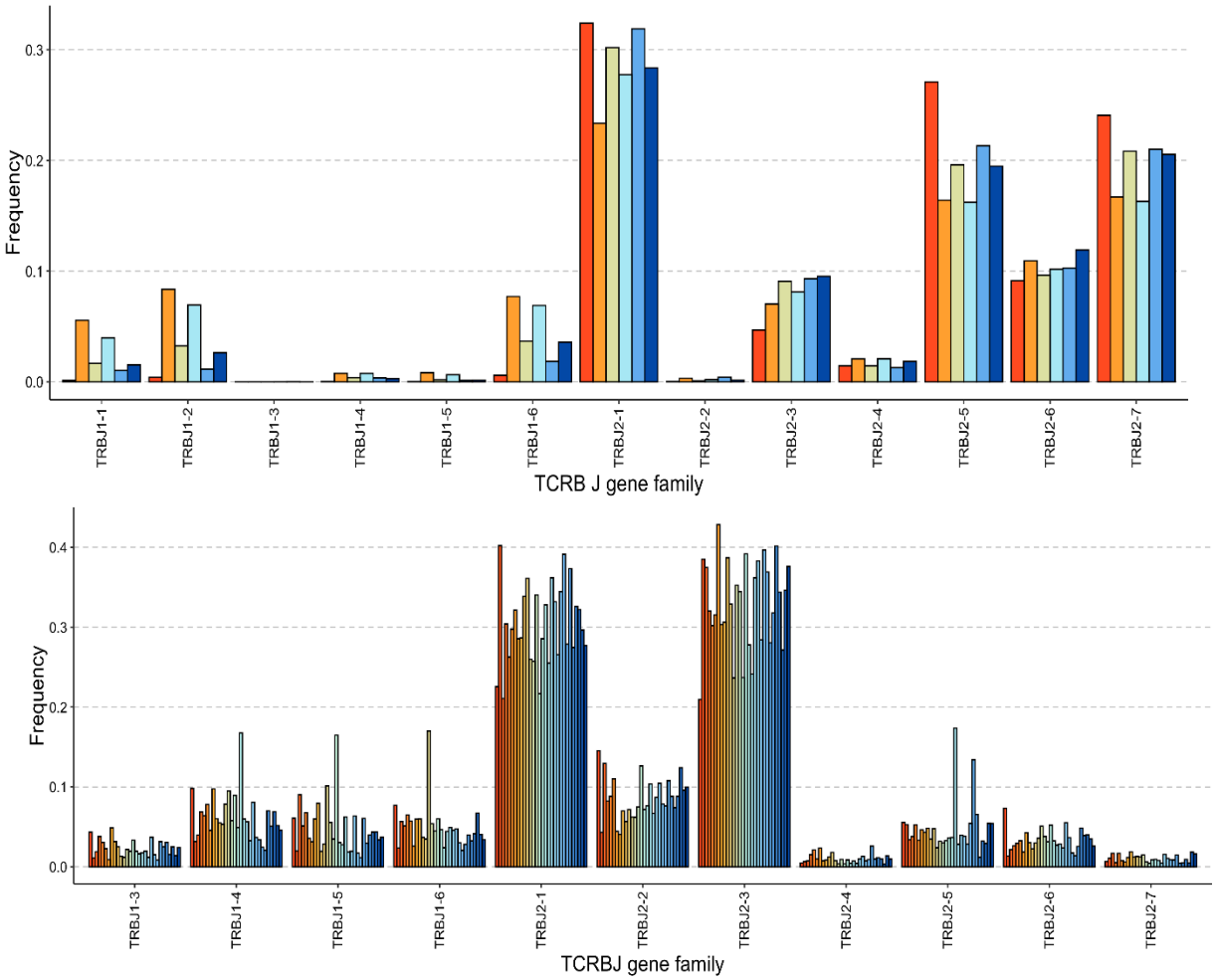


Figure 28. TCRB J gene usage in healthy controls.

Values from six randomly selected control samples using the *Robins et al. primer method* (**top**) and in 30 healthy controls using the *BIOMED-2 method* (**bottom**). Only six samples are shown for the *Robins et al. primer method* for simplicity showing the general trend observed across all samples. The y axis indicates frequency of the TCRBJ gene in the sample's TCRB repertoire.

4.2.2. Investigating the source of the TCRBV 7-2 skewing in the *Robins et al. primer method*

As there was no literature detailing a V7-2 skewing/bias when using the *Robins et al. primer method*, in order to optimise the method for use in this project, the source of the TCRBV7-2 primer skewing was investigated to assess whether it was biological or technical.

4.2.2.1. Type of DNA and concentration of DNA input did not lead to V7-2 skewing

The method was optimised using varying concentrations of DNA input using 100, 200 and 400 ng per 20μl PCR reaction. If there was too little or too much DNA input, the primers would either bind non-specifically or be exhausted and not perform. However, concentration of DNA did not affect the V7-2 skewing observed. Three sources of DNA input into the primer reaction were tested. One was gDNA, and the other two were cDNA generated using two different methods. One method used a poly T primer during synthesis, and another used a TCRB C region specific primer as detailed in **Section 2.5**. Regardless of the DNA input the extreme V7-2 skewing was still observed. Therefore, DNA concentration or type of DNA used was not causing the V7-2 skewing.

4.2.2.2. PCR conditions had no obvious effect on V7-2 skewing

The PCR reactions were optimised, using the brightness of the product band as an indicator of product yield. The PCR annealing temperature, Mg^{2+} concentration, and sequencing adapter concentration were optimised. Optimisation steps could not feasibly be carried out on all combinations of V/J primers so the two most common combinations, V20-1/J2-1, V5-1/J1-1 were tested in the optimisations. PCR product bands were observed in the optimisation rounds (data not shown) for both primer combinations. However, in the sequencing summarised in **Figures 27** and **28**, J1-1 was present in very low numbers, J2-1 had high usage, V20-1 and V5-1 were either present at a very low abundance or not present at all across the samples. The samples used in the optimisation to generate the DNA input were also used in the sequencing, potentially suggesting the V7-2 skewing is caused by a problem when all primers are combined in the multiplex PCR reaction to amplify TCRB. This can be supported by the observation that during the processing of the sequencing data, the insert length between forward and reverse primers varied dramatically, and some sequences were so short that they terminated before the end of the sequencing length and therefore were much shorter than expected to be as detailed in the original paper [**238**].

4.2.2.3. TCRBV7-2 skewing was not caused by a batch effect, sequencing or technical effects

Over the course of the optimisation using *Robins et al. primer method*, over 150 samples were sequenced across 4 sequencing runs from 40 different biological samples. Different primer aliquots were used across the samples, therefore, not having all primers present in the multiplex can be ruled out. The methodology was also used by a collaborator on a disease set who also observed the same V7-2 bias, ruling out batch effect, or sample specific biases. All optimisation experiments were performed in triplicate, and there was no difference in the results observed, V7-2 skewing was still present.

4.2.2.4. Reducing TCRBV7 family primers, or removing them altogether still caused skewing

In theory, if VB7 families for some unknown reason were causing the skewing alone in the system, when the primer concentrations were lowered, or primers left out completely, a more "normal" repertoire should be observed. However, this was not the case (**Figure 29.**).

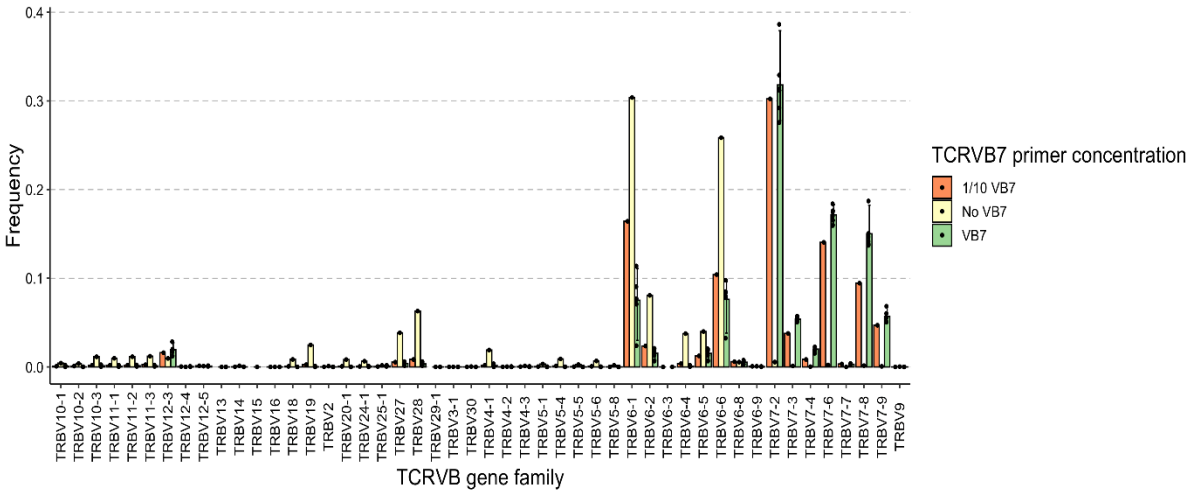


Figure 29. TCRBV gene usage with varying concentrations of TRCBV7 family primers.

Samples were amplified using the *Robins et al. primer method*. The samples either had normal concentrations of TCRBV7 families as described in **Section 2.5.** (green, n=6) a tenth of the original concentration (orange, n=1) or no TCRBV7 family primers present in the initial TCRB gene multiplex amplification PCR reactions (yellow, n=1). Not all sample data is shown for simplicity but the samples selected reflect the general trend across all samples.

When all TCRBV7 family primers were added to the PCR multiplex at a 10^{th} of the concentration originally used (**Figure 29.,** orange bars), high V7-2 usage was still observed. This accounted for over 20% of the overall repertoire and similar trends in overall TCRBV7 family usage were observed when compared with the samples with normal concentrations of TCRBV 7 families (**Figure 29.,** green bars).

Interestingly, higher usage across the TCRBV6 family was observed in samples with the lower TCRBV7 primer concentration than in the original. When observing results that contained no VB7 family primers, no VB7 family usage should be observed. However, as **Figure 29.** shows (yellow bar plots) there were very small levels of TCRBV7 family usage in these samples, potentially attributed to "tag swapping" **(Section 7.4.1.5.**). VB7-2 was significantly lower than in the other samples representing less than 2.5% of the repertoire. With no VB7 family primers, there was a significant increase in the usage across the TCRBV6 family. The TCRBV6 family was also present when the primers were included but at a second rate compared to VB7 family usage.

A lack of TRCRBV7 family amplification lead to an increase in the amplification of many other TCRBV families as to be expected in a normal repertoire, with TCRBV20-1 appearing, along with V19 and V28 amongst others. However, these levels of usage were still low compared to the *BIOMED-2 primer method*, with a skewing now being observed towards the TCRBV6 family, although not as severe a skewing as when using the V7 families and observing V7-2. In conclusion, varying the TCRBV7 primer concentration gave insight into why the skewing was occurring and potentially it is an issue inherent with the primer system.

4.2.2.5. *Robins et al. primer method* saw higher numbers of unique TRCB clonotypes despite similar sequencing depth

The *Robins et al. primer method* produced a large number of unique TCRB clonotypes in each sample, with the majority of values above 1000, and many above 2000. Some values were observed over 4000 per sample. These were consistently higher than those observed using the *BIOMED-2 method*. There were no significant discrepancies in sequencing depth between the two methods which might have been the cause for the differences. On average TCRVB7 primer produced most unique clonotypes, followed by 1/10 VB7 followed by noVB7 which saw much lower values, one of which was 651 clonotypes.

4.2.2.6. Decreasing the concentration of TCRBV7 family primers in the amplification TCRB PCR significantly decreased the number of shared clonotypes using the *Robins et al. primer method*

The "public" method was used to calculate the overlap of TCRB repertoires by calculating the number of shared TCRB clonotypes between two samples (**Figure 30.**). If both samples contained the same clonotype, the sums of this clonotype were added and the number represented in the grid. The higher the value, the higher the degree of sharing between the samples and the grid will be a darker red. The lowest values for shared clonotypes were observed when comparing the sample with no TCRBV7 families in the TCRB amplification PCR with the samples which have normal concentrations of TCRBV7, with values between 200 and 300 shared clonotypes. This was considerably higher than between individual control TCRB repertoires amplified using the *BIOMED-2 primer method* (**Figure 39.**)

Slightly higher shared clonotype numbers were observed between this sample and samples with a tenth of the concentration of TCRBV7 primer at 500 TCRB clonotypes. The samples with a tenth of the concentration of TCRBV7 saw a greater degree of shared clonotypes with the samples with normal concentrations of TCRBV7 at numbers between 700 and 1000 which was the highest value observed. When comparing the normal concentration samples, these had the highest degree of shared clonotypes with a range between 800 and 1000, with two thirds of the values being the maximum of 1000.

The general trend was that those using the normal TCRBV7 concentrations showed high similarity, this similarity in the shared number of clonotypes decreased as the concentration of TCRBV7 primer decreased. It was also significantly less when no TCRBV7 primer was present. A similar trend was observed when using the "overlap" method which measures the overlap between two samples. The size of the intersection between two samples is divided by the smaller sample.

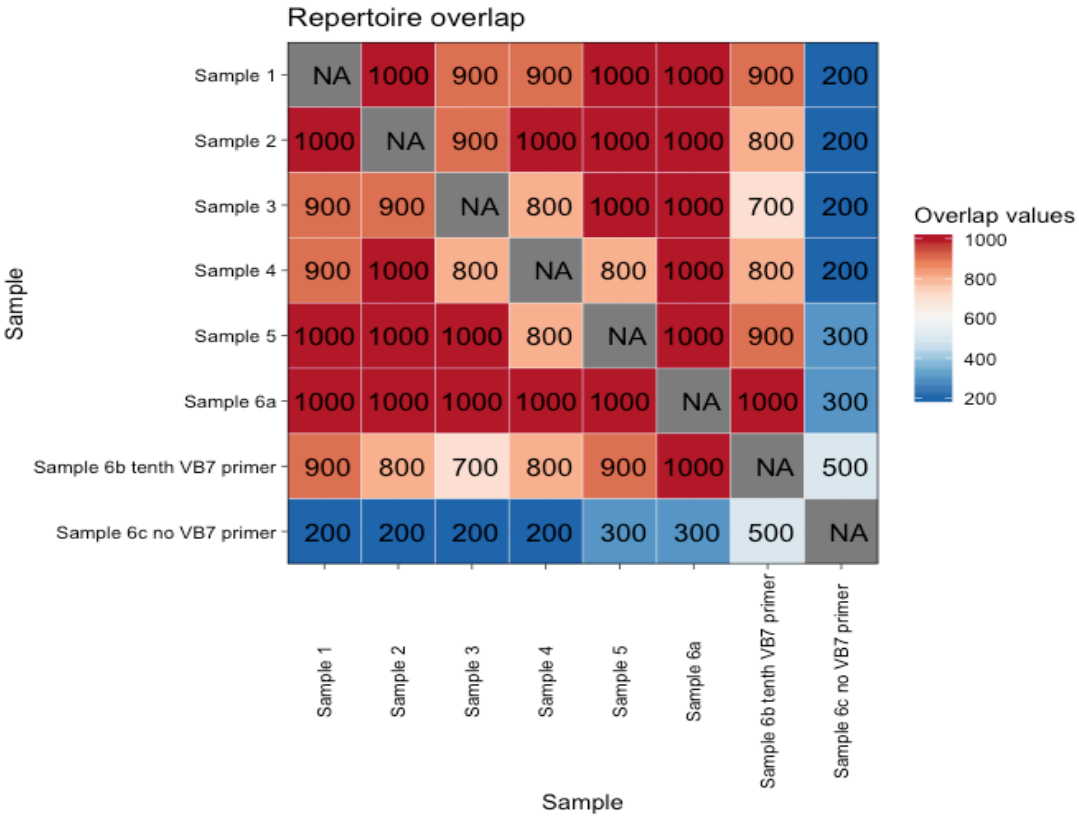


Figure 30. Overlap studies measuring clonotypes shared across samples using the *Robins et al. primer method*. The grid axis indicates the samples and the number in each square is the "public" overlap metric shared between two pairs of samples. The redder the scale the higher the degree of overlap. The grid uses the measure of shared TCRB clonotypes between two samples, where the clonal counts of the same clonotype in each group will be summed. The sample numbers: 1-5 and 6a were samples containing the standard concentration for all 45 TCRB V primers from 6 randomly selected "normal" gDNA samples representing the large number of samples sequenced to test this primer method. The last two samples used the same sample gDNA as sample 6a, but the TCRBs were amplified using a tenth of the original concentration of TCRBV7 family primers (6b) in the amplification TCRB PCR and using no TCRBV7 family primers (6c) in the TCRB amplifying PCR reaction.

4.3. Optimisation of *BIOMED-2 primer method*

As the *Robins et al. primer method* could not be optimised to reduce the V7-2 skewing, it was decided that the project would continue to solely use the *BIOMED-2 primer method*. Further optimisation to the method was needed before using disease samples. Factors including sequencing errors, PCR errors and differences in biological or technical replicates were investigated, the results of which are detailed below.

4.3.1. Increasing gDNA concentration increased the number of TCRB clonotypes

In the preliminary sequencing runs the concentration of gDNA inputted into the TCRB amplifying PCR reaction was tested using concentrations of 100ng/20µl, 200ng/20µl and 400ng/20µl per PCR reaction. As expected, doubling the concentration increased the number of clonotypes observed in the TCRB repertoire by at least double. Some diversity measures that used clonotype number in the calculations therefore also observed an increase with increased gDNA concentration.

One sample with a known clonal population had its repertoire measured using both 200 and 400 ng of genomic input. The 200ng sample had three replicates, one was a sequencing replicate taken from its parent sample just before the indexing reaction, and the other was a separately amplified sample run on a different sequencing run. TCRB clonotypes that passed processing steps were very similar ranging from 1050 to 1200. The d50 diversity measures ranged from 186-194, Chao1 estimates between 1048 and 1198 and Inverse Simpson estimations between 276 and 316. All samples had one clonal TCRB which was identical and at a percentage between 3.4 and 4.4 of the entire repertoires.

The same patient whose repertoire had been generated using 400ng, showed a greater number of clonotypes, over four times that of 200ng at 5020. D50 was higher at 349 and inverse Simpson was 553. Chao1 estimator was not applicable. There were no clonal TCRBs and the top clone was not identical to the other samples, at 1.2% of the repertoire. Another sample with replicates at 100, 200 and 400ng observed a similar trend. All replicates had the same TCRB clonotype as their top clonotype. This was not clonal for 200 and 400ng samples but was in 100ng, this transpired to have occurred due to the low number of TCRB clonotypes passing the processing steps (177) compared to 1150 and 2227 for 200 and 400ng respectively. A similar sample also saw the trend with one clone for the 100ng and none for 200 and 400ng, again due to the low TCRB clonotypes in the sample causing artificial inflation. However, this was not always the case as another sample with replicates at 100ng and 200ng, had double the number of TCRB clonotypes at 200ng, but both shared the same TCRB clonotype as the top clone varying between 1.5 and 1.7%. The second clonotypes were not identical.

4.3.2. Evaluating the pros and cons of gDNA concentrations

Although increasing the gDNA concentration in the PCR increased the number of TCRB clonotypes, there were pros and cons to increasing the gDNA concentrations in the *BIOMED-2 method*.

When using a higher concentration of 400ng per 20µl in the TCRB amplifying PCR reaction it was expected that a higher number of clonotypes would be captured potentially identifying a higher diversity in the repertoire. However, by using a higher concentration, the primers in the amplification may have been exhausted from overuse throughout the PCR cycles and failed to amplify all of the gDNA regions in the sample. This resulted in a diverse repertoire, but so diverse that any clonal TCRB may not be amplified in the reaction.

When performing repertoire studies, ideally all samples want to be amplified at the same concentration to reduce bias attributed to gDNA concentration. For some clinical samples, some of poorer DNA quality, extracting 400ng of gDNA would not be possible. Using the lower concentration of 100ng also encountered issues. If too low a number of TCRB clonotypes were captured in the sample, this would cause artificial clonality of some TCRBs. Clonality observed would appear due to low numbers of some of the rarer TCRB clones in the sample. Lower levels of gDNA could cause cross reactivity of primers producing greater amounts of primer dimer. Primer dimer can then interact with the TCRB sample during PCR producing non-selective binding products at different molecular weights to the desired TCRB library. Using 200ng was the concentration of gDNA that provided the most reliable results whilst capturing sample diversity. Chao1 estimators calculated from 200ng samples were below the number of TCRB clonotypes observed in the repertoire indicating sample diversity was being represented in the sequencing data. This concentration was also realistically attainable from clinical samples. Therefore, 200ng of each sample was used for all subsequent TCRB PCR reactions.

4.3.3. Buffy coat sample preparation lessens TCRB clonality

Twelve RTB samples from ten patients in this study were prepared using the buffy coat method. Four patients had both standard gDNA extraction (Lymphoprep®) and buffy coat samples sequenced for comparison and the rest had only buffy coat available for the specific time points taken. A number of the buffy coat samples were unable to be amplified for TCRB sequencing and only three patient samples that were only buffy coat were used in the analysis in **Chapter 5**.

Comparing buffy coat and gDNA preparations from the same patient taken at the same time showed that only up to a half of the number of TCRB clonotypes were observed in buffy coat samples, sometimes as a low as a fifth. Potentially this was because fewer T-cells were captured at the preparation stage, going into the first TCRB amplifying reaction, rather than at the sequencing stage and when normalising the data according to T-cell numbers. This meant that before the TCRBs were amplified, there were already fewer in the sample to be amplified by the TCRB primers. When using buffy coats samples, it might be assumed in future, that any clonal percentages detailed, could in reality be higher than observed. Buffy coat samples had top TCRB clonotypes at much lower percentages, generally below 2.42% (non-clonal) (**Table 15, Patient 1,2, and 4**) in the repertoire, than their standard gDNA preparation counterparts and the TCRB sequence tended to differ. Buffy coat and the matched standard preparation samples tended to share the same TCRB top clonotype when present at percentages above 2.42% (**Table 15, Patient 3**). This was clearly indicated by one patient who had a large TCRB clonotype above 46%. The gDNA sample was taken five years prior to the buffy coat sample, but because the TCRB clonotype was so large, it was picked up in the buffy coat sample as the top clonotype too (**Section 6.3.2**).

Table 15. Four patients who had Lymphoprep® and buffy coat samples taken at the same time. Each patient had a gDNA sample and a buffy coat sample. Number of TCRB clonotypes were significantly lower in buffy coat samples. Top clonotypes were shared between gDNA and buffy coat samples, when clonal TCRBs were identified in the buffy coat samples.

		Top clone		Second highest clone	
Patient	Number of TCRB clonotypes	% of TCRB repertoire	CDR3 (amino acid sequence)	% of TCRB repertoire	CDR3 (amino acid sequence)
Patient 1 -gDNA	2608	11	CSVGSGGTNEKLFF	1.8	CATSRTSESEGADTQYF
Patient 1 - buffy coat	749	1.8	CATSRESGGTDEQFF	1.8	CSVPRGTDTQYF
Patient 2 -gDNA	1016	5.7	CSVNSGSGNQPQHF	2	CATSRGEQFKQHF
Patient 2 - buffy coat	379	1.6	CATSRESGGTDEQFF	1.6	CSVPRGTDTQYF
Patient 3 -gDNA	892	21	CSVNWGSGNQPQHF	1.6	CASSPGDGGYEKLFF
Patient 3 - buffy coat(5 years later)	842	12	CSVNWGSGNQPQHF	1.5	CASSPGDGGYEKLFF
Patient 4 -gDNA	1596	14	CSVGSGGTNEKLFF	1.1	CSVDRADTQYF
Patient 4 - buffy coat	305	2	CSVPRGTDTQYF	1.9	CATSRESGGTDEQFF

4.3.4. Technical variations in TCRB repertoires

Technical variance was observed in TCRB repertoires when technical replicates from the same biological samples were produced from separate TCRB amplification reactions. The majority of technical variance was observed in TCRB sequences that were below 1.2% and were not clonal.

This could be attributed to the fact that in PCR reactions, if there are many TCRB sequences present at lower levels, they compete for the TCRB primers. This means each PCR reaction may amplify more, certain TCRB sequences that are rarer, in one PCR and less in another. However, the accuracy of the TCRB sequence percentages at these levels are not essential for answering the biological question of clonality.

4.3.5. Clonal TCRB sequences exhibit stability during TCRB repertoire sequencing

From sequencing library to sequencing library, clonal TCRBs are accurately identified. Clonal TCRB clonotypes were not biased by indexing reaction, sequencing process or downstream bioinformatics analysis. In order to ensure accurate representation of repertoires, it would be advised to run technical replicates of the same sample at the indexing stage, if practical (**Table 16.**).

In accordance with **Section 4.5.4.**, any TCRB sequence that accounted for more than 2.42% of the overall repertoire was considered clonal. When running different sequencing runs, to ensure that no differences in TCRB repertoires were observed due to being run on different sequencing runs, a couple of samples were run on every sequencing run for continuity. Initially, the commercially produced mixture of gDNA (Promega) from approximately 12 individuals, female and male, and of varying ages was used alongside an in-house sample. When looking at the mixed Promega sample, all technical samples showed no clonal populations with top TCRB clonotype percentages below 1.2%. These TCRB clonotypes, however were not identical. However, when using the in-house sample which had a top TCRB clonotype that was clonal between 3.4-4.4%, the top clonotypes were identical across different sequencing libraries. The second clonotypes were not clonal and were below 1.2%. These varied across sequencing libraries. This phenomenon was also observed across the patient samples that were technical replicates, sometimes biological replicates, or biological samples from the same patient but taken at different time points. This was another important reason for assessing a threshold for clonality at 2.42% or above rather than the standard 1% TCRB clonotype used in many TCRB studies, which if used would vary between the two samples from the same patient if between 1-1.2%.

4.3.6. Biases incorporated in the indexing and sequencing processes had no effect on TCRB repertoires

It was important to validate the *BIOMED-2 primer method* by assessing potential sequencing biases. To assess this, some samples were split at the indexing reaction stage and sequenced separately (**Table 16.)**

Table 16. Sample replicates to assess sequencing biases. Samples were split at the indexing stage and then sequenced separately to assess whether biases such as position on the sequencer or sequencing index had an effect on TCRB repertoire sequencing. The top clonal TCRB sequences along with the number of TCRB clonotypes were assessed.

		Top clone		Second highest clone	
	Number of TCRB clonotypes	% of TCRB repertoire	CDR3 (amino acid sequence)	% of TCRB repertoire	CDR3 (amino acid sequence)
Patient 1 - A	1050	3.9	CATQPGGGGNEQFF	0.9	CATSTIAGETQYF
Patient 1 - B	1200	4.4	CATQPGGGGNEQFF	0.9	CATSTIAGETQYF
Patient 2 - A	656	1.2	CSVEERGTGAEKLFF	0.9	PQARGPRGGKLFF
Patient 2 - B	610	1.2	CSVEERGTGAEKLFF	1	PQARGPRGGKLFF
Patient 3 - A	456	57	CSVGSGGTNEKLFF	1.2	CATGQDSNQPQHF
Patient 3 - B	311	61	CSVGSGGTNEKLFF	1.2	CATGQDSNQPQHF

When assessing these samples, TCRB repertoires from the same sample PCR reaction, split at the indexing stage, once sequenced, shared the same top clonotype and the second highest clonotype regardless of the clonotypes percentage in the entire repertoire (**Table 16.**). All paired samples showed similarity in the percentages of the TCRB repertoire that the top two clonotypes occupied with some being identical. Number of TCRB clonotypes passing the bioinformatics analysis also showed similarity between matched samples with most being within 150 TCRB clonotypes of one another irrespective of sequencing depth (number of sequencing reads) which naturally varied from one sequenced sample to the next. Overall, this highlighted that no biases were incorporated in the indexing reaction, sequencing reaction or bioinformatics pipeline that caused inaccuracies in TCRB repertoire analysis.

4.4. Case study analysis for the validation of the *BIOMED-2 primer method*

A number of studies were carried out using the *BIOMED-2 primer method* which have produced some results that verify the sensitivity and accuracy of the *BIOMED-2 primer method*.

4.4.1. *BIOMED-2 primer method's* sensitivity allowed for the identification of newly clonally expanded TCRBs

In this dataset, a group of patients with colorectal cancer or melanoma, had their TCRB repertoires tested at different points in treatment. The treatment involved the introduction of a genetically modified virus intra-tumour wise to evoke immune responses to treat these cancers. The time-points were at base level (no virus), a time point a couple of weeks after the virus was introduced to the patients and a time-point a couple of months after. **Figure 31.** showcases the top 24 CDR3s present in the peripheral blood of a patient prior to treatment. These were then tracked at the second and third time-points. After the virus was injected, two novel CDR3s emerged (**Figure 31.**, JX-01-L-F column) which were not present in the patient's repertoire prior to this. Potentially, this is a T-cell clone responding specifically to the virus which then remains in the repertoire after infection. This indicates the sensitivity of the BIOMED-2 method to identify novel T-cell clones.

All CDR3	JX-01-L-B	JX-01-L-F	JX-01-L-H
CASSYGADTQYF	0.026808038	0.017396366	0.019600027
CASAITTENTQYF	0.010371331	0.009959258	0.013686798
CASRKREATTDTQYF	0.007269192	0.009247882	0.006511195
CASSPGPGGGFNEQFF	0.00476896	0.009247882	0.005979669
CASSYLALGDTQYF	0.005139365	0.008083813	0.003986446
CASSAIQGGTDTQYF	0.005694972	0.007307767	0.003720683
CSVALAGGSDTQYF	0.000138902	0.007049085	0.008836622
CSVDQLAGSADTQYF	0.002916937	0.005044299	0.006311873
CSVGQGSADTQYF	0.005694972	0.004591606	0.002657631
CASNWGILDNSPLHF	0.001435318	0.004914958	0.005315261
CSVDRADTQYF	0.000740809	0.008277824	0.002192545
CSVPRSGGRGDEQFF	0.003055839	0.005432322	0.00259119
CASSLRWGETQYF	0.004444856	0.003039514	0.003322038
CSVDIGSSGANVLTF	0.003750347	0.002134127	0.003920005
CSVYQVSSYNEQFF	0.004074451	0.003168855	0.002126105
CSVVGNGANVLTF	0	0.002263468	0.006378314
CSVEQQGQADTQYF	0.006620983	0.001487422	0.000531526
CSASKTGIGTDTQYF	0.002176127	0.002069456	0.003920005
CASRRVGTGLFF	0.00314844	0.004138912	0.000730848
CASSRAEGNEQFF	0.003426243	0.00122874	0.002258986
CASSISPITQPQHF	0.001481619	0.002004786	0.003255598
CSALGTPRGNEQFF	0.002315029	0.002845502	0.001395256
CATSRDPGRPPYEQYF	0.001944625	0.001746104	0.002790512
CSVERADTQYF	0	0.000840717	0.005514584

Figure 31. Novel TCRBs detected using BIOMED-2 primer method. The top 24 CDR3s identified through TCRB sequencing using the *BIOMED-2 primer method*. The first column indicates CDR3 amino acid sequence. The second column shows the clonal frequency of the CDR3 in the entire TCRB patient repertoire before treatment. The third column was two weeks after the genetically modified viral infection. was introduced into the tumours. The final column was from samples taken a couple of months after treatment. The green in column two indicates two novel CDR3s that were introduced into the repertoire after the viral infection. Table courtesy of D. Newton.

4.4.2. Flow cytometry TCRBV gene identification method identified the same top TCRBV gene family as the *BIOMED-2 primer method* in a RAG deficient sample

A patient deficient in some recombination-activating genes (RAGs) that are important in gene rearrangement and recombination events for encoding T-cell receptors had their TCRB sequenced. Three different methods were used, previously generated data from a flow cytometry method (flow cytometry data kindly provided by the Savic laboratory/clinical immunology, St James's University Hospital) for TCRBV gene identification, the *Robins et al. primer method* and the *BIOMED-2 primer method*.

Using three methods on the same patient sample, allowed for the verification of the TCRBV family usage of this patient and for the reliability of the *BIOMED-2 method*. The TCRB sequencing methods used the same aliquot of patient gDNA. When comparing the two sequencing methods, the *Robins et al. primer method* showed the same pattern of results as with all other different samples mentioned in previous sections, with the majority of reads being from the TCRBV7 and TCRBV6 families. These results were very different to the *BIOMED-2 primer method*, which showed the top CDR3 clones as being TCRBV 27, 6-4 and 29-1. When looking at TCRBV usage irrespective of CDR3 sequence, TCRBV29-1 was the most highly used in the repertoire according to this method. When comparing this with the flow cytometry data, similarities were shared. In the flow cytometry data, TCRBV 27 was found to have high expression at almost three times the expression expected at a baseline normal expression. This supports the BIOMED-2 data but not the *Robins et al. primer method*.

Only 29 of the TCRBV gene families were identifiable using the antibody flow cytometry method. TCRBV6-4 and TCRBV29-1 were not represented families using the flow cytometry method and therefore were not able to be compared between methods. Having the flow cytometry data allowed the *BIOMED-2 primer* results to be verified as accurate and that the inherent primer bias in the *Robins et al. method* was causing inaccuracies. This was an important reason for taking the *BIOMED-2 method* forward and not the *Robins et al. method*.

4.5. Analysis of 31 healthy control TCRB repertoires to investigate natural TCRB repertoire population variance

In order to investigate PNH specific T cell responses in PNH patients, it was important to first establish a background for "normal" T-cell repertoires using 31 healthy controls. The average age of the controls was 37 years old, with a median of 35. Therefore, there is a slight skew towards the ages below 37 years old. The minimum age was 22 and the oldest person was 57 years old. When split into age range groups the breakdown of the ages is detailed in **Table 17**. Overall, 9 controls were male and 22 were female. When comparing healthy controls, it was important to investigate the possible effects of sex and age on the TCRB repertoire measures and statistics.

Table 17. Metadata for the 31 healthy controls sequenced using the *BIOMED-2 primer method*.

Age Range	Number of controls	Male	Female
20-24	1	-	1
25-29	6	3	3
30-34	8	2	6
35-39	4	2	2
40-44	4	1	3
45-49	4	-	4
50-54	2	1	1
55-59	2	-	2

4.5.1. Effect of age and sex on the number of unique TCRB clonotypes and their abundance in the repertoire

In this initial analysis, a TCR clonotype was any sequence with the same TCRB V and J gene combination, along with the same CDR3 sequence at the amino acid or nucleotide sequence level.

4.5.1.1. Number of unique TCRB clonotypes did not significantly differ between sex or age

When comparing the number of unique clonotypes for each control, there was no significant difference when using the nucleotide or amino acid sequence, regardless of age, sex or both. For both nt and aa, the mean number of clonotypes for all female controls was 971 and for males it was 973.

The minimum number for females was 424, males it was 681 and the maximums for each were 1793 and 1499 respectively (**Figure 32.**). Overall the range of unique TCRB clonotypes for the 30 healthy controls was between 424 and 1793.

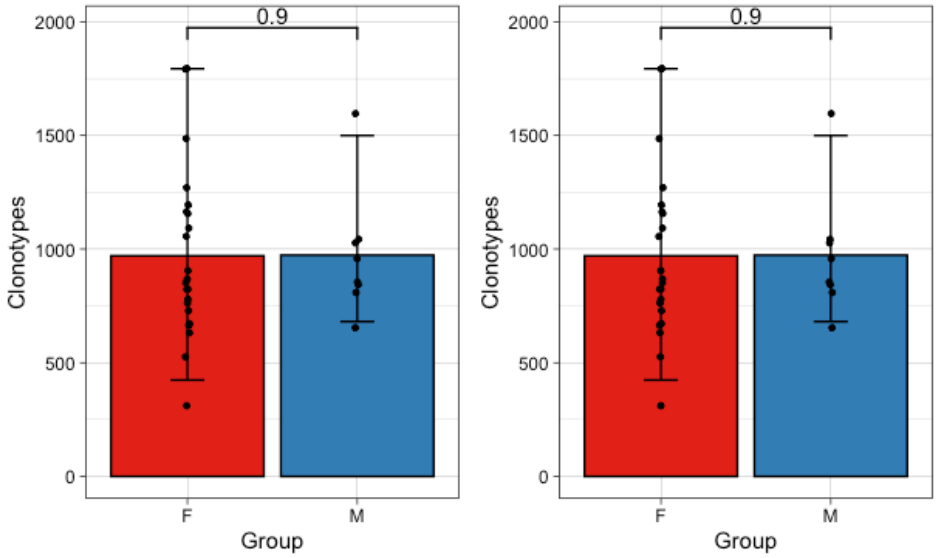


Figure 32. Number of unique TCRB clonotypes in 30 of the healthy controls.

One male control was removed due to a highly clonal population not indicative of a "normal" TCRB repertoire. The left graph used amino acid CDR3 sequence and the right used nucleotide sequence for CDR3. Each bar chart represents female controls (red) and male controls (blue). The P value of 0.9 shows no significant difference in unique TCRB clonotypes between sex (Wilcoxon rank sum test).

4.5.1.2. Abundance of TCRB clonotypes observed did not vary according to sex or age

When observing the relative abundance of TCRB clonotypes in the repertoires of the 30 healthy controls, no significant difference in the trends were observed when comparing age, sex (**Figure 33.**) or both. This was irrespective of using nt or aa as the CDR3 sequence when defining a TCRB clonotype. In general, the majority of clonotypes had a low abundance in each repertoire (not clonal) and very few if any, were present in medium to high abundance (clonal) showing high diversity in normal repertoires.

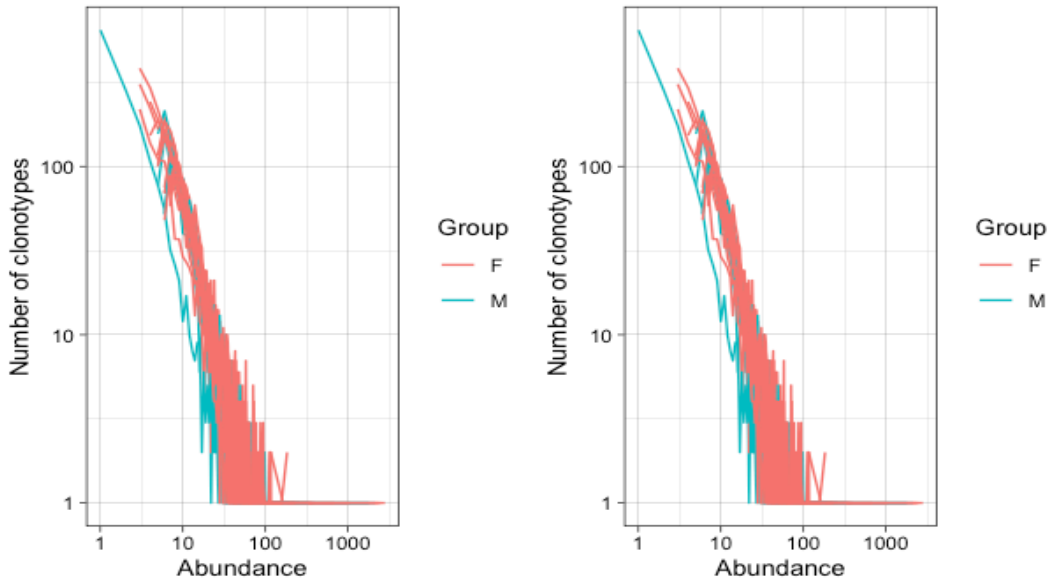


Figure 33. Distribution of TCRB clonal abundance in 30 of the healthy controls.

One male control was removed due to a highly clonal population not indicative of a "normal" TCRB repertoire. The left graph used amino acid CDR3 sequence and the right used nucleotide sequence for CDR3. Each line graph represents each sample. Red line graphs represent female controls and turquoise represent male controls. The general trend is the same for all of the healthy controls, with the majority of TCRB clonotype present at low abundance within the repertoires. This is indicative of "normal" diverse TCRB repertoires.

4.5.1.3. Number of TCRB receptors did not significantly differ irrespective of age or sex

When comparing the number of TCRB receptors in each sample, there was no significant difference when comparing age, sex or both (**Figure 34.**). This trend was the same when using nt or aa CDR3 sequence. The minimal number of TCRB receptors per sample amongst the female controls was 10779, the mean was 14285 and the maximum number was 18539. When comparing the results within the male controls, the minimum number of TCRB receptors per sample was 8706, mean was 13197 and the maximum was 16053. Overall, amongst all 30 controls, the range of TCRB receptors per sample was between 8706 and 18539. Most likely any subtle differences were biological rather than technical. Number of TCRB sequences could have been affected by factors such as sequence depth however, a number of measures were put in place to reduce this. The methods used to generate separate sequencing libraries were consistent. For example, all samples were measured using Pico Green™ to assess the concentrations of TCRB amplicon.

This allowed the same concentrations of each sample to be inputted into the sequencing library so that no sample was biased due to having a higher concentration in the sequencing bridge amplification reactions. PCR cycles were identical between sequencing libraries. All of these measures helped to lessen the potential differences in clonotype number attributed to variances in sequencing depth (read number) between samples and sequence runs.

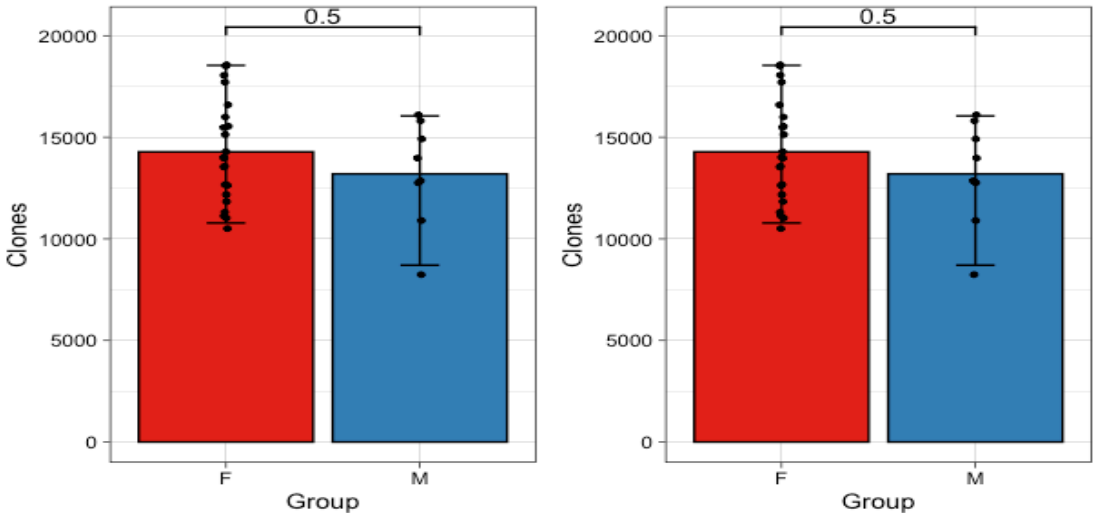


Figure 34. Number of TCRB receptors per sample in 30 of the healthy controls.

One male control was removed due to a highly clonal population not indicative of a "normal" TCRB repertoire. The left graph used amino acid CDR3 sequence and the right used nucleotide sequence for CDR3. Each bar chart represents female controls (red) and male controls (blue). The P value of 0.5 shows no significant difference in unique TCRB clonotypes between sex (Wilcoxon).

4.5.2. Healthy control TCRB repertoires showed no significant skew in CDR3 lengths

Within a normal TCRB repertoire, lengths of CDR3s would be expected to vary. Germline CDR3s with no insertion or deletions tend to be below 12 amino acids in length [**264**]. Lengths are normally distributed around a mean of approximately 14-15 amino acids. CDR3 lengths amongst the healthy controls followed this trend. There was no skew towards shorter or longer CDR3 lengths. This pattern was the same when comparing nucleotide (top, **Figure 35**) and amino acid CDR3 sequences (bottom, **Figure 35**). CDR3 nucleotide sequences had an average of 42-45 nucleotides (top), CDR3 amino acid sequence (bottom) had an average of 14-15 amino acid sequences. This trend was observed across all the age groups and there was no significant difference between sex (**Figure 35.**, male blue, female red).

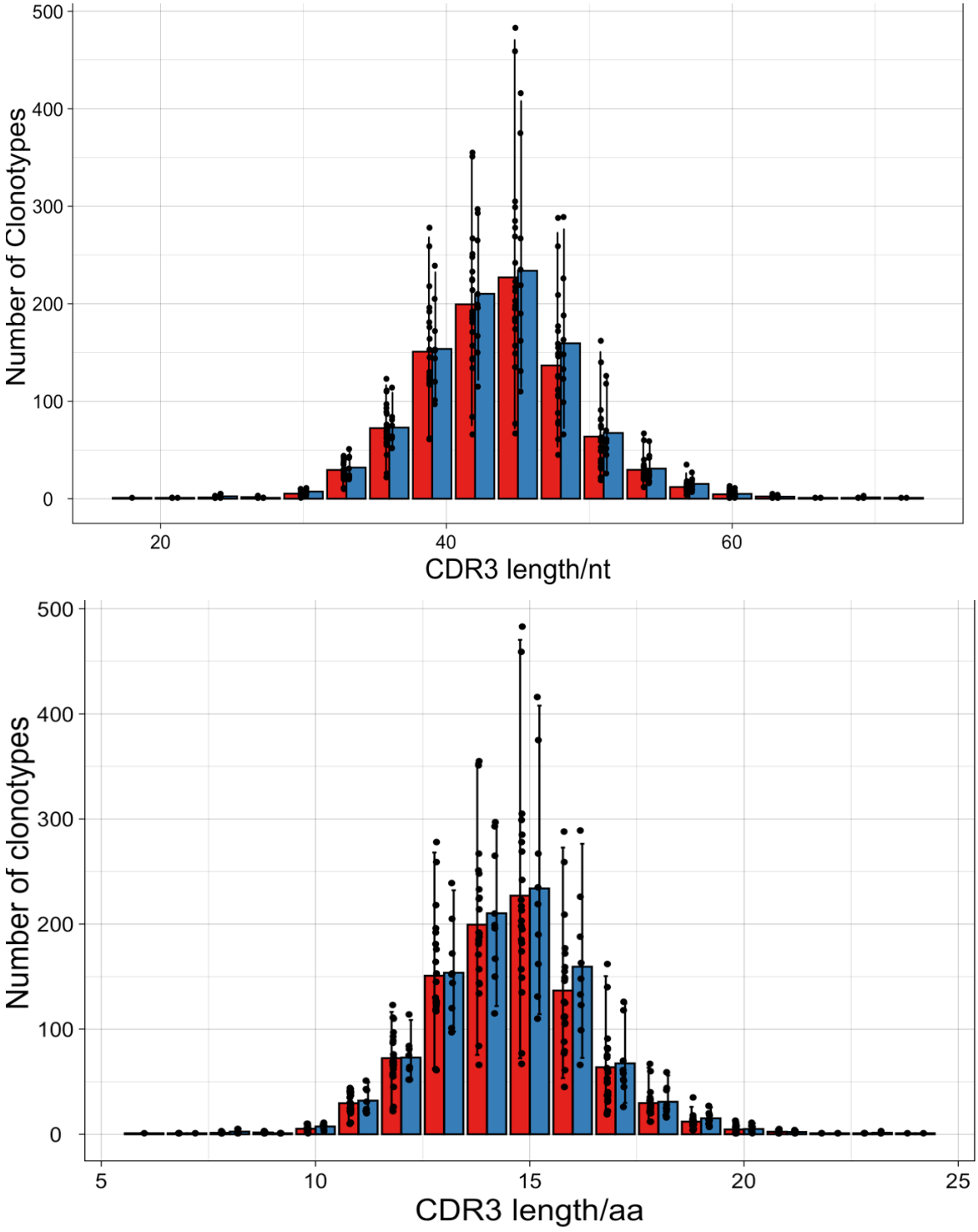


Figure 35. TCRB CDR3 length distribution in 31 healthy controls.

CDR3 nucleotide sequences had an average of 42-45 nucleotides (top), CDR3 amino acid sequence (bottom) had an average of 14-15 amino acid sequences. There was no significant difference between male and female groups (Wilcoxon). Male groups are indicated in blue and the female in red. Each dot is representative of a sample repertoire. Padj values were all 1 (Holm).

4.5.3. Healthy control TCRB repertoire showed no signs of decreased diversity after the age of 40

All of the analysis above was also performed when splitting the control groups into those who were above 40 and those below 40, the age at which it is thought that the thymus can no longer educate naïve T cells, leading to a drop in diversity and inability to produce new T-cell receptors [**99**].

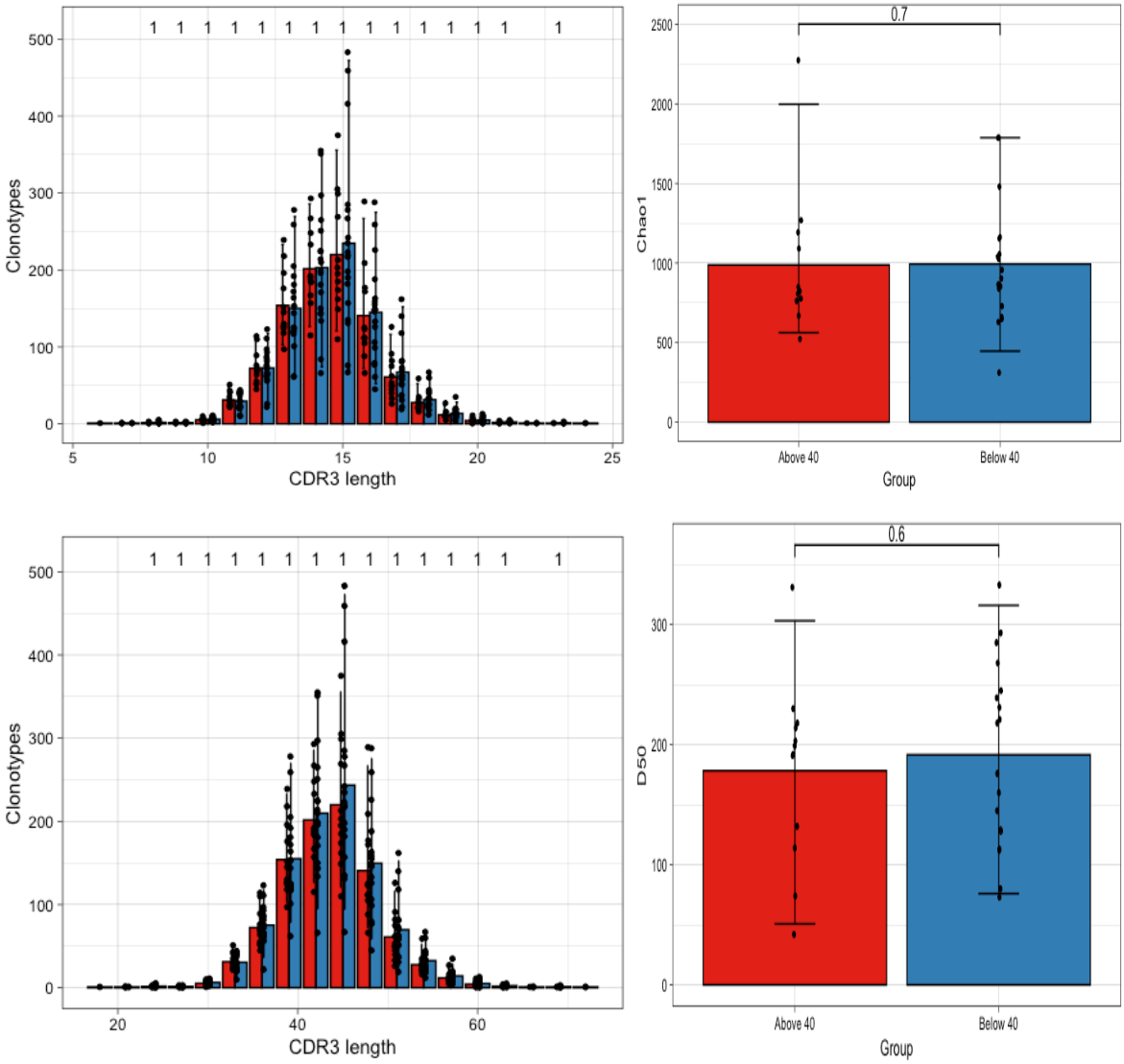


Figure 36. No significant differences were observed for basic TCRB repertoire statistics when 30 healthy controls were split into below and above 40 years old. Clockwise, CDR3 amino acid distribution of lengths, Chao1 estimating diversity, D50 measuring clonality and CDR3 nucleotide length distributions. A TCRB clonotype was a TCRB population with the same CDR3 amino acid sequence, TCRB and J genes. Blue indicates patients under 40 years old and red over 40 years old. P adjusted values above 0.05 indicated no statistical significance (Holm, Wilcoxon).

There was no significant difference between the age groups when investigating number of unique TCRB clonotypes, number of clonotypes per sample, CDR3 length distribution, TCRBV/J gene usage or diversity measures, such as d50 and Chao1 estimators (**Figure 36**).

Overall, there was no difference between using CDR3 amino acid or nucleotide sequence. However, in this project, TCRB clonotype is defined as the same TCRV gene, J gene and CDR3 amino acid sequence, to reduce convergent evolution bias. The majority of TCR repertoire studies also depict TCR clonotype as the CDR3 amino acid sequence, therefore using amino acid sequence over nucleotide sequence will be advantageous for later comparisons with literature.

4.5.4. Defining TCRB clonal expansions in the repertoire using *BIOMED-2 primers*

In order to define whether a clonal expansion is biological or as a result of PCR amplification, the following analysis was carried out. When assessing the clonal proportion of the top clone in each healthy control, the average proportion was 4.79%, the median was 2.2%. The difference in the median and the average would suggest a skew towards there being more low values for the top clonotype. For instance, if only a couple of samples had above average clonal expansions, which could be attributed to a current cold for instance, these will cause the average to look higher. This is why both the median and average were calculated for the majority of the analysis. One of the 31 healthy controls had an abnormally high T-cell receptor clone constituting over 50% of their repertoire, which is not characteristic of a "normal" TCRB repertoire and may be indicative of a chronic response to a common infection such as EBV. Therefore, this was excluded for the first part of this analysis. In order to assess the highest TCRB clone proportion in a healthy repertoire, the median value of the remaining 30 controls top clonotype proportion was calculated.

The second method plotted each of the 30 healthy controls' top clone proportion e.g. 2.2.%. A linear regression smooth curve was plotted through the points using the 'loess' method [**247**]. Tight grouping was observed along the line, remaining relatively flat at around 2.2% (**Figure 37, top**). The y intercept was plotted where the curve begins to increase exponentially. The y intercept was plotted at 0.0242 which is the equivalent of 2.42%. When the large 50% clone control was put back into the analysis, this value rose to 0.0485 equivalent to 4.85% (**Figure 37, bottom**). However, including the higher clone gave better grouping due to the scale for the healthy control data, with varying levels of T-cell response for their top clone.

The majority of normals had a top clonotype around the value of 0.0242 (**Figure 37, bottom**). There were a few controls with slightly larger clones between 0.0242 and 0.0485. Another cluster of controls appeared with values between 0.0485 and 0.200 and then the abnormally large clonal expansion from the control at >0.200.

Healthy controls in this project are defined as those who have not been diagnosed with Paroxysmal Nocturnal Haemoglobinuria. However, within a given population, it is highly likely a healthy control may have a cold or another type of infection. They could be suffering from a T-cell mediated disease such as gastritis [**265**] and as people get older, they tend see more clonal expansions specific to infections such as EBV and CMV [**266**] because of skewed repertoires towards memory T-cell subsets and re-stimulation. There is also the process of memory populations clonally expanding in response to re-stimulation from a previously "seen" antigen which can give rise to these fluctuations in clonality. In a given "normal population", there can be a variety of different levels of T-cell response dependent on some of these factors. Therefore, it is important to take these situations in to account as best as possible when establishing the background of "normal" T cell responses for comparisons against people with PNH [**267**].

Taking the majority of the normals into account and the groupings observed in **Figure 37.** bottom graph, in this project TCRB clones making up 2.42% or less of the repertoire were deemed non-clonal TCRB expansions and the expansion could be attributed to PCR amplification rather than biological. TCRB clones with a value greater than 2.42% but less than 4.85% are defined as possible TCRB clonal expansions but low response. Moderately responsive TCRB clonal expansion are those above 4.85% and in the range below 20% and those above 20% are highly responsive T-cell clone expansion.

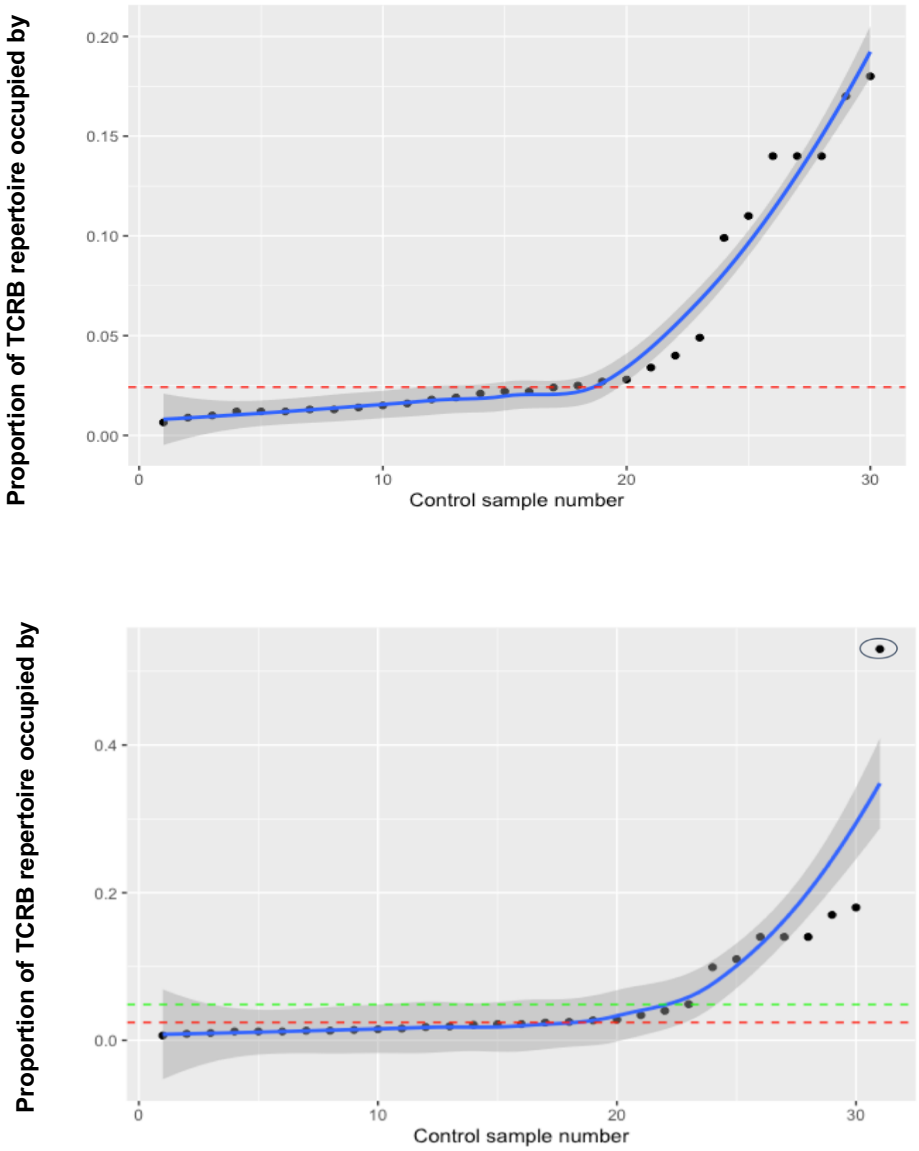


Figure 37. Investigating the proportion of the entire TCRB repertoire that the top clonotype in each healthy control contributes. The top clonotypes for 30 healthy controls (top) and 31 healthy controls (bottom) were plotted. For the 30 healthy control group, the healthy control with a single clone making up 50% of the repertoire was removed, as this was deemed abnormal. The proportional percentage is out of a maximum of 1. The line of best was plotted using the 'loess' method [**247**]. The y intercept lines represent the value at which the curve begins to increase exponentially. The y intercept for all 31 controls (green, bottom graph) value was 0.0485 and 30 controls (red) was 0.0242. The bottom graph, which includes the anomalous healthy control (circled in black) with a proportion of above 0.500, indicates that the majority of normals have a top clonotype around the value of 0.0242. There are a few controls with slightly larger clones between 0.0242 and 0.0485. Another cluster of controls appears with values between 0.485 and 0.200.

4.5.5. Monoclonal TCRB responses in comparison to polyclonal TCRB responses

When measuring TCRB clonal expansions it is important to assess whether they are monoclonal, clonal expansion is from one TCRB population, or polyclonal, consisting of more than one TCRB clonal expansion. In the 14 out of 31 healthy controls that showed a level of possible T-cell clonal expansion, the mono/poly clonal response was assessed. The clones that accounted for over 2.42% of the repertoire were selected and these were then split into response level as described above, and the number of different TCRB clones that made up each response level (**Figure 38.**). There was no association of monoclonality/polyclonality or T-cell clonal expansion level with a specific age range or sex in the healthy control cohort.

Figure 38. shows varying levels of T-cell responses across the healthy controls. As expected, the majority, 17 of the 31 controls showed no T-cell clonal response. Five exhibited low responsive monoclonal T-cell responses. Six of the controls were found to be producing both low and moderate T-cell responses which were either monoclonal or polyclonal and the one abnormal control, with the large T-cell clone over 50%, exhibited an monoclonal high response T-cell clonal expansion, along with monoclonal moderate and low responsive T-cell clonal expansions. This highlighted variability in T-cell clonal expansions that can be observed and considered natural variation in a population of people not diagnosed with PNH.

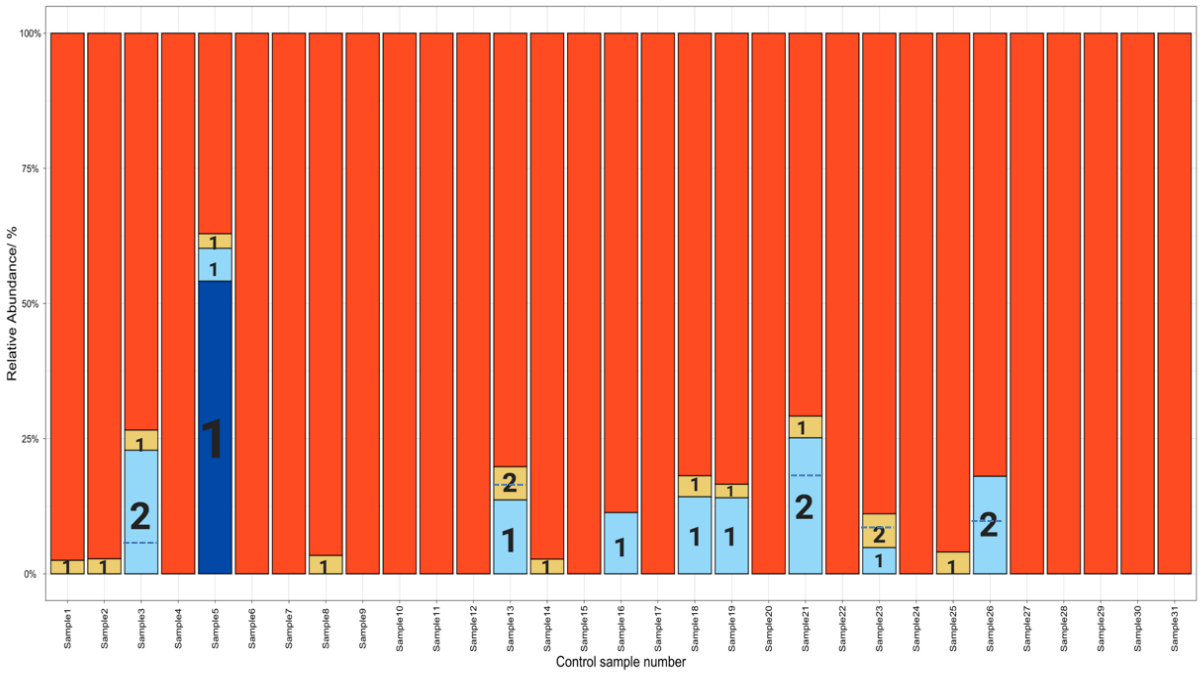


Figure 38. Monoclonal and Polyclonal T-cell receptor beta responses in healthy controls. For the 31 healthy controls, T-cell clonotypes were categorised into response levels. T- cell non-clonal response was any clonotype that represented 2.42% or less of the overall repertoire (red), low responders were clonotypes between 2.42 and 4.85% (yellow), moderate responders were between 4.85% and 20% (light blue) and high responders were clones above 20% (dark blue). The T-cell clonal expansion groups were then identified by the number of clonotypes in each group (number in the box) to investigate whether a response was monoclonal (one clonotype), or polyclonal (more than one clonotype). Each bar chart indicates the entire clonal space of one control repertoire, with each response size indicating the proportion of the TCRB repertoire it occupies. The blue dotted lines in boxes indicate the balance each TCRB clonotype in that response group occupies when multiple TCRBs are detected (polyclonal).

4.5.6. Identification of clonally expanded TCRBs in the controls from scientific studies and database cross validation

To test the theory that in a given non PNH population, clonal T-cell responses could be attributed to autoimmune disease or common infections for example, cross validation of the clonally expanded TCRBs across the 31 normals was investigated. The top clone of the non-clonally expanded controls was also analysed. Overall, 27 clonally expanded and 17 non-clonally expanded TCRB clones were searched for in scientific literature, in internet searches and the databases 'VDJdb' and 'McPAS-TCR' databases [**254,255**], which are manually curated and contain T-cell receptors from published work that are associated with pathological diseases. Overall, only 10 of the 27 TCRBs were found in this

analysis (**Table 18.**). Only one appeared in two controls, the rest occurred only once. The large clone from the control accounting for 53% of the repertoire was not found in the analysis. Two out of 14 low responsive TCRB clonal expansions were found in search and were linked to CMV and EBV. Interestingly, out of the 12 moderately responsive clonal TCRB populations in the healthy controls, only 2 were validated in a cross analysis. The TCRB attributing to 9.9% of a control repertoire was found to possibly be an effector CD8+ in a study, suggesting this control may be actively fighting an infection or responding to an antigen of some kind. However, the antigen was not identified. The other control repertoire had a sizeable clone at 17.6% of the repertoire, which was found to be associated with public CDR3 clones in studies. It was also found as the top clone of another healthy control but at a considerably lower level (non-clonal) of 1.9%. The TCRB can be public and pathogenic, perhaps a response to a common infection such as the common cold. The control with the high clone could be responding to a current infection, whereas the 1.9% clone could mean that the antigen is common enough to warrant a sizeable memory T cell clone, but the antigen is not currently present, so the clone is not activated and therefore not expanding.

Two TCRB clonotypes were associated with CMV and EBV which is to be expected, the controls were 25 and 39 years of old, one female and one male. There was a range of CD8+ and CD4+ T-cells identified, therefore, showing a spread of killing and protective immune responses across the population. Many of the low and no clonal response were found to be public CDR3s, TCRB clonotypes occurring in more than one individual (generally not related).

One of the TCRB clones was found in a study in two patients with IBD, but not in the control sets suggesting it could be IBD specific. This is an interesting finding, as it is estimated that 396 people per 100,000, annually, suffer from IBD [**268**]. If this is caused by T-cell responses, this supports the theory that T-cell clonal expansions of varying levels of response can be observed and it is very difficult to define a “normal response”. However, the other studies would suggest it could be public. Another example of this was in one control sequence that was identified as being specific in renal cell carcinoma patients, and in another study is thought to be a public CDR3 response, highlighting that it can be challenging to draw conclusions about the specificity of T-cells in disease, and the importance of having a good sample set of normals for comparison with the disease sets. When comparing the top 100 CDR3s across all of the healthy controls, according to proportion, only 2 had results in the analysis linking them to diseases such as cancer (melanoma), autoimmune disease such as ulcerative colitis, influenza and were CD8+.

Table 18. TCRB clones in the control set identified in published scientific literature.

From 31 healthy controls, 27 clonally expanded and 17 non clonally expanded (top clonotype for non-clonally expanded control repertoires) were cross validated in an internet search and using the McPAS-TCR and VDJdb databases to identify any known pathological associations particularly EBV and CMV common infections. Only 10 TCRB were identified in the analysis.

TCRB V gene	TCRB J gene	CDR3 sequence (aa)	Clonal proportion in repertoire/ %	Association
27	2-3	CASSFGGITDTQYF	3.4	-HLA type 7 CD8+ T cells -TRP specific CMV
27	2-3	CASSLTGDTQYF	17.6	-Common public TCRB appeared in 6 of 39 donors in a study -Found in lower % 1.9% in patient 33F
29-1	1-4	CASRKDSLTGELFF	2.6	-HLA-A2 restricted $BMLF1_{280}$ epitope from EBV
27	2-1	CASSLGGSYNEQFF	2.7	-Myocardial infarction patients and normal coronary artery controls – potentially common CDR3 -Narcolepsy patient -CD4foxp3-tumour subset -Conventional CD4+ T cell -2 patients with IBD but not in controls
13	2-3	CASSFQDTDTQYF	2.7	-Treg CD4+Tcell -Foxp3+ intra tumour
27	2-1	CASSLGGDYNEQFF	2.5	-TAIR database, public CDR3
27	2-1	CASSTYNEQFF	2.1	-Ageing study -CD4+foxp3-
27	2-1	CASSQTSGSYNEQFF	2.3	-CD8+ -Found in lung cd103-, not in blood or CMV patient in study -Public TCRB
5-6	2-1	CASSLDSPNEQFF	9.9	-CD8+ -Effector T-cell
19	2-3	CASSSQETQYF	2.3	-Renal cell carcinoma -Public tcr found in 100% normals blood

4.5.7. Repertoire overlap studies and generating public and private TCRB clonotype matrices

4.5.7.1. Low overlap observed between TCRBs of healthy controls

In order to investigate the overlap between healthy controls to assess how much similarity could be expected between TCRB repertoires, two methods for similarity were used in the analysis.

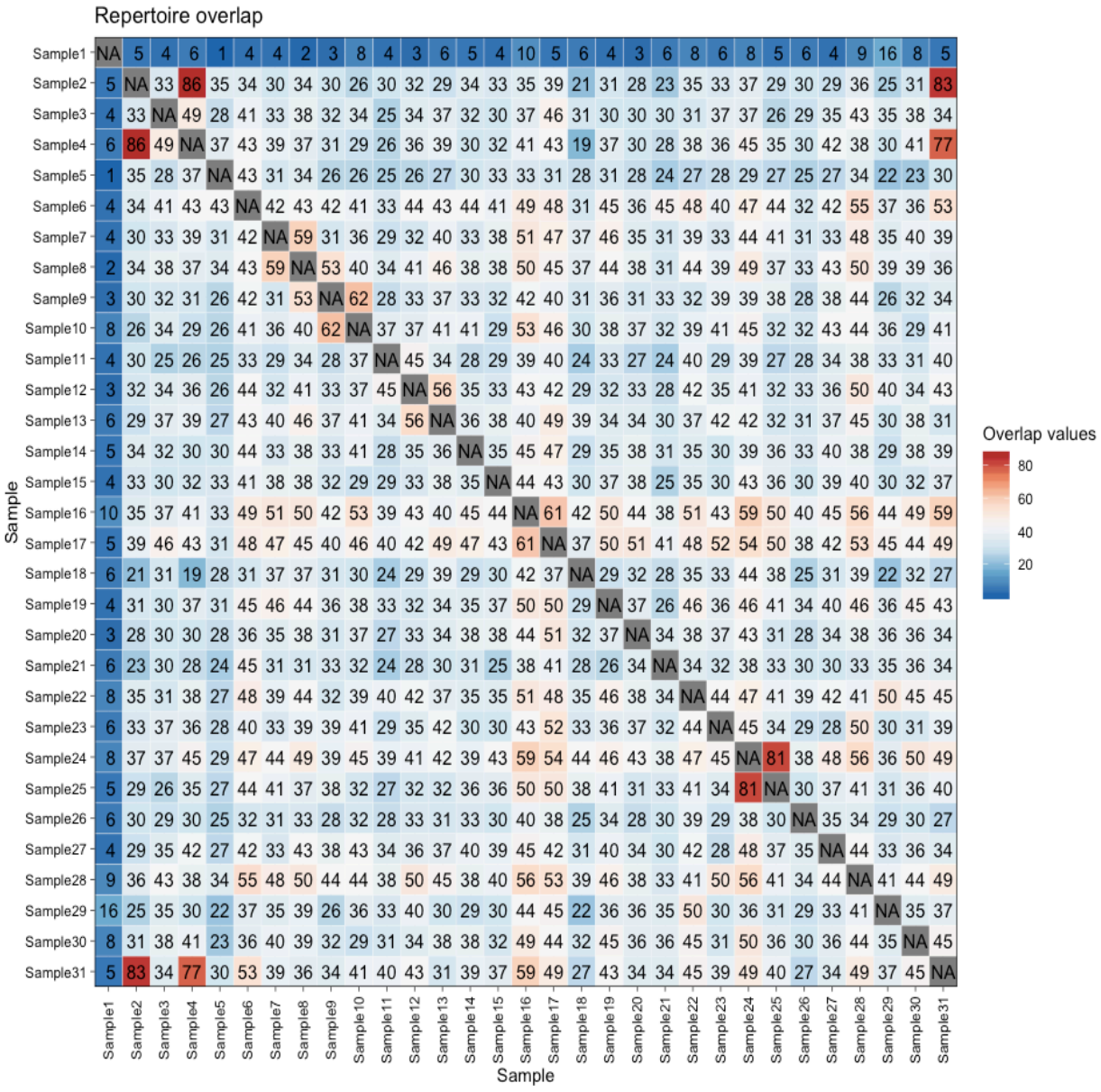


Figure 39. Overlap studies comparing TCRB clonotypes shared across 31 healthy controls

The grid axis indicates the samples and the number in each square is the "public" overlap metric shared between two pairs of samples. The redder the scale, the higher the degree of overlap. The grid uses the measure of shared TCRB clonotypes between two samples, where the clonal counts of the same clonotype in each group will be summed. TCRB repertoires were generated using the *BIOMED-2 primer method*. The majority of healthy controls showed low levels of overlap between repertoires.

One method, the “public method” (**Figure 39.**) used the measure of shared TCRB clonotypes between two samples, where the clonal counts of the same clonotype in each group were summed. The grid axis indicated the samples and the number in each square was the overlap metric shared between two pairs of the sample. The redder the scale, the higher the degree of overlap. The majority of the values for degree of similarity showed low numbers of shared TCRB clonotypes per sample.

One sample, a 32-year-old female, showed very low similarity with other repertoires ranging between 1 and 16 shared TCRB clonotypes. HLA is related to how an individual’s immune response presents peptides via MHC to T-cells [**269**]. Genetics influence HLA type. Perhaps this individual is of a different ethnicity to the majority of the samples, which could attribute to the low number of shared TCRBs. Ethnicity was not noted in the metadata. Sample 16 and 17 showed moderately high similarity between most samples especially between each other at values of 61 shared clonotypes, they were male and female and aged 27 and 32 respectively. Four of the 930 numbers were considered high similarity and were above 76 shared TCRB clonotypes. The highest values between repertoires were 77, 81, 83, and 86. Sample 4 and 31 shared 77 TCRB clonotypes and were both female, aged 33 and 32 respectively. The value of 81 was shared between samples 24 and 25 which were both male aged 34 and 51 respectively. Sample 2 and 31 which shared a value of 83, were both female and the age of 32. The highest value of 86 was observed between the samples 2 and 4, a 32-year-old female and a 33-year-old female. Considering the average number of TCRB repertoires in a sample was above 13,000, and unique TCRBs was around 1000, 80 shared TCRB clonotypes does not seem a particularly large number. Indicating that across healthy controls there is a low degree of overlap.

A similar trend was observed when using a second similarity method, the “overlap” method, which measured the overlap between two samples (**Figure 40.**). The size of the intersection between the two samples (shared TCRB clonotypes) was divided by the number of TCRBs present in the smaller repertoire. The higher the value the more TCRB clonotypes that were shared between the two samples. In general, there was a slightly higher trend of similarity than when using the “public” method, perhaps attributed to the fact that the size of the TCRB repertoires was taken into account to “normalise the values” irrespective of sequence depth.

The majority of the samples show lower levels of overlap between TCRB clonotypes. Two of samples show high levels, mostly those compared with sample 5 which was the youngest sample at 22 years of age and had the highest number of unique TCRB clonotypes in the female group.

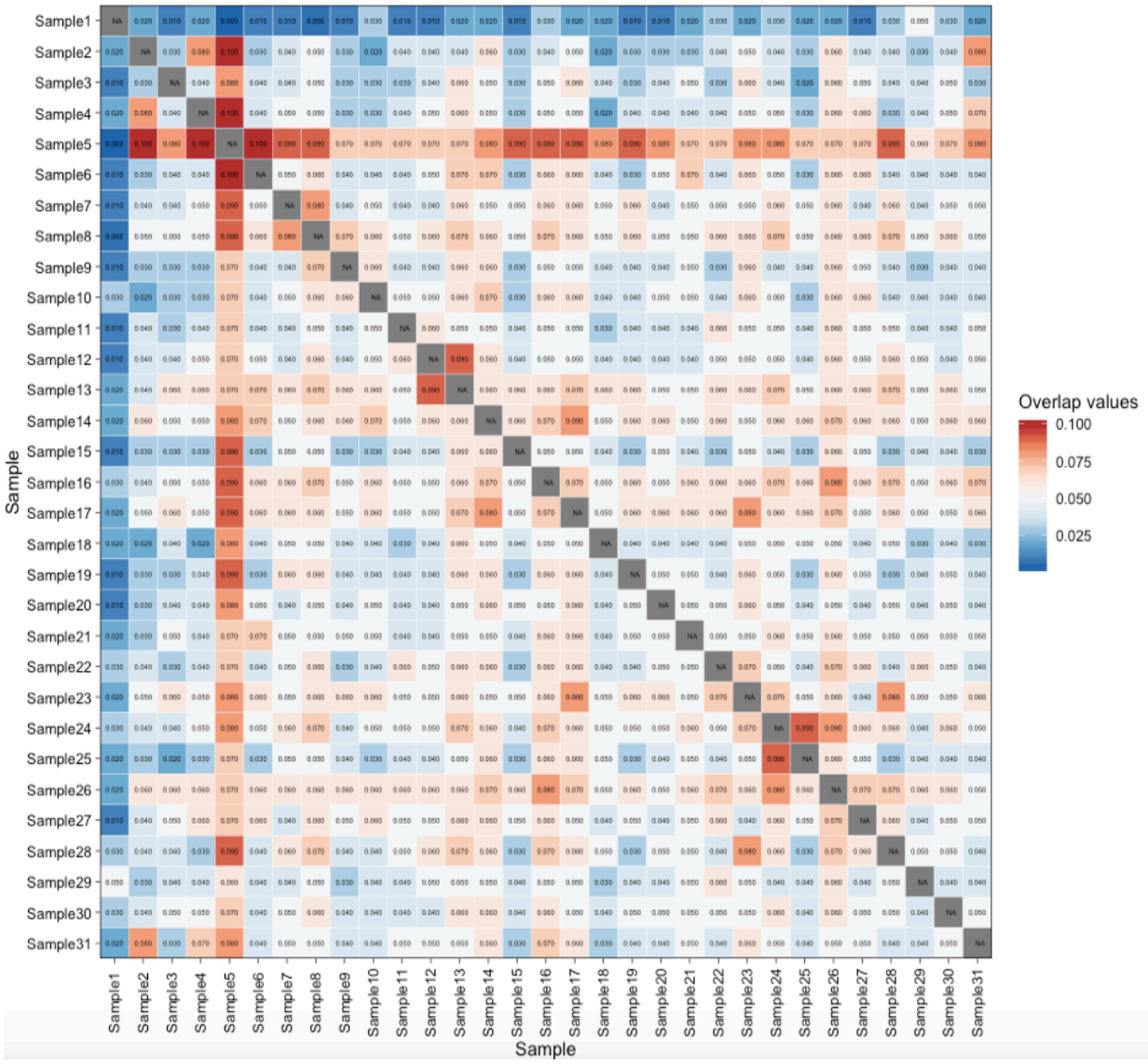


Figure 40. Comparing the "overlap" measure of TCRB clonotypes across 31 healthy controls.

The grid axis indicates the samples and the number in each square is the overlap metric shared between two pairs of samples. The redder the scale the higher the degree of overlap. The grid uses the "overlap" measure which is similar to the Jaccard index [**270**]. It measures the overlap between two finite sets in this case two samples' TCRB repertoires. The larger the overlap the higher the value.

4.5.7.2. Generating a public TCRB repertoire matrix

Public TCRB clonotypes in this project are defined as TCRB clonotypes with the same TCRBV/J gene combination and amino acid CDR3 sequence that appears in two or more healthy controls. These were combined across all 31 healthy controls to generate a public matrix for comparison with PNH samples.

Across the 31 healthy controls, only 1178 of the 25897 unique TCRB clonotypes were considered public clonotypes. 1178 clonotypes were cross referenced against TCRB clonotypes identified by searching scientific literature, internet searches and querying the "McPAS-TCR" and "VDJdb" databases. They were associated with a range of diseases including CMV, EBV, Influenza, Crohns, Diabetes, melanoma, ulcerative colitis, colorectal cancer and *M. tuberculosis* [**254**].

4.5.8. Amino acid characteristics of normal TCRB CDR3s

4.5.8.1. No significant difference in amino acid properties was observed between clonal response levels

CDR3s are the portion of the TCR that interacts most closely with an antigen. Investigating the amino acid characteristics of these CDR3s could help identify properties of the antigen it binds. In order to investigate the properties of amino acids that make up the CDR3 in healthy controls, all TCRB clonotypes were combined and then split into four groups according to clonal response level as defined in **Section 4.5.4**. CDR3 length, acidity, basic values and GRAVY (grand average of hydropathy) were the factors analysed.

When comparing these factors there was no significant difference at each property level between response levels (Kruskal Wallis) or when each response was compared to one another pairwise (Wilcoxon). The mean CDR3 amino acid length was 15 amino acids, as previously investigated in **Section 4.5.2.** (**Figure 35**, bottom). When investigating the percentage of acidic residues in normal control CDR3s the mean values were between 6.6% and 8.8% across the response levels. When looking at the percentage of basic residues in the CDR3, the mean value was between 0 and 5.8%. The zero percent was from the largest hyperexpanded clone in the normals at 53% which would not be regarded as a "normal" response. The final factor "GRAVY" measures the grand average response of hydrophobicity [**271**]. For all except the high responding clone, GRAVY was a negative number ranging between -0.218 and -0.436 indicative of more non-polar, hydrophilic residues in the CDR3. The high responsive cohort of one sample saw a mean GRAVY value of 0.287.

The other response levels saw positive values when including their maximum values, with no clonal expansion seeing the highest value of 1.19, which would indicate more polar residues in the CDR3 and more hydrophobic characteristics (**Figure 41**.).

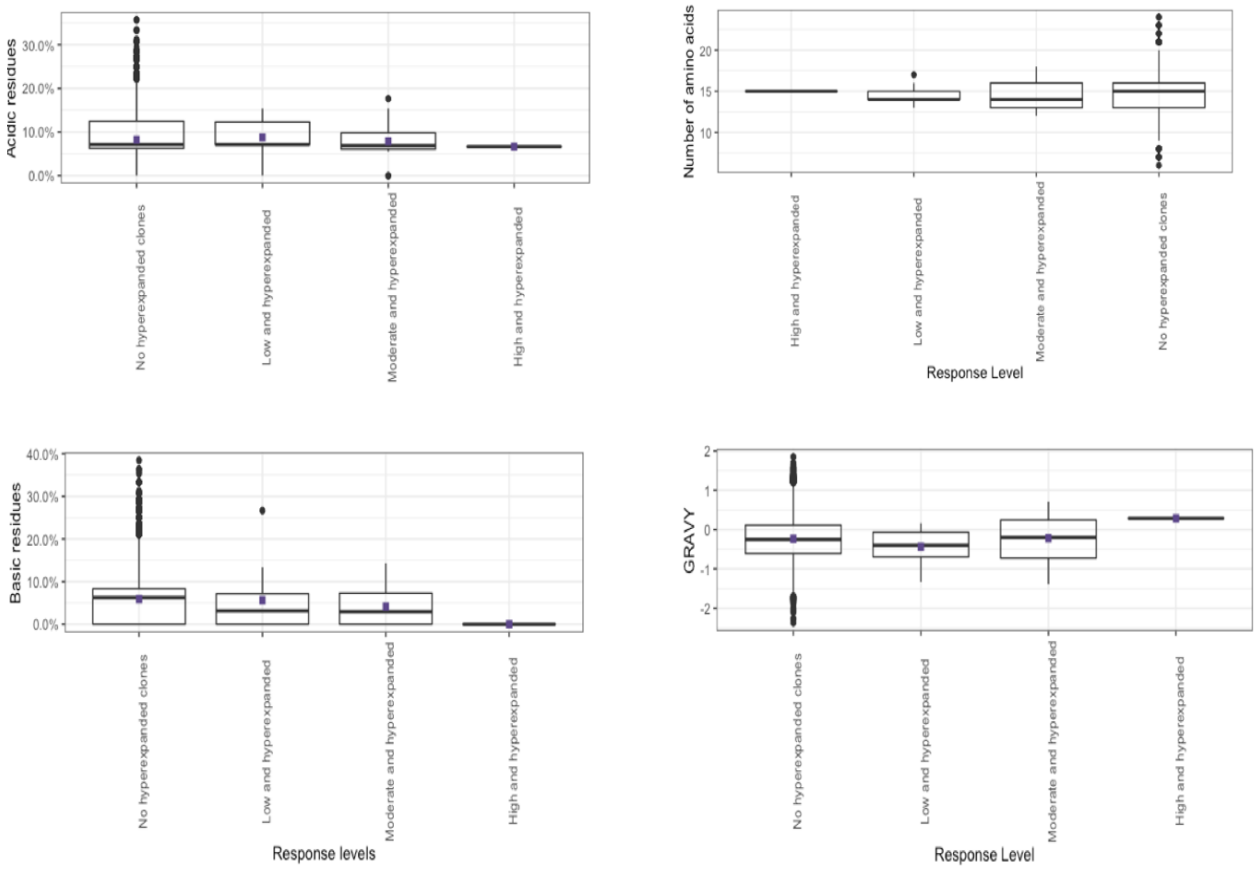


Figure 41. CDR3 amino acid characteristics in 31 healthy controls grouped by clonal response.

TCRB repertoires generated from 31 healthy controls were grouped together according to response level: no hyperexpanded clones (n=25,000+), low response hyperexpanded clones (n=14), moderate response hyperexpanded clones (n=14) and finally hyperexpanded clones with high response (n=1). The CDR3 amino acid sequences were then analysed and the following factors compared: CDR3 length (top right), acidity of residues (top left), basic value of residues (bottom left) and GRAVY index of residues (bottom right) (a measure of hydrophobicity). Outliers are indicated as black dots, the plots are standard box and whisker plots, with the mean value indicated by the purple square at each response level. Overall, there was no significant difference in values for each factor when comparing response level means (Kruskal-Wallis), or when comparing response levels pair wise at each level (Wilcoxon).

4.5.8.2. K-mer and positional residue analysis

The position and characteristic of an amino acid in a CDR3 has been linked to autoimmune disease. For instance, a hydrophobic residue in position 6 and 7 in mice, was found to be linked with self-reactive T-cells [**88**]. In order to evaluate as to whether there were any shared positional characteristics of clonal CDR3s in healthy controls, positional kmer analysis was used. Splitting CDR3s into kmers allowed CDR3 patterns to be compared irrespective of variations in lengths. Kmer analysis was performed on the CDR3s within their response levels using the whole CDR3 (k=14, maximum allowed), and then at two smaller levels (k=4,8) to analyse potentially more central sections of the CDR3 thought to interact directly with a peptide [**272**].

When observing kmers of length 4, no hyperexpanded, low and moderate response levels, showed similar trends in amino acid patterns. Serines and polar residues dominated most positions (**Figure 42.**). Whereas the one hyperexpanded clone saw a mixture of polar and hydrophobic residues, tyrosine and valine.

Octomer sequences contained the majority of polar residues at all response levels. The first three response levels showed similarities with the first three positions being serines and the final position being a hydrophobic phenylalanine. The hyper expanded kmers differed as they had strong positional hydrophobic residues in the first two positions followed by polar Y repeats (**Figure 43.**).

Analysing the CDR3 sequences when broken up into 14/15mers (**Figure 44.**) showed high similarity in the no and low hyperexpanded response levels. The first four bases being ‘CASS’ and the last three ‘QFF’. The pattern seemed to be first position polar residue, second hydrophobic, third and fourth polar. Positions 13-15 were neutral then two hydrophobic residues. Moderate responsive clones observed an almost identical pattern but replacing residue 2 with a polar serine, not a hydrophobic alanine. The hyperexpanded clone did not show much similarity to the other response levels and had a large hydrophobic component.

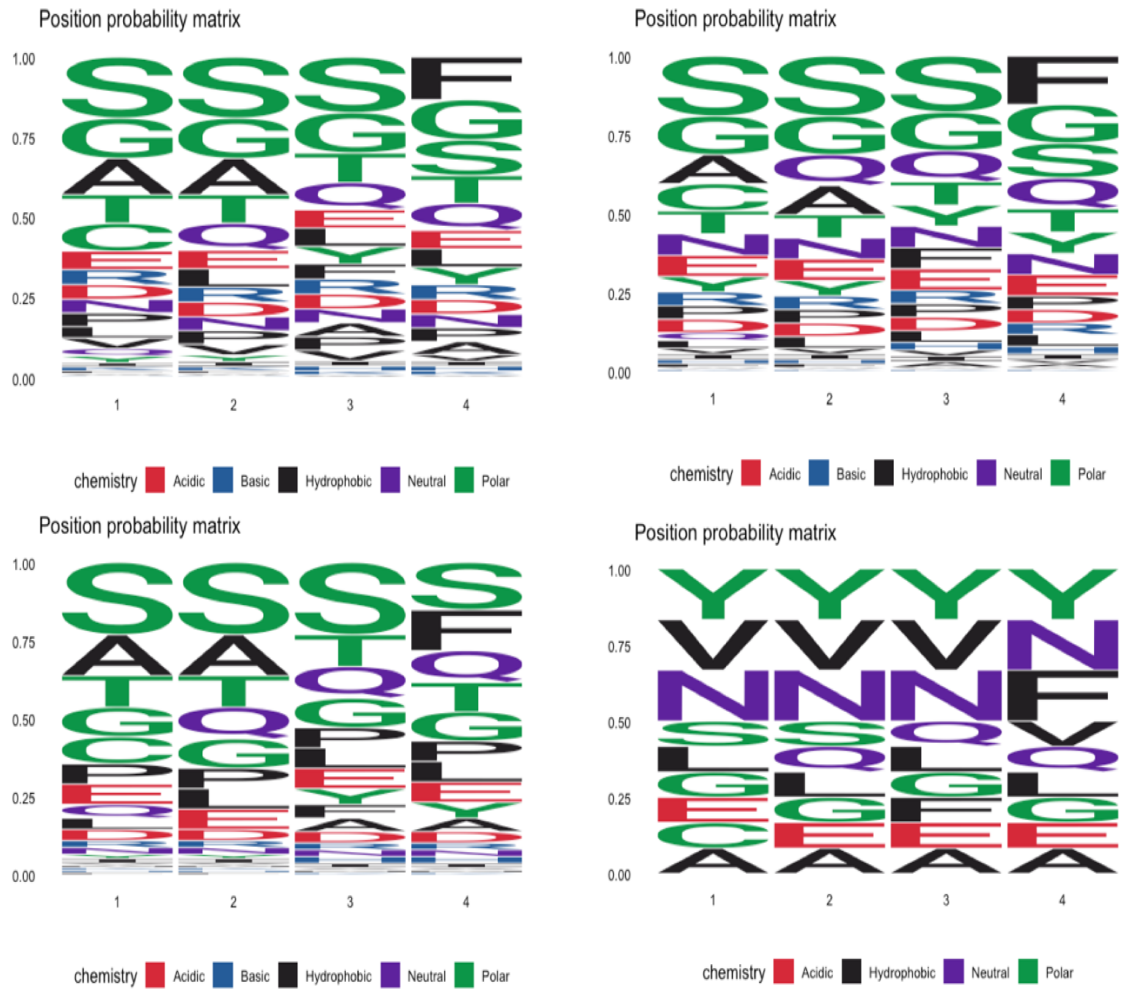


Figure 42. Positional characteristics of CDR3 amino acid residues in healthy controls as 4mers. Positional characteristics and properties of CDR3 sequences broken up into 4mers from 31 healthy controls were analysed and displayed as a position probability matrix. The results were grouped into TCRB clonal response levels: no hyper expanded TCRB (top left) (n=25,000+), low hyper expanded (top right) (n=14), moderately hyperexpanded (bottom left) (n=14) and highly hyperexpanded (bottom right) (n=1). The no, low and moderate groups observed similar trends with the majority of polar residues whereas the highly hyperexpanded residue contained polar and hydrophobic residues.

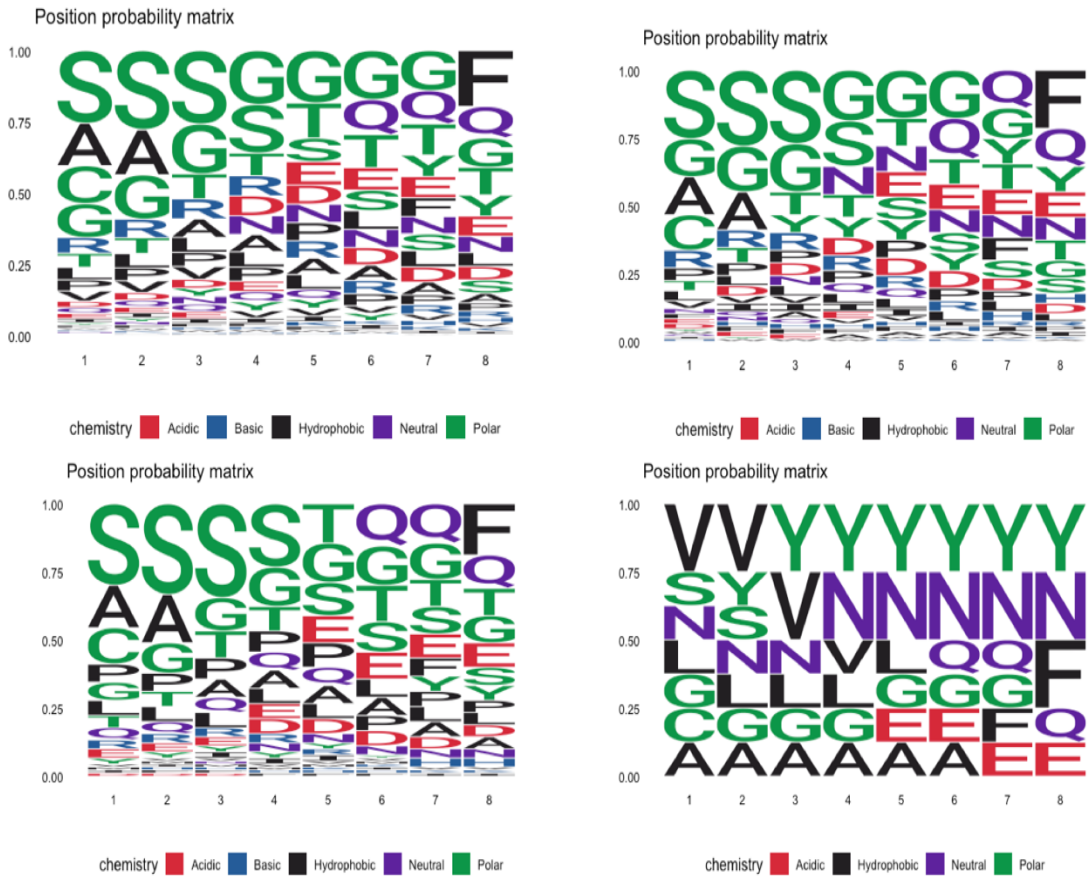


Figure 43. Positional characteristics of CDR3 amino acid residues in healthy controls as 8mers. Positional characteristics and properties of CDR3 sequences broken up into 8mers from 31 healthy controls were analysed and displayed as a position probability matrix. The results were grouped into TCRB clonal response levels: no hyper expanded TCRB (top left), low hyper expanded (top right), moderately hyperexpanded (bottom left) and highly hyperexpanded (bottom right). The no, low and moderate groups observed similar trends with the majority of polar residues, the first three bases being polar serines and the last position being a hydrophobic phenylalanine. Whereas, the highly hyperexpanded residue contained polar and hydrophobic residues only.

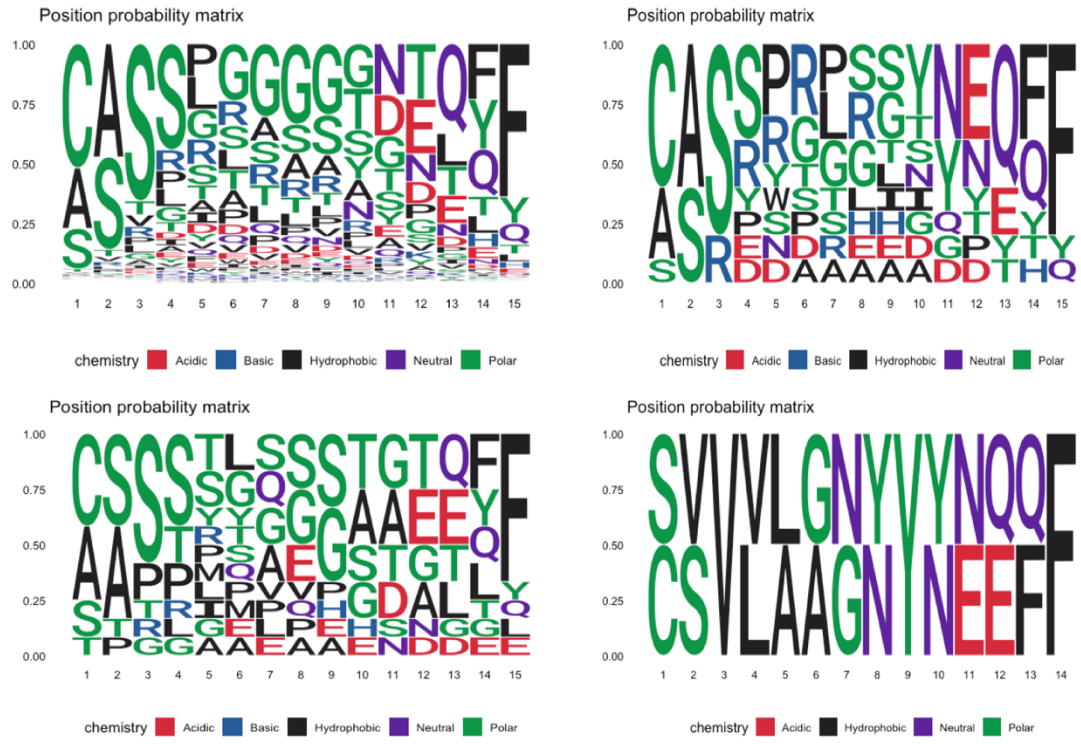


Figure 44. Positional characteristics of CDR3 amino acid residues in healthy controls as 15mers. Positional characteristics and properties of CDR3 sequences broken up into 14/15mers from 31 healthy controls were analysed and displayed as a position probability matrix. The results were grouped into TCRB clonal response levels: no hyper expanded TCRB (top left), low hyper expanded (top right), moderately hyperexpanded (bottom left) and highly hyperexpanded (bottom right). No and low hyperexpanded response levels showed high similarity with the first four bases being 'CASS' and the last three 'QFF'. The pattern seemed to be first position polar residue, second hydrophobic, third and fourth polar. Position 13-15 were neutral then two hydrophobic residues. Moderate observed an almost identical pattern but replacing residue 2 with a polar serine not a hydrophobic alanine. The hyperexpanded clone did not show much similarity to the other response levels and had a large hydrophobic component.

4.5.9. TCRB repertoire diversity was not affected by age or sex across 30 normals

Diversity measures were used to investigate the effect of age and sex on TCRB diversity in the normal population. This included d50, the number of TCRB clonotypes needed to represent 50% of the repertoire using the most abundant clonotypes first. None of the measures indicated a significant difference between age groups or sex for the healthy controls. The control with the large 53% clone was removed in the diversity analysis to prevent bias.

The Chao1 estimator had a mean value of 731 which was lower than the average number of unique TCRB clonotypes found in the healthy controls of 971 suggesting sequencing discovered the majority of the TCRB diversity in the repertoire.

Taking the average number of unique TCRB as 971 (30 healthy controls), the average d50 observed of 118, suggests the top 12% of the unique clonotypes account for half the proportion of the entire repertoire. The smaller the d50 the more clonally expanded the repertoire. The Gini index calculated the inequality of TCRB clonotype size across the control TCRB repertoires (maximum value possible would be 1, representing one clonotype occupying the whole repertoire, minimum value of 0, meaning all clonotypes are equal). The mean value across the normals observed was 0.417. The index could potentially be used to infer the stability of the TCRB repertoire as higher values are indicative of skewed repertoires. As on average, the controls did not have many clonally expanded populations, the Gini index and d50 values will be useful when trying to assess the clonality of PNH samples.

The Gini-Simpson index was used to calculate the probability that two TCRB receptors selected at random represented different TCRB clonotypes. The maximum value possible would be 1. All Gini-Simpson values were observed above 0.9 (**Figure 45.**) indicative of highly diverse TCRB repertoires.

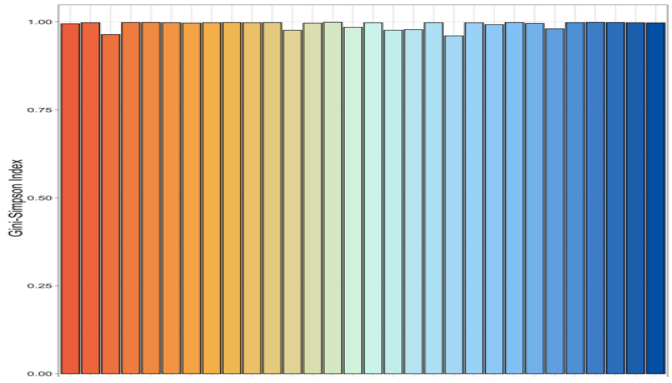


Figure 45. Gini-Simpson analysis of 30 healthy controls indicative of diverse TCRB repertoires. *BIOMED-2 primer method* was used to sequence the TCRB repertoires of 30 healthy controls. All Gini-Simpson values were above 0.9 indicative of highly diverse TCRB repertoires.

The Rarefaction method was used to assess species richness from the results of sampling using an extrapolation method. Generally, the curve for the controls increased rapidly as common clonotypes were discovered, then tailed off as the rarer clones were the final clonotypes to be observed. All but one sample showed saturation of diversity from the sequencing of their repertoires, suggesting adequate depth of sequencing was achieved ensuring an accurate representation of species diversity in each sample (**Figure 46.**).

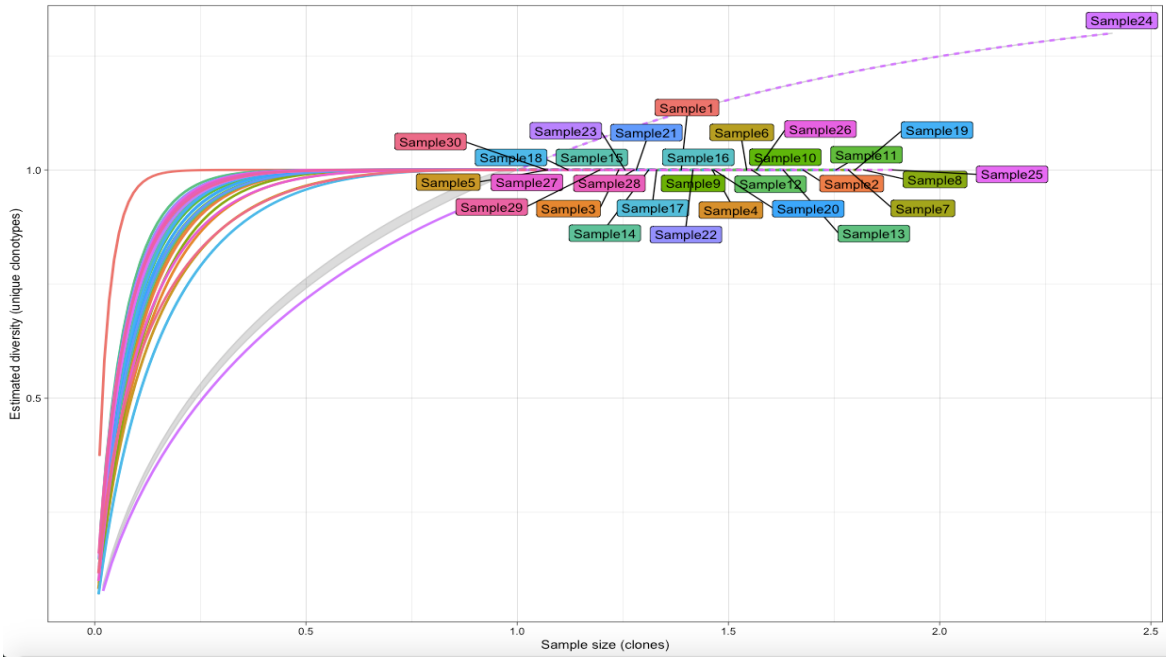


Figure 46. Rarefaction analysis of 30 healthy TCRB repertoires indicative of sufficient sequencing depth. *BIOMED-2 primer method* was used to sequence the TCRB repertoires of 30 healthy controls on which rarefaction analysis was performed. The solid lines were calculated using interpolation and the dotted lines are as a result of extrapolation. A sharp initial increase in the sample curves were suggestive of the detection of common TCRB clonotypes which slowed as the rarer clonotypes were identified. The plateau occurred when the analysis achieved “true diversity”.

In order to calculate both richness of TCRB species and the evenness of the populations, the Inverse Simpson method was applied to the control data. The minimum value possible would have been observed if only one clonotype was present in the repertoire. The maximum value would only have been achieved if perfect evenness had occurred across the repertoire and would be equal to the number of clonotypes. Higher indices values observed in the cohort were suggestive of stable repertoires with high evenness and richness of clonotypes.

The majority of the repertoires showed stability with values above 400 (**Figure 47.**). Those with lower richness and evenness could have been the more clonal repertoires with a few highly responsive TCRB clones. This did not include the individual with the hyperexpanded TCRB clone.

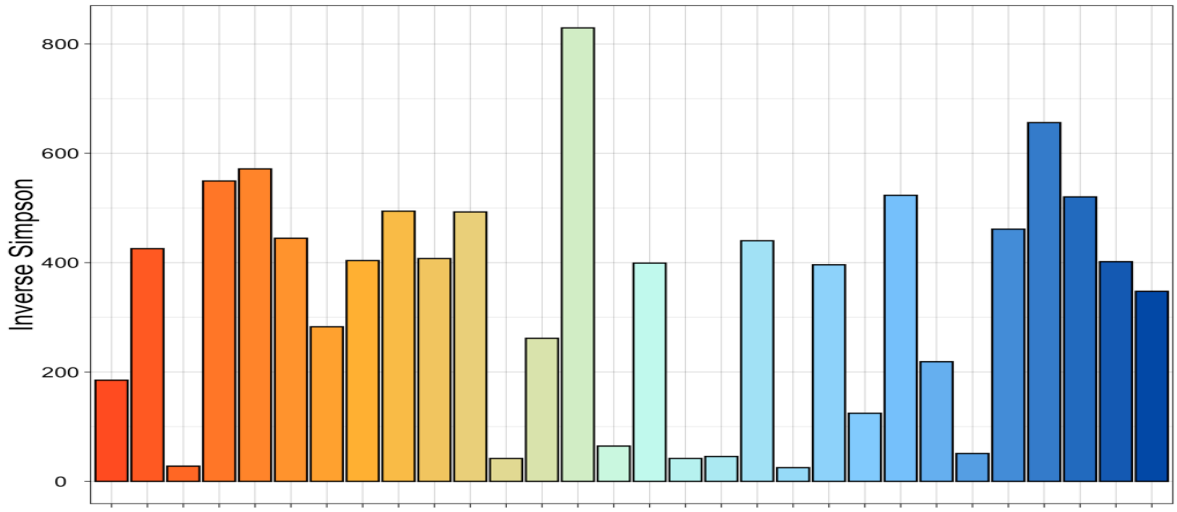


Figure 47. Inverse Simpson values across 30 healthy TCRB repertoires were indicative of variability in the natural population of TCRB repertories. The minimum value possible would have been observed if only one clonotype was present in the repertoire. The maximum value would only have been achieved if perfect evenness had occurred across the repertoire and would be equal to the number of clonotypes. Higher indices values observed in the cohort were suggestive of stable repertoires with high evenness and richness of clonotypes. The majority of the values were showing stability with vales above 400. Those with lower richness and evenness could have been the more clonal repertoires with a few high responding TCRB clones.

Hill values are a mathematically unified family of diversity indices used to allow the easy comparison of differences in diversity across multiple samples using different diversity measures. The values differed only by the exponent 'q'. Q =0 measured repertoire evenness, q =1 represented the Shannon index and q=2 was the Simpson diversity index. Although a variety of TCRB clonotype numbers were observed amongst the control samples, the general trend was similar (**Figure 48**.). There was a gradual decrease in diversity estimation with q value increase. The steeper the sample line the more indicative of unevenness in the repertoire. The flatter the line the more even the repertoire. The majority of normals appear to have some unevenness in the repertoire but all clonotypes are not completely uneven (clonal). Controls displayed a mixture of rarer and abundant TCRB clonotypes.

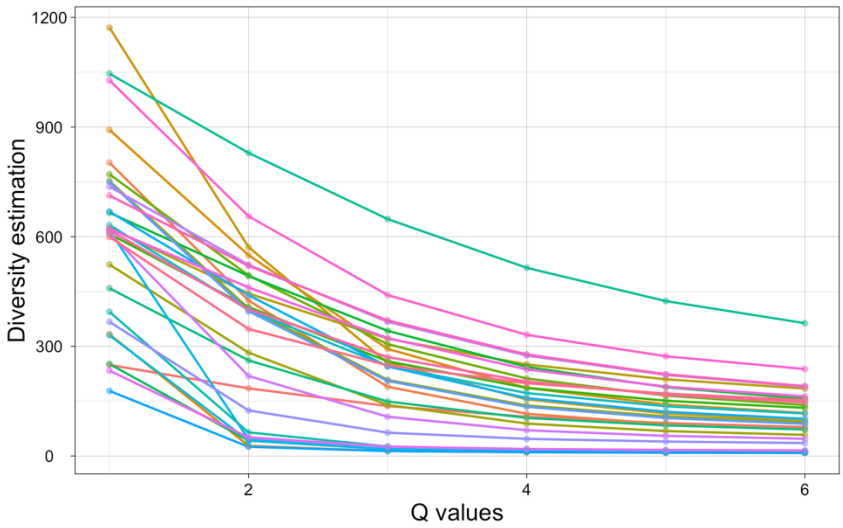


Figure 48. Hill indices highlighted variability in healthy control diversity. Hill family indices were applied to TCRB repertoires of 30 healthy controls. The *BIOMED-2 primer method* was used to generate sequencing producing TCRB repertoires. Q =0 measured repertoire evenness, q =1 represented the Shannon index and 2 was the Simpson diversity index. The general trend was similar across the normals with some unevenness observed but no repertoires were highly clonal. Rarer and more abundant TCRB clonotypes were present.

Using multiple diversity measures enabled deciphering of richness and evenness of TCRB clonotypes as well as abundance of rare and more common TCRB clonotypes. This identified a variety of diversity characteristics naturally observed amongst a population, along with similarities including high TCRB diversity. The results will serve as good baseline values when comparing diversity statistics amongst PNH samples in the next chapters.

4.6. Chapter summary

The work in this chapter was successful in achieving the aims outlined in its' introduction. An experimental TCRB method using an adapted *BIOMED-2 primer method* was successfully developed along with a robust bioinformatics pipeline. This enabled gDNA to be processed into TCRB sequencing libraries and successful sequencing reads were processed to recreate TCRB repertoires from blood samples (method outlined in **Chapter 2.**). This method was used to analyse 31 healthy controls and went on to be used in subsequent chapters to analyse TCRB repertoires in PNH and AA patients. By sequencing 31 healthy controls, a good base line for "normal" TCRB responses was created which will enable successful comparisons to be drawn in the next chapters as to whether any differences observed in PNH repertoires are disease specific or attributed to natural variation within the human population.

4.6.1. Important findings

4.6.1.1. Significant optimisation and testing of methods were required to ensure reliable TCRB repertoire data

Sections 4.2 and **4.3,** surprisingly emphasised that not all published TCRB methods re-create accurate representations of TCRB repertoires from blood samples. The *Robins et al. primer method* showed significant skewing in the repertoire for TCRBV7-2 usage in a normal sample that was not expected to be largely clonal. Problems with skewing of these primer sets have not been detailed in any papers. In fact, TCRBV7-2 was detailed as common in normals and patients with diabetes in one paper [**273**] that references the *Robins et al. primer method*, highlighting that the usage of V7-2 was expected to be prominent when using this method. Perhaps in a similar way that V29-1 is common in normals when using the *BIOMED-2 primer* method

Extensive work was undertaken to understand the reasons for the skewing and to try and counteract it in this project. It was however, only by developing a second method using the *BIOMED-2 primer method* that it became obvious that the skewing was attributed to the primer sets themselves. Interestingly, since the paper detailing the *Robins et al. primer* sets was published [**238**], the research group has published a new method related to these primers that now uses internal normalisers in the experiments to ensure accurate, unbiased results [**274**].

Being able to show that a sample which had had TCRBV flow cytometry performed on it, identifying V27 as most highly used, and then also finding this as the top V family usage when using the *BIOMED-2 primer method* on the same patient but not in the *Robins et al. primer method*, again strengthened the reliability of using the *BIOMED-2 primer method*. This was why the *BIOMED-2 primer method* was used for all subsequent experiments. In reality, multiple methods should not be needed to prove a TCRB repertoire data set is an accurate representation of the repertoire. The implications of this will be discussed in more detail in **Chapter 7**. In this chapter extensive work was carried out to optimise the *BIOMED-2 method* from the concentration of gDNA to use, accounting for sequencing run variation, to capturing as much TCRB diversity from a sample. In reality, more work could be performed to optimise this method further. For example, the rarefaction plots in **Figure 46.** suggested that the majority of the samples had achieved TCRB saturation. Exhaustive sequencing would not detect any new TCRB clonotypes. However, one study suggested that from one sample, around 1,000,000 unique TCRBs could be identified, much higher than the maximum values in this project [**121**]. This will fluctuate between individuals but serves as a reminder that more optimisation, testing and analysis can always be performed.

4.6.1.2. Defining a TCRB clonal population

One of the most significant findings was designing a method to define TCRB clonality. Often it is assigned an arbitrary cut off value of 1%. However, the calculation in this project takes into account artificial inflation of clonality attributed to PCR amplification. By using the assumption that the majority of normals would not have a clonal top TCRB, the 31 healthy controls grouped together into distinct response levels which aided the definition of clonality as a TCRB population with the same TCRBV/J, CDR3 amino acid present in a repertoire at a value of above 2.42%. Defining clonality was extremely important to test the project hypothesis that a single or series of TCRB clonal expansions are present and the same in all PNH TCRB repertoires. The measure also took into account the reality that amongst a normal population, some will have clonal TCRBs attributed to infection or other underlying autoimmune disease such as gastroenteritis that will not have been detailed in the RTB metadata.

4.6.1.3. Establishment of a “normal” TCRB repertoire baseline

Fundamentally, the work in this chapter wanted to establish a baseline of results related to normals that could be used to compare with PNH and AA repertoires to decipher disease specific TCRB responses from those that may be attributed to factors such as age or sex.

This was achieved by the findings in this chapter. The majority of normal repertoires were non-clonal, had high diversity, low clonality and had a low number of shared TCRB clonotypes between one another. A range of analysis was carried out to try and identify key areas of the TCRB repertoire that might be of most use in PNH studies. For example, the incorporation of CDR3 k-mer positional characteristics in normals allowed for an antigen specific analysis to be performed. The assumption was that CDR3s responding to the same antigen in multiple people would not necessarily have the same amino acid sequence due to events discussed in **Section 1.4**. However, the CDR3 may have similar positional characteristics for example hydrophobic residues. By analysing normals, the results acted as a baseline of what TCRBs may expect to look like in a non-antigen skewed repertoire. Using a range of diversity measures allowed different biological questions to be answered from the data. D50 tended to address clonality, whilst inverse Simpson focused on diversity.

Generating a public repertoire of TCRB clonotypes from these samples will allow for accurate inferences to be made about any TCRB clonal expansions in the PNH and AA patients. If the clonal expansions are also present in this public repertoire data set, then they are not exclusive to PNH and are likely as a result of a common infection, the origin of which, as of yet, has not been identified in research. By sequencing 31 healthy controls, it allowed for any anomalies such as the “normal” with the hyper expansive clone to be identified as an outlier and trends to be identified between individuals. Expanding the normal data set further will strengthen the conclusions able to be drawn from the data. Overall, the range of analysis, enabled a comprehensive dataset of base values for “normal” TCRB repertoires to be achieved.

4.6.1.4. Analysis results that influenced subsequent analysis of PNH patients

A number of findings enabled more meaningful analysis to be performed in the context of PNH patients in the next chapters. A lack of TCRB clonotypes being annotated with information such as disease associations in the cross-validation search emphasises that as more and more TCRB studies are carried out, more TCRB clonotypes will be able to be annotated in the context of disease and highlights current limitations in defining the pathogenicity of clonal TCRBs.

It also highlights the importance of the “normal” background of TCRBs generated in this chapter. These can serve as a comparison for PNH specific TCRBs or those appearing in normals in subsequent analysis.

Generating a background of normals showed that there was considerable variation between individuals naturally and that it was important to account for factors such as age and sex that could affect TCRB repertoires. Females are more prone to autoimmune disease, whereas males are more prone to infection [**275**]. This is why basic statistics were split according to age and sex, but no significant variances were linked to these factors. This could be due to the lack of normals over the age of 50. No patients were in their 70s or 80s where the effects of immune-ageing are represented in the TCRB repertoire [**276**]. Finally, the low number of TCRB overlap between samples agreed with the study referenced in **Chapter 1**. Perhaps it is unlikely that one or a series of TCRB clones will be present in all PNH patient repertoires. Many analysis methods will be needed to identify TCRB dynamics in PNH patients.

4.6.2. Conclusion

In conclusion the development, optimisation and testing of experimental and bioinformatic processes in TCRB sequencing allowed for a reliable method to be used on the analysis of 31 healthy controls. The findings in this chapter created an extensive collection of baseline values for measures that will be used to assess TCRB repertoire similarities and differences in PNH and AA patients.

Chapter 5- TCRB repertoire analysis of PNH and AA patients

5.1. Introduction

In this chapter the project progressed to using the developed experimental and bioinformatics method for analysing TCRB repertoires of PNH patients. Acting as another control type for TCRB repertoire values, samples from AA patients both with and without PNH were also sequenced and the TCRB repertoires analysed. The reason for using AA patients as a comparison was that AA is a bone marrow disorder known to be mediated by T-cells by means of an autoimmune attack [**277**]. AA TCRB repertoires should therefore differ in some way from normals. If TCRs are involved in PNH too, it would be expected that there would be differences in repertoires between them and the normals. If no differences are observed between PNH and normals, but there are differences between AA and normals then it might be that the hypothesis of TCRs being involved in PNH is not true. However, similar features in PNH and AA repertoires may indicate this T-cell involvement. PNH can develop in the background of AA and it is thought that the immune responses may be similar [**226**] and that GPI-targets in PNH can evade this response, but this is not confirmed. The autoantigens that the TCRs are responding to in each disease, for instance, may be the same or linked. Potentially the GPI loss in PNH could prevent the killing of these mutant HSCs by preventing T-cell interactions. Therefore, AA repertoires serve as the nearest to a positive control in this work.

PNH is caused by an acquired mutation in *PIG-A* that to date has no known cause or risk factors except AA and can occur at any point in a person's life [**278**]. This again adds complexity to the TCRB repertoire analysis. Although the hypothesis centres around a clonal TCRB in PNH, in reality, this could be multiple TCRBs, or a non-clonal TCRB repertoire with CDR3s with specific characteristics that differ from non-PNH repertoires.

Analysis methods included diversity measures, mono versus polyclonality studies, TCRB CDR3 amino acid profiling both positional and chemical, along with TCRBV/J gene family usages. Each repertoire had a countless number of analysis methods that could be carried out to investigate its composition and dynamics. However, methods in this chapter were selected to cover a sufficient range of repertoire descriptors in case specific clonal TCRBs were not identified exclusively in PNH.

Over 100 TCRB repertoires from 77 PNH and AA patients were analysed in this chapter but each one could not be discussed at length. Therefore, interesting cases in both PNH and AA, for instance, those with multiple associated diseases, were selected and analysed at the individual level in **Chapter 6.**

Other trends in the data, for example, CDR3 length, have been generalised by grouping AA and/or PNH patients together for comparisons with normals. Summary statistics have also been included, grouping patients into smaller categories to allow for a meaningful analysis.

The main aim of this chapter was to identify whether there was a specific TCRB or group of TCRBs involved in the pathogenesis or progression of PNH that were present in many PNH patients and exclusive to PNH.

Objectives:

- Identify and carry out a variety of analyses capturing all aspects of TCRB repertoire dynamics and clonality in order to not bias the study towards exclusively looking for TCRB clonality
- Analyse the TCRB repertoires of 77 AA and PNH patients
- Use the AA TCRB repertoires along with the normals analysed in **Chapter 4** as a comparison to identify PNH specific TCRB repertoire characteristics

Some of the data is not displayed in this chapter due to thesis page limits but is available on request.

5.2. Results

As the analysis in this chapter is extensive, some of the findings have been discussed in the context of wider research to help explain rationale for the analysis as it is presented and how it may have directed subsequent experiments in this project. The main findings will then be discussed as part of the chapter summary.

For the main analysis carried out in this chapter, the primary dataset consisted of 76 PNH and AA patient samples processed by the PNH RTB between 2017 and 2019. One patient with a large hyperexpanded TCRB clone (over 40%) was excluded from the primary dataset but included in clonality studies taking the dataset to 77 PNH and AA patient samples during that analysis. Where multiple samples were available for a patient, for instance, over short term time-points (>2 years) the most reliable samples were selected for comparison in the primary data set based on findings from **Chapter 4**. For example, where possible, if a gDNA sample using the standard method in **Chapter 2** and a buffy coat sample were available for the same patient, the standard method sample was used preferentially over buffy coat because of the results detailed in **Section 4.3.3**. Three buffy coat samples were present in this primary dataset as no alternatives were available.

5.3. Primary data-set basic statistics

In this section, basic statistics such as age and sex were evaluated in the context of category groups (**Table 19.**). This allowed for any variations in factors such as age and sex that could affect TCRB repertoires to be identified and assessed when drawing any conclusions as to whether a TCRB response was more likely PNH or AA specific, or, as a result of ageing for example. Basic repertoire statistics included distribution of CDR3 lengths in a repertoire, number of TCRB clonotypes that passed the bioinformatic quality control and clonal abundance which are excellent measures for assessing whether there are any abnormal TCRB dynamics happening within a repertoire. It also allows factors such as sequencing depth to be assessed early on in the analysis. Sequencing depth can affect the number of TCRB clonotypes in a population [**279**]. Any samples with low TCRB clonotypes were tracked back to the original experiments to assess factors such as original T-cell number inputted into the PCR and as to whether the sample was buffy coat. This meant that biological differences could be distinguished from technical.

The average age of the AA and PNH patients in this dataset was 46 with a slightly lower median age of 42, the maximum age was 85 and the minimum 19. The ratio of females to males was 1:1.24. When splitting the data into PNH and AA patients, the sample numbers were 43 and 33 respectively. As immune repertoires can be affected by factors such as age and some studies have linked differences between sex (**Chapter 1**) it was important to assess these in the dataset before drawing any conclusions from the data. The median age for PNH patients was 42 with an average of 47 years old. The youngest patient was 19 and the oldest 85. The ratio of females:males was 1:1.15.

For AA patients, the median age again was 42 with a lower average than PNH at 45 years old. The youngest patient was 23 and the eldest 77. The ratio of females to males was 1:1.36. Age and sex were then broken down into categories according to the clinical status of the patient (category defined in **Section 2.1.2.**) to assess how each category varied from the overall values (**Table 19.**) As there were more males in the study than females it would be expected that most categories had slightly higher numbers of males to females. Interestingly, in the AA patients, both AA with an increasing PNH clone and with a small PNH clone had broadly balanced ratios, with AA patients with no PNH clone seeing a higher skew to male patients with a ratio of 1:5, higher than the average trend across all datasets. AA patients with a small clone showed similar trends in age values as the general AA trend. However, AA with increasing clones were on average 3 years older than the norm. AA patients with no clone were 6 years younger than the average, perhaps suggesting an age element to the onset of PNH (**Section 7.3.2**).

Table 19. Category breakdown according to clinical status of the patient along with basic statistics of average age, median age and ratio of females to males in each patient category for the 76 patients used in the primary analysis dataset. Clone refers to the size of the PNH granulocyte.

Category	Total number in primary data-set	Average age / years	Median / years	Sex ratio Female:male
Aplastic; 60% PNH clone	1	70	70	0:1
Aplastic; increasing PNH clone	6	48	50.5	1:1
Aplastic; small PNH clone	19	44.9	42	1:0.9
Aplastic; no PNH clone	6	38.8	34.5	1:5
Haemolytic; PNH clone decreasing	7	56.8	58	1:1.3
Haemolytic; large PNH clone, stable	17	45.1	43	1:0.7
Haemolytic; new/increasing PNH clone	17	47.5	41	1:1.83
Other (Aplastic, PNH clone decreasing, complex case)	1	27	27	0:1
Other (Haemolytic, large PNH clone, complex case)	1	26	26	1:0
Other (Haemolytic, thrombotic, T-cell LGL)	1	36	36	0:1

PNH patients with a new or increasing PNH clones saw a higher number of males to females than the norm. Haemolytic PNH patients with a recovering (decreasing) clone saw values slightly above the norm and PNH with a large stable clone saw slightly more females in the category than males. PNH patients with a new or increasing clone were around the average age for PNH patients, patients with large stable clones were slightly below, but those with decreasing clones were significantly older in age with an average of 56.8 years, almost ten years above the average across PNH patients. Variations may also be attributed to variances in numbers within each category group.

Firstly, analysis was carried out across the whole of the primary dataset regardless of diagnosis or category. The average number of TCRB clonotypes identified in a TCRB repertoire post- bioinformatics processing was 18,514 with a median of 19,978. When identifying unique TCRB clonotypes, the average was 1217 with a marginally lower median at 1209 TCRB clonotypes. Clonal abundance plots (data available upon request) measuring the number of TCRBs that were more abundant as a stability metric, showed similar trends to those of the normals in **Chapter 4** with most TCRBs not appearing at highly abundant levels. Finally, when analysing the most common CDR3 clonotypes in the TCRB repertoires, the length distributions peaked at 15 amino acids and between 42 and 45nts. No statistically significant differences were observed for these values when comparing between age groups, sex or category. This helped to ensure that there was no obvious bias between factors that could affect TCRB diversity before considering PNH or AA (data available on request). It also highlighted that factors such as age were not skewing CDR3 length of TCRBs in a repertoire. Distributions that do not follow a Gaussian as expected (example in **Figure 49 C and D.**) are indicative of skewing in the TCRB repertoire.

The analysis above was carried out again splitting the data into PNH or AA patients and there were no statistically significant differences between the two patient groups. As at surface level, there seemed to be no significant difference between AA or PNH groups for TCRB clonotypes, CDR3 length, TCRB clonotype abundance or total number of clonotypes, it was important to perform a deeper analysis on these TCRB repertoires. For this, the groups were split further and compared at the type (AA, PNH or normals), diagnosis (AA with PNH, AA no PNH, PNH, normals, spontaneous remission from PNH) or clinical status category level (**Table 19.**).

5.4. Trends and differences in TCRB repertoires of AA, PNH patients and normals

The first part of the analysis involved comparing the patient samples with the normal repertoires analysed in **Chapter 4**. This would help assess if there were any obvious trends or differences observed between normal repertoires versus all patient samples with PNH or AA, helping to identify potentially PNH specific TCRB responses.

5.4.1. Number of TCRBs, unique TCRBs and CDR3 lengths

The 30 normals were compared with 27 AA patients with PNH clones, 6 AA patients with no PNH clone and 43 PNH patients. When comparing the number of TCRBs in each repertoire by clinical status category, statistical significance was achieved between two groups. This was the group "AA patients with a small PNH clone" and an AA patient annotated as a "complex case", but recovering from PNH (padj<0.05, Holm, Wilcoxon). However, the AA complex case was one patient which would mean statistical significance could be biased by this. However, no groups saw significant differences when assessing unique numbers of TCRBs (data available on request).

When splitting the data by diagnosis (AA no PNH clone, AA PNH clone, PNH, normal or spontaneous remission (recovered from PNH)) the number of TCRBs identified was statistically significant between the normals and all other diagnosis groups apart from PNH patients who had recovered (padj<0.05, holm, Wilcoxon) (**Figure 49A.**). Therefore, lower numbers of TCRBs were observed in normals and recovered PNH patients than those with AA or PNH. Although statistical significance was not achieved between diagnosis groups at the unique TCRB level (**Figure 49B.**) the trend is present and perhaps with higher patient numbers this would become significant.

Repertoires were also split into those with a disease (regardless of it being AA or PNH), normals or spontaneous remission from PNH to assess whether an analysis could distinguish between a repertoire from a patient with some degree of bone marrow disruption from normal. The padj value for unique TCRBs between the disease affected repertoires and normals was 0.1 (Holm) with normal repertoires having lower numbers. Although this is a large PNH study, due to the rarity of the disease, sample sizes are smaller meaning that calculating and inferring statistical significances can be difficult. The trend is present even though significance was not achieved.

There were significantly lower numbers of TCRBs (overall numbers of TCRBs irrespective of clones, not unique TCRBs/clones) in normal repertoires than disease affected repertoires, but not between disease affected repertoires and those who had recovered, or those who had recovered and normals. Only statistical significance was achieved between “disease” and “normals” at the TCRB number level. No difference in unique TCRBs was observed for any level of groupings above. No significant difference was observed between D50 values measuring clonality when looking at clinical status categories or diagnosis or for inverse Simpson values measuring diversity. As discussed in **Chapter 4**, TCRB responses are variable between individuals and although grouping patients allows for identifying trends, it could be hiding subtler differences in regard to overall clonality and diversity. This indicated the need for looking at TCRB repertoires on an individual basis some of which are detailed in **Chapter 6**.

Similar trends in CDR3 lengths were shared between diagnosis groups (**Figure 49C.**). The majority of the categories followed the same CDR3 length trend peaking between 14 and 15 amino acids long following a Gaussian distribution indicative of non-antigen skewed repertoires on the whole.

However, all of the “other” categories, referring to a complex PNH case, a complex PNH case with LGL and an AA with a decreasing PNH clone but complex case, peak in CDR3 earlier at 13-14 amino acids, with a slight skew towards shorter CDR3s (**Figure 49D.**). Again, this could be attributed to the low n numbers of the other cases, in total n=3, but shows that the methods are capable of identifying antigen skewed repertoires which could result from the patients having a complex diagnosis.

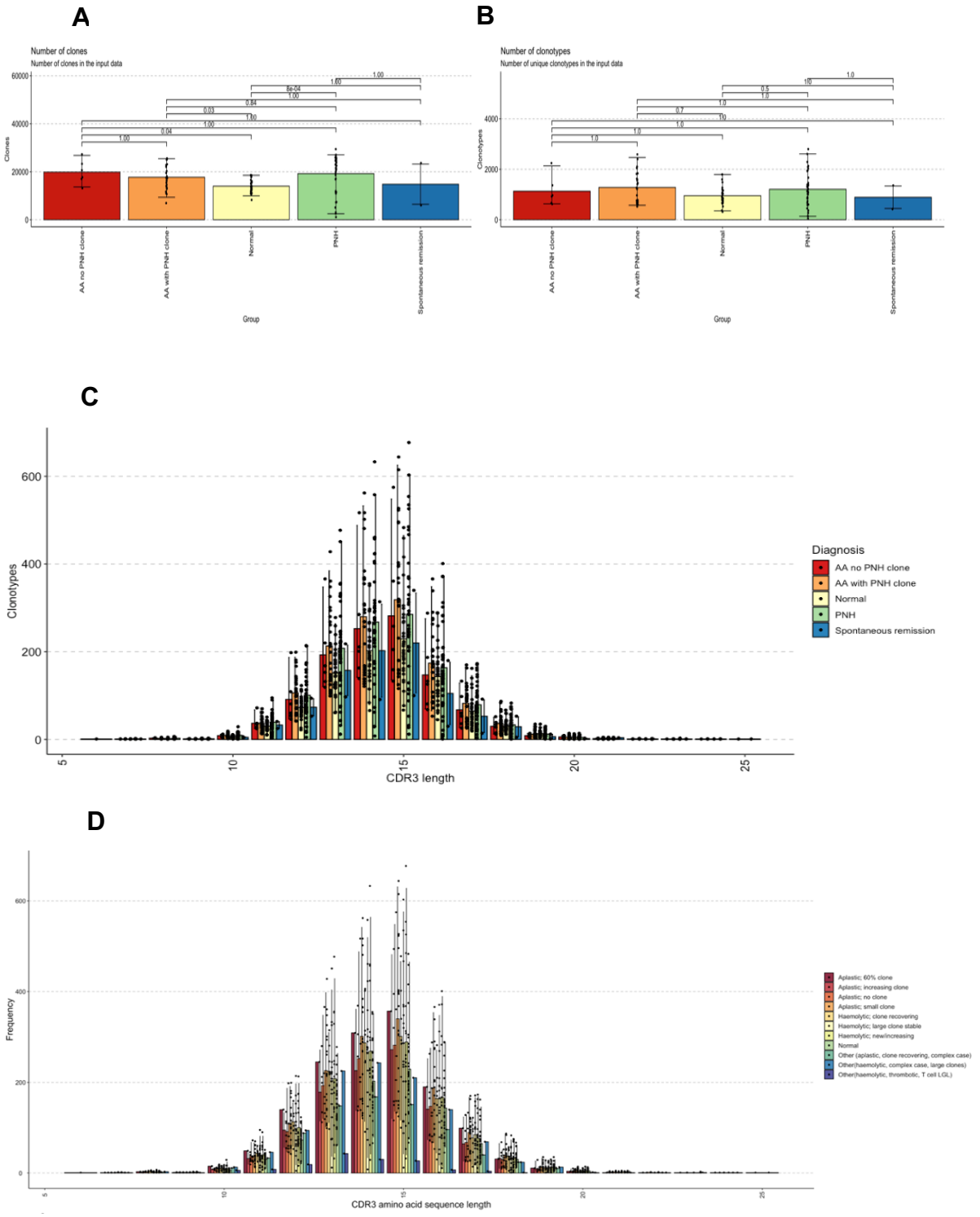


Figure 49. Basic TCRB analysis between diagnosis and category groups of PNH, AA and normals. Total number of TCRBs (**A**) number of unique TCRB clonotypes (**B**) and distribution CDR3 amino acid lengths in a repertoire (**C**) were assessed at diagnosis level, split into normals (n=30), AA no PNH (n=6), AA with PNH (n=27) and PNH(n=43). Distribution of CDR3 lengths was assessed at category level for AA, PNH and normal repertoires (**D**). P adjust values depicted were calculated using the Holm method from Wilcoxon paired testing.

5.4.2. TCRBV and J gene usage in PNH and AA patients

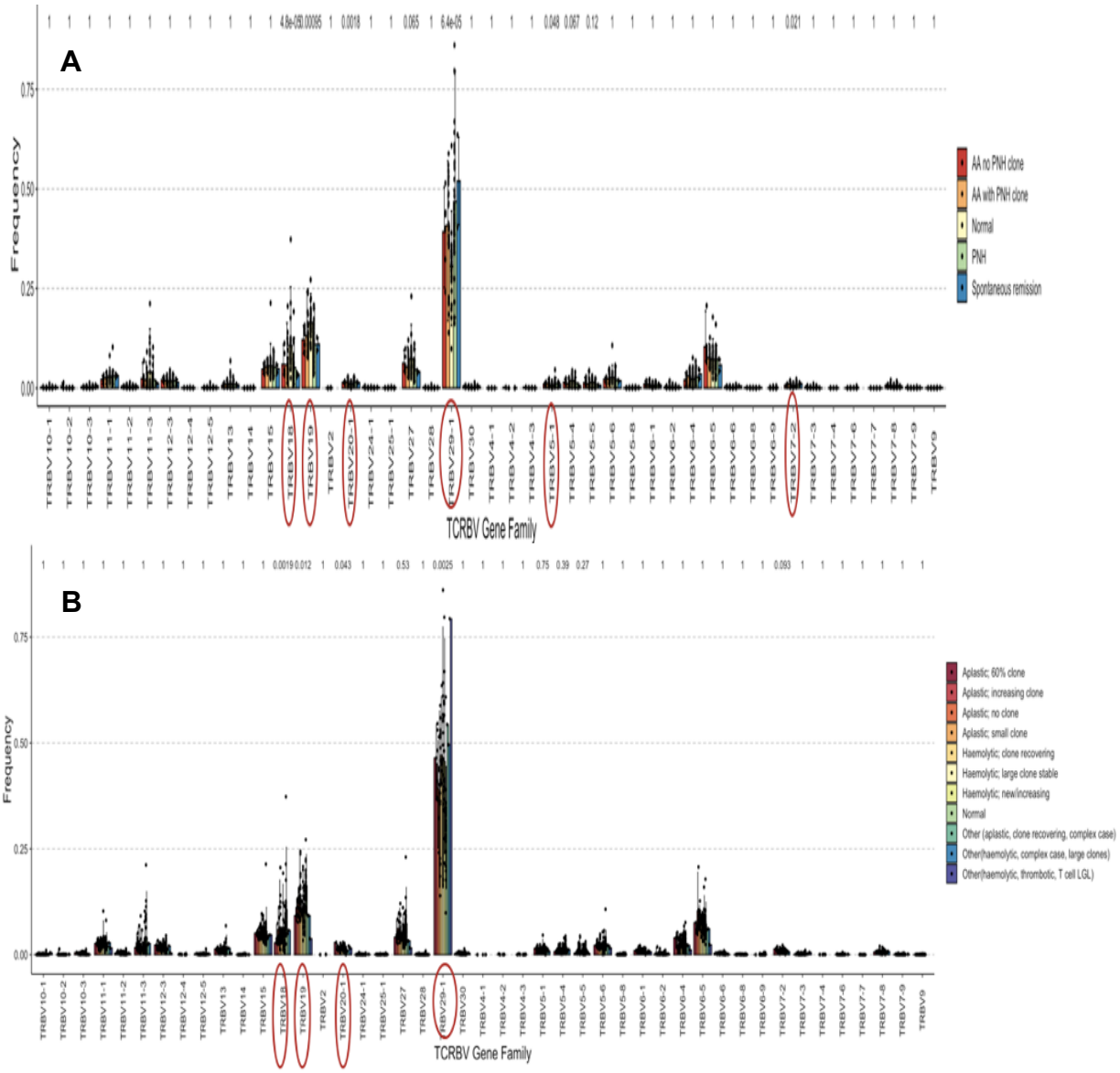


Figure 50. TCRBV gene family usage in TCRB repertoires. Data was split into diagnosis group (**A**) and clinical status category group (**B**). P adjust values depicted were calculated using the Holm method from Wilcoxon paired testing. P adj <0.05 was deemed statistically significant. Red circles indicate V gene families which were significantly different between groups. IMGT® nomenclature was used [**85**].

TCRBV and J gene usage can serve as an indicator as to whether there is any skewing towards particular V and J genes being used in TCRB repertoires that differ from normals. J genes are also thought to attribute to greater variance in the repertoire than TCRBV. This assesses shared properties of the TCRs irrespective of CDR3 clonality.

This means that the analysis is not biased towards looking for an antigen specific TCR in PNH or AA. As discussed in **Chapter 4**, V 29-1 was the most commonly used V gene and J2-1 and J2-3 for the J genes. Samples were first split into diagnosis, then clinical status category groups, in order to assess whether more specific groupings detected more changes.

When splitting the patient samples by diagnosis, TCRBV gene usage variations did occur (**Figure 50A.**). Although the categories and diagnosis groups shared similar patterns in V gene usage, significant p adj values indicate that between the groups there were differences within these families. Statistically significant differences were observed between the groups for TCRBV 18, 19, 20-1, 29-1, 5-1 and 7-2 at diagnosis level.

When observing these changes at the category level, V18, 19, 20-1 and 29-1 were significantly different between groups (padj<0.05, Holm, Wilcoxon) (**Figure 50B.**). This could be attributed to differences in group sizes rather than variances in gene usage between categories which most likely would only be observed when looking at the individual repertoires. This is supported by the finding that statistical difference for V 29-1 and V19 were because of the PNH LGL patient, with an n group of 1. V 19 was significantly lower than other groups for this patient and V29-1 significantly higher. This patient is discussed in more detail in **Section 6.4.6.3**. PNH patients generally saw V29-1 as the most abundant V gene at around 50%, followed by V19 at around 12%, then V15, V18, V27 and V6-5 at similar levels of between 7-10%. Similar trends were observed in AA with PNH patients and AA with no PNH. This trend did not differ much from normals either. However, some TCRBV gene usage variations did occur (**Figure 50A.**).

In regard to TCRBJ gene usage, the general trends in PNH, AA with PNH and AA no PNH patients, again, were similar to each other and to normals, with J2-1 and J2-3 appearing to be the most used in the repertoire (**Figure 51A.**). However, there was significant variation between the groups for some J gene family usage. For example, they all had J2-1 as the second most used J gene family, however, normals showed much higher levels than the other diagnosis groups (padj= 0.015, Holm). Whereas J2-6 and J2-7 was significantly lower in usage in normals than the other categories (padj <0.05, Holm). Potentially this was a result of grouping at the diagnosis level rather than category level. This meant that PNH patients who were recovering/decreasing PNH clone were grouped with those with active responses. There could be variances in TCR according to these groups which is why the data was also

split into diagnosis groups. Variances could also be due to differences in group numbers as discussed above.

At the category level, again similar J gene usage trends were observed for all categories, with J2-1 and J2-3 being the most highly used. This time, J2-1 and J2-6 showed a statistically significant difference between groups (padj<0.05, Holm, Wilcoxon) (**Figure 51B.**). J2-1 had the lowest usage in AA with no PNH clone and the highest in an AA complex case with a recovering/decreasing PNH clone. For J2-6, normals had much lower usage than the other categories. Potentially, again this could be attributed to low group numbers as the AA complex case, with a recovering/decreasing PNH clone, was a group of n=1. Whereas most other groups, the values were averaged across multiple individuals.

No significant differences were observed in TCRBV gene usage when comparing age, sex, AA with or without PNH or thymic involution (patient repertoires over 40 years old or under 40 years old) (data not shown) for PNH or AA categories or when compared to normals.

When comparing only AA categories, those with an increasing PNH clone or a large PNH clone had lower usage in J1-5 and higher usage in J2-5. When splitting males with females, J2-2 and J2-7 were slightly higher in males. Again, none of these findings achieved statistical significance at p<0.05. With a higher n number, the statistical significance may have been achieved and therefore, the findings should not be disregarded. No differences (p adj =1) were found when comparing TCRBJ usage with age, AA with PNH versus AA with no PNH, or thymic involution between AA or PNH patients, or within the groups. This helped to suggest that the J usage was more likely linked to changes in TCRB due to disease rather than factors such as age.

This section's findings suggested that potentially more variation in these patients was observed in the J gene usage than the V gene usage, attributed to more statistically significant differences being achieved between categories.

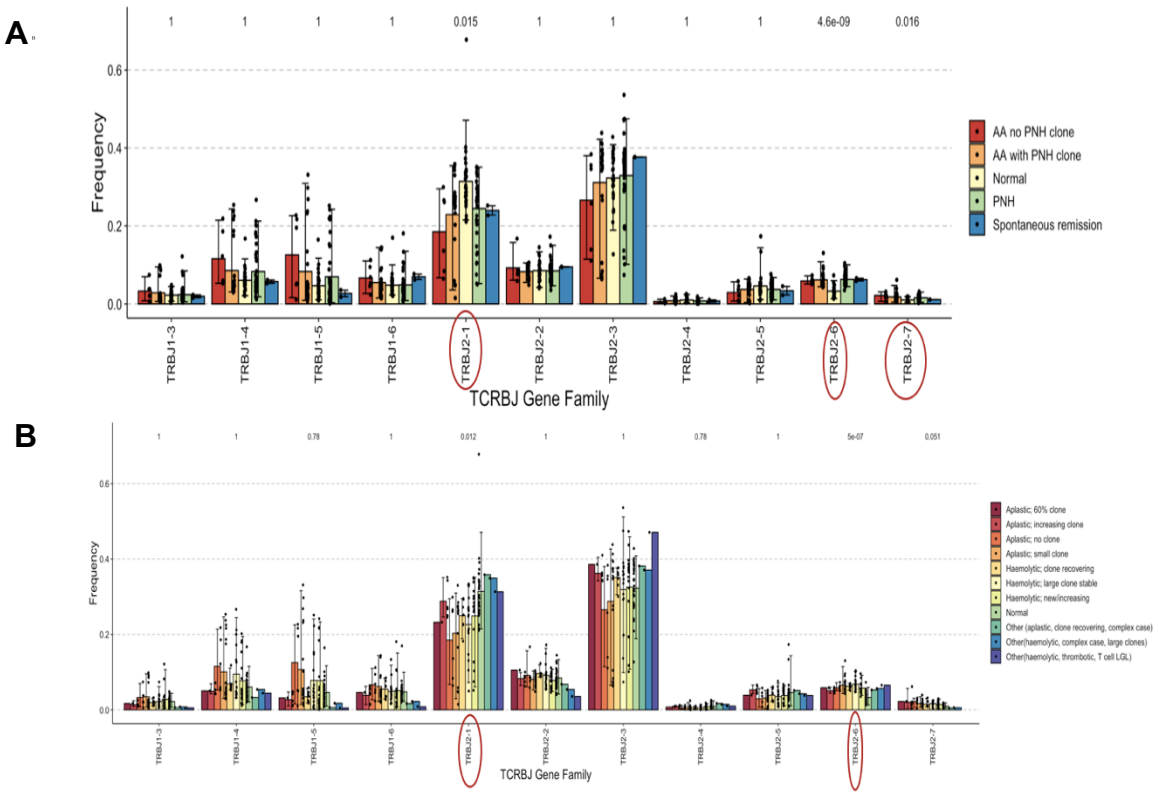


Figure 51. TCRBJ gene family usage in TCRB repertoires. Data was split into diagnosis group (**A**) and category group (**B**). P adjust values depicted were calculated using the Holm method from Wilcoxon paired testing. P adj <0.05 was deemed statistically significant. Red circles indicate J gene families which were significantly different between groups. IMGT® nomenclature used [**85**].

5.4.3. Diversity analysis of AA, PNH and normal TCRB repertoires

There are many different diversity measures used in TCRB repertoires, many have been incorporated from ecology research fields, comparing ecology species in an environment to T-cells in an immune repertoire [**280**]. Measures selected in this analysis allowed different biological questions to be asked from the data in regard to diversity. All diversity measures used the definition of a TCRB clone as having the same TCRBV/J and CDR3 amino acid sequence. The measures d50, inverse Simpson, Chao1 and “true diversity index” were used. D50 calculated the number of clonotypes, starting with the most abundant, needed to achieve 50% of the TCRB repertoire population [**248**]. This was used as an indicator for antigen skewed repertoires. If there was a large clone, or a polyclonal response with moderate responding clones for example, fewer TCRB clonotypes would be needed to achieve 50% resulting in a lower d50. Inverse Simpson measured the abundance of TCRB clonotypes in the repertoire, using the total number of reads and the total number of reads for a particular TCRB clonotype [**249**].

The minimum could be one and the maximum, the maximum number of TCRB clonotypes in the sample. For example, if there were one hundred TCRB clonotypes in the sample and an Inverse Simpson of 90, it would mean the TCRB repertoire is extremely diverse and not clonal. However, if the value was 20 it would indicate a decrease in diversity and more clonal populations. Chao1, is an interesting estimator used to calculate the number of TCRB clonotypes potentially missed from sequencing [**162**]. As discussed previously, many TCRB clonotypes are missed due to sampling biases. Here, the estimator tries to evaluate as to whether the sequencing performed well, detecting rarer clonotypes. Rarer clonotypes may contribute to PNH. The estimator works on the basis that if each clonotype is detected at least twice then likely, no more clonotypes would be discovered from the sample. In reality, exhaustive sequencing would be performed on a sample, but this is extremely costly. True diversity calculated the number of clonotypes needed to achieve equally abundant types (proxy for repertoire stability) [**281**]. Diversity values for d50, inverse Simpson, Chao1 and "true diversity index", saw no statistically significant differences between groups at either clinical status category or diagnosis level.

At the "type" level all AA and PNH patients (n=77) were grouped together versus normals (n=30) or spontaneous remission from PNH (n=2). AA/PNH patient groups had statistically significantly higher values than the normals (padj<0.05, Holm, Wilcoxon) for the "true diversity index", but for the other metrics, no statistically significant difference was observed. This significance was also achieved when the AA and PNH groups were compared separately to normals, with "true diversity" values being higher in the AA and PNH patients (padj<0.05, Holm, Wilcoxon). Lower values in normals could be indicative of more stable repertoires than the AA or PNH patients, as in order for, on average 1000 unique TCRBs per repertoire to be equally abundant, none could be present at a clonal percentage (defined previously as 2.42% or above) in each repertoire. This could mean that AA and PNH repertoires are more diverse than normals, especially in memory T-cell subsets attributed to previous infection or reactions that led to PNH or AA. It could indicate pathological memory T-cells waiting to re-activate. Clonality may be involved in these diseases but only at certain times of the disease. This will be investigated when analysing long term patient samples in **Section 6.3**. Again, what may be happening is that by grouping individual repertoires together, any significant differences in clonality that may occur in one patient, but not another, were being filtered about by the general trends in data. Grouping patients together assumes all TCRBs are reacting in the same way in all patients irrespective of clinical status. Implications of this will be discussed in **Chapter 7**, as factors such as individual treatment history e.g. immunosuppressants [**282**] could also affect these results.

5.4.4. TCRB homeostasis and clonal response levels in PNH and AA patients

Clonality and TCRB response levels in patients were a key focus for this project as it is hypothesised that a single clonal TCRB or series of clonal TCRBs are present in PNH patients. PNH may be linked with clonality, but it is unlikely that a clonal TCRB involved in PNH will be clonally expanded all the time unless linked with chronic infections. As discussed in **Chapter 1**, once an infection is cleared and the T-cells are no longer active, they shrink (become less clonal) and differentiate into memory T-cells which remain circulating in the blood in case of re-infection but at non-clonal levels. If PNH is linked to a particular autoantigen for example, the clonal expansions may be memory subsets at time of sampling waiting to be re-activated. Diversity in the repertoire may actually be in these memory populations due to multiple infections or immunological events leading up to PNH pathogenesis. Therefore, in this section clonal dynamics were investigated along with differences in the types of TCRB clones related to their abundance in the repertoire. AA patients, were included as the positive control for reasons detailed previously, relating to the autoimmune pathogenesis of AA known to be T-cell mediated.

In order to investigate the clonal dynamics of the groups, a number of analyses were performed. These included analysing clonotype abundance and the percentage of the repertoire that the top specified number of TCRBs occupied. The abundances used on the x axis in the plots in **Figure 52.** were used as they were part of the standard method used by *immunarch*, a leading software for immune repertoire analysis [**185**]. At diagnosis, clinical status category and type level (disease, normal or spontaneous remission) (**Figure 52. A, B, C**) clones that appeared between 31-100 times in a repertoire, were statistically significantly lower in normals than in the other groups (padj<0.05, holm, Wilcoxon). As these TCRBs were present, not at clonal levels but were amongst the higher of the non-clonal TCRBs in a repertoire, this could be indicative of the memory T-cell subsets circulating from recent infection. This would explain as to why PNH patients who had recovered also shared this difference to normals. As even though they had recovered from PNH, their immunological memory would be similar to those who still had PNH in terms of memory T-cells and past antigen exposure. This analysis is particularly interesting given that non-clonal TCRBs are hard to detect in sequencing and are often different between technical sample sets as described in **Section 4.3.4.** When assessing the varying numbers of top clonotypes, at the category and diagnosis level, no statistically significant differences were observed between groups with the numbers of TCRBs that occupy a given space in the repertoire. When the x axis was set to 1 this calculated the space that the most abundant TCRB occupied.

Increasing this number gradually allowed the calculation to act as a proxy for clonality in the repertoire as some patients may have polyclonal responses with low responding TCRB clones compared to other patients with one large clonally expanded TCRB (**Figure 52. D, E, F**).

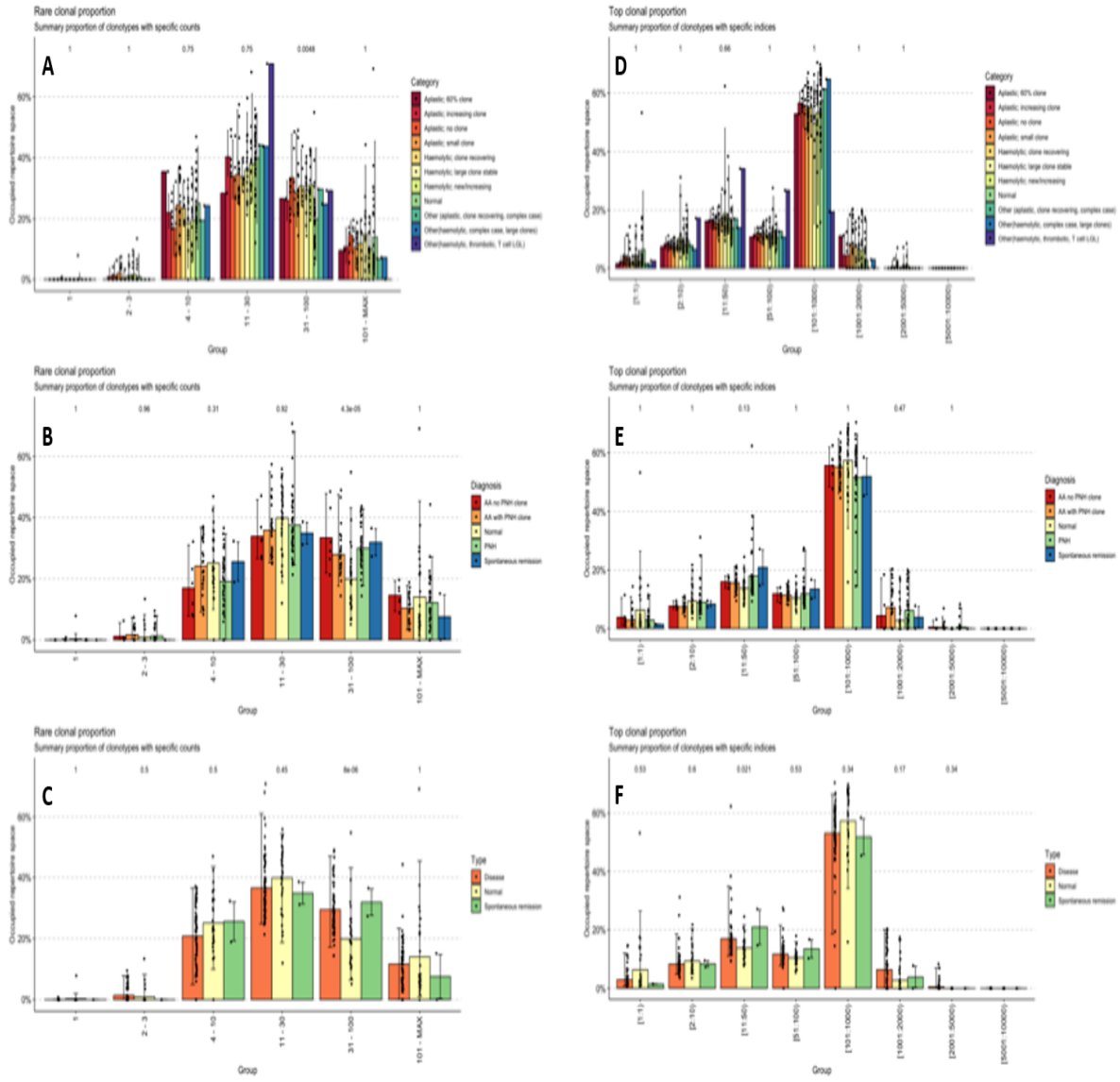


Figure 52. Clonal homeostasis analysis in 43 PNH, 33 AA, and 30 normal TCRB repertoires.

Rare TCRB clonotype analysis within a repertoire was analysed at category (**A**) group (**B**) and type levels (**C**). This calculated the relative abundance of clonotypes split into lower counts and the percentage of repertoire space occupied by them. The number of specific TCRB clones needed to occupy repertoire space was calculated at diagnosis (**D**) category (**E**) and type levels (**F**). The x axis indicates the number of TCRB needed to occupy the percentage of repertoire space that number of TCRBs occupied, acting as a proxy to assess the clonality of repertoires. **.** P adj values were used, calculated using the Holm method from Wilcoxon testing, $p<0.05$ deemed significant).

The top 10 clonotypes accounting for the same repertoire space in normals and the other groups at type level is indicative of clonal TCRBs perhaps to more common infections such as Influenza and CMV rather than TCRBs involved in PNH. Statistical significance between groups was found at the 'type level' between the repertoire space taken up by the 11^{th}-50^{th} most abundant clonotypes, which was slightly lower in normals than the other groups. None of the patients or normals had over 10 clonal TCRB clonotypes present in their repertoire which will be discussed in the next sections. Therefore, it can be assumed that these differences are in non-clonal TCRBs. Again, this could suggest that in AA, PNH and patients recovering from PNH, that the TCRBs that are non-clonal but most abundant, most likely are responding to a recent infection, take up a greater amount of the repertoire than the same number of TCRBs in normals, as they may not have fully contracted. There could me more immunological responses occurring in PNH and AA patients in response to treatment as well. It could also mean that there are subsets of TCRs in PNH and AA that are not clonal in regard to CDR3 but are more abundant than in normals.

However, the trends were generally consistent between groups with most of the repertoire space being occupied by the 101th to 1000^{th} TCRBs, indicative of less clonal and more stable repertoires overall in groups. Clonal TCRB repertoires were not restricted to a single category or group. As expected, the response was variable and will be discussed in more detail in the following sections.

Overall TCRB response levels were analysed at group levels (**Figure 53.**). When analysing TCRB response at the clinical status category, type and diagnosis level, there were no statistically significant differences between the groups at each TCRB response level, non-clonal, low, moderate or high (defined in **Table 13.**). The majority of TCRB populations were present at non-clonal percentages for all groups, PNH, AA or normals (**Figure 53.)**. However, normals appeared to have more moderate and high responding clones with fewer non-clonal TCRBs compared to the other groups. PNH have more non-clonal and low responders. This perhaps indicates that in PNH, the TCRBs that may later be identified as "disease specific", occupy repertoire space that in normals is occupied by moderate and low responders. Again, this supports the homeostasis plot findings, suggesting that it could be non-clonal populations implicated in PNH. Perhaps they are non-clonal because of the definition of clonality in this project which will be discussed in **Section 7.2.1.3**). They may not be antigen specific, therefore defining clonality using a CDR3 amino acid may not be informative. It could also be that it is not clonality that is an issue in PNH, more so the overall make-up of the TCRB repertoire, for example, CDR3s may be more negatively charged. This theory was investigated further in the next section.

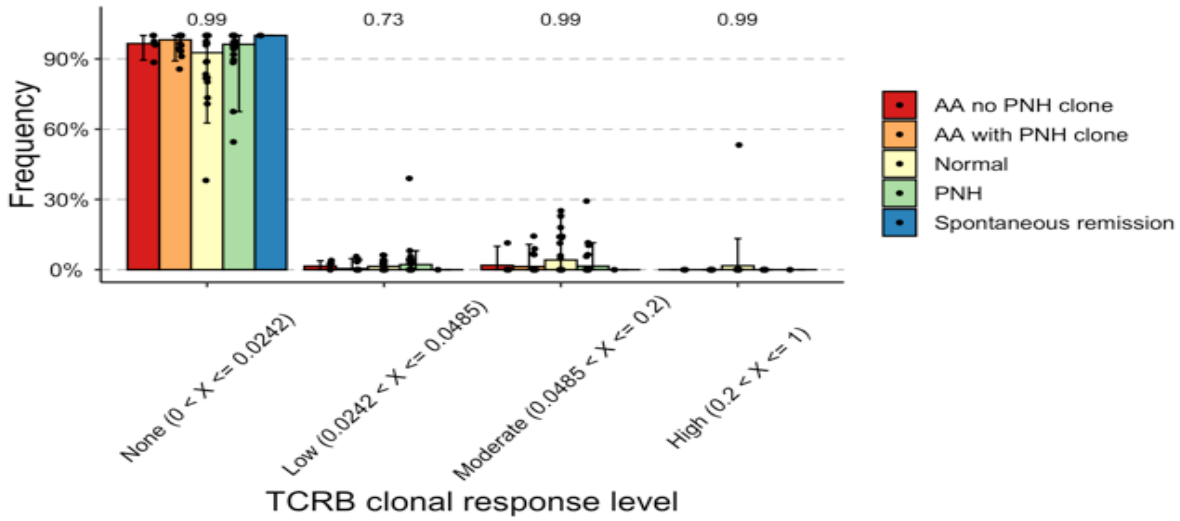


Figure 53. TCRB clonal response levels in PNH, AA and normal TCRB repertoires.

TCRB repertoires were grouped according to AA no PNH, AA with PNH, PNH patient, normal or spontaneous remission from PNH. The TCRBs in each group were then collated and annotated with a response level, non-clonal, low, moderate or hyper defined in **Table 13**. The analysis calculated the percentage of the space per group, occupied by each clonal response. P adj values are shown calculated using the Holm method from Wilcoxon.

5.4.5. CDR3 amino acid properties in PNH and AA patient TCRB repertoires

Individual characteristics of the amino acids that make up the CDR3 will affect features such as charge, size of the protein and its interactions, which could result in different interactions with peptides presented via MHC molecules [**283**]. General trends in CDR3 amino acid sequence lengths did not appear to differ significantly between diagnosis groups. However, it was important to assess other CDR3 amino acid properties. As discussed above, CDR3 amino acid sequence was part of the definition for identifying TCRB clones. CDR3s are the portion of the TCRB that come into contact with an antigen acting as a good indicator as to whether in PNH, the immune response is responding to a particular antigen for example, GPI. T-cell populations may not share the same CDR3 due to processes detailed in **Chapter 1**. that contribute to private TCRB clonotypes, but CDR3s between patient repertoires may share certain characteristics.

Properties of CDR3 amino acid sequences, such as hydrophobicity and length, have been linked to autoimmune disease, vaccinations and stages of immune response [**284**]. Steric recognition of the TCR with the MHC:peptide complex is essential to generate an immune response [**245**].

Interactions such as Van der Waal forces, hydrogen bonds and disulphide bonds between amino acids in the CDR3s and peptides are essential for these interactions. Some amino acid properties, such as those that have bulky side chains and others that are large and aromatic, can make these interactions more difficult or for instance longer CDR3s can make the interaction less specific and easier to interact with multiple peptides including self-peptides [**286**].

In order to investigate this, a number of CDR3 amino acid properties were analysed in the context of diagnosis categories to assess if there were any general trends observed between normals, PNH, AA with PNH or AA with no PNH in TCRB repertoires. Both the average and median were calculated for each group across 9 CDR3 amino acid properties: GRAVY (grand average of hydropathy), length, bulk, aliphatic index, basic residues, acidic residues, aromatic residues, polarity and charge (**Table 20.**). GRAVY measured the ratio of hydrophobic to hydrophilic residues [**271**]. Bulkiness of amino acid side chains are associated with promoting hydrophobic interactions, glycine has a very simple structure and has a bulkiness value of only 3.400 compared to Trp which has the highest at 21.670 [**287**]. The aliphatic index referred to the relative volume of the CDR3 that was taken up by aliphatic side chains, for example isoleucine and valine [**288**). Basic and acidic values measured the percentage of each type at informative positions of the CDR3) [**245**]. Aromatic residue percentages were measured at informative positions as well [**245**] along with polarity of amino acids [**289**]. Charge referred to the overall net charge of the CDR3 calculating the ratio of negative to positive residues. As will be observed below, net charges were observed as being slightly negative in PNH which could indicate more of the negative amino acids D and E being present [**290**]. Some patients' characteristics varied from the norm (**Figure 56.**) and are discussed in **Chapter 6**.

The mean values for each property were compared between groups to assess whether there was any significant difference. Medians were not used to test statistical significance but were detailed in **Table 20.** as an indicator of potential biases from varying n numbers in groups that arise from using mean values. Differences in the mean values of properties in **Table 20.** may seem small to achieve statistical significance. However, the statistical tests took into account overall distribution of the data in each group around the mean, when comparing statistical differences between means of groups. Summaries of trends in some of the interesting CDR3 properties are in **Figure 54**.

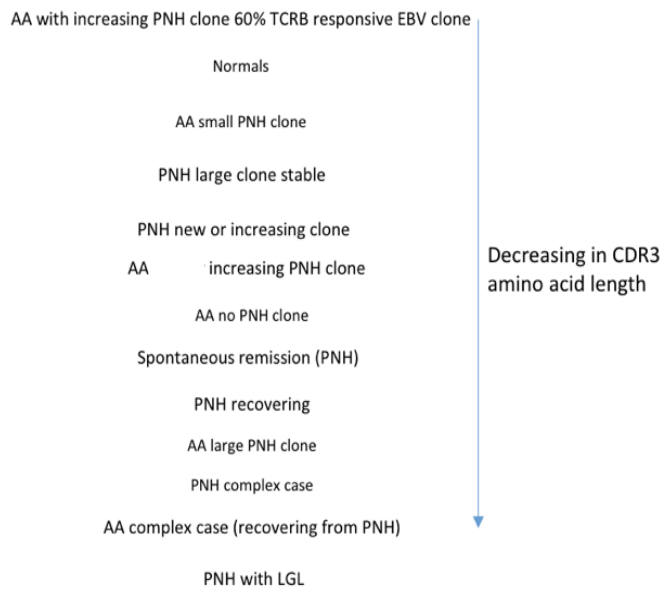
5.4.5.1. CDR3 properties of PNH TCRB repertoires differ from normals

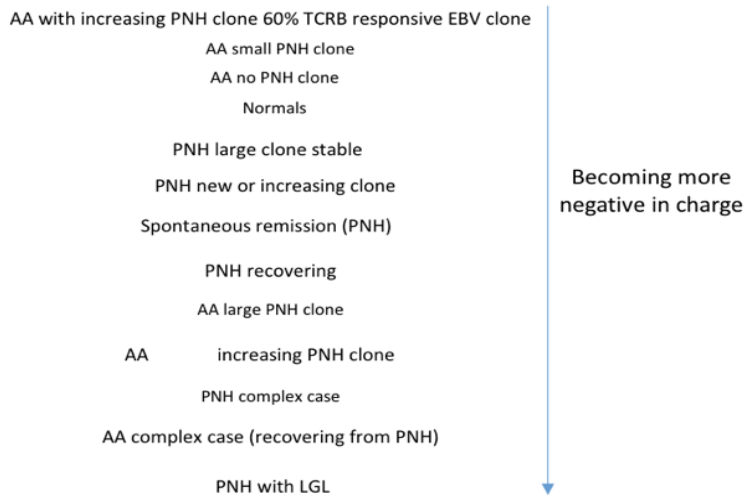
To assess whether overall PNH CDR3 TCRBs were significantly different to normals CDR3 properties were compared between the groups.

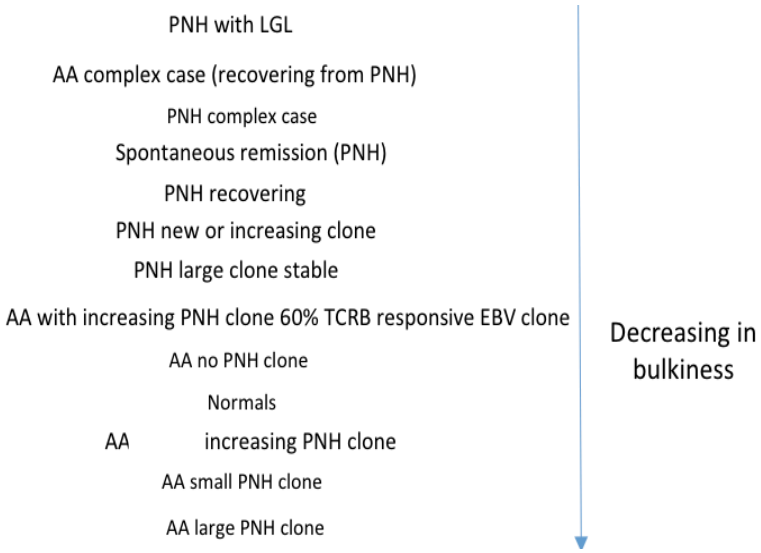
TCRB repertoires showed statistically significant differences for the properties: length (padj<0.0001, Holm, Wilcoxon), GRAVY (padj<0.001, Holm, Wilcoxon), percentage of acidic residues at informative positions (padj<0.0001, Holm, Wilcoxon), charge (padj<0.05, Holm, Wilcoxon), aliphatic index (padj<0.05, Holm, Wilcoxon) and finally, percentage of aromatic residues at informative positions (padj<0.05, holm, Wilcoxon). There was no statistically significant difference for basic residues, polarity or bulk. CDR3s in PNH TCRBs were shorter than in normal repertoires. They also contained slightly more hydrophobic residues than normals but overall CDR3 repertoires were still hydrophilic like normals. Compared to normals, PNH CDR3s also had more negatively charged CDR3s, with higher percentage of acidic residues, higher aliphatic index, and, finally, lower percentages of aromatic residues.

Table 20. Average and median CDR3 amino acid property values for 30 normals, 43 PNH patients, 27 AA patients with a PNH clone and 6 AA patients with no PNH clone.

CDR3 property	Average/median	Normals	PNH	AA with PNH	AA no PNH
Length	Average	14.5000	14.4000	14.4566	14.4067
	Median	15.0000	14.0000	14.0000	14.0000
Gravy	Average	-0.2330	-0.2157	-0.2297	-0.2290
	Median	-0.2530	-0.2380	-0.2533	-0.2528
Basic	Average	0.0588	0.0589	0.0610	0.0634
	Median	0.0625	0.0625	0.0667	0.0667
Acidic	Average	0.0824	0.0844	0.0820	0.0810
	Median	0.0714	0.0714	0.0714	0.0714
Aliphatic	Average	0.3900	0.4009	0.3963	0.4026
	Median	0.3690	0.3769	0.3688	0.3769
Polarity	Average	8.3540	8.3526	8.3594	8.3576
	Median	8.3590	8.3583	8.3643	8.3625
Charge	Average	-0.5560	-0.5884	-0.5527	-0.5184
	Median	-0.9610	-0.9632	-0.9613	-0.9610
Bulk	Average	13.6400	13.6439	13.6209	13.6354
	Median	13.6200	13.6283	13.6113	13.6107
Aromatic	Average	0.1680	0.1667	0.1667	0.1663
	Median	0.1540	0.1538	0.1538	0.1538


AA with increasing PNH clone 60% TCRB responsive EBV clone
Normals
AA small PNH clone
PNH large clone stable
PNH new or increasing clone
AA increasing PNH clone
AA no PNH clone
Spontaneous remission (PNH)
PNH recovering
AA large PNH clone
PNH complex case
AA complex case (recovering from PNH)
PNH with LGL
Decreasing in CDR3 amino acid length


AA with increasing PNH clone 60% TCRB responsive EBV clone
AA small PNH clone
AA no PNH clone
Normals
PNH large clone stable
PNH new or increasing clone
Spontaneous remission (PNH)
PNH recovering
AA large PNH clone
AA increasing PNH clone
PNH complex case
AA complex case (recovering from PNH)
PNH with LGL
Becoming more negative in charge


PNH with LGL
AA complex case (recovering from PNH)
PNH complex case
Spontaneous remission (PNH)
PNH recovering
PNH new or increasing clone
PNH large clone stable
AA with increasing PNH clone 60% TCRB responsive EBV clone
AA no PNH clone
Normals
AA increasing PNH clone
AA small PNH clone
AA large PNH clone
Decreasing in bulkiness

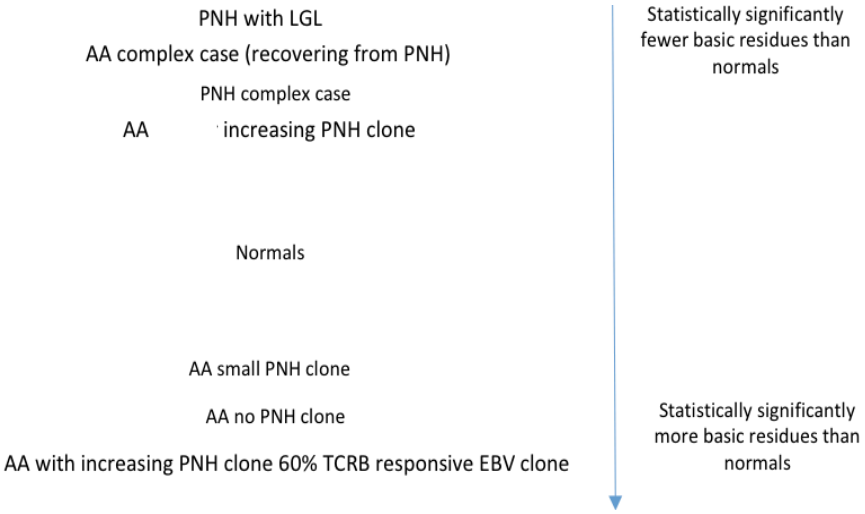


Figure 54. CDR3 property trends of TCRB repertoires from PNH, AA patients and normals. Summary of CDR3 length, bulkiness, charge and percentage of basic residues between TCRB repertoires at the category level.

5.4.5.2. CDR3 properties of AA patients with a PNH clone differ from normals

AA TCRB CDR3s would be expected to differ from normals, due to the T-cell mediated mechanisms underlying AA pathogenesis, however, it wanted to be assessed whether with the addition of PNH clones, differences were still apparent. When comparing normal TCRBs with AA patients with PNH clones, significant differences were observed for CDR3 length (padj<0.0001, Holm, Wilcoxon), basic residues (padj<0.0001, Holm, Wilcoxon), aromatic residue percentages (padj<0.05, Holm, Wilcoxon), aliphatic index and charge (padj<0.01, Holm, Wilcoxon). Overall, in AA patients with a PNH clone, the CDR3s were shorter than in normals, they had a higher percentage of basic residues, higher aliphatic index, were less negative in charge but still negative overall and had a lower percentage of aromatic residues.

5.4.5.3. CDR3 properties of AA patients with no PNH clone differ from normals

As AA patients are known to have T-cell dysfunction, if CDR3 properties are good indicators of immune dysfunction, AA patients with no PNH should see differences with normals. This was the case for some important CDR3 characteristics.

When comparing AA patients with no PNH clone to normal TCRB repertoires, statistically significant differences were seen for CDR3 length (padj<0.0001, Holm, Wilcoxon) , basic residues (padj<0.0001, Holm, Wilcoxon), aliphatic index (padj<0.0001, holm, Wilcoxon) and charge (padj<0.0001, Holm, Wilcoxon). AA patient repertoires generally contained shorter CDR3s, CDR3s with higher percentage of basic residues and higher aliphatic indices. Similar to PNH repertoires, CDR3s in their repertories were also less negatively charged than normal TCRB CDR3s.

5.4.5.4. PNH TCRB repertories differed from AA no PNH clone TCRB repertoires

As PNH CDR3 TCRB characteristics differed from normals highlighting immune dysfunction, evaluating whether these responses were related or different to AA pathogenesis when there is no PNH was essential. In the PNH TCRB repertoires, there were statistical differences between the AA no PNH group for percentage of acidic and basic residues along with overall charge (padj<0.0001, Holm, Wilcoxon). PNH repertoires had lower percentages for basic residues but consequently higher values for acidic and were overall more negatively charged than AA with no PNH.

5.4.5.5. PNH TCRB repertoires differed from AA patients with PNH

Deciphering further, as to whether there were CDR3 TCRBs involved in PNH, which were different from those in AA, even if the two diseases were co-existing, was important. Any shared characteristics that were not shared between AA and AA no PNH may indicate properties of PNH TCRBs. Alternatively, TCRB mechanisms in PNH in the AA context may differ from those who had solely PNH. For this, CDR3 characteristics between PNH patients and AA patients with PNH clones were compared. The only property to have no statistically significant difference between PNH and AA with PNH repertoires was percentage of aromatic residues. PNH repertoires generally had shorter CDR3s, and slightly more hydrophobic residues in their CDR3s than AA with PNH. PNH repertoires had lower percentages of basic residues, consequently higher percentages of acidic residues and CDR3s with a higher aliphatic index. Finally, PNH patient CDR3s had fewer polar residues, were more negatively charged overall, and were structurally bulkier than in AA patients with PNH (padj<0.05, Holm, Wilcoxon). This highlighted the properties of potentially distinct TCRBs involved in PNH when in the presence or absence of AA.

5.4.5.6. CDR3 properties differed between AA patients with and without PNH

By comparing CDR3s in AA patients with and without PNH, it was thought that characteristics of CDR3s involved in PNH in the context of AA could be identified. Patients with both AA and PNH saw statistically significant differences in length (padj<0.05, Holm, Wilcoxon), basic residues (padj<0.01, Holm, Wilcoxon), aliphatic index (padj<0.05, Holm, Wilcoxon) and charge (padj<0.01, Holm, Wilcoxon) compared to AA with no PNH.

Overall, AA with PNH had longer CDR3s, lower number of basic residues, lower aliphatic index and more negatively charged residues than in AA with no PNH. Potentially, this could mean that AA patients who go on to develop PNH could be identified from their peripheral blood TCRB repertoires. In order to investigate this more, it would be interesting to track AA patient repertoires who have no PNH clones and see if they begin to develop a PNH clone, tracking their TCRB repertoires to see whether the changes detailed above occur.

Overall this section highlighted that although repertoires may not be clonal, there are significant differences in CDR3 properties between normals and the disease groups as well as between the disease groups themselves, providing insight into differences in TCRB repertoires. Differences were observed between PNH patients and those with AA and PNH, perhaps suggesting different T-cell subsets are involved in PNH when in the context of AA. Differences in CDR3 properties between AA with and without PNH, such as those with PNH having fewer basic residues, could indicate populations of T-cells with more acidic residues relate only to PNH when AA is absent. The differences observed did not support the PNH T-cell hypothesis in terms of clonal populations but has suggested there are changes in CDR3 TCRBs between PNH, normals and AA patients.

5.4.5.7. Significant differences in TCRB CDR3 properties were observed at clinical status category level

As it was indicated that there were differences in TCRB repertoires when investigating CDR3 characteristics between AA, PNH patients and normals, it was important to repeat the analysis at the clinical status category level. This is because those with new or increasing PNH, for example, are assumed to be in an active, progressive stage of disease. Whereas PNH patients who are recovering or have large stable clones will not have active disease or it will be stable and not progressing.

TCR repertoires will differ under these different conditions and therefore it was important to assess the CDR3 characteristics in the context of each category. CDR3 property analysis between diagnosis groups showed some significant differences as to be expected. Some of the interesting findings were detailed below. All p values detailed are p adj values calculated using the Holm method from Wilcoxon pairwise mean comparisons. Visual trends between categories for some of the interesting CDR3 properties are detailed in **Figure 54**.

CDR3 average length varied between patient category groups

Firstly, as studies have suggested that autoimmune diseases can be attributed to TCRB repertoires with CDR3s skewed towards the shorter size, CDR3 length was analysed [**91,273**] (**Table 21.**).

Table 21. Average and median CDR3 amino acid lengths of AA and PNH patients split by category.

Category	Average CDR3 length /aa	Median CDR3 length /aa
Aplastic; no clone	14.25	13.5
Aplastic; small clone	15	15
Aplastic; increasing clone	13.7	14
Haemolytic; clone decreasing	13	13
Haemolytic; new / increasing	13.7	14
Haemolytic; large clone stable.	13.4	14

When excluding the AA patient with the hyper-expanded EBV specific TCRB clone, all categories except AA patients with a small clone, had significantly shorter CDR3s in their TCRB repertoires than normals ($p<0.001$). AA small clone were longer but not statistically significant from normals. AA patients with no PNH had significantly shorter CDR3s than AA patients with small PNH clones. AA patients with increasing clones had longer CDR3s than recovering PNH patients/those with decreasing PNH clones and shorter than AA patients with a small PNH clone ($p<0.05$). AA patients with small clones had significantly longer CDR3s than all other categories except normals and the AA patient with the large EBV clone. PNH patients with large stable clones or new and increasing clones had longer CDR3s than recovering PNH patients ($p<0.05$). Patients who had spontaneously recovered from PNH had shorter CDR3s than AA patients with small clones ($p <0.05$).

Bulkier CDR3s were observed in the LGL patient repertoire

In this section bulkiness of CDR3s was assessed. Bulkiness refers to whether the amino acids that make up a CDR3 have bulky side chains that can change the way in which CDR3s interact with antigens [**291**]. A previous study showed that introducing an amino acid with bulky side chains such as tryptophan restored reactivity of T-cells against the well characterised autoantigen, myelin basic protein which has been linked with multiple sclerosis [**292**]. Therefore, it was important to assess whether CDR3s in PNH had amino acids with bulkier side chains that may alter the plasticity of T-cells in how they interact with antigens via MHC complexes. When comparing the bulkiness of CDR3s in TCRB repertoires, the lowest bulkiness value was for AA with a large PNH clone at 13.58. When comparing the categories with normal CDR3s, only the AA patient with a large PNH clone had a less bulky CDR3s than normals ($p<0.05$). The other main categories (excluding complex cases with n=1) showed no statistical differences with the normals. AA with the large clone and AA patients with small clones had significantly less bulky CDR3s than recovering PNH patients/those with decreasing PNH clones and PNH new or increasing clones ($p<0.05$).

Little variation was observed between repertoires when investigating polarity

Non-polar amino acids tend to be found in regions that are in contact with membranes [**293**]. TCR signalling involved in T-cell activation, like many processes occurs across membranes [**294**]. Therefore, this property was investigated. Much fewer statistically significant differences were seen between categories when calculating polarity of CDR3s. AA patients with small clones had significantly more polar residues than PNH patients who were recovering (decreasing PNH clones) and PNH patients who had a large stable clone ($p<0.05$).

Significant differences in net charge between normals and diagnosis categories

CDR3s that have an excessive number of positively charged amino acids are thought to impair MHC-dependent TCR signalling and in general undergo death by neglect during positive selection and are removed [**295**]. In PNH, as T-cells are thought to target normal HSCs, it was interesting to assess whether these T-cells had more positive residues and potentially had escaped positive selection. PNH with new or increasing clones, spontaneous remission PNH and PNH with large stable clones had no difference in net charge compared to normals. The AA patient with the large EBV clone, AA patients with no PNH and AA with small PNH clones had more positive residues in the CDR3 although still negative overall.

AA with increasing PNH clones, AA with a large PNH clone and PNH recovering patients all had more negative residues in the CDR3 than normals ($p<0.05$). The AA patient with the large EBV clone had an almost neutral charge at -0.173, which was the least negative. Some studies have suggested that PNH could be nature's cure, trying to re-populate a bone marrow in a toxic immune environment in AA patients [**230**]. AA patients could have more positive residues in their TCRB CDR3s indicative of issues with positive selection and T-cells in PNH have fewer positive residues evolved to try and restore some form of "normal" immune response to improve the bone marrow environment.

Percentages of basic and acidic residues differed between categories

As best to current knowledge, no studies specifically assess basic and acidic CDR3 interactions with peptides in humans for TCRB. However, as this may change as TCRB studies advance the analysis was still carried out. Normals showed significant differences in percentages of basic residues between AA patients with no PNH, those with increasing PNH and AA with small PNH clones. AA increasing had lower percentages of basic residues, whereas AA patients with small PNH clones or no PNH clones had higher basic residue percentages. Spontaneous remission repertoires of patients who had PNH only, had significant differences in basic residues when compared with AA increasing PNH clone patients. No differences were observed within PNH groups, or the AA patient who had a large PNH group. When assessing all acidic residue percentages in the repertoires, only AA patients with no PNH did not have differences when compared to normals highlighting a potential link of acidic residues in CDR3s and PNH. AA with small PNH clones had fewer acidic residues than the normals, with the other categories having more. Spontaneous remission patients showed differences with AA no PNH clones, normals and AA small PNH clone. AA no PNH clones, normals, and AA small clones had fewer acidic residues.

Few differences were observed for aromatic residue percentages

TCR:MHC interactions rely on Van der Waals interactions and hydrogen bonds between amino acids in CDR3s and peptides [**296**]. Large aromatic residues binding to other large aromatic residues have high Van der Waal scores which can be both good for forming immunological synapses but can result in too strong a bond being created, which is problematic if the interaction is with a self-peptide [**297**]. In order to investigate if CDR3s in PNH contained more aromatic residues an analysis was performed. Only AA small PNH clone and PNH large clone stable repertoires showed a significant difference from normals for aromatic residue percentages ($p< 0.001$) and were lower than in normals.

CDR3 property changes in response to stage of PNH

CDR3 properties were analysed in the context of the stage of PNH the patient was at. The aim being to assess whether changes in the TCRB repertoire are dependent on whether a patient currently had PNH, was stable or was recovering (seeing a decrease in PNH clone size). This could potentially identify a biomarker for PNH progression or perhaps remission, allowing for predictions to be made about a patient's future PNH state or severity.

In this section the PNH categories: spontaneous remission, PNH new or increasing, PNH large clone stable and PNH recovering were compared with the normal datasets. In this section, mainly the significant differences are discussed.

PNH patients recovering from PNH had the shortest CDR3s

All categories (spontaneous remission, PNH new or increasing, PNH large clone stable and PNH recovering) were significantly shorter in CDR3 length than normals ($p<0.0001$). PNH recovering repertoires had significantly shorter CDR3s than PNH large clone stable ($p<0.001$) and PNH new or increasing clones ($p<0.05$). Interestingly, this highlights differences in TCRB repertoire lengths even between those recovering (seeing a decrease in PNH clone sizes) and those that have recovered. This could indicate immune processes involved in the recovery of PNH which then share more "normal" characteristics once fully covered. Perhaps patients recover due to as shift in immune response but go on to get autoimmune diseases linked with the shorter CDR3s. These ideas will be discussed further in the context of T-cell exhaustion and autoimmunity and PNH later in the thesis.

PNH patients with active disease had fewer hydrophobic CDR3s

Both PNH recovering and large clone stable had more hydrophobic residues than normals ($p<0.0001$). They both also had more hydrophobic residues (less negative GRAVY value) than the new or increasing PNH repertoires. ($p<0.01$). Interestingly, this finding is converse to theories of autoimmunity which suggest that hydrophobic residues play a role in the development of autoimmune disorders [**298**]. However, studies tend to suggest it is the position of these hydrophobic residues that have more of an impact than the number in the CDR3 [**88**]. Positional analysis of amino acids will be discussed in **Section 5.5.3**. It does however highlight that with time, as generally recovering PNH patients and those with large stable clones have had the disease 10+ years, the immune response perhaps becomes more autoimmune or at least changes occur that are different to normals and those with progressive PNH.

PNH CDR3s had higher acidic residue percentages than normals

No significant differences were observed for percentage of basic residues. All PNH groups had significantly higher acidic residue percentages compared with normals ($p<0.05$). Both PNH new or increasing and large clone stable had lower acidic residues than PNH recovering ($p<0.05$) which had the highest, followed by spontaneous remission repertoires. This could indicate that CDR3s with lower acidic residues are involved in the pathogenesis and progression of PNH. If these characteristics in CDR3s decrease in the repertoire, potentially they attribute to ongoing recovery. Acidic and basic amino acids form salt bridges or electrostatic interactions between one another. These are important forces when a peptide interacts with a CDR3 [**299**]. The differences in TCR properties here could relate to the change in peptide that the TCRs are interacting with at different stages of disease. Potentially, as the active PNH and stable PNH groups have CDR3s with fewer acidic residues, this may mean that the peptides they are binding to, perhaps involved in PNH pathogenesis, have more acidic residues, therefore the CDR3s need to have more basic residues for interaction. This would explain that as these TCRs contract, and the repertoire contains more CDR3s with acidic residues, interacting with peptides that have more basic residues that are not involved in PNH, recovery may occur.

PNH patients had more aliphatic index residues in CDR3s than normals

Few studies have looked at aliphatic index values in CDR3s in humans. One study in mice, found that enriched usage of aliphatic residues at position 5 of a CDR3 beta was indicative of Tregs [**300**]. All PNH patients had higher aliphatic index values than normals ($p<0.05$). PNH new or increasing had a lower aliphatic index than PNH recovering, whereas PNH large clone stable had higher ($p<0.05$). It would be a stretch to suggest that the variations in these groups was due to changes in Treg populations with the stages of the disease, but it would be interesting to split the TCRB repertoire further into subsets such as Tregs to see if this was the case.

PNH patients with decreasing PNH clones had more negative CDR3s than normals

Only PNH patients with decreasing PNH clones had a significantly more negative net charge than normals($p<0.05$). TCRB CDR3s in CD4+ T-cells have been found to be associated with more positively charged amino acids, whereas CD8+ are associated more often with negatively charged CDR3s [**301**]. This does not mean that recovering PNH patients have more CD8+ T-cells necessarily, but again, it would be interesting to split the T-cells into subsets such as CD4+ and CD8+ and perhaps supports the theory of autoimmunity in PNH recovery.

Aromatic residues lower in PNH large clone stable than normals

Aromatic residues were significantly lower in PNH large clone stable than normals ($p<0.01$) and when compared with PNH new and increasing ($p<0.05$). As stated before, more aromatic residues could be indicative of steric interactions with TCR and MHC: peptide. Differences compared with normals may indicate steric changes that hinder interactions or lead to pathological activation of T-cells for example with self-peptides [**296**].

No differences were observed for polarity of bulkiness between PNH and normals

No significant differences in polarity or bulkiness of CDR3s were observed between groups or normals.

CDR3 property differences between active PNH in AA and PNH patients

In order to assess whether there are similar characteristics in TCRB repertoires when PNH is active, the two categories with active PNH were compared, AA and PNH patients with increasing PNH clones. If features were found to be similar in both they might be linked to progression of PNH. It would also provide insight as to whether there are differences in TCRB repertoires dependent on whether the patient has active PNH on its own, or active PNH in the context of AA. GRAVY, length, polarity, bulkiness and percentage of aromatic residues had no significant differences. Therefore, these factors may be linked to active PNH pathogenesis/progression. The differences below, highlight changes occurring in active PNH TCRB repertoires dependent on an AA context or solely PNH.

Basic residue percentages were significantly higher in the PNH new or increasing patients than in the AA ($p<0.001$). Acidic residues were significantly higher in the AA increasing than the PNH ($p<0.001$). The reverse to the trend previously observed. This could be indicative of differences in the peptides that TCRBs are responding to and may relate more to differences in PNH clone progression in an AA context or solely PNH. In AA with PNH, the peptide may have more basic residues, whereas solely in PNH, more acidic residues. The aliphatic index was higher in PNH than AA ($p<0.05$) perhaps loosely related to PNH repertoires having more Tregs [**300**]. Net charge was significantly more negative in AA ($p<0.0001$), again, perhaps loosely related to AA repertoires having more CD8+ T-cells which would agree with current research into the immune responses in AA detailed in **Chapter 1**. These factors may link more to AA versus PNH pathogenesis and are interesting factors to assess as pathogenesis leads to progression. PNH clones may have different mechanisms for immune evasion dependent on whether it is in an AA or solely PNH context.

CDR3 TCRB repertoire changes in active versus recovering PNH patients

In order to assess whether changes occurred at the CDR3 level between PNH patients with active disease versus those who were recovering, CDR3 characteristics of TCRB repertoires were compared in this context. Potentially this could identify a biomarker for PNH that changes according to status of the disease. Having multiple time sets taken, from patients who had PNH to then when they have recovered would also enable a biomarker to be identified. These comparisons and differences in TCRB repertoires at the different stages will help aid further research into TCRs and PNH. PNH new or increasing clones were compared with PNH recovering, and the AA complex case who was recovering from PNH. To note, the complex case was one patient, statistical inferences may not be strong, but it was included as an interesting comparison. Acting as backgrounds, spontaneous remission from PNH were compared along with normals, to see whether active, recovering cases or both, were differing from the norm.

Shared characteristics between normals and spontaneous remission patients could highlight potential CDR3 markers in PNH activity

GRAVY and polarity, bulkiness and percentage of aromatic residues had no significant differences between normal repertoires or PNH spontaneous remission patients, indicating that when PNH is active. TCRBs with those characteristics could be implicated if differences are found in other groups.

Recovering patients/falling PNH clones had more hydrophobic residues in CDR3s

PNH patients with decreasing PNH clones along with the AA, patients with falling PNH clones, contained significantly more hydrophobic residues in CDR3s than spontaneous remission, normals and PNH new and increasing ($p<0.05$, Holm, Wilcoxon). Interestingly, PNH new or increasing repertoires were similar to normals and spontaneous remission repertoires. As hydrophobic residues have been linked with autoimmunity and dysfunctional immune responses [**88,298**], it is interesting that recovering PNH patients had more of these residues and will be discussed in **Chapter 7**. This could be linked to factors such as immune ageing, as recovering patients tend to be older and with age repertoires tend to become more clonal [**302**].

All PNH categories had shorter CDR3s than normals including spontaneous remission

When comparing CDR3 amino acid sequence length, all categories were shorter than normals even if they had recovered from PNH ($p<0.001$, Holm, Wilcoxon). AA complex had significantly shorter CDR3s than the other categories ($p<0.01$, Holm, Wilcoxon). PNH new or increasing repertoires had CDR3s significantly longer than PNH recovering repertoires ($p<0.05$, Holm, Wilcoxon). However, both categories were not significantly different from spontaneous remission categories, potentially attributed to a lower sample number (n=2) for spontaneous remission patients. Spontaneous remission having shorter CDR3s than normals, indicated that although recovered from PNH, the patient still had a dysfunctional immune system. This may take some time to recover. It is less likely to be an effect of immune ageing as the spontaneous remission patients were aged 39 and 40. Both had however, had PNH for over 19 years at time of sampling and therefore their immune system will have been dysfunctional for some time. Shorter CDR3s tend to be linked with autoimmunity and antigen skewed repertoires [**92,273**], both of the spontaneous remission patients were also female, perhaps more prone to autoimmunity. It would be interesting to see if spontaneous remission occurs in male patients who are less prone to autoimmunity and more likely to get infections.

Higher acidic residues compared to normals indicative to responses to basic peptides

All of the groups had significantly higher acidic residues than normals ($p<0.05$, Holm, Wilcoxon). As stated previously, this could indicate antigen skewed responses to peptides with more basic residues. AA increasing PNH clones had significantly lower basic residues than PNH new or increasing, spontaneous remission and normals ($p<0.001$, Holm, Wilcoxon) and AA increasing clones had significantly more acidic residues than PNH new or increasing PNH clones ($p<0.001$, Holm, Wilcoxon). This finding was interesting. It highlighted that perhaps some TCRBs in the patient were responding in the context of PNH to more basic peptides, but there were also TCRB linked to AA that may have been responding to more acidic peptides too. This trend was stated above when looking at AA versus PNH. PNH tended to have more acidic CDR3s than in AA. Normals had significantly lower basic residue percentages than all other categories ($p<0.05$, Holm, Wilcoxon). PNH recovering had significantly higher percentages than PNH new or increasing ($p<0.05$, Holm, Wilcoxon) but neither category differed from spontaneous remission patients.

PNH TCRB could be associated with higher aliphatic index

In terms of aliphatic index, spontaneous remission patients, PNH new or increasing patients and PNH recovering patients, had significantly higher aliphatic index results than normals and the AA increasing clone ($p<0.05$, Holm, Wilcoxon). AA complex saw no differences between any categories.

PNH recovering had a significantly higher index than PNH new or increasing ($p<0.05$, Holm, Wilcoxon). This could highlight a role for Tregs in PNH recovery [**300**].

More negative CDR3s associated with PNH and AA than in normals

AA with increasing PNH saw significantly more negative charge than all other categories, which were all similar in value ($p<0.01$, Holm, Wilcoxon). PNH recovering had significantly more negative CDR3s than normals ($p<0.05$, Holm, Wilcoxon), however, not when compared with PNH new and increasing. This could implicate CD8+ in AA and PNH repertoires.

Finally, there were no significant differences between categories for polarity or percentage of aromatic residues suggesting these characteristics are not involved in TCR responses in PNH. Differences in bulkiness were only observed for the AA complex which was bulkier than all other categories.

CDR3 property changes in response to PNH status of AA patients

In order to investigate whether TCRB CDR3 properties changes with AA pathogenesis, the AA patient categories were assessed in the context of normals. No significant difference in GRAVY or polarity values were found between the categories and normals. This analysis may help decipher factors that affect the TCRB attributed to AA specifically (AA no PNH) and those AA repertoires with stages of PNH. It could also help predict in future, whether these AA no PNH could go on to develop PNH.

Shorter CDR3s in AA groups than normals indicate antigen skewed repertoires

AA repertoires were significantly shorter in regard to CDR3 than normals ($p<0.001$) indicative of autoimmunity, serving as a good control in this study. AA small clone CDR3s were significantly longer than AA no clones, AA large clones or increasing clones.

Acidic CDR3 TCRBs higher in PNH patients with increasing PNH clone than normals

AA small clone and AA no PNH had significantly more basic residues than normals ($p<0.0001$) whereas AA PNH increasing had lower basic residues ($p<0.0001$). AA no PNH had significantly more basic residues than AA increasing ($p<0.001$) and AA large clone ($p<0.001$). AA small PNH clone also had higher basic residues than AA increasing and AA large clone ($p<0.001$). AA small clone had significantly fewer acidic residues than normals, AA large PNH clone and AA increasing clone ($p<0.01$). AA no PNH had fewer acidic residues than AA large clone and increasing ($p<0.001$). AA large clone and increasing had more acidic residues than normals ($p<0.0001$). These findings are in line with the PNH specific TCRB repertoires who all had higher acidic residues in their repertoires than normals, detailed above. This suggest that TCRBs in the repertoire as a result of or causing PNH may have more acidic residues, and if antigen specific in response, could be interacting with peptides with more basic properties. It also suggests that as the AA patients would be most likely on immunosuppressants to suppress the AA specific CD8+ T-cells, that although not present at clonal levels necessary, this analysis can pick up changes in the repertoires and those that could be related to PNH rather than AA in these patients.

Aliphatic index higher in AA patients with fewer or no PNH clones

AA small clone and no PNH clone had higher aliphatic index values than normals, AA increasing and AA large clone ($p<0.01$). Interestingly, this finding is the opposite to what was expected based on the PNH aliphatic index values calculated above. PNH seemed to be associated with higher aliphatic index values than normals, but in the context of AA this was not the case. Perhaps AA immune responses contribute to this change, this would also support the reverse trends being observed when comparing the smaller categories of AA and PNH previously.

Negative charges of CDR3s varied with PNH clone sizes

AA small clone and no PNH were significantly less negative in charge but still negative overall compared to normals, AA large clone and increasing clones ($p<0.0001$). Both AA large clone and increasing were more negative than normals ($p<0.01$). As negative charged CDR3s are linked with CD8+ T-cells and more positive residues with CD4+ it could indicate that as the PNH clones increase, more CD8+ populations appear in the TCRB repertoire. The AA with no PNH can still have a more dominant CD8+ response but this is increased with PNH clones potentially.

Few categories saw differences in bulkiness and aromatic residues

Only normals and AA large PNH clone saw a difference in bulkiness with AA large clone being less bulky. Only AA small clone and normals saw a difference in percentage of aromatic residues with small clones having a lower percentage($p<0.01$).

Overall, analysing CDR3 characteristics in the context of diagnosis and clinical status, was a good indicator of changes in TCRB responses in PNH and the findings will allow for further work into how different amino acid properties could affect the TCRB repertoire.

5.5. TCRB clonal responses in AA and PNH patient repertoires

In order to investigate the hypothesis that PNH patients had TCRB clonal expansions related to PNH pathogenesis, clonality studies were performed on the AA and PNH cohorts. One sample was excluded from the primary analysis set of 76 patients to reduce bias as it contained one single TCRB clonal expansion representing over 40% of the entire repertoire. This sample was looked at on an individual level (**Section 6.3.2.**) and was included in the following clonality analysis.

Out of the 77 AA and PNH patients in the clonality study, 23 repertoires showed monoclonal immune responses and 5 showed polyclonal responses (2 or more clonal expansions, TCRB clonotype accounting for 2.42% and above of the TCRB repertoire) (**Table 22**.). The mean age of patients with monoclonal TCRB responses was 49.7 years old with a median of 49.5. This was slightly above the average age of AA and PNH patients overall of 46 and median of 42, so potentially monoclonality was skewed by age but not significantly. Interestingly, for both the monoclonal and polyclonal responses, the female: male ratios showed a higher skew towards males, with 1:1.75 and 1:1.5 respectively, both higher than the 1:1.25 calculated across the 76 patients. The non-clonal subsets had a ratio of almost 1:1 with 25 males to 24 females. Studies have discovered that immune dysfunction has a strong sex bias [**98**], with men being more susceptible to infection and some cancers. Women tend to develop autoimmune diseases due to strong immune and have also been found to have more CD4+ T-cells than men [**303**]. However, in this study TCRB clonality did not appear to have a significant sex bias.

Previous sections have analysed repertoires irrespective of clonal responses so as not to bias analysis towards searching solely for clonality in case PNH does not involve clonal TCR expansions as hypothesised. In this section, analysis was now carried out in the context of TCRB clonality in order to explore the hypothesis, whether clonality is linked with PNH or not.

For example, TCRBV and J gene usage comparison was performed between repertoires that were clonal and non-clonal to assess whether clonal repertoires had higher usage of certain genes than non-clonal.

Table 22. TCRB clonal responses in each patient category from the primary data analysis of 76 PNH and AA patients. The TCRB response that categorises the most patients for each status category is depicted in bold and red text.

Category	Non-clonal	Monoclonal	Polyclonal	Total number in primary data-set
Aplastic; 60% PNH clone	**1**	0	0	1
Aplastic; increasing PNH clone	**4**	1	1	6
Aplastic; small PNH clone	**15**	4	0	19
Aplastic; no PNH clone	2	**4**	0	6
Haemolytic; PNH clone decreasing	**6**	1	0	7
Haemolytic; large PNH clone, stable	**8**	6	3	17
Haemolytic; new/increasing PNH clone	**12**	4	1	17
Other (Aplastic, PNH clone decreasing, complex case)	**1**	0	0	1
Other (Haemolytic, large PNH clone, complex case)	**1**	0	0	1
Other (Haemolytic, thrombotic, T-cell LGL)	0	**1**	0	1

To investigate whether PNH or AA patients had a skew towards mono or polyclonal responses or whether the patient's clinical status affected the TCRB response, clonality was broken down into categories (**Table 22.)** In a similar trend to the normals analysed in **Chapter 4**, the majority of categories favoured non-clonal responses, as 49 out of the 77 patients had non-clonal TCRB responses. Poly-clonal responses were the rarer response with only 5 patients exhibiting them, the majority being PNH patients with large stable clones. Only one AA patient had a poly-clonal response, and this was a patient with an increasing PNH clone. Despite AA patients with small PNH clones having the most patient numbers, nearly three quarters showed no clonal TCRB responses. Haemolytic large clone stable had seemingly similar numbers between non-clonal and monoclonal. Active PNH patients and those with AA and no PNH clone tended to have more clonal repertoires. The clonality exhibited could be linked to separate infections, for example, CMV. As in previous sections, normals and PNH/AA patients tended to observe similar levels of clonal TCRB abundances in the repertoire, whereas normals had fewer TCRB clones in the compart consisting of TCRBs clones occurring 31-100 times (**Section 5.4.4.**). Another possibility could be that, the TCRB repertoire in AA patients with no PNH is being disrupted and a year or so later, PNH clones may appear. It would be interesting to re-sample the AA no PNH clone patients over multiple time points and compare this with any changes in diagnosis.

To understand whether PNH patients with clonal expansions, had T-cells interacting with similar type peptides, the characteristics of these clonal TCRBs were investigated at a deeper level. TCRBs for all PNH patients were categorised into non-clonal(<2.42%), low responders (2.42-4.85% clone), moderate responders (4.86-20%) and high responders (20%+). The percentages refer to the space that the clone takes up in the entire TCRB repertoire. Response level is a term coined to describe the size of a TCRB clonal expansion in a TCRB repertoire. Low responders are smaller clones. The high responders are large clonal TCRB expansions, often attributed to chronic infection and are generally rare in normal repertoires (**Chapter 4.**).

In order to investigate TCRB clonal expansions in patients with AA, the above analysis was repeated on just the AA related cohort consisting of 33 patients. For AA patients with PNH, who had monoclonal TCRB responses, four had low response clonal expansions and 4 were moderate. In patients with AA and no PNH clones, 3 were low response and one moderate. When comparing clonality in terms of response level, there was no significant difference between the response levels when split by clinical status category or diagnosis.

5.5.1. No significant differences in TCRBV gene usage in clonal versus non-clonal repertoires

When grouping all the AA and PNH patients according to clonal response, non-clonal, monoclonal or polyclonal, or when splitting them into PNH or AA and then the clonal response level, no statistically significant difference was seen between any of the groups or response levels (data available on request). It was important to investigate this to understand as to whether some responses were antigen specific in case PNH is caused by a super-antigen for example. Antigen specific clonal populations would be expected to be present at some point during the PNH pathogenesis and progression.

5.5.2. Potentially significant differences in TCRBJ gene usage were observed between clonal and non-clonal patient TCRB repertoires

When comparing non-clonal versus monoclonal and polyclonal TCRB repertoire responses, regardless of having AA or PNH, although statistical significance was not achieved, J1-4, J2-3 and J2-6 showed differences. Monoclonal repertoires tended to have higher J1-4 usage and J2-6 usage but lower J2-3 usage (**Figure 55. left.**). With a higher n number, the statistical significance may have been achieved and therefore should be taken into account although not a significant finding.

When comparing clonal versus non-clonal repertoires irrespective of AA or PNH, again J1-4 and J2-6 were more used in clonal repertoires and J2-3 was used less in clonal TCRB repertoires. Again, the findings were not statistically significant (**Figure 55. right**). When comparing AA categories, those with an increasing PNH clone or a large PNH clone had lower usage in J1-5 and higher usage in J2-5.

When splitting males with females, J2-2 and J2-7 was slightly higher in males. Again, none of these findings achieved statistical significance at p<0.05 but with a higher n number, the statistical significance may have been achieved and therefore should not be discarded. In conclusion, again more variation was likely when analysing TCRBJ usage than TCRBV gene usage, particularly in the context of clonal repertoires.

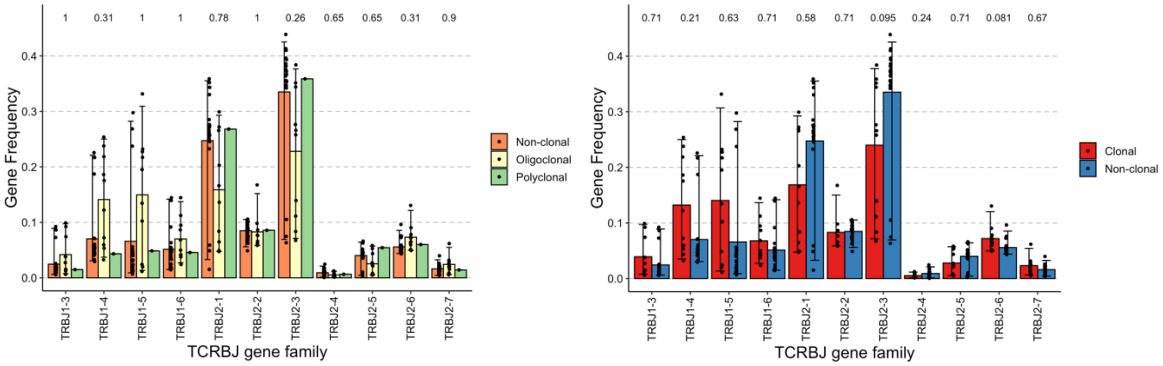


Figure 55. TCRBJ gene usage in 43 PNH and 33 AA patients according to clonality. The analysis compared clonal versus non-clonal TCRB responses in the primary data cohort of 43 PNH patients, 6 AA with no PNH and 28 AA with PNH. TCRBJ gene usage was compared between non-clonal n=49, monoclonal n=22 and polyclonal n =5 TCRB responses. Four out of 5 polyclonal repertoires were PNH, and 12 out of 22 were PNH for monoclonal responses. P adj values shown were calculated using the Holm method with pairwise Wilcoxon mean comparisons. P<0.05 was considered significant.

5.5.3. No obvious specific CDR3 patterns were associated with clonal TCRB

CDR3 characteristics were assessed at both the positional pattern level and amino acid property level in regard to immune response. However, due to the low n number of the clonal populations, and high n number for non-clonal populations of 50,000 +, any statistical tests would be biased. **Figure 56**. details CDR3 properties that were investigated in PNH patients, showing that the majority had similar profiles. Some of the patients with abnormal results in the context of PNH, were analysed on an individual level in **Chapter 6** in more detail.

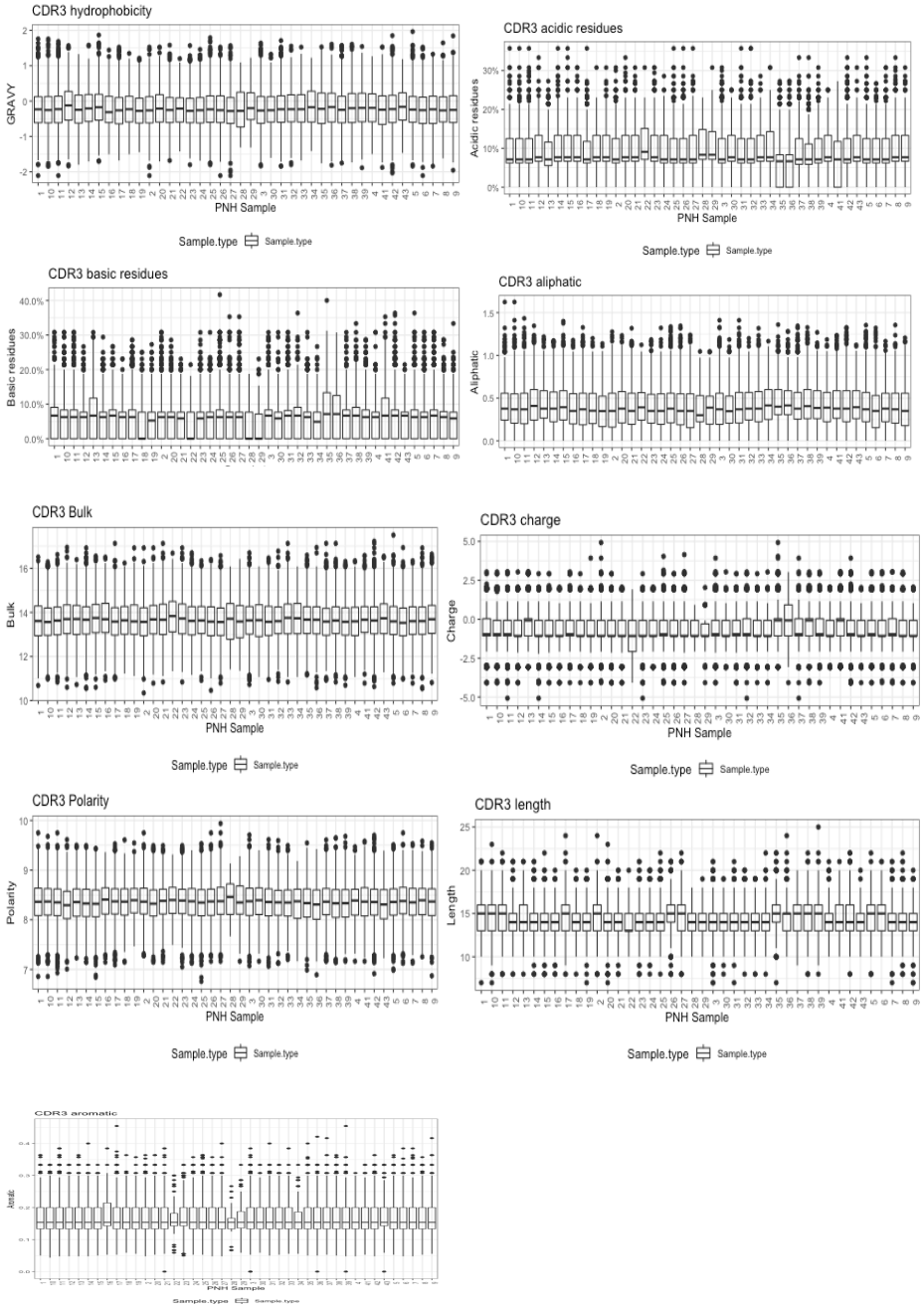
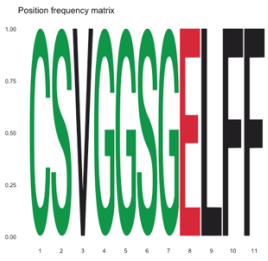
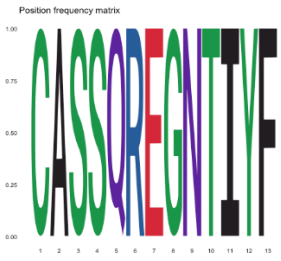


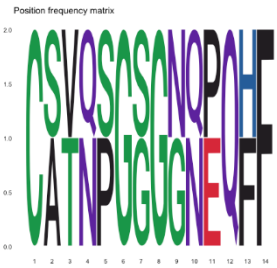
Figure 56. Forty three TCRB repertoires from PNH patients measuring the overall CDR3 amino acid characteristics for all TCRB clones in an individual's repertoire.

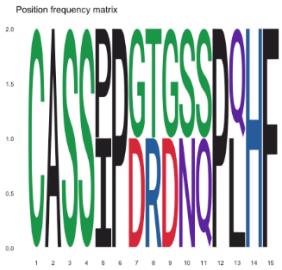
For each box plot the central line indicates median. GRAVY, bulkiness and polarity are calculated as averages of the scores across informative positions. GRAVY measures the grand average of hydrophobicity across a CDR3. Bulkiness calculates the average bulkiness and polarity is a calculation of the average polarity of the amino acids . Aliphatic values are calculated using the aliphatic index and charge is overall net charge. Basic, acidic and aromatic residues are calculated as the fraction of informative positions that have these characteristics. For basic residues these are the amino acids Arg, His and Lys. Acidic residues are Asp or Glu. Aromatic are His, Phe, Trp or Tyr. [**245**]

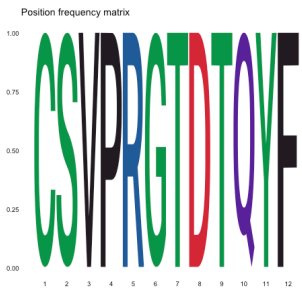
To investigate as to whether there were any positional amino acid properties that could be contributing to the TCRB clonality, the TCRB clonal expansion's CDR3 was analysed using a positional weight matrix. This assessed whether particular sections of the CDR3 had certain amino acid characteristics. The analysis assessed clonal TCRB in AA or PNH categories and then non-clonal expansions in some normal repertoires on the basis that they were not linked to a chronic infection at least in that individual. A variety of residues were used in each position of the CDR3 with no particular patterns being linked to a clonal TCRB in a specific clinical status category. CDR3 is defined as the region between a cysteine and ending with a phenylalanine-glycine. Therefore, it would be expected that all CDR3s would start with a C and end with an F residue [**304**]. One clonal CDR3 from an AA small PNH clone patient and one from a PNH patient with a large stable clone showed hydrophobic residues at positions 6 and 7 in the CDR3, which in mice have been linked to self-reactivity [**88**]. As only one is present in two patients, it may be an outlier.

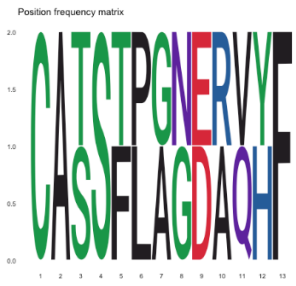


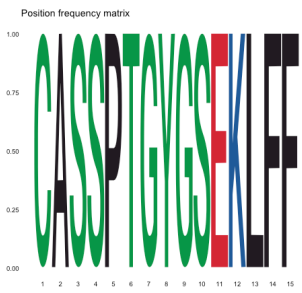


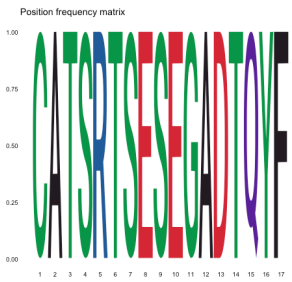


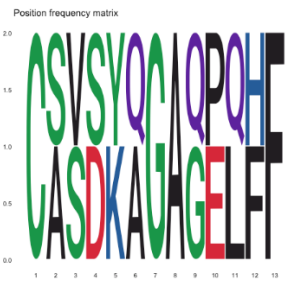


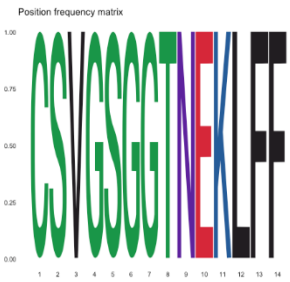












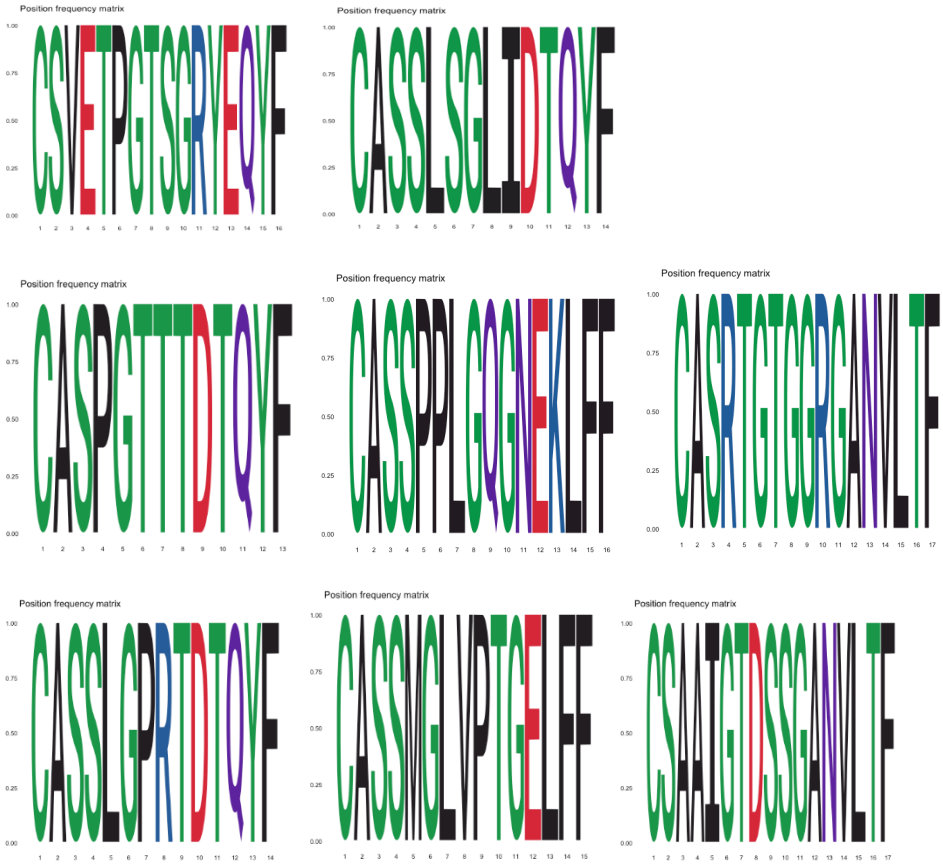


Figure 57. Position frequency matrix analysis of clonal CDR3s in AA and PNH patients. From top to bottom, by row, the CDR3s are examples of clonal CDR3s from AA patients progressing PNH with a large EBV TCRB clone, PNH new or increasing, PNH large clone stable, AA no clone, AA increasing clone, AA small clone and finally three examples from non-clonal normal repertoires, of non-clonal CDR3 TCRBs. CDR3s were taken from multiple patients as examples of structures. Red is acidic residues, blue basic, black hydrophobic, purple neutral and green polar.

5.5.4. Thirty-four unique TCRB clonal expansions were identified in AA and PNH repertoires

TCRBs that were annotated as accounting for more than 2.42% of an individual's repertoire were defined as clonal based on findings in **Chapter 4**. Thirty-four unique clonal TCRBs were found in AA and PNH patient repertoires. Twenty-six of these were novel and unique to AA or PNH patients at the time of analysis, as they were not identified in the 29,140 TCRBs (irrespective of clonality) identified in the normal cohort or when searching the TCRB databases or literature searching (**Section 2.9.2.3.**) (**Table 24.**). Only two of the TCRB clonal expansions were found in literature searching or searching TCR databases. 'CASKGGNQPQHF' was found in three other patients in a study so is not unique to AA. 'CSVGSGGTNEKLFF' was found with multiple hits linking to EBV and Influenza A [**254,255**].

This CDR3 was also one of four TCRB clonal expansions found to be persistent (clonal at multiple time-points for one patient) and appeared in five PNH and AA patients spanning a range of the patient categories along with 5 normals. Persistent clonotypes are important for understanding chronic immune responses. As PNH is a chronic disease, if T-cells have a significant involvement, TCRB clones may expect to be persistent and correlated with PNH clone size. These dynamics are best investigated using longitudinal TCRB studies (**Section 6.3.**). Two well represented clonotypes 'CSVETPGTSGRYEQYF' and 'CSVDKAGAGELFF', whose origin is currently unidentified by other research, were found in 23 and 24 of the 30 normals respectively. They were, however, only found in one AA patient with an increasing PNH clone and one AA no PNH patient respectively (**Table 24.**).

In the AA increasing PNH clones, 3 out of 7 patients had clonal repertoires. One, patient 00563 (**Section 6.3.2.**) had a hyperexpanded clone at 61%, V29-1/J1-4. All clonal responses were seen in males aged 65 and above in this category. The other four non-clonal repertoires were, one male aged over 65, and the rest were female and aged between 22 and 36, considerably younger. This again, could be linked back to the fact of women being more prone to autoimmunity and men infections. With ageing, the repertoire tends to represent more infections such as EBV which could be the cause of clonality in the older male TCRB repertoire [**305**].

In AA no clone, the group was made up of mostly males of varying ages, and four being monoclonal. Three low responsive TCRB clones and one moderate at 11%. TCRBV15, 18, 10-2 and 6-5 were highlighted as interesting as they routinely emerged in the top 10 clonotypes in this cohort and were families not usually observed at high frequencies in normals. The moderate TCRB clone was V6-5/J1-5 followed by a non-clonal V10-2/J1-4.

In the AA small clone (which included those who were recovering from PNH or clone was falling) there were 4 monoclonal repertoires. Three had moderate responders and one low. Of the two patients who had falling PNH clones, one was non-clonal, and one was a low responder at 3.5% with a novel TCRB 'CASPGTTTDTQYF', patient 004V3, detailed in **Section 6.3.1.** Interestingly, this clone was persistent but decreased over the six years between samples with the decrease in PNH clones size.

Five out of the 17 PNH with new or increasing clones were clonal with one being polyclonal in response. Three of the patients were originally diagnosed with AA but now only PNH, and two of these were monoclonal with low or moderate responses.

Nine out of 17 PNH patients with large clones but stable had clonal repertories. Three were polyclonal, however, one was polyclonal due to low clonotype number. Five had low response clones and 3 were moderate. Interestingly, the majority of the clonal repertoire patients had had PNH for longer periods of time (10+ years) compared with some of the non-clonal (2+ years).

Table 24. TCRB clonal expansions in PNH and AA patients.

CDR3 amino acid sequences from 77 AA and PNH patients present at clonal TCRB frequencies (occupied 2.42% or above of the TCRB repertoire). Each TCRB CDR3 sequences was annotated with the number of times it appeared in the normals cohort along with the category and clonal percentage it was present at in the AA and PNH cohorts. The two CDR3s in italics indicate CDR3s identified in other studies. The four CDR3s in bold indicate persistent clones, those found at clonal frequencies in multiple time points from the same patient. 26 were novel to AA or PNH at time of the analysis.

TCRBV	TCRBJ	CDR3 amino acid sequence	Number of times CDR3 appears in normals	Category	% of TCRB reperotire occupied
V5-3	J1-4	CARSHEGGLDEKLFF	0	HLCS	6.2
V11-1	J2-2	CASGDRVTGELFF	0	HLCS	8
V19	J1-5	CASHPFRGNQPQHF	0	HLCS	3.2
V19	*J1-5*	*CASKGGNQPQHF*	*0*	*aasc*	*6.5*
V5-5	J2-3	CASPGTTTDTQYF	0	aasc	3.5
V19	J2-6	CASRTGTGGRGANVLTF	0	aasc	9
V19	**J1-4**	**CASSARTGHEKLFF**	**0**	**HLCS**	**3.2**
V12-3	J1-5	CASSFPGNDRVHF	0	HLCS	2.8
V19	J1-5	CASSIGVPSGNQPQHF	0	HLCS	2.5
V19	J1-5	CASSIPDRGSQPQHF	0	PNH new inc	2.5
V27	J2-3	CASSLSGLIDTQYF	0	AA new inc	3
V6-4	J1-6	CASSPPGTDNSPLHF	0	PNH new inc	3.7
V18	J1-4	CASSPPLGQGNEKLFF	0	aasc	6.8
V18	J1-4	CASSPTGYGSEKLFF	0	HLCS	15
V11-3	J2-1	CASSPWENEQFF	0	AA new inc	2.6
V5-4	J1-3	CASSQLGDGNTIYF	0	HLCS	8.7
V6-5	J1-3	CASSQREGNTIYF	0	PNH new inc	4.4
V6-5	J1-5	CASSYQGAQPQHF	0	AA no PNH	11
V19	J1-4	CASTRTGGGPNEKLFF	0	HLCS	3.19
V15	J2-3	CATSRTSESEGADTQYF	0	AA no PNH	2.6
V15	**J1-4**	**CATSSQAGEKLFF**	**0**	**PNH recovering**	**4.2**
V29-1	J2-6	CSAAIGTDSSGANVLTF	0	LGL	2.5
V29-1	**J2-6**	**CSAHADAGANVLTF**	**0**	**HLCS**	**9.8**
V29-1	J2-3	CSVDRADTQYF	0	HLCS	4.7
V29-1	J1-5	CSVNSGSGNQPQHF	0	PNH new inc	5.7
V29-1	J2-3	CSVPRGTDTQYF	0	HLCSx2	6.4, 2.7
V15	J2-3	CATSTLAGEAQYF	1	HCLS	4.4
V29-1	J1-5	CSVNWGSGNQPQHF	1	HLCS	12
V19	J2-1	CATQPGGGGNEQFF	2	PNH new inc	3.9
V29-1	J2-2	CSVGGSGELFF	2	PNH new inc	2.6
V29-1	***J1-4***	***CSVGSGGTNEKLFF***	***5***	***AA no PNH, HLCSx2, aasc, AA new inc***	***4,4.1,11,14, 46***
V29-1	J2-7	CSVETPGTSGRYEQYF	23	AA new inc	4.3
V29-1	J2-2	CSVDKAGAGELFF	24	AA no PNH	2.7
V15	J2-1	CATSRESGGTDEQFF	2	HLCS	3.9

5.5.5. Polyclonal TCRB responses in PNH and AA patients

Table 25. TCRB repertoire clonality study highlighted 5 patient samples that had polyclonal T-cell responses. The clonal TCRB responses for each patient are in **bold**, these are TCRB clonotypes comprising 2.42% or more of the overall TCRB repertoire. Some non-clonal TCRB were included to indicate the sharp decline in clonal to non-clonal percentages.

Sample	% of TCRB repertoire	CDR3 sequence (aa)	TCRBV	TCRBJ	Number of clonotypes	Diagnosis notes
502	**3.03**	**CASSLSGLIDTQYF**	**TRBV27**	**TRBJ2-3**	1220	Aplastic progressive disease 3 years, increasing/variable clones now
	2.60	**CASSPWENEQFF**	**TRBV11-3**	**TRBJ2-1**		
	1.77	CATADWNNEQFF	TRBV6-2	TRBJ2-1		
004VR	**3.17**	**CASSARTGHEKLFF**	**TRBV19**	**TRBJ1-4**	1002	Haemolytic, large stable clone
	2.50	**CASSIGVPSGNQPQHF**	**TRBV19**	**TRBJ1-5**		Interesting case. Grans all type III, mono all type II, red cells almost indistinguisable from normal but likely 100% clone. Thrombotic
	2.35	**CATSRESNDHGQPQHF**	**TRBV15**	**TRBJ1-5**		
	1.49	CASSSNSRVGKLFF	TRBV27	TRBJ1-4		
0053Z	**6.43**	**CSVPRGTDTQYF**	**TRBV29-1**	**TRBJ2-3**	43	Haemolytic, large stable clone Thrombotic with large (>99%) clones. Stable disease for 7 years
	4.70	**CSVDRADTQYF**	**TRBV29-1**	**TRBJ2-3**		
	3.94	**CATSRESGGTDEQFF**	**TRBV15**	**TRBJ2-1**		
005DE	**14.74**	**CASSPTGYGSEKLFF**	**TRBV18**	**TRBJ1-4**	1198	Haemolytic, large stable clone
	8.09	**CASGDRVTGELFF**	**TRBV11-1**	**TRBJ2-2**		
	6.57	**CASTRTGGGPNEKLFF**	**TRBV19**	**TRBJ1-4**		
	3.19	**CASHPFRGNQPQHF**	**TRBV19**	**TRBJ1-5**		
	1.85	CASSKWGQGDQPQHF	TRBV19	TRBJ1-5		
005DY	**4.41**	**CASSQREGNTIYF**	**TRBV6-5**	**TRBJ1-3**	848	Haemolytic new/increasing
	3.72	**CASSPPGTDNSPLHF**	**TRBV6-4**	**TRBJ1-6**		Sample taken a couple of months after diagnosis
	0.82	CASIQRQAASSNSPLHF	TRBV6-5	TRBJ1-6		

Five polyclonal TCRB repertoires were identified in this study, 4 PNH and one AA (**Table 25.**) Each patient repertoire was analysed at an individual level.

Patient 0053Z, who was in the category of PNH, with a large stable clone, had 13 out of 43 TCRB clonotypes in the repertoire present at clonal levels. However, on analysis of the T-cell count data, the sample only had 15% T-cells before being processed. The repertoire was unusual in that the majority of the CDR3s had no basic residues. The patient was thrombotic which may have attributed to issues with sampling separation. Therefore, most likely the clonality observed was due to low T-cell numbers which could have a biological origin, or, been artificially produced by the blood taking process.

Interestingly, for the remainder of the polyclonal patient responses, the inclusion of the first non-clonal TCRB (**Table 25., not in bold type**) highlights how rapidly the percentage of a TCRB population dropped after the clonal responses.

For example, patient 005DY had two clonal TCRBs at 4.41% and 3.72% of the TCRB repertoire so were considered moderate levels of response. The next clonotype is non-clonal at a much lower percentage of 0.82%. This was almost five times lower than the second top clonotype in the repertoire. This suggested that the polyclonal responses were biological rather than technical. Patient 005DY was diagnosed only a couple of months prior to sampling, suggesting that they were in a known active stage of PNH, progressive state which may indicate why there was two clonal expansions.

The majority of the five patients had TCRB clonotype numbers near to the average of 1217 in patients in the primary analysis set. 005DY had a slightly lower number as it was a buffy coat sample, as was 005DE, but not significantly low enough for TCRB responses to be attributed to this. Three out of the five patients were PNH with large stable clones, one was aplastic with an increasing clone and one was haemolytic PNH with a new and increasing clone. The ages ranged from 28 to 83 so the response was not linked with age, and the sex was 3:2, male to female, again not showing an obvious sex link.

When carrying out TCRBV and J gene usage on just the polyclonal subsets, interestingly, the usage was different to that of the primary analysis when looking at individual polyclonal responses (**Figure 58.**). When observing patterns in TCRBV and J gene usage in the repertoires of these polyclonal responses, all patients except 005DE shared the most common TCRBV family as 29-1 in agreement with non-clonal responses and the normal samples in **Chapter 4**.

005DE shared V29-1 as its third most common TCRBV, close in number to TCRBV18 and V19 which were top. Generally, a similar pattern usage was observed across the samples, 0053Z had an abnormally high TCRBV29-1 usage possibly attributed to the low T-cell sample counts. The usage differed slightly to the usage observed in the PNH or AA subsets. V18, 19, and 27 were amongst the higher usages but not across all five samples.

005DE also saw V11-1 at higher usage levels than the other patient samples. V11-1 to date has not been annotated in a particular disease. Overall, this highlighted that there may be subtler changes in TCRB clones in individual repertoires not identified when averaging results of all patients' repertoires across a group.

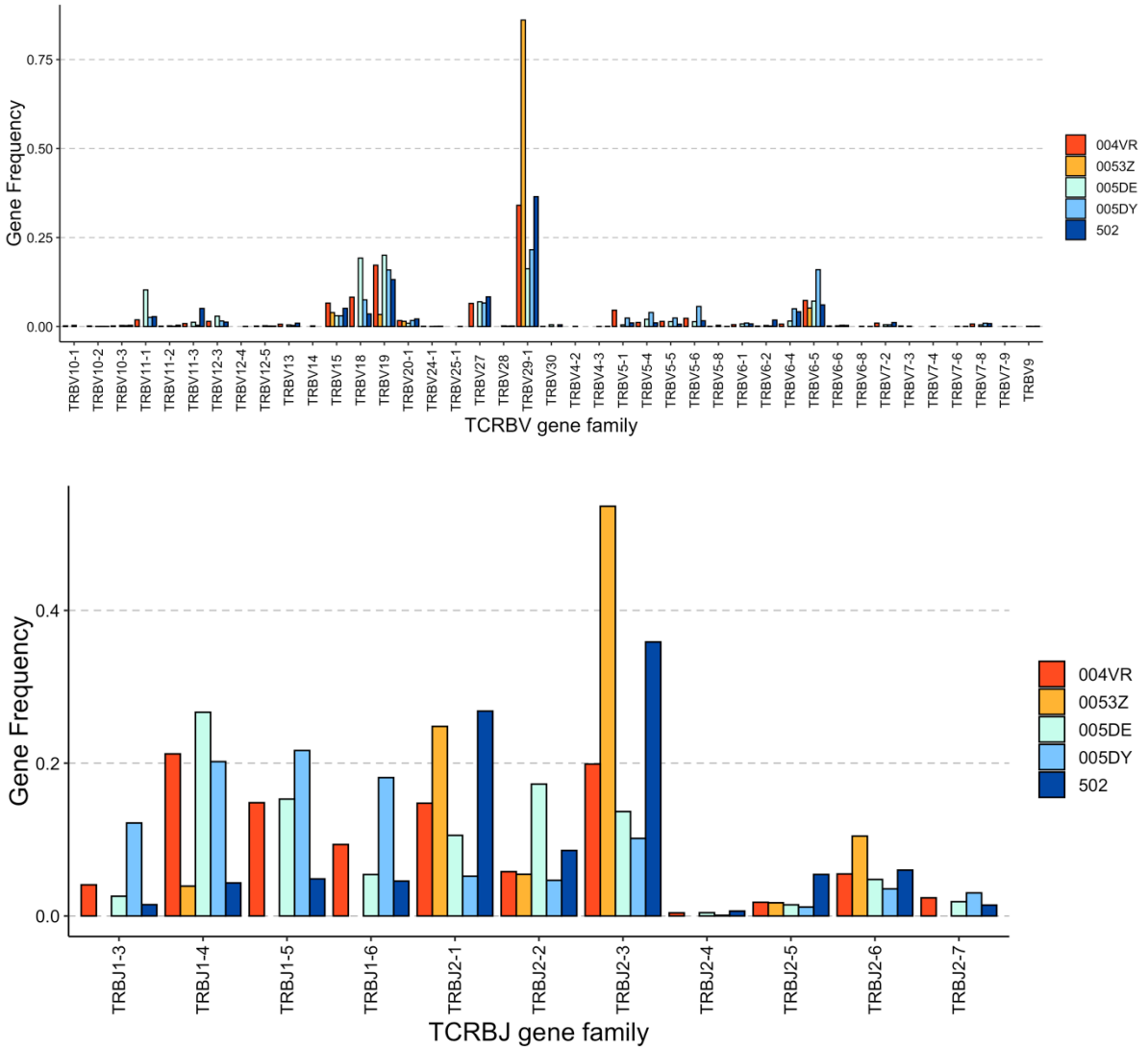


Figure 58. TCRBV and J gene usage across the 5 patient samples that produced polyclonal TCRB responses. Polyclonal TCRB responses were defined as TCRB repertoires having 2 or more TCRB clonal expansions. A TCRB clonal expansion was a TCRB clonotype occupying 2.42% or more of the individual's TCRB repertoire. 004VR, 0053Z and 005DE were PNH patients with a large stable clone, 502 was an AA patient with an increasing clone and 005DY was PNH with a new or increasing clone.

5.6. GPI- versus GPI+ T-cells

Having analysed TCRB repertoires in the context of PNH, it was important to understand whether clonal TCRB might be involved in the pathogenesis of PNH (GPI+ T-cells) or the progression (GPI- T-cells). When PNH first emerges the majority of the T-cells will be GPI+ and will be responsible for the initial immune responses [**306**]. Potentially, the emergence of GPI- T-cells, allows for the mutated PNH HSCs to expand and progress the disease further.

Flow cytometry TCRBV data (kindly provided by Dr S. Richards) was analysed for 8 patients, chosen for their high GPI- T cell populations (over 25% considered large), large PNH clones or the length of time they had had PNH. The majority of patients had normal CD4+ to CD8+ ratios and the T-cells were not generally memory or viral specific e.g. for CMV when analysing cell specific markers for these factors. CD57+ for example, was used to differentiate terminally senescent T-cells [**307**]. One patient had mainly GPI+T-cells (GPI- 3.13%), but an inverse ratio of CD4 to CD8 with almost double the number of CD8:CD4 T –cells than expected. The majority of the V beta frequencies were lower than in normals in the GPI- fraction and many in the GPI+ fraction. When higher than normal expression of V beta gene families was seen in patients, the majority of patients saw higher than normal expression of V beta gene families in the GPI- fraction than in GPI+ fraction. This was not true in two patients who coincidentally had inverse CD4:CD8 ratios compared to the other patients.

The majority of the T-cells were naïve, to be expected especially in the GPI- compartments as these will have been most recently produced in patients as they developed PNH. Two patients saw slightly higher numbers of memory T-cells than their counterparts at 26.54% and 33.42% compared to the rest of the patient values being under 12%. On both occasions these were GPI+ T cells. V27 was exclusively overrepresented in only GPI – T-cells. V 4-1,4-2,4-3, was highly expressed exclusively in one patient's GPI- population, accounting for 21% compared to the mean value in normals of 2.56%. V 12-3 was exclusively down regulated in one patient in the GPI- compartment. V18 was exclusively highly expressed in one GPI- compartment. V19 highly expressed in one patients GPI+ at 9.58% compared to the norm at 5.15% (data available on request). This again highlights the variety of TCRB responses in PNH patients and identifies V27 as of interest from this data.

5.7. Chapter summary

The work in this chapter was successful in achieving the main aim of assessing as to whether there was a specific TCRB clone involved in the pathogenesis or progression of PNH. The findings suggested that there was not one single TCRB clone involved in PNH pathogenesis shared between all patient samples. However, a number of interesting observations are summarised below, some of which are discussed in more detail in the context of the research field as a whole in **Chapter 7**. A range of analysis was performed to ask different biological questions from the TCRB repertoire data in regard to PNH and T-cells. The 77 AA and PNH patient TCRB repertoires were successfully analysed and compared with normals to identify some potentially PNH specific TCRB characteristics.

Relating back to the original project hypothesis, a single TCRB clone or series of TCRB clones were hypothesised to be present in the repertoire of all PNH patients. The findings in this chapter do not support this. However, they do provide evidence for disrupted TCRB repertoires in PNH patients that differ from normal subsets.

5.7.1. TCRB clonal responses in PNH and AA

5.7.1.1. Twenty-six novel TCRBs identified in PNH and AA patients

The analysis investigating the hypothesis that identical TCRB clones are shared amongst PNH patients and prevalent in the repertoire, led to the conclusion that this was not the case. Mainly attributing to natural variation in TCRB repertoires, due to infections, it made it very hard to decipher PNH specific responses.

However, 34 unique TCRBs were identified in the 77 PNH and AA repertoires, 26 which were novel (at time of the analysis) to PNH or AA as they were not present in the public TCRB repertoire generated in **Chapter 4**. The importance of having the public repertoire was highlighted by two well represented TCRB CDR3s ‘CSVETPGTSGRYEQYF’ and ‘CSVDKAGAGELFF’ in the patient cohort, whose origin is currently unidentified by other research. They were, however, found in 23 and 24 of the 30 normals respectively. This meant that they were public TCRBs and unlikely to be linked exclusively to PNH. Five samples showed polyclonal TCRB responses (more than one TCRB clonal population) and the majority were PNH with large stable clones which could have clinical relevance.

When breaking the clonality of repertoires into categories, age or sex, some distinctions could be made. AA with no PNH had more clonal than non-clonal TCRB repertoires within a group, which was as expected, as AA has been linked with clonal CD8+ T-cell expansions [**308**]. Although, clonality in general was not exclusive to one category group, it varied across the groups. Clonality was slightly skewed towards males, as males are more prone to infections than females, the clonality may not be related to PNH but rather a different infection or virus. One group showed a distinct separation in the metadata of clonal repertoires. This was AA patients with an increasing PNH clone. All of the clonal repertoires were either male and over 65 or young females. This would need to be investigated further.

No specific CDR3 pattern was shared between the PNH and/or AA cohorts for clonal TCRBs not indicating any particular pathological amino acid sequences. However, the method could be improved by applying machine learning techniques to assess trends and differences rather than manually, to draw accurate conclusions from the data.

5.7.1.2 Are PNH specific TCRBs in memory states in some repertoires?

Interestingly, when analysing the number of unique TCRBs that occupied a specific percentage of repertoire space, TCRBs that appeared 31-100 times in a repertoire (non-clonal) were higher in the disease subsets than normals. This suggested that perhaps it is the non-clonal populations in some PNH subsets, dependent on stage of disease, that are implicated in PNH. This would link roles for T-cell subsets such as memory T-cells that are present at non-clonal percentages in the repertoire. Perhaps in active stages of disease, the TCRB will be clonal. As the PNH status remains stable, the TCRBs may contract and differentiate into memory subsets. In order to investigate this further, PNH newly diagnosed patient repertoires were analysed in **Chapter 6**. If TCRB clonal populations are present in these patients it may support this theory. Equally, if spontaneous remission patients, also analysed in **Chapter 6**., have non-clonal TCRBs, this would support the theory. For further work, T-cells could be sorted into subsets prior to sequencing based on surface markers. For example, central memory T-cell subsets express CD45RO [**309**].

5.7.1.3. Diversity was not decreased in PNH patients compared to normals

Diversity of TCRBs in a repertoire can be an indicator of stability in immune response, the more diverse the repertoire, the greater the ability to recognise more infections. Equally, if there are more clonal populations responding to infection, diversity will be decreased and therefore, decreased diversity could be pathological [**310**]. It was therefore expected that PNH patients would have significantly lower diversity than normals. The only diversity measure to differ between PNH/AA patients and normals was "true diversity" which was indicative of stability in normals compared to AA and PNH repertoires. It might have been expected for more diversity measures to be different between the groups. However, it could be due to diversity measures using the definition of a TCRB clone as the same V/J gene, CDR3 amino acid usage that lead to the insignificance (**Section 7.2.1.3**.). The diversity may be related less to antigen infection and more to do with selection processes in the thymus during T-cell development (**Chapter 1**). This is why TCRBV and J gene usage was also assessed irrespective of CDR3 but again no significant differences were observed at group levels.

5.7.1.4. TCRBV/J gene usage in PNH and AA patients

Interestingly, it might have been expected that TCRBV and J gene usage in the repertoires would be skewed towards particular families in PNH and AA in response to clonal expansions. However, this was not the case. When observing the trends in AA and PNH cohorts, the general trend was similar to that of the normals with V29-1 being highly used followed by V-19. J2-3 and J2-1 were highly used J genes also seen in normals. When splitting the V and J gene usage by clonality response, (non-clonal, mono or polyclonal) no differences were observed in V usage but in monoclonal responses J2-3 usage fell and J1-4 and J2-6 increased. This may suggest that clonal populations usage differs from normals, however it could not be inferred if this was specific to AA or PNH or in response to an infection for example. This highlighted that perhaps V and J gene usages were subtler and unique to an individual's response. This is why in **Chapter 6.** Some individual's data were analysed in depth and showed variances in TCRBV/J gene usage.

5.7.2. Disease types could be distinguished by overall TCRB CDR3 repertoire amino acid properties

5.7.2.1. PNH TCRB clones had more acidic and more negative residues in TCRB CDR3s than normals and AA patients

As clonality did not appear to be a significant find in the data sets, CDR3 amino acid properties in the TCRB repertoire as a whole were analysed irrespective of clonality. This helped assess whether there was immunological dysfunction in the TCRB repertoire. Non-surprisingly, AA with PNH, AA no PNH and PNH patients could be separated from normals based on some CDR3 amino acid properties. PNH patients had more acidic residues, more negatively charged and shorter lengths of CDR3s in their repertoires than normals. They also had lower aromatic residues, higher aliphatic and more hydrophobic residues. Some of these characteristics have been linked to autoimmunity, and others with Tregs, which would suggest an attempt to re-balance the TCR repertoire [**300,311**].

Interestingly, AA with PNH or showed to have lower basic residues, more acidic and more negative CDR3s than AA with no PNH, which again highlights that in the context of PNH, TCRB CDR3s seem to have more acidic residues and more negative residues. This will relate to the type of antigens they respond to e.g. antigens with more basic residues. This would also suggest that the antigens that PNH related TCRB clones are either interacting with or unable to interact with, depending on the mechanisms, are different to those in AA. It would be interesting to track the AA no PNH repertoires over time and assess whether they change and whether they develop PNH. CDR3 properties could serve in future as a biomarker to predict perhaps whether PNH will present in some AA patients.

5.7.2.2. Are recovering PNH TCRB repertoires becoming more autoimmune?

PNH patients who were recovering were significantly older than the average PNH patient by around 10 years. These patients also had the shortest CDR3s (excluding complexes cases and groups of n=1) and more negative CDR3s than the other solely PNH groups. These factors could suggest some antigen skewing in the repertoire and have been linked with CD8+ T-cells [**312**]. It would be interesting to see if these patients had or go on to develop any known autoimmune diseases leading to the recovery or as a result of recovering from PNH. The older age may suggest mechanisms of immune-ageing (**Section 7.3.2**) as factors in recovery. Active PNH groups had fewer hydrophobic residues in CDR3s. Hydrophobic CDR3s have been linked with autoimmunity [**88**]. Fewer of these would suggest it is less likely that there are autoimmune mechanisms involved with active PNH but could link with autoimmunity in recovery.

5.7.2.3. PNH clones increasing in the context of AA had the most negative CDR3s

Out of the AA groups, AA increasing PNH clone had the most negative CDR3s and a higher number than the other solely PNH groups. As these are indicated in CD8+ populations, it could mean that TCRB responses are attributed to AA, or responses differ in PNH when AA is also present. It could also be that in AA with increasing PNH, the bone marrow environment is extremely toxic due to AA, allowing PNH clones to expand, which either could be as a result of or result in the change in these TCRB characteristics.

5.7.3. Chapter conclusion

In conclusion, the successful production of 77 AA and PNH TCRB patient repertoires and subsequent bioinformatics analysis allowed for TCRB dynamics to be assessed in the context of disease, by comparison with the large normals data set. Although no shared TCRB clone specific to PNH was identified, the results identified novel TCRBs in AA and PNH along with disruption of TCRB repertoires present in PNH patients that share common CDR3 amino acid properties. This immune dysfunction was not observed in the normal datasets. The analysis in **Chapter 6.** will expand on these findings and assess TCRB repertoire dynamics in patients over multiple time points, along with interesting cases and some BM and PB matched samples.

Chapter 6 – In depth analysis of TCRB dynamics in AA and PNH patient TCRB repertoires

6.1. Introduction

This results chapter follows on from the work in **Chapter 5**. However, rather than focussing on trends between categories and diagnosis sets, a number of patients were selected to analyse on an individual level. Experiments used to improve the reliability of conclusions drawn from the TCRB data are also discussed in this section including comparing TCRB repertoires from bone marrow and peripheral blood matched TCRB repertoires. **Chapter 4** and **5** highlighted how dynamic TCRB repertoires are between different individuals rather than over-time for the same person. This chapter investigates the TCRB dynamics over time and in the context of other factors such as BM versus PB. In depth analysis of the samples in this chapter allowed for more conclusions to be drawn about the dynamics of TCRBs in PNH in relation to the BM, at the start of the disease, during spontaneous remission, and changes over-time in the context of changes in clinical status.

6.1.1. Analysing TCRB repertoires over time to understand TCRB and PNH clone size dynamics

TCRB repertoires are dynamic and many factors can have an effect on a repertoire both short-term, for example acute viral infection, or over the long term, for example a chronic infection like EBV (**Section 1.4.**). A number of AA and PNH patients had multiple short-term time point samples (< 2 years) and long-term time points (>2 years). Analysing these TCRB repertoires helped identify persistent TCRB clones, those that appear clonal at multiple time points, often indicative of chronic disease. Immune mechanisms linked with chronic disease such as T-cell exhaustion will be discussed in the context of these results in **Chapter 7**. Fundamentally, TCRB repertoire studies are challenging, as realistically the whole immunological history of the patient prior to the sample being taken has led to the current diversity of the repertoire. This makes inferring information from the TCRB repertoire of one sample at one time-point with accuracy, difficult. However, multiple time points can help strengthen these inferences and assess changes over-time.

6.1.2. Establishing whether “interesting cases” can be identified by abnormal TCRB repertoires

A number of individual PNH and AA patients were annotated as “interesting cases” by Dr S. Richards. For example, one patient did not haemolyse as expected despite having 50% Type III red blood cells. It was therefore interesting to analyse these patients on an individual basis to assess whether there were changes in TCRBs that could be linked to their specific cases. Included in this cohort were a number of newly diagnosed PNH patients to assess whether TCRB repertoires are markedly different when first diagnosed with PNH, when pathogenesis would be expected to be at a height. An AA patient with progressive disease was also assessed here. Due to clonal TCRBs not being present in all AA repertoires in the previous chapter, perhaps due to immunosuppressive treatment, it was important to analyse the repertoire of this patient. If AA is progressing, it would be expected that a clonal expansion of TCRBs would be present and that immunosuppressant treatments either had not been administered yet or were not involved in the suppression of TCRB clones. Two spontaneous remission patients were analysed in this section. They had recovered from PNH having had it for over 19 years each. They were assessed on an individual level and in the context of normals and PNH to identify any TCRB changes occurring linked to recovery.

6.1.3. Bone marrow TCRB repertoires versus peripheral blood TCRBs

PNH, like AA, is a bone marrow disorder and ideally bone marrow samples would be used in this project. There are many practical problems associated with obtaining bone marrow samples, in particular, the invasive nature of the procedure. Healthy control bone marrows are difficult to obtain and therefore any bone marrows considered “normal”, are usually ones from patients with another haematological disease or are under investigation for an undiagnosed haematological disease. This would result in limited comparisons of repertoires with normals and would make identifying PNH specific TCRB responses in the BM difficult. It would also be unethical to put patients through unnecessary bone marrow sampling. Often bone marrows are taken at the beginning of treatment or when something has gone wrong which would also bias the results. For this reason, peripheral blood samples were used as the default. However, 3 patient samples had bone marrow and peripheral blood matched samples. These were analysed to discover whether peripheral blood TCRBs can be used as a representation of the repertoire in the bone marrow environment. On the other hand, the peripheral blood may also show something occurring that is unrelated to the bone marrow environment but still related to the disease. For example, infections that show in the blood such as EBV specific clonal expansions, could drive changes in the bone marrow or vice versa [**313**].

There could be a population of GPI- T-cells in the peripheral blood driving the progression of the disease but not present in the bone marrow.

6.1.4. Experimental bone marrow TCRB repertoires versus patient bone marrow TCRB repertoires

As detailed in **Chapter 1**, this project work follows on from original experiments by Dr. R. Kelly. These experiments used PNH and normal BM samples in a laboratory setting to assess the fitness of *PIG-A* mutated HSCs compared to normal HSCs in the context of PNH. They were also used to assess the ability of the bone marrow to generate haematopoietic cells/colonies. The *in vitro* long-term bone marrow culture model was designed to mimic the *in vivo* conditions of a PNH patient's bone marrow using a stromal cell line. Preliminary data showing that when T-cells from patients with PNH were present in the artificial bone marrow, *PIG-A* mutated HSCs had a selective growth advantage over normal HSCs, implicated T-cells as an extrinsic cell factor that help the clonal expansion of mutated HSCs. When these T-cells were not present, normals HSCs had the proliferative advantage as the T-cells were inhibiting the function of the normal HSCs. This was indicated by more progeny that were GPI+ [**234**].

These experiments have shown that normal bone marrow cultures last 8 weeks compared to the PNH which lasts around 4 weeks. When T-cells are removed, the PNH culture also lasts 8 weeks. Both a normal bone marrow and a PNH bone marrow culture were sampled at 45 days still containing T-cells. The PNH BM, still containing T-cells and still functioning at 45 days was an unexpected finding and identifying the T-cell populations was of great interest. Therefore, these samples were sequenced and analysed to assess the TCRBs in the bone marrow repertoire investigating as to whether the effect of T-cells is linked to the TCR which would be indicated by skewed clonality. The immune escape of PNH HSCs due to their GPI loss is an accepted hypothesis [**314**]. It therefore would be expected that there are a number of clonal mature T-cells involved in the pathogenesis or progression of PNH residing in the bone marrow. These will have migrated from other secondary lymphoid organs, such as the spleen, so would potentially be matched in the peripheral blood [**315**]. Another possibility is that the response is not clonal in response to superantigens. The symptoms could be as a result of cytokines, stress responses (NK receptors) or inflammatory mediators [**316**].

6.1.5. UMI- adapted TCRB data to improve TCRB repertoire studies

Throughout the project a UMI TCRB sequencing method was developed. This involved the tagging of each biological TCRB in a PCR reaction with a unique 12bp code split at the beginning and end of a sequence. UMI techniques are becoming more routinely used as it is thought they can reduce technical biases associated with TCRB sequencing [**166**] previously discussed in this thesis. A summary of the results from successfully sequenced UMI libraries will be discussed.

6.1.6. Chapter aims and objectives

The aim of this chapter was to investigate TCRB repertoires in AA and PNH patients on a more individual basis to assess subtleties in responses that may not be picked up when grouping patients together. By using samples from scenarios detailed above, the analysis hoped to strengthen the reliability of the project findings and improve understanding of the links between TCRBs and PNH clone sizes.

Objectives:

- Assess patient TCRB repertoire dynamics over time
- Determine whether TCRB repertoires from patients annotated as “interesting cases” differ from other patient repertoires
- Draw parallels between bone marrow and matched peripheral blood TCRB repertoire dynamics
- Compare experimental LTBMC TCRB sequencing data with a PNH patient BM sample
- Analyse UMI adapted TCRB repertoire data

As research into TCR dynamics, particularly CDR3 amino acid characteristics, in PNH advances, the biological findings of the work in this chapter will become more apparent. The chapter is structured so that the sections will act as case studies that can be read independently from one another as points of reference and interest in future.

6.2. Short-term time points showed changes in TCRB repertoires

Seven patients had multiple short-term time points, 4 of which also had longer term time points and will be discussed in the next section.

6.2.1. Patient 00551 – short-term samples

The first patient 00551, was 51 at time of the first sample, male and had PNH with a large stable clone (100%) that had been stable for over 15 years. Both samples were monoclonal taken 6.5 months apart. However, the clonal TCRB was not the same. The first sample had a clone at 10.4% which was moderate with a V29-1/J1-4 and CDR3 'CSVGSGGTNEKLFF' linked with EBV and Influenza A as discussed in the next section. The second sample had a top TCRB clone at 3.1% which was low response, V6-5, J1-5 and CDR3 'CASSQRAGYQPQHF' and no hits in databases or literature. The samples shared 167 clonotypes and there were no significant differences between CDR3 characteristics. This number of shared TCRB clonotypes was above the maximum found to be shared between two normals in **Section 4.5.7**. This is to be expected in a sample from the same patient. A number of shared memory T-cells should be picked up at each sample stage. The buffy coat sample had under half the number of TCRBs of the gDNA sample with only 11,238 to 25,952, as to be expected with buffy coat samples.

The number of unique TCRBs for the first sample was almost six times the number for the buffy coat sample at 2608 to 459. The inverse Simpson value for the first sample was lower at 81 compared to 222 6.5 months later. However, d50 was higher at 260 to 96. V usage for the first sample was V29-1, accounting for almost half the repertoire, then V19 and V 6-5 at just above 10% each. This was similar to the trend observed in normal TCRB repertoires. Six and a half months later V usage had dropped slightly to around 35% and V19 and V6-5 were just under 10% as the most common. J usage for the first sample was J2-3 followed by J2-1 for the most common. Six and half months later this was J2-3 followed by J1-5 closely followed by J2-1.

The first sample's top clone was not present in the sample taken 6.5 months later. The second sample was buffy coat whereas the first was not which could have had an effect on the clone percentages and the TCRB clones identified (**Section 4.3.3.**). Most likely the EBV related clone was not picked up in the BC sample rather than completely disappearing as even if it was not active it would be expected to be circulating in a memory state. These memory states however would be non-clonal and perhaps if multiple technical replicates were taken, it would have been detected (as detailed in **Section 4.3.**).

The second sample's top clone was present in the first sample but only at 0.28% showing clonal expansion over the 6.5 months perhaps due to re-infection. Half of the top ten TCRBs in the first sample were present in the second sample. Eight of the top 10 TCRBs in the second sample were present in the first. This would infer that, assuming the majority of the top 10 TCRs were at detectable levels in each patient, that a number of new memory T-cells could have entered the repertoire over the 6.5-month period.

6.2.2. Patient 004WZ

Sample 004WZ was a patient whose diagnosis went from AA to haemolytic and was aged 43 and male. At the time of sampling they had a large PNH clone, greater than 95% and had stable disease for over 2 years. The second sample was taken 17.5 months later and was buffy coat. The first sample showed no clonal response. However, the second sample showed a polyclonal (3 clonal TCRBs) response, with clones at 9.65% ('CSAHADAGANVLTF'), 8.4% ('CASSQLGDGNTIYF') and then 6.5% ('CATSTWDREGANVLTF'). The next clone was non clonal at 1.6%. Both samples shared the top TCRB, V29-1, J2-6, 'CSAHADAGANVLTF', however, it was only present at 1.7% in the first sample. The other two clonal populations in the second sample were present in the first sample, but at very low populations. This indicated that they had clonally expanded over the 17.5 months. It would suggest that these were memory T-cells, and due to reinfection for example, have clonally expanded. It would be interesting to sample the patient again to see if these populations had contracted since or whether the infection or chronic immune response remained.

There were no significant differences between CDR3 properties for both repertoires inferring general stability over the time points for the majority of TCRBs. However, when looking at CDR3 property distributions of patient 004WZ in the primary cohort (**Figure 56**), the majority of the CDR3s had no basic residues. Both samples saw on the lower end of TCRB clones expected at below 12,000 with the buffy coat sample seeing about a third less. The second sample had about half the number of unique TCRBs decreasing from 602 to 379. Diversity in the first sample was greater than the second with an Inverse Simpson value of 318 to 43 and d50 went from 138 to 51, showing greater clonality 17.5 months later. The two samples only shared 26 TCRB clonotypes which was lower than expected. Both samples shared V29-1 as the highest usage, but the first sample had more than double the usage. V19 was the second highest usage at about half of V29-1. For J usage, J2-3 and 2-1 were the highest for the first sample, 17.5 months later this had changed to J2-6 closely followed by J-1-3, J1-4 and J1-5.

Six of the top ten for the first sample were present 17.5 months later, but only 5 for the later sample were present in the first sample. The discrepancies could be to do with the buffy coat sample, which may cause technical variation, or potentially linked to the disease or a separate infection. The repertoire suggested a drop in diversity over time and it would be interesting to link this with future changes in clinical status.

6.2.3. Patient 004UV

004UV was 58, male and had PNH with a decreasing PNH clone of 45% at time of sampling the first sample. Samples were taken five months apart, both of which were non-clonal with top clones below 1.25%. As discussed in **Section 4.3.4.**, with levels this low it was unlikely top clones would be identical. V29-1 had the highest usage for both samples nearing 40% of the repertoire followed by V19 at 10%. J2-1 and 2-3 were most used for both samples in agreement with more "normal" TCRB repertoires. There were no significant differences in CDR3 characteristics between TCRB clones. Both samples showed good numbers of TCRBs at over 24,000. The number of unique TCRBs increased slightly over time from 1318 to 1931. Inverse Simpson values remained fairly consistent at 513 and 529 with d50s of 232 and 251, so there was a slight increase in diversity and reduction of clonality although no TCRBs were clonal in response level, inferring stability in the repertoire. Over a third of the TCRB clonotypes were shared between samples at 547, again inferring stability. All top 10 clonotypes for each sample were found in the other sample's repertoire. Tracking of top 10 clonotypes from one time point to 5 months later showed them all at non-clonal levels, again inferring stability. This recovering TCRB repertoire indicated that recovering TCRBs could be more stable than those with active or progressing PNH disease. All of the analysis inferred a stable TCRB repertoire over time in line with the recovering PNH status.

6.2.4. Patient 0054O

Patient 0054O was 38 at the time of the first sample, male and was AA with a small PNH clone at less than 1%. The samples were taken five months apart. The majority of the TCRBs in the first sample were clonal as very few TCRBs were produced in the final sequencing reads. Usually the sample would have been disregarded but due to an additional time point it was analysed. The sample 5 months later was not clonal. Only 216 TCRBs were identified compared to 26097 five months later, 10 were unique compared to 2601. The later sample, had high diversity with an inverse Simpson of 786 and d50 of 381 inferring stability again. V29-1 was the most used V gene and J2-3 and J 2-1 for the J gene usage in

agreement with more “normal” TCRB repertoire usage. All of the TCRBs from the first sample were present in the second.

Five of the top ten TCRBs found 5 months later were also present in the first sample, which was a good finding considering there were only 10 unique TCRBs identified. This highlights similarity over the time points despite there being a technical issue leading to a lack of TCRB reads in the first sample. This could indicate that in both patient 004UV and 0054O, stability in the TCRB repertoire shown by similar diversity measures and shared clonotypes, could be linked to stability of clinical status.

6.3. Long-term TCRB repertoire analysis identified persistent TCRB clones

For nine of the patients, there were backdated gDNA samples from 2013 available. They were used to track changes over time in patient TCRB repertories and to determine whether responses varied with some of the changes in clinical status (**Table 26.**). This analysis allowed the identification of any persistent clones. Persistent TCRBs could be indicative of chronic immune responses. These were defined as any clonal TCRBs (same CDR3 amino acid sequence, TCRBV and TCRBJ, occupying more than 2.42% of the TCRB repertoire) that were present at clonal levels in multiple time points for a patient. Three patients had persistent clones identified.

Table 26. Metadata for 9 PNH or AA patients with samples at least 4 years apart. Detailing any changes in clinical status of PNH and the time frame between each patient sample. Red samples indicate patients whose clinical status varied between time points.

Patient Sample	Clinical status in 2013	Clinical status at recent sampling	Years between samples
00551	Haemolytic (100% stable)	Haemolytic (100% stable)	6-6.5
004VG	Haemolytic (90% stable)	Haemolytic (90% stable)	4
004VN	Haemolytic (70%)	Haemolytic (70%)	4
004VR	Thrombotic, large WC clone, small RC clone, strange type II's	Thrombotic, large WC clone, small RC clone, strange type II's	4
00567	Haemolytic (100% stable)	Haemolytic (100% stable)	6
0053E	Haemolytic (90% stable)	Haemolytic (90% stable)	5
004VH	Haemolytic (90% stable)	Haemolytic (90% stable)	4
004V3	Haemolytic (70%, falling)	Haemolytic (>50%, falling)	4
00563	Aplastic (>30%, variable)	Aplastic (>50%, slow rise)	6

6.3.1. Patient 004V3

Patient 004V3, aged 52 and female, back in 2013 had a diagnosis of PNH with a large clone of 70% but this was falling. Another TCRB repertoire sample was analysed 4 years later where the patient still had

PNH and the clone was falling, but the clone was estimated to be lower at above 50%. Circos plots (**Figure 59.**) are used extensively in the subsequent analysis.

They are excellent for assessing whether certain TCRBV/J gene usages and combinations are more common in a repertoire. They investigate clonality irrespective of CDR3 which is important when it is not known if the immune response is antigen specific. The circle represents the TCRB repertoire.

One half represents the V genes, the other the J genes. The ribbons joining them represent clones that share that TCRBV and J gene combination. The wider the ribbon the more repertoire space that VJ combination occupies and can be an indicator of clonality. In normals the most common usages were V 29-1, with J2-1 and J2-3.

Interestingly, when analysing the changes in TCRB repertoires for patient 004V3, TCRBV/J pairings (**Figure 59**) significantly changed over 4 years. In the 2013 sample, there was significant usage of the pairing TCRBV15/J1-4 (**Figure 59, left**). When looking at the patient's repertoire 4 years later this pairing had significantly dropped in usage and the more "normal" pairing V29-1/J2-3 became the most abundant. The change in the VJ pairings coincided with the fall in PNH clone size. Both samples had similar numbers of TCRB clones in the TCRB repertoire at 21417 for 2017 and 23598 for 2013. However, the sample in 2017 had almost double the number of unique TCRB clonotypes at 1138 to 2013's 620. As the repertoire appeared to be returning to "normal" it could be indicative of fewer autoimmune mechanisms or re-population of memory T-cells from previous pathogenic responses.

When calculating diversity metrics, the sample from 2013 was less diverse, and appeared more clonal. Sixty-three TCRB clonotypes were needed to make up 50% of the TCRB repertoire space compared with 161 in 2017. The inverse Simpson metric also indicates a higher diversity in the 2017 sample with a value of 247 compared to 2013's 28. The repertoires shared 99 TCRB clonotypes, more than between two normal individuals in **Chapter 4**. Both repertoire samples remained monoclonal and the top clone remained constant, therefore annotated as 'persistent'. This TCRB clonotype was V15, J1-4, with a CDR3 'CATSSQAGEKLFF'. The size of the TCRB clone decreased in line with the decrease in the PNH clone over time.

In 2013 it was 18.3% which was a moderate TCRB clonal response, almost at the large hyperexpanded stage (20%) down to a low-level responder at 4.15%. Querying this sequence across the TCRB databases and literature searching returned no results. From the top 10 TCRBs in 2017, 3 were not

present in the 2013. Tracking the clonotypes highlights how the top clone has decreased over the time.

The other top clones were not clonal and were present either at very low levels or not at all between samples. No significant differences were observed between CDR3 characteristics for each sample. It would be interesting to isolate this T-cell population further and perform single cell sequencing and paired chain sequence on the clonal CDR3 that is varying according to PNH status.

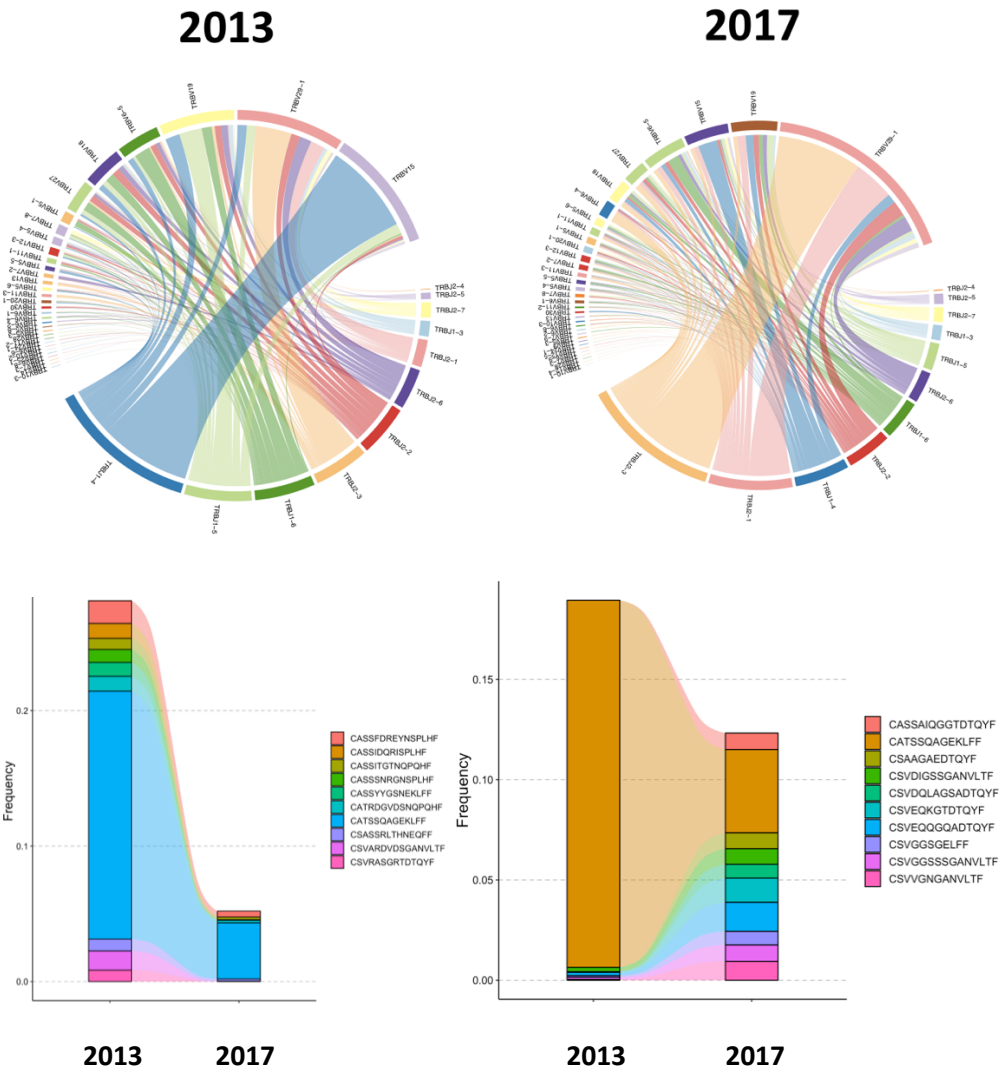


Figure 59. Patient 004V3 TCRBV/J pairings and top 10 TCRB clonotypes over 4 years.

The left column was the sample in 2013 and the right column from 2017. Circos plots showed the most common TCRBV/J pairings in the repertoire. The bottom row tracked the top 10 clonotypes from the 2013 sample through to 2017, the larger the box the more abundant the clone. The right-hand side tracked the top 10 from 2017 back to the 2013 repertoire.

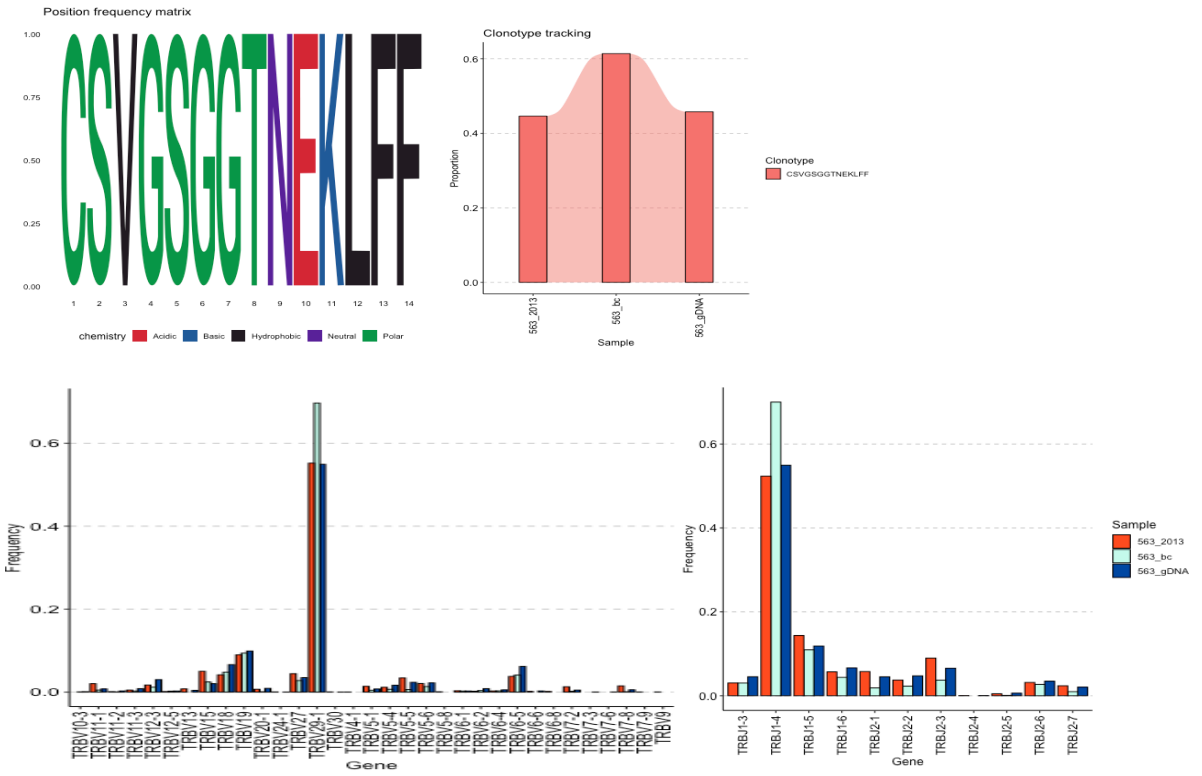
6.3.2. Patient 00563

Patient 00563 was male and aged 64 in 2013 and diagnosed as AA with an increasing PNH clone. Patient 00563 had a gDNA sample taken back in 2013 and another sample taken six years later. One of the recent samples was buffy coat which was also split at the second PCR stage to run as a sequencing bias test (**Section 4.3.6.**) and one was a Lymphoprep® sample.

In 2013 and then six years later the patient had the same diagnosis of having AA with an increasing PNH clone. At the time of the most recent sample the patient had had AA for ten years and a slowly increasing PNH clone currently at 52%. The sequencing test on the buffy coat sample showed the same TCRB clone for the top and second clone in the TCRB repertoire at similar percentages. All samples, from 2013 to the current time, had the TCRB clone with CDR3 sequence 'CSVGSGGTNEKLFF' as their top clone at very high percentages of over 45%. This was considered a hyperactive TCRB clonal response and possibly due to chronic infection or persistent disease such as AA. Unique TCRB sequences were considerably lower than the average in the AA and PNH datasets and can be attributed to having a large clonal population in the TCRB repertoire. High skewing was observed across all samples for both TCRBV and J gene usage (**Figure 60.**) with TCRBV29-1 and J1-4 accounting for more than 40% of the repertoire, representing the large clonal TCRB expansion. Seventy-seven TCRB clonotypes were shared between the buffy coat and gDNA samples, and the 2013 sample shared 35 with the buffy coat and 62 with the Lymphoprep® gDNA. Inverse Simpson values were the lowest observed in the study at 5 in 2013 decreasing to 4.7 and 2.6 six years later for gDNA and buffy coat samples respectively. D50 followed the same pattern of 8, 7 and 1 respectively showing incredibly clonal repertoires.

In order to analyse the monoclonal CDR3, both a positional frequency and probability (data not shown) matrix analysis was performed. This showed that over half of the CDR3 consisted of polar amino acids, concentrated in the first half of the CDR3, four hydrophobic residues were present, generally at the end of the sequence and there was one each of neutral, acidic and basic residues next to each other (**Figure 60 top left**). The CDR3 was tracked across the sample time points over six years (**Figure 60 top right**). This showed that both gDNA samples, one from 2013 and the other six years later, had similar levels of the CDR3 at 45% and 46% respectively. The buffy coat sample six years later had a higher abundance of the TCRB clone. This could be attributed to the fact that fewer T-cells were captured in buffy coat samples (**Section 4.3.3.**). In some respects, this shows that this person's repertoire was stable, but in a chronic state of infection or persistent disease.

The patient was 63 at the first sampling so it could be an age-related infection such as CMV or EBV but it is a very abnormal response. In order to assess whether the CDR3 was found in other diseases, a literature and database search was performed. The CDR3 was found in a number of publications and databases as specific to EBV. The MHC was annotated as HLA-A2, antigen protein was BMLF-1 in CD8+ T-cells [**254,255**]. Other autoimmune diseases such as multiple sclerosis have been linked to EBV infection [**317**] which could be relevant to this patient.



Sample	Age	Top clone / %	CDR3 sequence /aa	TCRBV family	TCRBJ family	Second top clone /%	CDR3 sequence /aa	TCRBV family	TCRBJ family	Unique TCRB
gDNA	70	46	CSVGSGGTNEKLFF	TRBV29-1	TRBJ1-4	1.2	CASSSTNEKLFF	TRBV12-3	TRBJ1-4	593
Buffy coat	70	57	CSVGSGGTNEKLFF	TRBV29-1	TRBJ1-4	1.2	CATGQDSNQPQHF	TRBV19	TRBJ1-5	456
Buffy coat-sequence bias check	70	61	CSVGSGGTNEKLFF	TRBV29-1	TRBJ1-4	1.2	CATGQDSNQPQHF	TRBV19	TRBJ1-5	310
gDNA-2013	64	45	CSVGSGGTNEKLFF	TRBV29-1	TRBJ1-4	1	CASSIDSNQPQHF	TRBV19	TRBJ1-5	392

Figure 60. TCRB repertoire analysis of an AA patient with a slowly increasing PNH clone from 2013 to 6 years later. Positional frequency matrix analysis of the clonal CDR3 and clonotype tracking. Patient 00563 had an monoclonal response both in 2013 and six years later at a frequency of over 45% of the entire TCRB repertoire. Analysing the properties of the amino acids that made up the CDR3 (top left) and tracking of the size of the clone over time (top right) helped the understanding of this patient's TCRB repertoire dynamics. Two samples six years later were taken as buffy coat and gDNA and compared to the sample from 2013 to assess whether any changes in the TCRB repertoire had occurred over time. The graphs show TCRBV and J usage overtime. The table shows the top two clones for each sample, including a sequencing bias repeat of the buffy coat sample, which highlighted that the TCRB clone is not a technical replicate and was biological.

6.3.3. Patient 004VR, PNH with interesting monocyte and red cell populations

Patient 004VR was a PNH patient with a large stable clone in 2013 and 2017. Granulocytes were all type III, monocytes were all type II and red cells were almost undistinguishable from normal, but it is likely the clone was 100%, the patient was also thrombotic. This patient was annotated as an interesting case and was 40 years old and female.

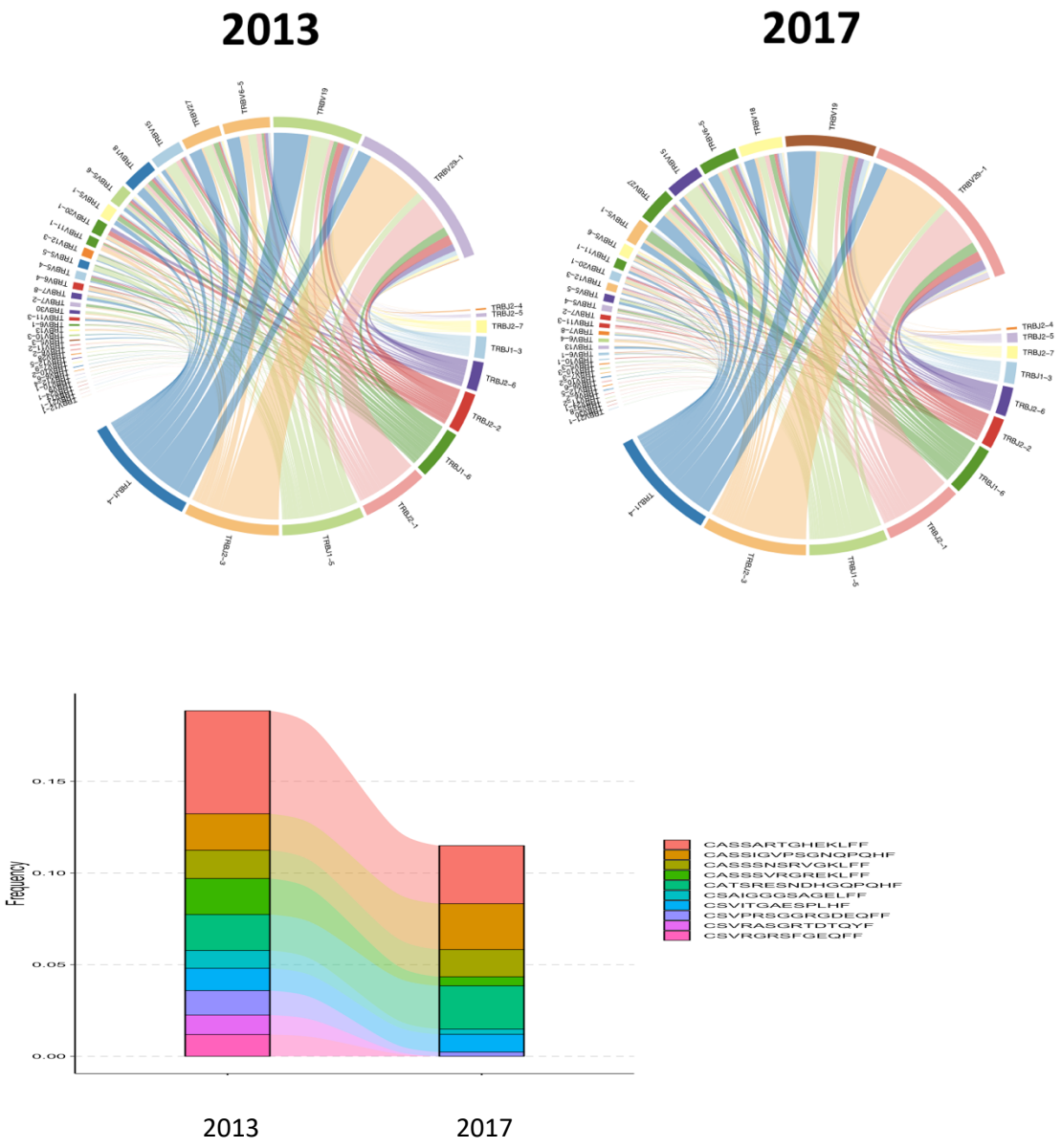


Figure 61. TCRB repertoire analysis for patient 004VR over 4 years.

TCRBV/J pairings for patient 004VR from 2013 to 2017 (top). Clonotype tracking (bottom) tracked the top 10 most abundant clonotypes in 2013 in the 2017 repertoire. The larger the square the more repertoire space occupied by the CDR3.

Between 2013 and 2017 overall TCRBV/J pairings were stable with 29-1/J2-3 being the most common indicative of more "normal" TCRB V/J gene usage. Even the small pairing between V29-1/J1-4 was consistent between samples. The circos plots were similar between long term samples perhaps indicative of the stable PNH status, unlike in the previously presented patients where their clinical status was changing (**Figure 61., top**). On investigating clonality, in 2017 the patient had a polyclonal TCRB response with two low responding clones. V19, J1-4 at 3.2%, 'CASSARTGHEKLFF', and V19, J1-5, CDR3 'CASSIGVPSGNQPQHF' at 2.5%. Looking at TCRBJ gene usage, J1-4 and 2-3 accounted for over 20% of the repertoire each. J1-5 and 2-1 accounted for around 15% each. Neither of the CDR3s returned any matches when querying the TCR databases. TCRBV15, 18 and 6-5 accounted for around 6-8% of the repertoire each, V19, around 18% and V29-1 the most at over 30%. V29-1 paired with J2-3 and J2-1 were the most common pairings. The same top two TCRB clones were present in the 2013 sample. However, the response was monoclonal, and the top clonotype V19/1-4 was present with a moderate response of 5.62%. Over the years the T-cell clone may have decreased slightly but more likely remained at a similar level, giving allowance for slight technical variances in percentages of TCRBs. When looking at diversity metrics the more recent sample was slightly more diverse with a higher inverse Simpson value of 225 to 137. D50 values increased over time from 83 to 142, indicative of less clonal TCRB repertoire in 2017. It would be interesting to see if this stability and increase in diversity would mean that in a few years the patient starts to recover. There were 142 shared TCRB clonotypes which was much higher than observed between two healthy individuals. When tracking the top 10 clonotypes of each sample into the next, 2 of the 2013 sample TCRB were not in the 2017 sample. However, all 2017 TCRB top 10 clonotypes were present in the 2013 sample (**Figure 61., bottom**). Both returned similar clone numbers of over 22,000 and relatively similar unique TCRB clonotypes 711 in 2013 and 1002 in 2017. Compared to patient 004V3 these values seem relatively stable over time.

In 2017, when looking at CDR3 length, 004VR saw no significant differences when compared to normals or its PNH large clone stable group. 004VR had significantly more basic residues in CDR3s than normals or the PNH LCS group ($p<0.0001$) and significantly fewer acidic residues ($p<0.0001$) which was different to the general trend that PNH TCRB clones had more acidic residues (**Section 5.7.2.1**.) The aliphatic index was significantly higher than in PNH LCS group or normals ($p<0.05$). There was no significant difference in polarity. 004VR overall net charge was significantly less negative than PNH LCS and normals ($p<0.0001$). 004VR had significantly fewer aromatic residues than normals but not than the PNHLCS group ($p<0.05$). It would be interesting to monitor this patient to see if the TCRB clone begins to contract and if so PNH may decline or vice versa.

6.3.4. Patient 00567

00567 was a patient with PNH, haemolytic 100% large stable clone back in 2013 and when sampled 6 years later in 2019. This patient had had stable disease for over 30 years. The patient was male and was 66 years of age at the time of the most recent sampling. The circos plots indicated relatively stable V/J pairings in line with PNH clinical status, with a slight decrease in V19 and increase in 6-5. V29-1/1-5 was most common at both time points.

Both samples showed an monoclonal response and shared the same clonal TCRB sequence of V29-1, J1-5 with a CDR3 of 'CSVNWGSGNQPQHF'. In 2013 this clone was a hyperexpanded clone at 20.7%, by 2019 the TCRB clone was only a moderate responder at 11.5% but still a significant size. The second TCRB clone was also consistent across both samples although not clonal with the CDR3 'CASSPGDGGYEKLFF', V5-4/ J1-4. These TCRBs were persistent and could be indicative of persistent TCRBs in some patients with stable PNH and a stable PNH clone size.

Neither TCRB clonotypes returned any known results for origin, or disease pathology. Both samples had good TCRB numbers of over 23,000 and unique TCRB clonotypes of 842 and 892 for the 2019 and 2013 samples respectively. Diversity metrics were d50s of 110 and 88, along with inverse Simpson indices of 64 and 22 respectively. Lower diversity was shown, as to be expected with these larger monoclonal populations. When tracking the clonotypes for the top 10 in the 2013 sample, only half were found in the 2019 sample perhaps contracting to memory T-cell populations over-time. Three of the top 10 2019 sample were not found in the 2013 sample (**Figure 62**). No difference was observed between CDR3 characteristics between the samples.

Further work could include isolating the clonal TCRBs and sorting them according to cell markers before sequencing. T-cell markers associated with T-cell exhaustion and senescence could be used to infer whether the TCRB is chronic. The drop in the persistent TCRB clone could be linked to PNH recovery in a few years, tracking the patients TCRB could help determine this, as the patient has had the disease for a considerable length of time.

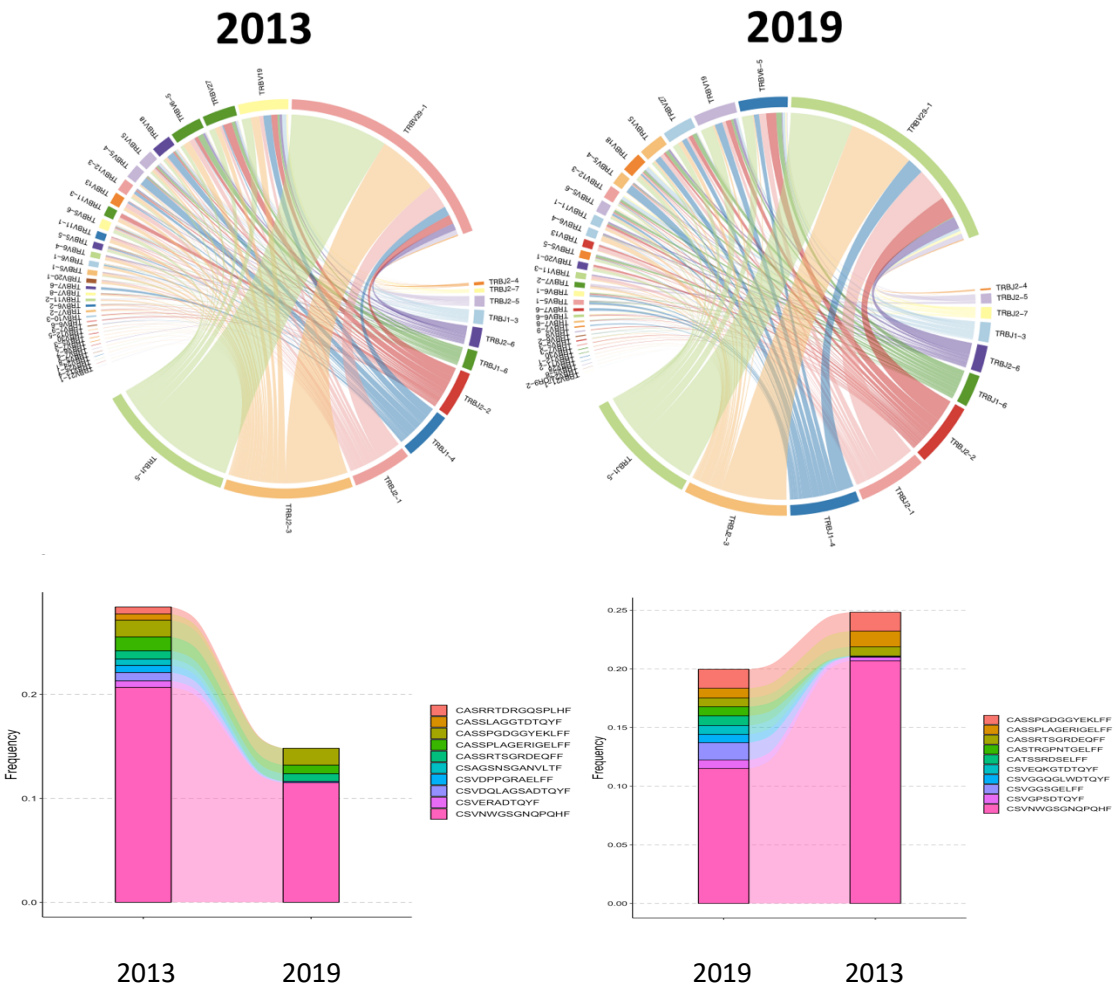


Figure 62. TCRB repertoire analysis in patient 00567 over 6 years.

TCRBV/J pairings for patient 00567 from 2013 to 2019 (top). Clonotype tracking (bottom left) tracked the top 10 most abundant clonotypes in 2013 in the 2019 repertoire. The larger the square the more repertoire space occupied by the CDR3. Bottom right tracked the 10 most abundant clonotypes in the 2019 repertoire back to 2013.

6.3.5. Patient 004VN

Patient 004VN had had haemolytic PNH for over 25 years. Both at sampling in 2013 and four years later, the diagnosis was a large 70% clone that was remaining stable. Two samples were available from 2013 with slight variation and were taken four months apart. According to their V/J pairings (**Figure 63.**) V29-1/J2-3 was the most common pairing indicative of a "normal" TCRB repertoire. However, in one 2013 sample, V29-1 seemed to have a higher usage in the repertoire than another sample which appeared to be more diverse. By 2017, the V29-1/J2-3 usage seemed to have increased slightly, but generally similar pairings to 2013, showing alignment with the consistency of the diagnosis and stability. No significant difference was observed between CDR3 characteristics between each sample.

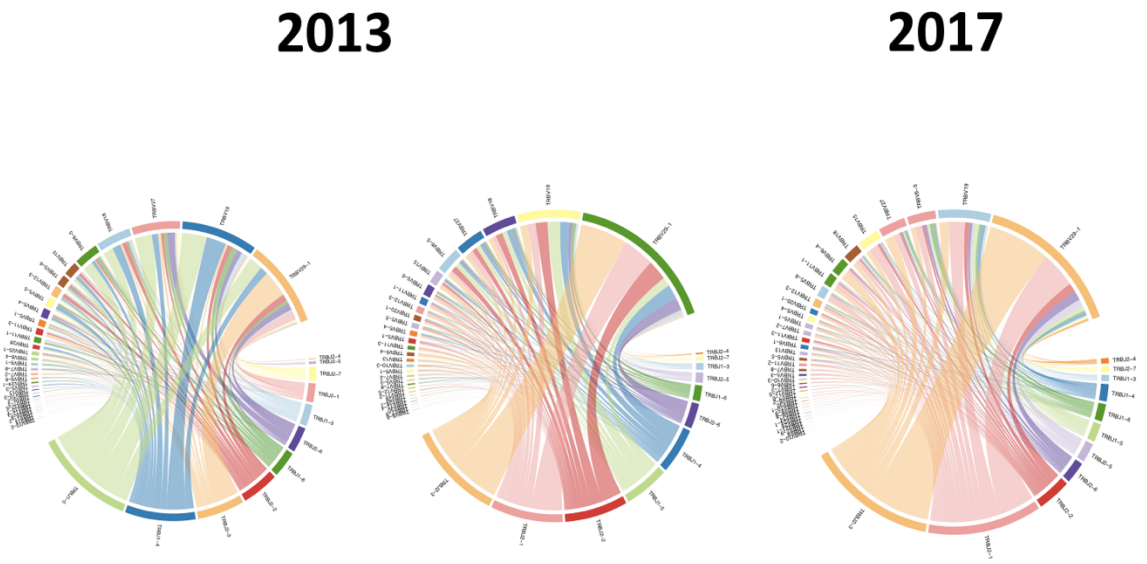


Figure 63. Patient 004VN TCRB repertoire analysis of TCRBV/J gene combinations

Two circos plots from 2013 and one from 2017 tracking how TCRBV/J pairings changed overtime.

None of the samples had clonal TCRB populations and the two samples from 2013 did not share the top TCRB. However, one sample's top TCRB was at 0.88% which, as previously shown (**Section 4.3.4.**) is most likely so small in abundance that even re-sequencing the same sample would result in considerable variation and it may not be detected each time. The 2017 sample did not share a top TCRB with either sample. When assessing the number of unique TCRB clonotypes in each repertoire, the 2013 samples had considerably lower numbers than the 2017 sample with 714, 1391 and 2129 clonotypes respectively. All repertoires had total number of TCRB between 22000 and 26200 Therefore it is unlikely the difference in unique TCRB clonotypes was technical and most likely it was biological.

It could be, as the percentages are not clonal, an indicator of increased diversity in memory T-cell populations. The 2013 sample with the lowest number of unique TCRB appeared to have a more diverse repertoire when not taking CDR3 into account (**Figure 63. first circos plot**). Using CDR3 specific calculations showed lower diversity. Inverse Simpson and d50 calculations had lower diversity at 293 and 113 respectively. It also differed in TCRBJ usage having J1-4 and 1-5 as highest usage. It shared 92 TCRB clonotypes with the other sample from 2013 and 116 with the 2017 sample.

The other 2013 sample had a higher inverse Simpson value at 666 than the 2017 sample at 570, but a lower d50 at 284 to 290. These numbers were similar and would show consistency and stability of the repertoire. Their TCRBJ usage was highest for J2-3 and J2-1 and they shared 197 TCRB clonotypes over double the number observed between two healthy individuals (**Section 4.5.7**). All top ten TCRB clonotypes for the 2017 sample were found in the 2013 sample that shared the same TCRBV/J usage and similar diversity values again showing stability over time with clinical status. Only two of its top 10 TCRB clonotypes were not in the other sample from 2013 indicating similarity. Although no additional metadata was available for the 2013 samples in order to explain the variances, it could be that a treatment changed or there was an alternative infection leading to the differences.

6.3.6. Patient 0053E

Patient 0053E was diagnosed as haemolytic PNH with a large stable clone (>90%) which remained the same between the five years of the samples and had been stable for 11 years. Two samples were sequenced from 2013 and one from 2018. When considering the V/J parings (**Figure 64.**) all samples had V29-1 as the most common with J2-3 but there was some variance in J family usage between all samples. One of the samples from 2013 was polyclonal with a low and moderate TCRB response at 3.14% and 4.75% respectively. Both clones shared V29-1 and J2-3 as the TCRB gene families. The other 2013 sample and the 2018 sample had no clonal TCRB populations. The clonal 2013 sample was taken only four months after the other sample. As this expansion, 'CSVRASGRTDTQYF' was not observed in 2018, it could be attributed to an infection at time of sampling. However, there were no known pathologies at the time of the analysis.

When comparing TCRBs and unique clonotypes in the repertoires, the polyclonal repertoire had 555 unique TCRBs, lower, as expected, than the other 2013 sample at 839 and the 2018 sample which had 1395 clonotypes. Interestingly, this was not as a result of lower TCRB numbers in the repertoire as all samples saw values in the range of 21000 to 26000. The diversity measures showed a similar trend with the clonal 2013 sample having a lower inverse Simpson and d50 at 159 and 113.The other 2013 sample had values of 336 and 133 with the 2018 repertoire having the greatest diversity with 506 and 232. When tracking TCRB clonotypes, almost all of the top TCRBs in the 2018 sample were present in both 2013 samples. The reverse trend was observed for the 2013 sample that had middle diversity of all samples. The 2018 sample only had two of the top 10 TCRB clonotypes and the other 2013 sample only 3 (**Figure 64.**).

For the polyclonal 2013 sample, 8 of its top 10 were found in the other 2013 sample and only two in the 2018 sample. The clonal populations were present in the other 2013 sample but had non-clonal and very low counts and were not present in the 2018 sample. The 2013 samples shared 72 TCRB clonotypes in line with numbers shared between two healthy individuals which is seemingly low for two samples taken in the same year from the same patient, indicative of immune dysfunction. The clonal TCRB sample shared 76 TCRB clonotypes with the 2018 sample and the other 2013 sample shared 127. There was no difference in CDR3 characteristics between samples. No matches were found for the clonal TCRB sequences.

Although the diagnosis did not change over time, results suggest repopulation, with a greater number of non-clonal TCRBs entering the repertoire over time, generating a greater diversity in the repertoire and inferring stability. The patient was 31 at the time of the first sample. Therefore, thymic involution was not quite at its peak, so newer, potentially more “normal” TCRBs could be generated. Perhaps this indicates attempts to restore more “normal” immune responses This might explain why TCRB clones present at higher frequencies in the older samples were not found at all in the newer sample’s TCRB repertoire.

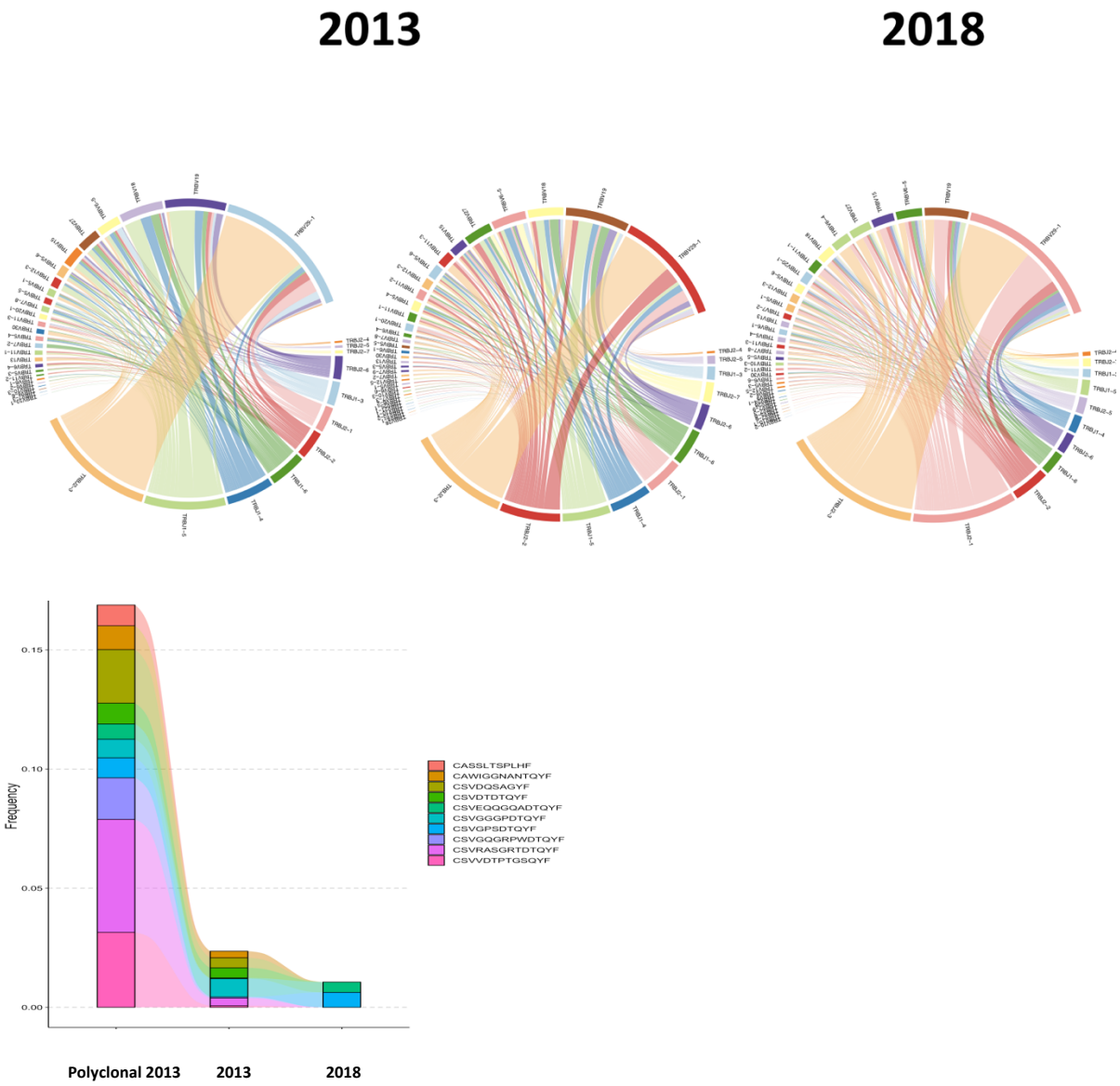


Figure 64. Patient 0053E TCRB repertoire analysis.

Circos plots depicting TCRBV/J pairings in two samples from 2013 and one in 2018 (top). The top ten most abundant clones from the polyclonal 2013 repertoire were tracked in the other 2013 sample and the 2018 sample (bottom).

6.3.7. Patient 004VH

Patient 004VH was diagnosed with a large stable PNH clone at above 90% and had been stable for over ten years. In 2013 the sample showed non-clonal TCRB populations. In 2017, both samples showed an monoclonal TCRB response. The two samples were taken two weeks apart. This was because the patient was on the Eculizumab stage of a clinical trial. The monoclonal response was on the border between low and moderate at 4.4%, increasing to 6.24% two weeks later.

Strangely, the CDR3 was not the same for these repertoires with the TCRB clones being V15/J2-3 and V5-3/J1-4 respectively. All samples had similar TCRB numbers between 19888 and 21697. Unique TCRBs varied between the 2013 and first 2017 sample having values of 1034 and 1130, with the second 2017 sample having 868. Over time, the inverse Simpson index decreased from 513, to 223, to 180. The 2013 sample had a d50 of 221. The first sample had a value of 163 and two weeks later 192. The 2013 sample shared only 37 TCRB clonotypes with the first 2017 sample followed by 74 with the next which was considerably lower than expected. The two 2017 samples only shared 47 TCRB clonotypes, which is lower than expected. The 2013 and first 2017 sample share V 29-1 as the most common V family gene and J 2-3. Two weeks later, V 19 seemed to be equal with 29-1 with the J 1 gene families highly represented (**Figure 65.**). No significant differences were observed between CDR3 characteristics for each sample.

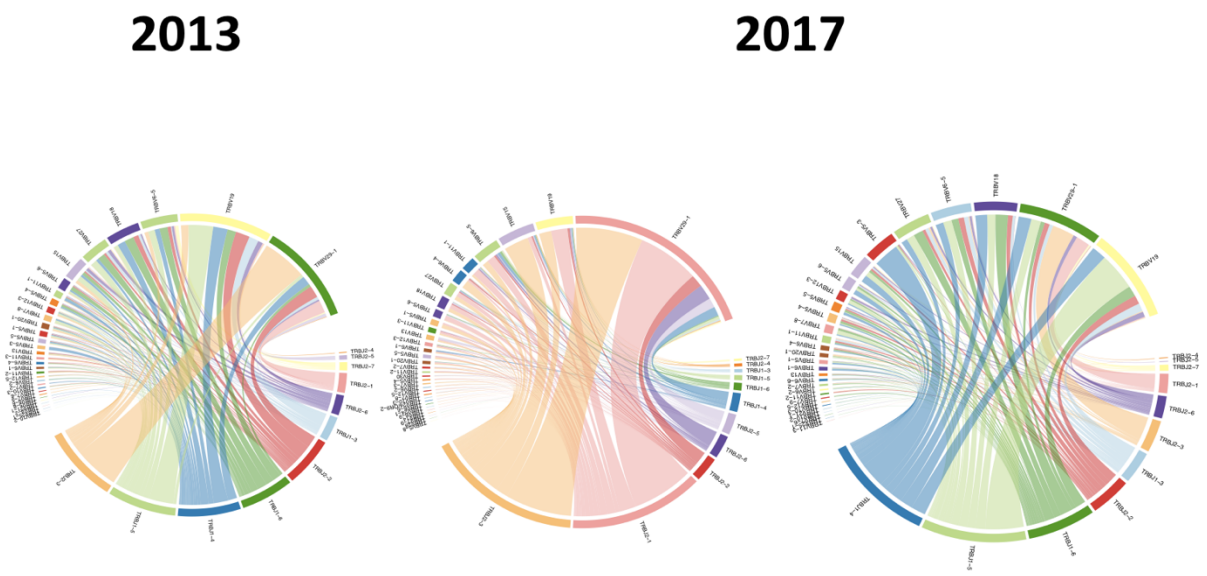


Figure 65. Patient 004VH TCRB repertoire circos plots showing TCRBV/J parings from 2013 to two samples in 2017.

When tracking the clonal populations observed in 2017 to see whether they were present back in 2013 just at low frequencies, the top clone in the first 2017 sample, 'CATSTLAGEAQYF' was not found in the other samples. The sample two weeks later had a top clonotype, 'CARSHEGGLDEKLFF', which was the third top clone in 2013 but at only 0.8% and not found in the sample two weeks before. If it was present in the sample two weeks before but at a low frequency, it may not have been picked up in that specific sequencing run. In order to assess whether these TCRB clones were linked to infection rather than PNH, their origin was queried in specialised databases, but no matches were found. The new clonal populations could be attributed to infections which may not have been present or had cleared two weeks later. The patient was also entering a new clinical trial, perhaps inferring a change in stability which was not noted in the metadata as the generation of an monoclonal response may or may not be directly linked with PNH status. Low sharing of TCRBs could indicate immune dysfunction and it would be interesting to assess if there was a specific reason for the patient being selected for the clinical trial. These theories will be discussed further in **Chapter 7**.

6.3.8. Patient 004VG

Patient 004VG was 59 in 2017, female, and had haemolytic PNH with a PNH clone above 90% and had been stable for 9 years. From 2013, the V/J pairings of V29-1/J2-1, J2/3 increased over the four years. TCRBV/J pairings in 2013 appeared more diverse than in 2017, despite the 2013 sample having an monoclonal low responding TCRB population of V18, J1-3 with a CDR3 'CASSPPGAAGNTIYF'. The 2017 sample was not clonal. Therefore, this monoclonal TCRB contracted over the four years. Both samples had good numbers of TCRBs, above 21500. The 2013 sample had a significantly lower number of unique TCRBs at 465 to the 2017 sample's 1340 along with a lower inverse Simpson of 189 to 330. D50 was also lower in 2013 at 85 compared to 174 in 2017. The 2013 repertoire was less diverse, more clonal but used a wider range of V/J parings (**Figure 66.**). Only 13 TCRB clonotypes were shared which was very low and only one of the 2017 sample top ten clonotypes was present back in 2013. None of the top ten samples from 2013 were present in the 2017 sample, including the clonal TCRB, suggesting it had contracted over time to undetectable levels. No significant differences were observed in CDR3 characteristics between samples. Over the 9 years since diagnosis, the repertoire may have become more stable with a more 'normal' TCRB repertoire being sustained. Most patients do not recover, they remain stable. Although these TCRB changes are not indicated with the PNH patient entering recovery, monitoring this patient over time perhaps would see an eventual recovery in line with the shift in TCRB response.

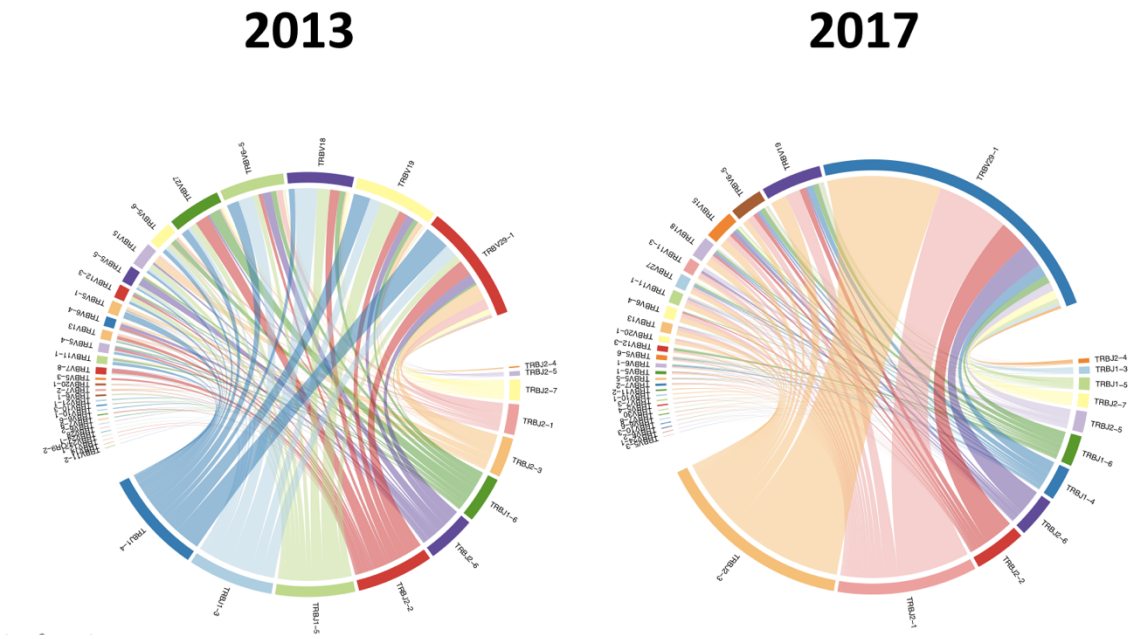


Figure 66. Patient 004VG TCRB repertoire circos plots showing changes in TCRBV/J parings from 2013 to 2017.

6.3.9. Patient 00551 – long-term samples

The patient 00551 had haemolytic PNH with a large stable clone at 100% and had had stable disease for over 15 years. The V/J pairings between 2013 and 2018 remained the same supporting this with V29-1 being the most common V family and J1-4,2-1 and 2-3 being the common J families (**Figure 67.**). Both samples saw high numbers of unique TCRBs at 2289 and 2608 increasing with time. TCRBs decreased slightly over time from 27,986 to 25,952. Inverse Simpson values were low, attributed to the clonal expansions with values of 28.5 increasing to 81 over time. D50 also increased from 162 to 260, which was to be expected as the clonal population decreased over time.

The two samples shared 657 TCRB clonotypes which was one of the largest numbers of shared TCRB clonotypes seen in this study. Both were monoclonal and the top clone in both samples was V29-1, J1-4 with a CDR3 'CSVGSGGTNEKLFF'. In 2013 it was close almost considered a hyperexpanded TCRB clone at 18.4%, which decreased over the years to 10.4%, which was a moderate response. This was a persistent TCRB clone. On searching for data on this CDR3 it became apparent that it was linked with EBV, CD8 T cells, HLA-A*02 and MHC1. One hit also returned Influenza A [**254-255**]. The persistent TCRB clone could be indicative of chronic stable disease in the PNH patient and chronic EBV infection.

As the TCRB clone is falling it would be interesting to see if it continues to contract over-time and whether this has an effect on TCRB clinical status. If the EBV specific clone decreases, and the PNH patient recovers, it could provide a strong basis for EBV infection as a factor in PNH, similar to AA and other immune disease [**318**]. This was the same CDR3 found in patient 00563. All of the top 10 TCRBs for the 2013 sample were present in 2018 and vice versa showing considerable stability in the repertoire irrespective of clonality. The immune response had not changed drastically over the five years in line with the stable PNH diagnosis. There were no differences in CDR3 characteristics between samples.

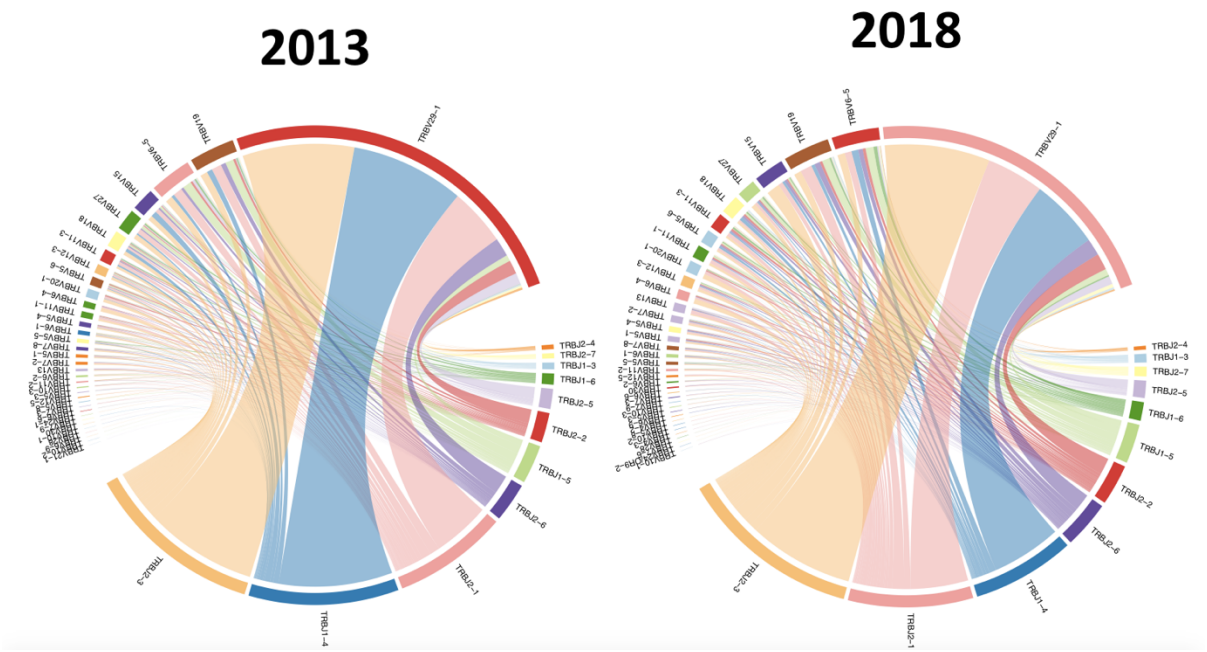


Figure 67. Patient 00551 TCRB repertoire circos plots showing changes in TCRBV/J parings from 2013 to 2018.

6.4. Case studies

A number of patients were annotated as interesting cases by Dr S. Richards. Therefore, these samples were analysed on an individual basis in the context of their diagnosis. Statistical analysis included padj values using the Holm method having compared means pairwise using the Wilcoxon method.

6.4.1. Patient 004YD, PNH with LOH

Patient 004YD was in the PNH new or increasing clone category, male and aged 61. In 2010 they were diagnosed as AA but now PNH with an 85% clone. They had a LOH from 6p25.3 to p21.32 which could lead to partial or full loss of MHC [**319**]. Therefore, analysis of this patient's repertoire could indicate alternative immune evasion in PNH. HLA genes are essential for producing MHCs which are needed for MHC mediated presentation of peptides [**320**]. Partial or complete loss could mean that immune responses are generated via T-cell subsets such as the CD1d restricted groups (**Section 1.2.2.**) that do not rely on MHC. The repertoire response was monoclonal with a low responsive TCRB.

The clone was present at 3.9% and was TCRBV19/J2-1 with the CDR3 amino acid sequence 'CATQPGGGGNEQFF'. The CDR3s were significantly shorter than normals ($p<0.05$) but not the PNH new or increasing group as a whole. CDR3s had significantly fewer basic residues than PNH new or increasing clone ($p<0.05$) but not fewer than normals. No significant differences were found between 004YD CDR3 properties and normals or the PNH new or increasing clone group when looking at acidic residue percentages, aliphatic residue percentages, degree of polarity, bulkiness and percentages of aromatic residues. However, 004YD had more negatively charged CDR3s than the other groups ($p<0.05$). TCRBV gene usage saw over 40% of the repertoire using V29-1, over 15% was V19 followed by V6-5 at just under 10%. J2-3 followed by J2-1 both over 20% were the most highly used J genes followed by J1-4 at around 11% and J1-5, 9%. Irrespective of CDR3, TCRBV29-1/J2-3 or J2-1 were the most common parings. In order to draw conclusions about the mutation and the TCRB response more patients would be needed, as there may be other factors in this individual's TCRB repertoire causing variations. Extracting the clonal populations and identifying cell markers on the T-cells as a whole would allow identification of CD1d restricted T-cells and a comparison could be made between this patient and other PNH patients as to whether more CD1d restricted T-cells were present in this patient as a result of the mutation.

6.4.2. Patient 005A5, AA with new PNH clone

Patient 005A5 was an AA patient with a new PNH clone at less than 10%. It was selected because of its recent diagnosis of PNH which could indicate differences in TCRBs attributed to being at the beginning of disease progression. V29-1, J2-3 was the most common pairing similar to normals. TCRBV gene usage analysis showed V29-1 accounting for over 40% of the repertoire followed by V19, which was above 10% and V6-5 at 9%. J usage saw 2-3 at over 30% followed by J2-1 at over 25%.

The immune response was not clonal which was surprising. CDR3 characteristics were compared between normals and the AA increasing PNH group for reference. There were no significant differences observed for GRAVY, percentage of aromatic residues, polarity or aliphatic index. Both 005A5 and the AA increasing PNH clone group had significantly shorter CDR3s than normals ($p<0.001$) but not than one another. Both 005A5 and normals had significantly more basic residues than the category group, significantly fewer acidic residues and overall less negative charge (but still negative) ($p<0.001$). 005A5 had CDR3s overall that were less bulky than the other groups ($p<0.05$). It would be interesting to have multiple samples as the patient progresses through the stages of PNH to see if there are variances in the TCRB response linked with PNH clone size and PNH progression.

6.4.3. Patient 005A9, AA with falling PNH clone

Patient 005A9 was an AA patient who had a large PNH clone originally, but it was falling, counts were improving, and it was now regarded as a small clone. Although multiple time points were not available to track TCRB clone sizes with falling PNH clones, it was interesting to assess the falling PNH clone in the context of PNH. The repertoire showed an monoclonal TCRB response at low levels (3.5%), V5-5/J2-3, 'CASPGTTTDTQYF'. The second clone was non clonal at 1.5%, V6-5/J2-2.

Despite this, the most common pairings were still V29-1/J2-3 and J2-1. The repertoire was compared to normals and AA with small PNH clones for CDR3 characteristics. 005A9 had no significant differences with the groups in regard to lengths, number of acidic residues, polarity, charge, bulk or aromatic residues, GRAVY or aliphatic index. It had significantly fewer basic residues than the AA small clone group more in line with normal repertoire values ($p<0.05$). This supports **Chapter 5** findings that PNH patient TCRBs have more acidic residues even in an AA context when analysed on an individual basis. Multiple time-points would allow inferences to be made as to whether the TCRB clone is falling with the PNH clone.

6.4.4. Patient 005D7, PNH, sample did not haemolyse as expected

Patient 005D7 was a PNH patient with a large stable clone. Interestingly, the sample did not haemolyse despite more than 50% type III red cells. It was analysed to assess whether these clinical observations were perhaps also shown in the TCRB repertoire in reference to *PIG-A* mutations (**Section 7.3.3.4.**). The repertoire showed an monoclonal response, low responding at 4.1%. The TCRB clone was V29-1, J1-4 and 'CSVGSGGTNEKLFF' which was the same as patients 00551 and 00563 linked to EBV.

TCRBV gene usage was highest in V29-1 at 40% of the repertoire followed by V19 and V6-5 at 10% each. J2-3 accounted for about 30% of the repertoire, followed by 2-1 at 20%, J1-4 at 12% and J1-5 at 10%. The most common pairings were V29-1, J2-3, J2-1 followed by J1-4. The second most common grouped pairings were the same J genes but with V6-5. When comparing the sample with the PNH large stable clone group and normals the sample had no significant differences for CDR3s properties GRAVY or bulk. It had significantly more basic residues than normals and the PNH group ($p<0.0001$) and fewer acidic residues, which is the reverse trend found in **Chapter 5** but most likely caused by the EBV specific TCRB clone. It also exhibited a higher aliphatic index ($p<0.0001$), lower negative charge (but overall still negative) ($p<0.0001$) than both groups but only a lower polarity than normals ($p<0.05$). Finally, the sample had lower percentages of aromatic residues than both groups ($p<0.01$). Identifying the numbers of TCRBs in the repertoire that were GPI+ or GPI- would be important to assess any links between the different PNH type cells and TCRs as the similarities would likely be attributed to the *PIG-A* mutations in this case.

6.4.5. Patient 005CY, newly diagnosed with AA

Patient 005CY was diagnosed with AA less than a year before the sample was taken and sequenced. Originally the patient had a tiny PNH clone, but no clone at sampling. The patient was a good example of active AA disease. The repertoire showed a moderate TCRB monoclonal response (11.4%), V6-5, J1-5, 'CASSYQGAQPQHF'. V29-1/J2-3 was the most common pairing followed by V6-5/1-5, which was the clonal population. V usage saw V29-1 accounting for over 40% of the repertoire, followed by V6-5 and V19 at 20% and <10% respectively. J 2-3 accounted for 25% of the repertoire followed by J1-5 at over 20%. When analysing CDR3 characteristics, the patient was compared to normals, AA no PNH and AA small PNH clone (as at one stage the patient had PNH). No differences were seen for GRAVY values, bulkiness, aromatic residues, polarity or percentage of acidic residues.

005CY saw shorter CDR3s than in normals ($p<0.05$) like the other groups and more basic residues than all the other groups ($p<0.05$) in line with AA TCRBs having more basic residues from **Chapter 5**. The patient also had a higher aliphatic index and significantly less negative CDR3s (although still negative overall) than normals ($p<0.01$). It would be interesting to see the diagnosis of the patient at present, whether PNH came back or not and monitor the changing TCRB repertoires over-time. As the patient seemed to have PNH then not, they may be likely to present with PNH again than another AA no PNH clone patient, therefore would be the first sample to be assessed for longitudinal sampling out of the AA no PNH patients when investigating if TCRB repertoire changes over-time could be linked with PNH presentation.

6.4.6. Complex cases

6.4.6.1 Patient 004WF, AA complex case, decreasing PNH clones

Patient 004WF was the complex AA case mentioned previously. The PNH clones were decreasing but it was annotated as a complex case. Common pairings were V29-1 with J2-3 and J2-1. V29-1 usage accounted for over 50% of the repertoire, V19, 10% and V6-5 5%. J2-3 accounted for 35% of J gene usage followed by J2-1 at 33%, all other J genes were below 10%. No clonal TCRBs were found. CDR3 characteristics differed from the norm as detailed in **Figure 54**. The normal repertoire V/J usage and no TCRB clonal populations would be indicative of a recovering patient. Multiple time-points would be necessary to evaluate whether TCRBs are linked with the recovery. As there are no clonal TCRBs it would be difficult to track contracting PNH clones with TCRB clones, unless the top clonotype was persistent even at non-clonal abundances.

6.4.6.2. Patient 004XV, PNH complex case, large PNH clone

Patient 004XV was the PNH complex case mentioned previously with large PNH clones. The most common pairings were V29-1, J2-3 and J2-1. Similarly, to the AA complex case, common pairings were V29-1 with J2-3 and J2-1. V29-1 usage accounted for over 50% of the repertoire, V19, 10% and V6-5 5%. J2-3 accounted for 35% of J gene usage followed by J2-1 at 33%, all other J genes were below 10%. No clonal TCRBs were found. CDR3 characteristics differed from the norm as detailed in **Figure 54**. Multiple time points would be necessary to assess as to whether the repertoire was stable over time, or TCRB PNH dynamics as there are no clonal TCRBs to track.

6.4.6.3. 004XX, PNH patient with LGL

Patient 004XX was a PNH patient that was annotated as a complex case with an 85% PNH clone. The patient was thrombotic and also suffering from LGL as mentioned in **Chapter 5**. The response was monoclonal, with a low response TCRB clone at 2.5%.

Although perhaps not as clonal as expected. The TCRB was V29-1/J2-6 and CDR3 amino acid sequence 'CSAAIGTDSSGANVLTF'. A number of TCRBs were almost above the clonal threshold and were appearing at very similar abundances, which was not usually seen in repertoires perhaps suggesting the potential for a polyclonal response also suggested in the literature [**321**]. These could all be similar and responding to super-antigen.

Over 80% of TCRB V gene usage was 29-1 with the next highest usage, V15 and V19 being under 10%. J2-3 occupied over 40% of the repertoire, followed by J2-1 at 30% and J2-6 at below 10%. The majority of the repertoire was occupied by V29-1/J2-3 and J2-1 pairings (**Figure 68.**). This was the highest usage of V29-1 so although indicative of normal responses the abundance was abnormal and likely pathological. CDR3 characteristics differed from the norm as detailed in **Figure 54** and **Section 5.4.5**. In particular, the repertoire was unusual in that the majority of the CDR3s had no basic residues. This was only seen in three other patients in the primary dataset. Identifying the subsets of T-cells would be interesting to assess whether the higher V29-1 usage was linked to a particular subset of T-cells such as Tregs trying to restore immune function as the clonal response was not large considering the skew in TCRBV usage.

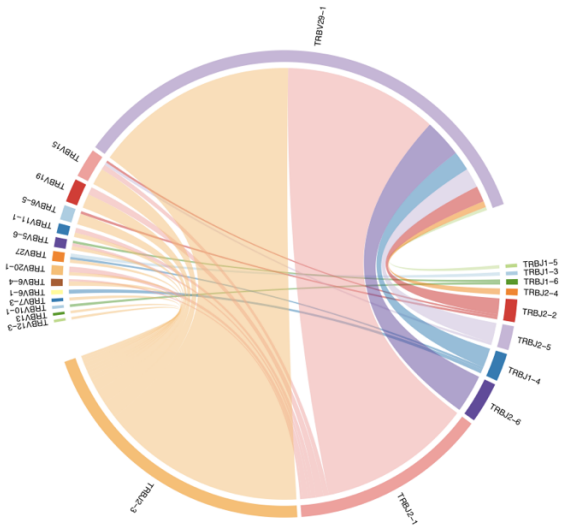


Figure 68. Circos plot indicating TCRBV/J pairings in the TCRB repertoire of a patient, 004XX with PNH and LGL.

6.4.6.4. Recently diagnosed PNH patient TCRB repertoires

Three patient samples, 005BO, 005BY, and 005DY were all new PNH patients. TCRB repertoires were sequenced from samples one month, seven months and three months after diagnosis respectively. It was assumed that they were in active states of PNH. However, this may depend on the time taken to receive a diagnosis. As no obvious TCRB clones were identified in all PNH patients in **Chapter 5**, it was important to analyse these newly diagnosed patients to assess whether perhaps the TCRB clones decrease with time from diagnosis. All samples had median CDR3 acidic residue percentages in line with other PNH patients. However, the box plot for 005BO, the most recently diagnosed patient, showed an opposite distribution to all but one of the other PNH patients. 005BO also had a slightly higher median value than the general PNH CDR3 length at 15 (**Figure 56.**). For percentage of basic residues, 005BO showed equal distributions around the median along with three other PNH patients. The rest of the PNH patients, including 005BY and 005DY had positively skewed data above the median (**Figure 56.**)**.** For overall net charge of CDR3s, 005BO was one of five PNH patients that did not have a median value of around negative one. 005BO had an overall net charge of zero (**Figure 56.**) whereas 005BY and 005DY had a charge around negative 1. Overall, this highlighted that the patient just diagnosed had abnormal distributions of characteristics in the CDR3 compared to other PNH patients. 005BO and 005BY were non-clonal. However, 005DY was polyclonal, with two moderately responsive TCRB, V6-5, J1-3 at 4.41% and V6-4, J1-6 at 3.72% (**Table 25.**).

When comparing the samples to the normals for changes in CDR3 characteristics, 005BO had significantly more hydrophobic residues in the CDR3s (less negative GRAVY value) than the normals ($p<0.01$), 005BY ($p<0.05$), 005DY($p<0.05$), and the entire PNH new or increasing clone group ($p<0.01$), suggestive of potential autoimmune related CDR3s. The other categories had no significant difference to normals. 005BY, 005BO and normals had longer CDR3s than the whole PNH new or increasing category and 005DY ($p<0.05$). Number of basic residues was significantly higher in 005BO than both 005BY and normals ($p<0.0001$). 005BY had significantly higher basic residues than normals ($p<0.01$). 005DY had the highest number of basic residues and significantly higher than all but 005BO ($p<0.0001$). PNH new or increasing were the only category to not differ significantly from the normals but was significantly lower in basic residues than all other of the PNH patients. 005BO saw significantly lower acidic residues than 005BY and normals ($p<0.0001$), PNH new or increasing ($p<0.0001$) but no significant difference with 005DY. 005BY and normals did not show a difference but 005BY had significantly more than 005DY but fewer than PNH new or increasing ($p<0.05$).

005DY had significantly lower values than all values apart from 005BO ($p<0.05$). PNH new or increasing had significantly higher values than all other categories ($p<0.05$). This was an interesting find. All the newly diagnosed patients had more basic residues in their CDR3s and fewer acidic residues than in PNH patients. This is the reverse trend to PNH patients as a whole which saw more acidic CDR3s. This might suggest that there are different subsets of T-cells involved in the pathogenesis and then progression of PNH. To investigate this theory, sorting the T-cells in GPI+ and – before sequencing for all PNH patients would allow variances in these populations to be assessed over-time. Perhaps either GPI+/- are biased towards acidic or basic. These percentages will vary with progression of PNH.

The order of aliphatic index from lowest to highest was normals, PNH new or increasing, then samples from 7 months to 1 month after diagnosis. All level saw significant differences ($p<0.05$) apart from between 005BY and PNH new or increasing, and 005BY and 005DY. 005BO had the highest aliphatic index. Polarity was not previously identified as a CDR3 characteristic with variances in PNH from **Chapter 5**. However, 005BO, had CDR3s with polarity significantly lower than all the other groups ($p<0.05$) indicative of more non-polar residues. The other groups showed no significant difference. Again, when investigating overall net charge, 005BO had a significantly less negative charge, nearing neutral compared to all categories except 005DY, (005BY and normals($p<0.0001$)). 005BY and 005DY had a significantly less negative charge than normals and PNH new or increasing ($p<0.05$). There was no significant difference between normals and the PNH new or increasing category.

There was no difference in bulk of amino acids between the groups. When observing aromatic residue percentages, again 005BO was found to have significantly fewer of these residues than all the other groups ($p<0.05$). No other categories had significant differences. When comparing TCRBV gene usage, 005BO and 005BY had over a third of the repertoire using V29-1, with just over 10% being V19, followed by V6-5 and V11-3, then V18. This was a different trend to normals which usually have a third of the repertoire as V29-1 too, but then around 15-20% V19, 10% V18, 5-8%

V6-5. When comparing TCRBJ gene usage, 005BY and 005BO had high J1-4, J1-5 and J2-3 in the repertoire, compared to normals which were J2-1 and J2-3. 005DY 3 months after diagnosis, had J1-4, 1-5 at around 20% followed by 1-6, then 1-3 at above 10%. then J2-3 at around 10%, V29-1 around 20%, V19 and V6-5 at around 15%, V18 at 7.5%. This trend was not observed when grouping all PNH new or increasing clones together, “newly diagnosed” could be a sample up to 2 years from diagnosis.

Ideally, more newly diagnosed patients would have their TCRB repertoires sequenced to strengthen the trends.

Tracking these patients as they progressed through PNH would firstly allow TCRB sizes to be linked with PNH clone sizes, persistent or chronic response TCRBs to be identified and assess whether more basic CDR3s are present at the start of the disease, becoming more acidic over-time.

6.4.6.5. AA patient with progressive disease

Patient 0059M was a patient with AA annotated as progressive, with a less than 1% PNH clone. As it was one of the only patients annotated as progressive and active AA, the repertoire was investigated at an individual level to assess the expected T-cell involvement in AA. This patient had an monoclonal TCRB response, with a moderate responsive clone at 14.4%, V29-1, J1-4, 'CSVGSGGTNEKLFF' which was the EBV specific clone identified previously in this analysis. The CDR3 had no hits for origin. TRCBV29-1 usage was very high with most other V family genes not accounting for much of the repertoire. Both J2-3 and J1-4 accounted for over 20% each of the repertoire. Interestingly, this again associated EBV with AA, and this time progressive AA. EBV did appear in normals too, so inferences need to account for this. Tracking this patient over-time would allow for dynamics of the TCRB to be assessed, whether the EBV specific TCRB clone increases with AA progression and possibly PNH clone size.

6.5. Assessing if peripheral blood TCRB repertoires relate to matched bone marrow

PNH is a bone marrow disorder and therefore ideally, bone marrow samples would be sequenced to assess whether TCRB repertoires in the bone marrow change in context of PNH. However, bone marrow samples were difficult to come by and therefore peripheral blood samples were used. For three patients, two had AA with a small PNH clone and one had spontaneously remitted from PNH, matched peripheral blood and bone marrow samples were analysed to see how the TCRB repertoire varied and whether taking blood was a good proxy for conditions in the bone marrow. Tissue resident T-cells in the BM may be responsible for PNH pathogenesis and disruption of the BM [**322**]. These will not be found in peripheral blood, so it was important to see whether the top TCRBs in the BM were also present in the PB, if PB is being used to evaluate PNH pathogenesis.

6.5.1. Patient 0054M, AA with a variable PNH clone

The first AA patient was 0054M, female aged 55, annotated as having a variable PNH clone, possibly falling and neutropenic. There was a peripheral blood and matched bone marrow sample and then another blood sample taken 7 months later. The majority of T-cells in the BM are activated and memory subtypes [**323**]. The bone marrow sample had a very low number of T-cells (14%) resulting in a repertoire that looked very clonal regardless, with 18 out of the 34 TCRB being clonal. The BM sample therefore had a polyclonal response with a top clone being moderate in response at 8.4%, 'CSVPTGVSYNEQFF', V29-1, J2-1. whereas the matched PB was non clonal.

The PB sample 7 months later was monoclonal with a clone TCRBV19/J1-5, 'CASKGGNQPQHF' present as a moderate clone at 6.4%. This CDR3 was previously identified in a study in two individuals and it was annotated as recurrent or public [**254,256**]. Both PB samples had good numbers of T-cells between 78% and 85%. Both peripheral blood samples showed a good number of TCRBs with similar values of around 25,000. However, the BM sample saw much lower values at only 824. Unique TCRBs were very low as stated for BM at 34, the matched peripheral blood had 2266 but 7 months later this had almost halved to 1365. Inverse Simpson generated a similar trend with low diversity in the BM at 28, the highest diversity observed in the matched peripheral blood sample at 511 and medium in 7 months later at 114. D50 values were 12, 254 and 204 respectively, again highlighting the stark difference in diversity between BM and PB potentially linked with the differences in cell numbers.

BM and the matched PB sample showed the same trend in J family usage with J2-1 and J2-3 being the most used. The PB 7 months later showed J1-5 followed by J1-4 as the most common. In the BM V29-1 was the most used gene family followed by V19 at almost ten times less. The matched PB showed a similar trend to the BM but at lower frequencies, about 2-fold lower, most likely attributed to higher TCRB numbers. For the non-time matched PB, V19 was the most common, closely followed by V29-1 at much lower levels, almost 5 times lower than in the BM (**Figure 69.**). BM shared all 34 unique TCRB clonotypes with its time matched PB sample and 25 with the PB sample 7 months later. The two PB samples shared 168 clonotypes. This indicated excellent consistency between the matched samples despite low T-cell numbers and then between the PB samples.

The top ten TCRB in the BM were all present in the time matched PB sample and only one was not present in peripheral blood 7 months later (**Figure 69.**) again showing good consistency. All the top ten TCRBs in the non-time matched PB were present 7 months before in the PB, however, only one TCRB was present in the BM sample.

This TCRB was the top clone in the PB matched sample, present at a similar percentage in the BM as the second top clone and present at 1.1% in the PB 7 months before, highlighting an expansion in the peripheral blood and then contraction 7 months later. This emphasised that PB samples could be used as a proxy for BM taken at the same time.

When tracking the top 10 clonotypes in the time matched PB sample, only 3 were not present in the other PB sample and only one not present in BM. There were no significant differences in CDR3 characteristics between the samples. Although the trends in TCRBs e.g. V/J usage were shared between PB and BM, the TCRB clonotypes were not necessarily present in the samples at the same time, for instance if they were memory T-cells undetected using the methods. In the BM there will be a proportion of tissue resident memory T-cells present that will not be in the PB. However, in this sample all TCRBs in the BM were identified in the PB.

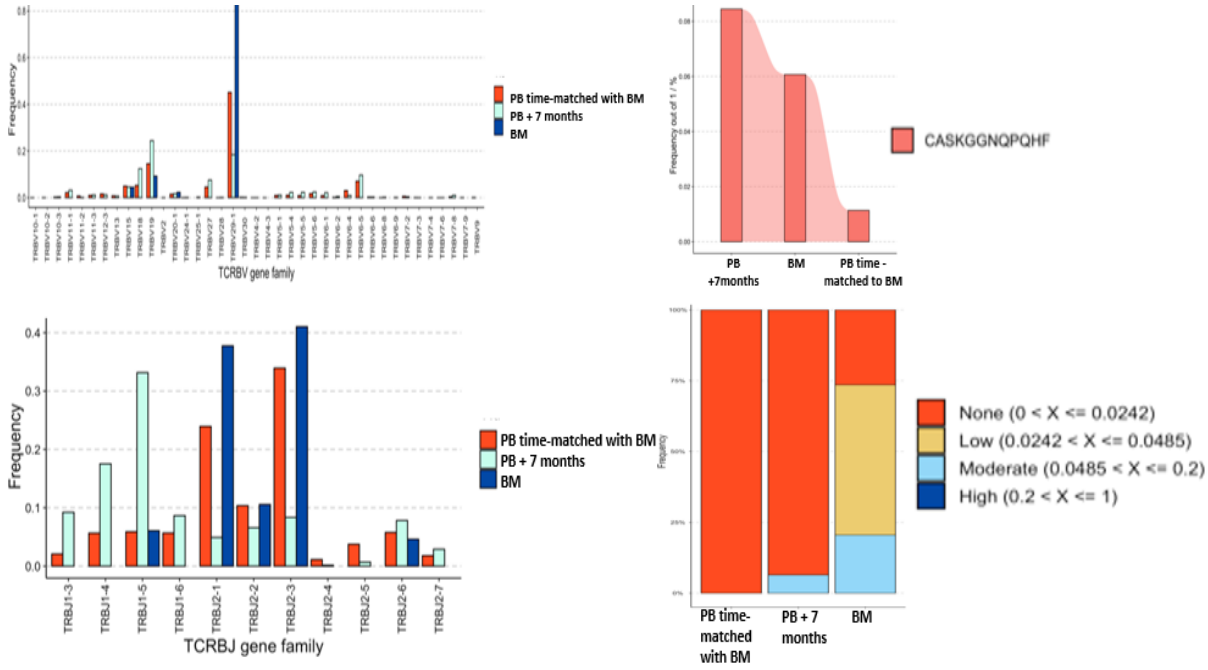


Figure 69. TCRB repertoire analysis of BM and PB blood matched patient 0054M.

TCRBV usage (top left), TCRBJ usage (bottom left) red is the PB and blue time matched BM, turquoise was PB sample taken 7 months later. Clonotype tracking of the top TCRB clone in the PB 7 months later, in the bone marrow and then in the BM time-matched PB (top right) tracked the CDR3 of interest, present in all 3 samples, expanding in the PB over 7 months. Clonal homeostasis in samples (bottom right) PB time matched, PB 7 months later, BM time matched highlights clonal expansion of the moderately responsive TCRB CDR3 of interest in the PB over 7 months. Red is non clonal TCRBs, yellow is low response and blue moderate response based on the parameters outlined in **Chapter 2** and shown on the plot.

6.5.2. Patient 0052F, AA with 10% PNH clone

Patient 0052F was an AA patient with a 10% PNH clone. The patient was 22 and female. The BM had 1045 unique TCRB while the matched PB, 1365 and the PB sample 11 months later had 966. When looking at overall numbers of TCRBs, the BM had considerably lower numbers at 13763 compared to the PB samples that had levels above 22000 as to be expected. Both the BM and PB matched samples were non-clonal with a shared top clone of V15/J2-1, 'CATSRESGGTDEQFF'.

One reason for the BM appearing surprisingly non-clonal was the higher number of T-cells observed in the sample at 42% than in patient 0054M. At the V/J paring level (**Figure 70.**) the time matched BM and PB sample showed similarities with V29-1 being the most common with J2-3 and J2-1. All these findings are good indicators for use of PB representing the BM repertoire. The BM sample had a d50 of 204, the matched PB, 243 and the PB 11 months later had the lowest at 186 indicating that clonality was increasing over time. This trend was true for inverse Simpson with values of 461, 601 and 155 respectively which showed that diversity was decreasing over time. Eleven months later, diversity in the PB was under a third of what it was previously. The BM shared 426 clonotypes with its matched PB sample, and only 91 with PB 11 months later. The reason for the increase in clonality 11 months later in the PB was an monoclonal TCRB response, a moderate responder at 6.8%, V18, J1-4 and a CDR3 amino acid sequence of 'CASSPPLGQGNEKLFF'. V18 and V19 were most commonly paired with J1-4 and J1-5 in the repertoire in this sample. The clonal expansion was present at low frequency in the BM at 0.05% and not at all in the matched PB, suggesting the clone was present in the bone marrow but not in the PB blood previously, but over the 11 months circulated in the blood and became clonally expanded (**Figure 70.**). Another possibility is that it was present in the earlier PB but at non-clonal undetectable levels, either way still expanding over time. It would be interesting to monitor the AA patient and assess whether this TCRB clone persists over time and increases with PNH clone size.

The two PB samples shared 98. All of the top 10 clonotypes in the BM were present in the matched PB and 8 in the later PB sample (**Figure 70.**). This would suggest good concordance between peripheral blood and bone marrow samples. The matched PB had all top 10 clonotypes in the BMs repertoire and all but two in the PB sample 11 months later. The top ten for the PB sample 11 months later were only present twice in both the BM and earlier PB sample. There were no significant differences in CDR3 characteristics. Overall this patient supported the use of matched PB as a proxy for BM TCRBs and identified a potential clone linked to AA or PNH.

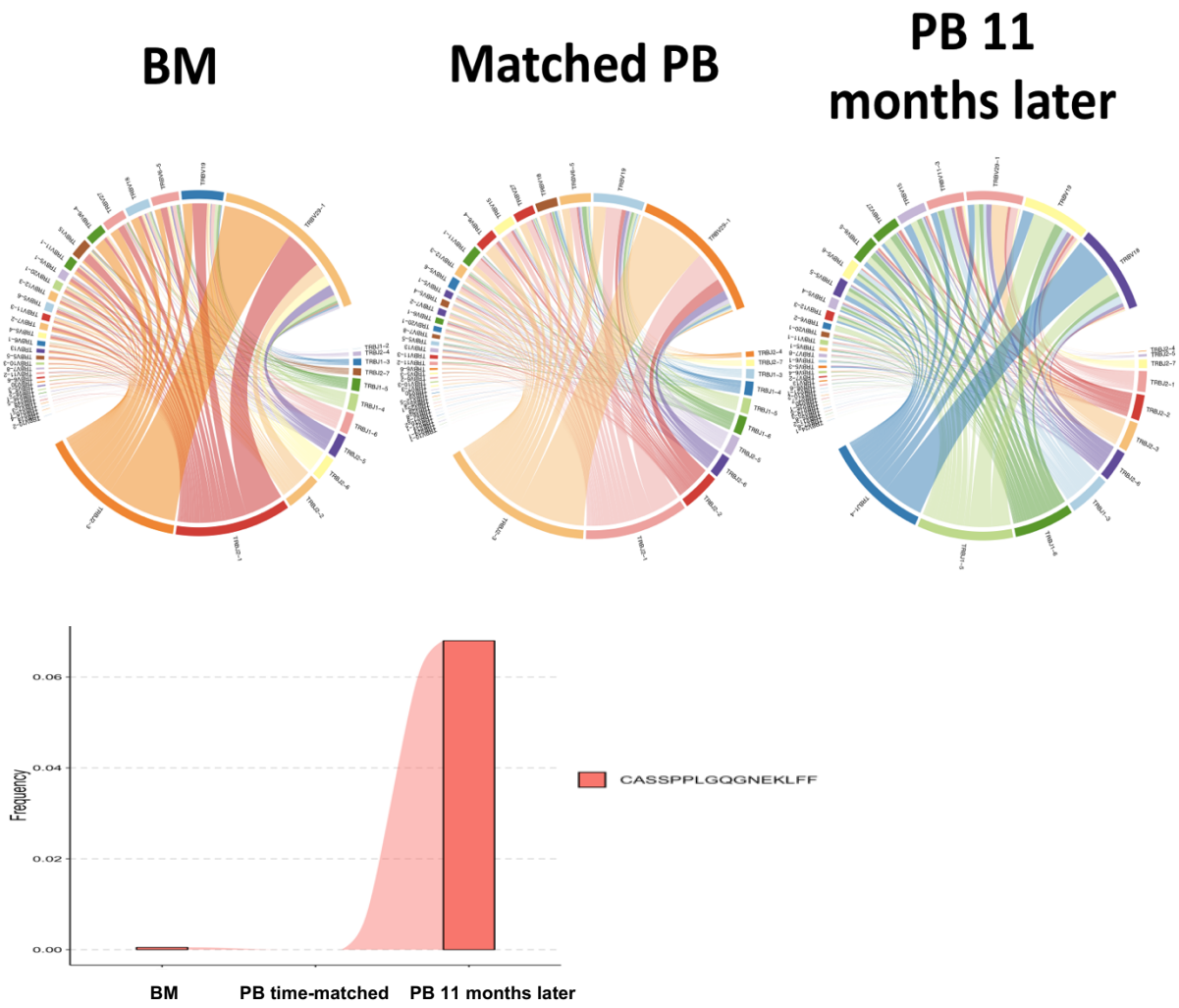


Figure 70. TCRB repertoire analysis over 11 months in patient 0052F

Circos plots of TCRBV/J pairings between BM, PB time matched and PB 11 months later for patient 0052F (top). Bottom graph - tracking of the CDR3 'CASSPPLGQGNEKLFF' overtime in patient 0052F. The patient had AA with a 10% PNH clone. The first sample was the BM, the second the time matched PB and finally a PB sample taken 11 months later.

6.5.3. Spontaneous remission from PNH

PNH is a rare disease and therefore patients that have spontaneously remitted are extremely rare [**394**], with two patients being included in this study. Due to the small numbers of patients at this clinical stage it is often hard to assess whether the mechanisms are individual based on factors such as genetics or a general feature of these patients. Theories have suggested that PNH clones arise in order to re-populate a toxic bone marrow environment, as a "natural gene-therapy" [**230**]. However, whether remission occurs in terms of a healthy, normalising response, recovering from PNH, or whether remission is a transition to other malignancies is not fully understood. One study investigating 6 spontaneous remission patients, found that 4 fully recovered but 2 went on to develop other malignancies [**395**]. This highlights that even in recovery, overall responses of an individual can vary. The diversity of the immune response along with other factors that affect TCR repertoires could be a cause. In order to investigate this further two spontaneous remission patients, one with PB and BM matched TCRB repertoires were analysed. The final PB BM matched sample progressed the TCRB analysis on to patients who had spontaneously recovered from PNH. There were two patients in this category. The first, 004SO, was 39 and female. The patient was thrombotic, had had PNH for over 23 years and at the time of the sample, a 7.34% clone. The metadata for the patient showed that they were taken off Eculizumab 2 months before the sample was taken and had a 10% PNH clone in peripheral blood at the time the bone marrow was taken. The patient also had had a viral infection a month before the sample was taken. The bone marrow was made up of 15% T-cells, 60% were CD4+ and 35% CD8+T cells. CD34+, the marker for HSCs, made up only 0.29% of total cells. The morphology of the BM was regarded as normocellular.

The PB sample had no clonal populations with over 23,000 TCRBs of which 1358 were unique. Likely by the time of sampling, the viral infection had cleared and any clonal TCRBs had contracted to non-clonal levels. The BM sample was polyclonal with 11 TCRBs clonal, one moderate (6%) and the rest low responders. This was expected due to the lower number of T-cells present which resulted in 945 TCRBs of which 44 were unique, suggesting a more technical produced clonality than biological. However, even though it may bias clonality measures, the order of the TCRBs should be as present in the bone marrow. As expected, diversity measures showed the PB sample to be over 100 times greater in diversity compared to the BM with d50 and inverse Simpson values of 455 and 223, to 38 and 16 respectively. Forty-three out of the 44 TCRB clonotypes in the BM were shared between the samples.

Both samples shared V29-1 as their most common usage, in agreement with more 'normal' repertoires, followed by V19 at more than four times less. J2-3 then J2-1 were the J family genes most common for both samples, again indicative of more "normal" repertoires (**Figure 71.**).

Only two of the top ten TCRBs in the PB sample were not present in the BM and all but one of the top 10 were present in PB showing good alignment. PB and BM shared the same CDR3 as top, 'CSVPRGTDTQYF', despite it only being present at 1.5% in the PB. All these findings showed good concordance of PB and BM matched samples and support the use of PB in place of BM.

The patient 004QR had also spontaneously recovered from PNH and was 40 at the time of sampling and female. The patient was diagnosed with both AA and PNH 19 years before the time of this sample. The patient had stopped Eculizumab a month before the sample was taken. Although there were no bone marrow samples available for this patient, diagnostic notes about the bone marrow mentioned that the trephine was 'traumatised' and appeared 'markedly hypocellular' in keeping with the original AA diagnosis. However, there was normal production of blood cells in the BM.

There was no evidence of myelodysplasia and therefore it was thought that the biopsy might not be representative of the entire BM as it appeared to have normal counts. Forty nine percent of the T-cells in the BM were CD4+ and 47% were CD8+ with only 0.38% being CD34+ HSCs. No clonal expansions were observed in the PB, with a lower number of TCRBs than most samples at 5979, with 421 unique TCRBs. An inverse Simpson value of 221 along with a d50 of 82 suggested lower diversity than perhaps expected for a non-clonal TCRB but some stability in the TCRB generated by the diversity. Again, high usage of V29-1 and J2-3/2-1 suggest a more 'normal' TCRB representation. When comparing the two spontaneous remission TCRB repertoires, each shared 190 TCRB clonotypes, which was higher than values of overlap observed between two normals in the **Chapter 4**. Interestingly, all of the 004S0's top 10 TCRB were in 004QR's repertoire and all but one vice versa, unusual for two individuals. This perhaps is indicative of a commonality in response resulting in ultimate remission.

Both patients were female and therefore more prone to autoimmune diseases. Molecular mimicry could be a factor at play. Molecular mimicry is where antigens structurally resemble self-peptides [**324**]. TCRs bind with the antigen and are activated subsequently generating memory T-cells. These T-cells may then go on to recognise self-antigens, thinking that it is the antigen, leading to auto immune responses. Tregs have been identified to raise the threshold needed to trigger autoreactive T-cell responses reducing the risk of autoimmune disease resultant from molecular mimicry [**325**]. Spontaneous remission occurred after many years, and generally PNH patients who had large stable PNH clones had the disease for years too. The non-clonal populations in both suggest recovery but it was important to observe CDR3 properties to try and establish perhaps shared properties of their non-clonal TCRBs. As these could be memory T-cells it may elude to previous immune dysfunction.

These large stable repertoires had higher aliphatic index values than normals potentially linked with Tregs [**300**]. TCRB repertoire findings suggesting re-populations observed in a number of PNH recovering/decreasing PNH clone populations in this project combined with the findings of these spontaneous remission populations, may suggest the re-population is in the Treg subset and that after a time, the Tregs may reduce the effect of the autoimmune TCRB responses resulting in eventual remission. Perhaps the high number of shared TCRB at non-clonal percentages between remission patients are memory T-cells from previous infections that could be an indicator of autoimmune responses attributing to PNH. The expansion of Tregs may have dampened this process down to aid remission. However, this is only two samples and it would be interesting to explore this theory if more samples become available. Longitudinal studies for PNH patients, tracking TCRB changes in time, possibly with fluctuations in PNH clone sizes and diagnosis will help support these claims. Sorting T-cells into subsets such as Tregs and memory T-cells would be important further steps and tracking the patients after recovery to assess whether they develop other malignancies or recover would be essential in deciphering these mechanisms.

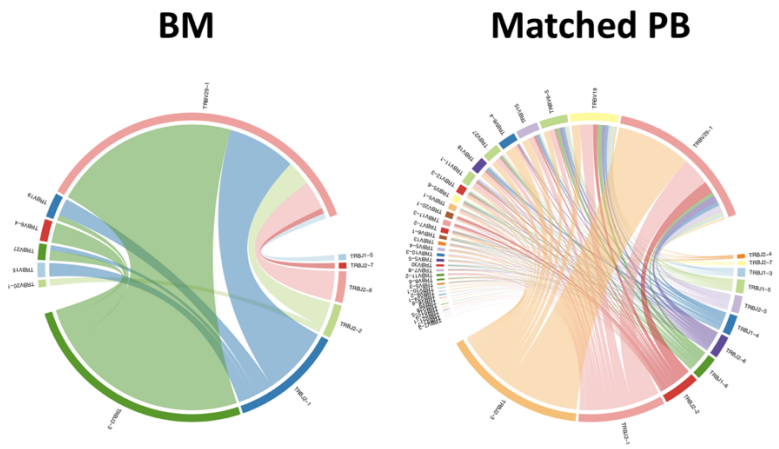


Figure 71. Differences in TCRB repertoires of BM and matched PB samples in a spontaneous remission patient, patient 004SO, who had PNH previously. Circos plots indicating TCRBV/J pairings, width of band indicates abundance in repertoire.

6.6. Experimental bone marrow versus human bone marrow samples

The following data was sequenced from experimental bone marrow samples generated using methods for the experiments that led to the hypothesis of T-cells being involved in PNH. These were carried out by R.Kelly *et al.* [**234**] in the Section of Experimental Haematology**.** These methods involved culturing normal and PNH bone marrows with a supporting stromal cell line and removing the T-cells and then re-introducing them into the experiments in order to determine the effect on the function of HSCs.

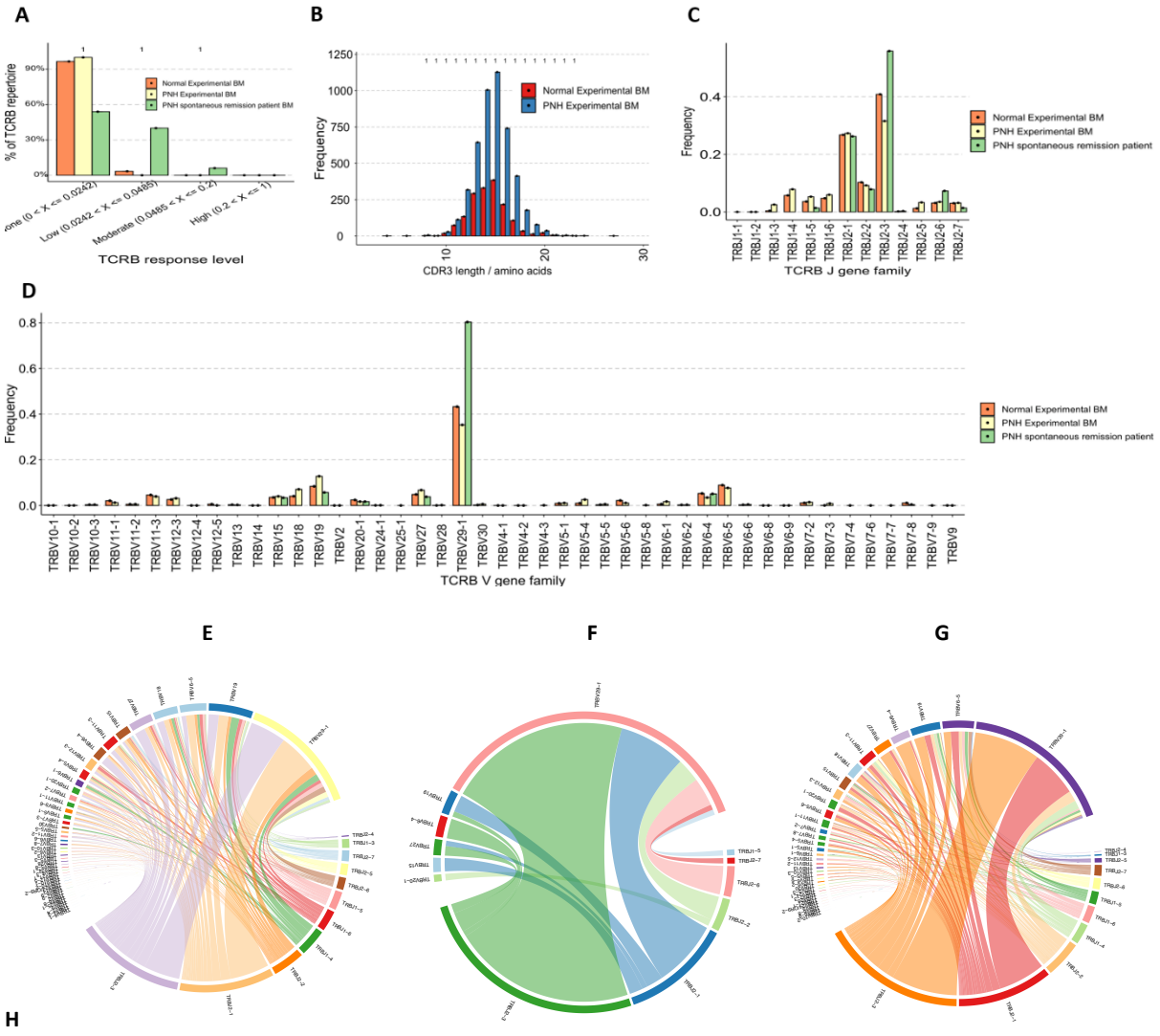
It was unusual for the PNH bone marrow culture system to still have functional T-cells alive in culture after 45 days. The first test was to see whether there were enough T-cells to be able to generate TCRB amplicons and successfully sequence the samples. This was achieved and will allow for future analysis of the T-cells in this culture to be carried out. One PNH BM culture system and one normal were sequenced. These TCRB repertoires were compared with a BM sample from a patient who had undergone spontaneous remission from PNH at the time of sampling to investigate whether there were any similarities or differences observed between the different samples. If similarities were observed between the patient PNH repertoire and the experimental, it could help support future findings using the experimental bone marrow system. Another reason for performing the analysis was to evaluate as to whether there were similarities in the PNH and normal bone marrow cultures. As it was unusual for T-cells to last so long in the culture, it was interesting to investigate whether, over time, the PNH and normal TCRB repertoires from the experimental bone marrows shared similarities. The analysis was also used to assess whether TCRB repertoires could be recreated from these T-cells and if there were any obvious differences being experimental rather than natural.

When comparing the experimental and the spontaneous remission bone marrow sample, it should be noted that the patient sample saw a lower number of TCRB clonotypes than the experimental samples. This was because T-cell percentage data was available for this sample and T-cells were present at only 15%. The experimental samples were unable to be adjusted for T-cell count numbers due to lack of data. Consequently, their samples may have an element of "TCRB noise" attributed to sequencing errors, but this should still only be present at low levels. This may lead to differences in mean values in some of the analysis, due to fewer cells than the experimental numbers. However, using p adj values for measuring significant differences, reduced false positive rates and allowed for distribution around the mean to be taken into account.

The spontaneous remission sample had only 44 unique TCRB clonotypes with 945 clones present in the repertoire. The experimental samples saw much higher numbers of unique TCRB clonotypes at 1643 and 4712 for normal and PNH respectively, and overall TCRB clone numbers of 379188 and 366692. These high numbers were as a result of no T-cell calculation thresholds. The normal experimental BM and the PNH experimental BM shared 371 clonotypes which was much higher than levels of TCRB clonotypes shared between normal peripheral blood samples from **Chapter 4**. Much lower levels of overlap were observed between the PNH spontaneous remission blood sample and the normal and PNH bone marrow experimental samples with 23 and 26 shared TCRB clonotypes respectively.

When assessing TCRBV/J usage and TCRB clonal level responses, there were no significant differences between the three samples. Similar patterns were observed in TCRBV/J gene usage with PNH PB samples, with V29-1 being the most common. However, a slight difference was observed with the second most common TCRBV being present at lower levels than in PNH PB samples, with V6-4, 6-5, 11-3, 12-3, 15,18,19 and 27 having similar levels (**Figure 72. D**). TCRBJ usage showed a similar trend to PNH PB samples with TCRBJ 2-3 being most common followed by 2-1 (**Figure 72. C)**.

CDR3 amino acids trends were also shared with peaks at around 15 amino acids, However, when assessing the mean values, the CDR3 length means were statistically significantly different between all groups (padj<0.05, Holm, Wilcoxon). Mean values were 14.4, 14.75 and 13.5 for the PNH experimental BM, normal experimental bone marrow and spontaneous remission bone marrow respectively. The spontaneous remission TCRB repertoire, therefore, had shorter CDR3s than the experimental bone marrows. When assessing other CDR3 characteristics for each repertoire, significantly statistical differences were observed. Lower GRAVY values were observed in the experimental normal TCRB repertoire compared to the PNH (padj<0.05, Holm, Wilcoxon). The PNH experimental BM repertoire was significantly more basic than the spontaneous remission, but less basic than the normal TCRB repertoire padj<0.05, Holm, Wilcoxon). Consequently, all groups showed significant differences for acidic residues, with spontaneous remission having the most and the normal BM the least (padj<0.05, Holm, Wilcoxon). Again, all categories showed differences in charge, with spontaneous remission being the most negatively charged and the normal the least (padj<0.05, Holm, Wilcoxon). PNH experimental had significantly more aromatic residues than the normal (padj<0.05, Holm, Wilcoxon). No differences were observed between samples for bulk, aliphatic index or polarity (**Figure 72.**). The experimental bone marrows saw very few TCRB clonal expansions. If they were clonal, they were at low response levels (**Figure 72. A**.) The circos plots showed diverse V J pairings with the more common between V29-1 and J2-3, followed by 2-1. However, the spontaneous remission circos plot, indicated a more clonal TCRB repertoire (irrespective of CDR3 with the majority of V/J pairings as V29-1/J2) (**Figure 72.**). This could be differences between patient samples and those generated in a laboratory. Overall there was no obvious changes in the experimentally generated TCRBs to the ones from patient samples. Enough T-cells were present at the end of the LTBMC to be sequenced and recreated into repertoires. This means that future experiments can be designed to sample T-cells from multiple time points during the BM culture experiments before 45 days to add more depth to TCRB analysis in PNH and the changes in TCRB repertoires over these time-points.



BM sample	Acidic / %	Aliphatic	Aromatic / %	Basic / %	Bulk	Charge	Gravy	Length / aa	Polarity
PNH_BM	8.5	0.387	17.2	5.9	13.6	-0.597	-0.208	14.4	8.34
Normal_BM	8	0.389	16.9	6.4	13.6	-0.46	-0.258	14.75	8.36
Spontaneous remission_BM	11	0.378	15.9	3.8	13.6	-1.05	-0.262	13.5	8.43

Figure 72. TCRB repertoire statistics between two experimentally cultured bone marrows, one PNH (yellow and blue) and one normal (orange and red), and a bone marrow sample from a patient who had entered spontaneous remission (green). A) TCRB response levels in TCRB repertoire, B) CDR3 length between the two experimentally cultured bone marrows , C) TCRB J gene usage D) TCBV gene usage, E) TCRBV/J pairings in experimentally generated PNH bone marrow F) TCRBV/J pairings in spontaneous remission sample G) TCRBV/J pairings in experimentally cultured normal bone marrow sample. H) Mean values for CDR3 characteristics for each of the BM sample TCRB repertoires. P adjust values in A and B calculated using Holm method with Wilcoxon paired testing of means.

6.7. Use of UMIs in TCRB repertoire sequencing

Detailed below are the results from extensive development of UMI adapted TCRB methods to assess their use in this study to help counteract amplification biases in the TCRB sequencing data. Low PCR cycle numbers of one to three cycles were used to reduce PCR bias when developing the UMI TCRB sequencing methods in **Chapter 2**. Genomic DNA was lost in the cleaning steps necessary to remove primer dimers before sequencing. The majority of the UMI samples did not have enough TCRB amplicon left to pass the final cleaning stage. Consequently, no or low concentrations of TCRB amplicon were left at the end of the experimental process for inclusion in the sequencing library. During the project, over 100 UMI:TCRB samples were created using both the *BIOMED-2* and *Robins et al. primer methods* using variances in first and second round PCR cycle numbers. UMIs did not alleviate the biases observed in the *Robins et al. primer method*. Five samples using the *BIOMED-2 method* were successfully sequenced (**Table 27.**) These were samples of normals previously successfully sequenced in **Chapter 4** to serve as a comparison.

Table 27. UMI sequencing results using the *BIOMED-2 primer method*.

Details of the number of PCR runs, sequencing reads produced and % GC content.

Sequencing sample number	UMI sample by cycle number for TCRB amplifying PCR	Number of TCRB sequencing reads	GC % content each read
1	1 cycle	722	50 R1 49 R2
2	1 cycle	56388	50 R1 and R2
3	2 cycles	81	51 R1 50 R2
4	3 cycles	84	51 R1 52 R2
5	3 cycles	38	52 R1 51 R2

Table 27. details the number of cycles used in the first TCRB amplifying round and how many reads were generated from the sequencing reads. The lack of reads in samples 1,3,4 and 5 were due to low concentration of TCRB amplicons in the final sequencing library. The samples that had only one PCR TCRB amplifying cycle surprisingly performed the best with 722 sequencing reads and 56,388 respectively. GC content percentages were in normal ranges. However, only sample number 2 (**Table 27.**) passed quality control.

Only 600 sequences passed all the quality control stages and of those, 396 were collapsed into unique reads. 392 of these were used for analysis. Eighty sequences were productive. The top clone was TCRBV29-1/J1-4 which had 7 biological clones and a frequency of 1.9% which was extremely low and could not be used for further analysis. Sample runs 3-5 which had 2 PCR TCRB amplifying cycles or more were poor quality and did not pass quality control. Extensive work was carried out to get UMIs working on both of the TCRB sequencing methods discussed in **Chapter 4**.

A large number of Ns were found in the sequences which did not happen with non-UMI sequences suggesting issues with the sequencing of the UMI TCRB amplicon. UMIs will be fully discussed in **Chapter 7**. As of yet, no methods successfully use UMIs in TCRB repertoire studies using a double ended gDNA approach. A recent paper has suggested that UMIs can actually introduce more bias than the non-UMI techniques [**327**]. This is why methods are moving towards more single cell techniques and why some studies choose to use cDNA, however, as discussed in **Chapter 2**, this was not applicable to this project work.

6.8. Chapter summary

The analysis in this chapter was successful in achieving the chapter aims. By analysing in depth, a selection of PNH and AA patients on an individual basis, it allowed for subtleties of TCRB changes in the repertoire to be identified including in TCRBV and J gene usage which were not as statistically significant when grouping patients by category. Short and long-term patient samples, especially those whose diagnosis changed over time, strengthened links between TCRBs and clone sizes. Cases annotated as “interesting” mostly showed differences in the repertoire to their category group counterparts and normals with unusual TCRB characteristics. BM and PB matched samples proved the reliability of using PB in PNH repertoire studies despite it being a bone marrow disorder and as PB had many benefits over using BM, such as ease of sample taking, this will enable more studies to be carried out in future. LTBMC culture sequencing showed that there were T-cells left in the culture after 45 days and that the profiles were similar in TCRBV/J usage to a patient bone marrow of someone who had recovered from PNH. However, differences were observed at CDR3 property level particular for percentage of basic/acidic residues and net charge. Therefore, this approach will be invaluable in assessing changes in the TCRB repertoire as the LTBMC experiments progress over time.

Finally, although managing to obtain UMI TCRB products at a high enough concentration to be sequenced was a success in itself, despite best efforts to develop a working method, the data after quality control did not provide enough information to continue the use of this method. UMIs will be discussed in **Chapter 7** and how methods for their use could be improved.

6.8.1. Important findings

Some of the main findings in this chapter will now be summarised. Others will be discussed in more detail in **Chapter 7** in the context of multiple chapter findings and wider research in the field.

6.8.1.1 TCRB repertoires as identifiers of clinical status

If TCRB repertoire data could be used as a biomarker for disease, or a predictor for the severity of PNH or potential progression/remission of a patient, it could aid treatment processes. To evaluate this, TCRB repertoires on an individual level were analysed in the context of diagnosis. Changes in TCRB repertoires were markedly different between some patients with different diagnosis.

Differences were observed in overall TCRB repertoire properties in newly diagnosed PNH patients

Analysing TCRB repertoires when patients were in active PNH states can provide insight into T-cell dynamics when pathological PNH processes are occurring. A PNH sample taken a month after diagnosis, when the pathological inducing responses would still be expected to be present and getting progressively worse, had a TCRB repertoire that was more hydrophobic, had more basic residues than PNH patients in the new or increasing group, had the highest aliphatic index, were longer, had a neutral net charge which was unusual, showed differences in polarity and had high J1-4 and J1-5 usage. It was, however, not clonal which would have been expected based on the project hypothesis. It would be interesting to see if these changes were still apparent a year to 2 years after diagnosis, especially once treatment had begun. A few months delay in diagnosis could have changed this response. As PNH is rare, there can be a delay getting to the PNH diagnosis, it is hard to assess when the disease began, and therefore accurately linking TCRB changes at the start of the disease is challenging. The PNH new or increasing clone group contained patients who had been diagnosed or were progressing up to 2 years since diagnosis when TCRB clonal responses may no longer be prevalent.

This could mean that subtleties found on an individual level could be missed when grouping the patients by diagnosis. Three patients who were diagnosed within 7 months of sampling showed a variety of TCRB repertoire responses, which indicates how challenging it is to decipher PNH specific responses, as like immune responses in general the patterns vary considerably from patient to patient. Interestingly in contrast to the main findings in **Section 5.7.2.1**. that PNH patients had more acidic TCRB CDR3s, all of the newly diagnosed patients had more basic CDR3s. With an increase in newly diagnosed PNH patients' repertoires being sequenced and longitudinal studies it would be interesting to assess whether these are anomalies or if in fact different TCRBs are involved with the initial pathogenesis and then subsequent progression of PNH moving from more basic to more acidic over time. Relating the TCRBs to GPI+/- would help study this theory further.

Spontaneous remission repertoires had high sharing of TCRB clonotypes and were non-clonal

The two spontaneous remission from PNH patients both showed non-clonal and relatively stable repertoires. Their profiles appeared to share attributes with the "normal datasets", with V29-1/J2-3 or 2-1 usage. The high diversity and low clonality measures indicated stability in their repertoires. A high number of shared TCRBs, higher than the average between two individuals, was extremely interesting. It could be an indicator of shared responses in the non-clonal memory T-cell subsets from previous infection or perhaps in Tregs involved with autoimmune responses [**327**] either prior, during or post recovery. Tracking PNH patient repertoires from diagnosis would help improve these findings. However, as many of the recovering/decreasing PNH clone patients have had PNH 15+ years, sampling all PNH patients and performing longitudinal analysis in response to PNH diagnosis, will help identify changes in TCRB responses before diagnosis changes.

Recovering patients saw increases in TCRB stability over time

Patients who were recovering from PNH or had decreasing PNH clones also showed stability in their repertoires with more normal repertoire features increasing with time. The majority were non-clonal. Those that had clonal populations, had TCRB responses that decreased over time with diagnosis. For example, patient 004V3 had a PNH clone falling over 6 years from 70 to 50%. Both responses were monoclonal for V15/J1-4 which decreased from 18.3% to 4.15% in line with the shrinking PNH clone which was a major finding for linking TCRBs with PNH clones. Interestingly, the patient's repertoire went from having V15/J1-4 as most highly used to V29/J2-3, which was the most common combination in the normals.

The number of unique TCRBs doubled and diversity increased almost by ten times. There were no significant differences in CDR3 characteristics between the samples indicating stability in recovery. The increase in non-clonal population overtime and in recovery, supports the finding in **Chapter 5** that non-clonal TCRBs occurring 31-100 times in a repertoire were more frequent in PNH than normals. This may mean re-population of the TCRB repertoire aided recovery. It would be interesting to isolate the TCRB clonal population contracting with recovery to see if it was GPI+/-. It would also be interesting to take a sample from this patient again, to see if the PNH clone and TCRB clone continues to fall.

Patient 004UV was non-clonal but also a recovering/decreasing PNH clone patient. The long-term points were both non-clonal with VJ gene usage similar to the normals and no significant differences in CDR3 characteristics. Patient 004UV had diversity measures that remained similar over the time points. Stability was also indicated with the sharing of all top 10 TCRBs between the two samples which was not common for all samples from the same patients. This highlighted that recovering PNH patients or those with falling PNH clones had highly diverse and stable repertoires.

PNH large clone stable TCRB responses varied but were similar to those recovering

By analysing some of the patient repertoires on an individual level, it highlighted how even those with the same diagnosis, the TCRB repertoire could be different. It could also depend on the point that the sample was taken from diagnosis. For example, PNH large clone stable patients tended to have had PNH over 10 years and recovering PNH patients/decreasing PNH clones 15+ years. Depending on when the long-term time points were taken, could depend on what point of change in diagnosis the PNH patient could be seeing. For example, 004VN and 00567 both had PNH with a large stable clone. Both had had PNH over 25+ years each but had slightly different TCRB responses. Patient 004VN was not TCRB clonal, had V29/J2-3 as a common combination and no significant difference in CDR3 characteristics over time inferring stability similar in response to the patient with a decreasing PNH clone, 004UV. Patient 00567 shared a response similar to that of the PNH patient with a falling PNH clone, 004V3. Patient 00567 had an monoclonal response with each of the long-term points sharing the top two clonotypes. The monoclonal response was V29-1/J1-5 at 20.7% decreasing with time to 11.5%. The second TCRB remained non-clonal but was an interesting V family not often observed of V5-4. Patient 00551 shared a similar monoclonal response. This was to an EBV specific peptide that decreased over the years from 18% to 10%, possibly linking EBV with PNH, in a similar way to AA. With the TCRB shrinking over-time, it would be interesting to see if in another 5 years, it continued to shrink and whether the patient started to recover from PNH.

It could be that once the TCRB clone begins to fall, the PNH clone falls and the patient recovers. On the other hand, the PNH clone could fall, leading to the fall in TCRB clones and recovery.

AA active and progressive disease exhibited moderate monoclonal TCRB responses

As AA is known to have clonal CD8+ TCRBs involved in its pathogenesis, and the fact that the majority of AA patients in **Chapter 5** were not clonal, it was important that monoclonality was observed in an AA patient known to be in a progressive stage of the disease. This patient had a moderate TCRB clone at 14.4% and again the TCRB clone was linked to EBV. The newly diagnosed AA patient, again who would be expected to have a clonal TCRB as the disease is active, resulting in symptoms, had a moderate clonal TCRB response at 11.4%. This time the TCRB clone was V6-5/J1-5. In conclusion, this showed that the point at which the patient sample was taken in diagnosis was important when observing TCRB repertoire changes. Newly diagnosed patients were most likely not on treatments such as immunosuppressants which could be as to why expected clonal TCRBs were identified. This is discussed in greater detail in **Chapter 7**.

6.8.1.2. Differences in the LGL patient and EBV infected patient TCRB repertoires may help identify immunological factors involved in PNH progression or pathogenesis

LGL is a condition linked with autoimmunity whereas EBV is an infection [**328-329**]. With men being prone to infections and women more to autoimmunity, it was interesting to assess whether TCRB repertoires in AA or PNH were more in line with LGL or EBV infected repertoires which tended to polarise at opposite ends of the spectrum when investigating CDR3 property characteristics.

In reality the EBV infected repertoire and LGL patient were only one patient each, so statistical tests would not be reliable but still provides valuable insight. LGL is large granular lymphocytic leukaemia, a chronic lymphoproliferative disease. Twenty percent of LGL patients have autoimmune diseases prior to the LGL [**330**]. The PNH patient with LGL in this project would be on immunosuppressants as treatment, so clonal populations of TCRB may be suppressed. A study investigating a PNH patient with LGL who had their TCRB repertoire analysed using a method “Immunoscope” [**331**], found V7-2/ J1-5 to be a dominant clone. This paper is almost 20 years old and methodologies have improved since then. The PNH, LGL repertoire in this thesis was monoclonal with a very low responding TCRB clone, around 2.5%, V29-1. TCRBV7-2 was not prevalent in the repertoire**.** However, the repertoire had a few almost clonal populations at similar levels.

Findings of polyclonal TCRB responses have been reported in other studies too [**321**]. Supporting the suggestion that the LGL patient may have an autoimmune disease, along with PNH, the TCRB repertoire clearly showed differences from normals and other PNH patient categories when looking at CDR3 characteristics. The LGL repertoire had the shortest CDR3s, the bulkiest, most negatively charged at -1, fewer basic residues than normal and the highest percentage of acidic residues of all the categories, indicating dysfunction in TCRBs irrespective of clonality. Interestingly, PNH patients tended to have more acidic residues than normals and the AA with no PNH category, like the LGL repertoire. The majority of the LGL repertoire CDR3s actually had no basic residues which was unusual. Net negative charges of CDR3s have been associated with CD8+ biased TCRB repertoires [**301**]. It has been hypothesised that the origin of LGL is transformed CD8+ T cells expressing a clonal TCRB [**332**]. Although the TCRB clone may have been altered by immunosuppressants, a CD8+ skewed repertoire may be evident. It would be interesting to perform single cell, paired chain sequencing on this sample to assess whether the top 5 TCRB clones at similar abundances were responding to a superantigen for example. The repertoire saw the highest V29-1 usage in the study at over 80% of the entire repertoire. Despite V29-1 being highly expressed in normals, it was never observed at such high levels highlighting abnormalities.

There could also be a distinct difference in TCRB repertoires dependent on the mechanisms of immune dysfunction. LGL in some patients may have autoimmune responses, chronic stimulation of T-cells. The EBV specific T-cell clone found in the patient sample below, could highlight mechanisms involved in chronic infection events. The patient with a persistent EBV specific T-cell clone was found to have the longest CDR3 of all patients, it was midrange for CDR3 bulkiness values and had more basic and lower acidic than normals, in contrast to LGL. The EBV repertoire was almost neutral in net charge unlike all other repertoires which were much more negative for overall net charge. Potentially, CDR3 analysis can identify changes in TCRB repertoires that are associated with immune responses dependent on whether the response is auto immune or is responding to an infection such as EBV. Machine learning approaches that can analyse and identify trends in these datasets would be beneficial in deciphering differences between more autoimmune led responses and infections. Testing data sets that the models could learn from could be patients with identified T-cell exhaustion, chronic infection and then compared with T-cell mediated autoimmune disease patient TCRB repertoires such as Rheumatoid Arthritis [**333**].

6.8.1.3. Multiple time points highlighted sensitivity of the TCRB sequencing method

In a number of ways, using multiple time points helped strengthen the reliability of the TCRB data and highlighted changes that occur over time. The findings that many patients had the same top CDR3 clonotype 6 years later, highlighted that the TCRB primers were correctly amplifying blood samples. It showed that the method is capable of detecting changes in TCRB repertoires. Many patients had clonal populations detected in recent samples that were non-clonal previously in the repertoire indicative of re-infection and vice versa. For example, in patient 004WZ repertoires, both samples shared the top TCRB clonotype. However, it had expanded from non-clonal at 1.7% to clonal at 9.65%. Most likely, the expansion was a memory T-cell subset in response to reinfection and less likely attributed to PNH as the diagnosis did not change overtime.

However, it shows that with multiple time points, more meaningful analysis can be performed, and inferences made about TCRB clonal populations. TCRB repertoires potentially could detect changes in response to treatment. An example of this was patient 004VH who was PNH with a large, stable clone. Two samples were taken 2 weeks apart as the patient moved to the Eculizumab stage of a clinical trial. Both samples were monoclonal but for different TCRB clonotypes. Such a rapid change in response perhaps was not expected over two weeks. Interestingly, the new TCRB clone was traced back to non-clonal percentages in a 2013 sample at 0.8%, suggesting re-infection. This could be also be linked to a change in treatment. It also highlights the capacity of the method to detect progressive immune responses over time.

6.8.1.4. Long term studies identified persistent TCRBs and fluctuations in line with PNH clone size

The problem with using one sample point, is that it gives no indication as to whether clonal populations are shrinking, stable or expanding, which can provide important information about the repertoire. The longer term TCRB studies showed insight into the dynamics of TCRBs in PNH and AA over time and provided great insight. Two patients had changes in diagnosis between 2013 and when the most recent samples were taken. 004V3 was a PNH patient with a falling PNH clone over the four years. In line with this, the same persistent TCRB clone, V15/J1-4, in the monoclonal repertoire remained the only clonal TCRB years later and was falling from 18% to around 4% with the PNH clone.

This was an important find for TCRB repertoires in the context of PNH. It clearly indicates correlations between TCRB and PNH clone sizes. Isolating this TCRB clone, deciphering its T-cell surface markers, tracking its further changes over-time and potentially isolating it for single cell sequencing, can provide invaluable input into T-cells and PNH, especially potential antigens. Patient 00563 was AA with a slow rising PNH clone over 6 years. The patient was male and 64 years old at time of the first sample. Again, a persistent TCRB clone was identified, V29-1, J1-4 and at a hyperexpanded level of above 40% remaining stable across the time points. This TCRB was EBV specific. AA has been linked with EBV infection, so this was an interesting finding [**334**]. The repertoire was responding to a chronic EBV infection and was susceptible to T-cell exhaustion. Phenotyping the clonal TCRB population would help prove this. Markers such as CD44, LY6C and KLRG1 are highly expressed on CD8+ effector T-cells but at low levels in exhausted T-cells [**335**]. These markers could discriminate as to whether this EBV specific TCRB has deteriorated in T-cell function leading to "exhaustion". Multiple time points also helped to identify stable TCRB repertoires that shared a high number of TCRBs and shared top TCRB clones, whether clonal or not and had consistent diversity and clonality values over time. Identifying stable repertoires and comparing with diagnosis served as much as an indicator of TCRBs in PNH as seeing TCRBs shrink with PNH clone sizes. If TCRBs vary with time, but diagnosis does not change, it could be an indicator of infection not related to PNH or AA, further emphasising the importance of longitudinal studies. It could also predict future changes in diagnosis.

6.8.2 Chapter conclusions

Overall, the findings in this chapter helped to strengthen the hypothesis of T-cells in PNH along with strengthening the inferences that could be made from previous chapter results by producing reliable results over multiple time points. This chapter highlighted the need for multiple time points in TCRB repertoire studies to be able to assess the overall stability of the repertoire, and whether any TCRB clones vary with diagnosis, which was found to be true. To improve upon this work it would be useful to analyse multiple time points in the normals data in **Chapter 4**. This would allow for natural variations in TCRB repertoires over-time to be assessed and to investigate clonal expansions and contractions in TCRBs that fluctuate with time. Analysing older normals over time too, would help identify whether EBV specific T-cell clones are clonally expanded and common amongst the normals as well as the AA and PNH patients. In turn, this would help to strengthen findings that are exclusive to PNH and/or AA repertoires. Tracking patient TCRB repertoires over-time, especially those with changes in diagnosis, will improve the comparisons that can be made about TCRB repertoires and PNH in future.

Chapter 7 – Discussion

7.1. Project achievements

In this project a number of areas were developed that allowed main findings to be identified, this section will highlight these before moving on to the main findings of the project.

7.1.1. Collating a PNH/AA cohort based on important clinical parameters

One of the main achievements of this project that allowed for clinical inferences to be made about the patients, was the clear identification of PNH and AA cohorts characterised by clinical parameters and being able to break these down into factors such as sex and age to assess these factors on TCRB repertories. This was achieved by the work and expertise of Dr. S. Richards. By breaking the larger than average cohort of PNH and AA into diagnosis and category groups, it allowed changes in TCRB repertoires to be made in reference to PNH clone size. The large cohort of PNH/AA patients samples allowed for a comprehensive analysis of TCRBs in PNH/AA.

7.1.2. Evaluating and developing successful TCRB sequencing methods

In order to generate the data in these results chapters, a major challenge was developing a TCRB sequencing method from gDNA that produces as accurate as possible recreation of an individual's TCRB repertoire. In order to do this, all steps in the methods were optimised. Experiments were designed to test what concentration of genetic input worked best, to evaluating changes in repertoires attributed to sequencing errors. By evaluating a number of TCRB sequencing methods, with varying reliability in results and investing a considerable amount of time in developing methods, it allowed for reliable TCRB data to be generated. When differences that were not expected in TCRB repertoire results were observed, it also meant that it was unlikely that the discrepancies were attributed to technicalities between samples. By developing a method that could be used on both cDNA and gDNA, it allowed for the historical gDNA samples to be analysed in the context of more recent patient samples. These samples were invaluable in assessing TCRB repertoire changes with PNH and identifying two patients, whose changes in clinical status were linked with changes in TCRB clones (patient 004V3 and patient 00563).

7.1.3. Developing a large panel of normal TCRB repertoires

A key achievement that allowed for PNH and AA specific inferences of TCRB repertoire changes to be made was the decision to collect normal control repertoire samples. Having a cohort of 30, allowed for sufficient trends and variations in normals to be identified in the context of age and sex. For example, the majority of the normals had non-clonal TCRBs but some had clonal populations attributed to infections such as flu. By having a large cohort, it allowed anomalies for example the normal with a large hyperexpanded TCRB to be identified as so. It also highlighted the fact that TCRB repertoires are dynamic and that although differences could be linked to disease, the repertoire contains the immunological history of the patient before the disease as well. By generating a healthy background dataset using the same method as the disease repertoires, it meant that there would be no differences between cohorts linked to biases in sequencing techniques. Comparing clonal TCRBs in PNH and AA with the normal datasets also allowed for inferences to be made about whether the TCRBs could be disease specific or also present in normals.

7.1.4. Developing a robust pipeline capable of processing HTS reads

Another main aim for this project, was developing from scratch, a bioinformatics workflow that was capable of processing large volumes of HTS data whilst at the same time generating accurate, reliable TCRB repertoire data. Each stage was optimised, from deciding which Phred score was optimal for high quality sequencing data, to using algorithms to remove PCR attributed errors. Another considerable issue was deciding which analysis methods to choose that would allow insight into TCRB repertoires in PNH. As not much is known about T-cells and PNH, it was important to not bias the results in terms of looking for clonality for example. Diversity measures were selected to answer different biological questions, from Chao1 looking at unidentified rare clones, to d50 assessing clonality. The pipeline successfully processed over 150 million reads in this project and produced reliable TCRB repertoire data.

7.1.5. Generating an in-house method for determining TCRB clonality

The TCRB clone definition was established in this project having looked at both the options of using nucleotide or amino acid sequence of the CDR3. As large differences were not observed for basic statistics such as unique TCRBs between normals, CDR3 amino acid sequence was used in the definition. This meant that a TCRB clone was defined as TCRB reads that had the same TCRBV/J genes and CDR3 amino acid sequence.

Amino acid sequences were used to reduce biases attributed to convergent recombination events. When reading published studies on TCRB repertoires, it became evident arbitrary values often of 1% and above were used to define a clonal TCRB, however, non-clonal TCRBs could be present above 1% due to factors such as PCR amplification creating artificial clonality. In this study it was decided to use the panel of normal TCRB repertoire data to calculate this value based on the assumption that most normals should not have clonal TCRBs. When comparing the top clones in the normals, distinct groups in response levels were formed. This led to the definition of a clonal TCRB being a TCRB read with the same V/J gene usage, CDR3 amino acid sequence and present at more than 2.42% of an individual's repertoire once the reads were filtered by T-cell counts. This allowed for a stringent approach to clonality to try and assess truly whether PNH had clonal TCRBs or not. Using arbitrary values such as 1% in previous studies could have led to TCRB sequences identified as clonal perhaps wrongly. This is something that wanted to be avoided in this project.

7.2. Main project findings

The main findings from this project will be discussed in the context of current research in the field and any future work that would help improve and further the findings. This project revealed many significant findings related to the development of TCRB sequencing methods, the bioinformatics workflow and generating a background of "normal TCRB repertoire data", but for clarity the main findings will be discussed in more detail in the context of the AA and PNH repertoires, with the main focus on PNH.

7.2.1. Are PNH patients' TCRBs responding to a superantigen?

The main hypothesis of this project work was that there was a single TCRB clone or a series of TCRB clones present in the TCRB repertoire of PNH patients at clonal levels responding to an antigen, potentially a superantigen involved in the pathogenesis of PNH. The main aim of the work in this project was to provide evidence for T-cells in PNH by identification of changes in TCRB repertoires in PNH patients when compared to normals and/or AA. Following on from this the question of clonal TCRBs shared between PNH patients was considered. In conclusion, the hypothesis was rejected. There was no single, or series of shared TCRB, appearing at clonal levels, shared across PNH patients irrespective of disease stage or diagnosis. However, multiple other findings of T-cell involvement in PNH were made.

It is not just a clonal CDR3 that could indicate a shared antigen or superantigen that is responsible for PNH. Disruption of TCRB repertoires were found in PNH patients, and, by analysing TCRB repertoire dynamics in the context of disease stage, allowed some degree of T-cell involvement to be assessed. A summary of the project findings supporting the overall conclusions are detailed below.

7.2.1.1. No specific TCRB was persistently clonal and also associated with all PNH or AA patients

The hypothesis of this work was that a large TCRB clone or series of TCRB clones would be associated with PNH in line with the hypothesis that the presence of T-cells from PNH patients allowed for the expansion of PNH HSCs in a LTBMC by limiting the effectiveness of the normal GPI+ HSCs. When assessing the TCRB repertoires across PNH and AA patients the level of mono and polyclonal TCRB repertoires was also observed in the normals. Research has suggested that in healthy adults, CD8+ clonal expansions are present from youth, can be persistent and expand with age [**336**] despite not causing immediate disease. A recent study by Shi et al. (2020) [**337**] discovered that T-cell clones of "uncertain significance" (TCUS)(term first introduced in this paper) showing phenotypic similarities to those in T-LGL were present in other disease repertoires (diagnosed T-cell malignancies) and a number of normal samples. The finding in this project of non-clonal TCRB clones that occur 31-100 times in the repertoire were higher in PNH than normals also highlighted the difficulty of assessing specific TCRB clones in this study, highlighting the necessity for longitudinal studies. The finding of clonal populations in normals and, combined with the lack of disease associated, publicly available, TCRB data, highlights the difficulty in deciphering TCRB clones specific to disease.

No persistent TCRB clones were found amongst all patients who had a PNH clone. Given the dynamic nature of TCRB repertoires, "public" versus "private" TCRB responses and genetic recombination events, such as insertions and deletions, that go on during T-cell development (**Chapter 1**) it would perhaps be naïve to think a TCRB clone associated with PNH pathogenesis or progression would be identical in all patients or at a persistent clonal level throughout the disease. However, 26 novel TCRB clonal expansions were found in AA or PNH patients when compared with the normals and publicly available datasets. A number of persistent TCRB clones were found to fluctuate in the same direction as PNH clones strengthening the project hypothesis.

Further sorting the TCRB clones into T-cells subtypes would allow more understanding about their importance in the patient TCRB repertoires and whether they are linked to pathogenesis/progression or an infection directly or indirectly linked to the disease. Supporting the reliability of the findings, progressive AA patients, where T-cells are known to be clonal and mediate autoimmune response [**338**], were found to have clonal TCRBs acting as a positive control. However, this could prove challenging for memory T-cells as they tend to be a rare sub-set of T-cells and therefore small in number.

TCR plasticity studies have also shown, building on the previous "Lock and Key theory" that one TCR binds to one specific antigen, that TCRs have levels of plasticity meaning that they can interact with a range of peptides, allowing detection of a wider range of pathogens [**339**]. This all supports the theory, that even though PNH patients did not have a or a series of defined TCRB clone(s), similar TCRs in patients could be present reacting to a superantigen. It was the reason for designing a sophisticated approach in this thesis to look at the total TCRB repertoire at multiple time points to establish at times when the disease is changing, if TCR clones were expanding or contracting.

Karadimitris *et al.* found an invariant, autoreactive, CD1d restricted, TCRValpha21 sequence thought to be involved in PNH pathology [**236**]. The paper suggested it was paired with VB19. Interestingly, this invariant alpha chain was also present in the majority of normals in the study. The method used 454 sequencing and was based on 11 patients. The project work in this thesis also detected a number of TCRB clones in normals and PNH patients. A direct comparison could not be made with Karadimitris *et al.'s* findings as TCR alpha was not tested. VB19 did appear as TCRB clones in some PNH patients, but across normals as well, V19 had the second highest V family gene usage in the repertoire, so the findings are most likely, not significant. This work can add to the initial findings from these studies as considerably more PNH patient TCRB repertoires have been analysed. The TCRB method has been adapted to more high throughput methods which allows greater sequencing depth of the repertoire and estimations of TCRB clonality to be made.

As stated, when considering the hypothesis, fewer clonal TCRBs appeared in PNH repertoires than perhaps expected. This does not necessarily mean that TCRB clonal expansions and PNH HSCs expansions are not correlated. A number of reasons could attribute fewer clonal TCRBs being identified. At a given time-point it is impossible to assess whether a TCRB clone is expanding or shrinking based on these methods. If a T-cell responsible for PNH pathogenesis is in memory phase, with the changes in disease stage, it will be present at levels difficult to detect in one patient sample especially in the periphery.

This is why, when possible, BM matched samples were sequenced, as if the sample was taken when in an active state of disease, especially early in progression, it would be expected to have clonality in the BM. The project findings found that PB was a good proxy for capturing the diversity of the BM repertoire. Clonality was harder to compare, due to BMs having much fewer T-cells in samples which can artificially inflate clonality of TCRBs.

If a PNH clone is falling, the TCR response could have dampened down a couple of months prior to sampling or prior to detection of the falling PNH clone. In support of this idea, Kelly *et al.* reported this theory in 2009 [**340**], on the observation that two patients on Eculizumab treatment for longer than 12 months saw a dramatic decrease in their PNH clone. They thought the decrease was not linked to Eculizumab but more likely, the immune response selecting PNH HSCs, "expired" over time allowing for more normal HSCs to re-populate the BM and resulting in the PNH patient beginning recovery. A cause could also be that T-cells are involved in either pathology or progression rather than both processes, so may not be present in recovering PNH patients/patients with decreasing PNH clones or only present at clonal levels in active PNH cases which is why this was investigated in **Chapter 6**. Perhaps there is an initial TCR response, some form of dysfunction in regulation perhaps, leading to activation of other TCRs. It could be that TCRs that occur in the later response phase of the disease are those common to other autoimmune diseases and have features of autoimmune TCRs such as more hydrophobic residues, due to disruption of the immune response earlier in the disease. This is why, annotating the TCR subsets, combined with longitudinal samples would be beneficial to future studies, deciphering whether there are fluctuations in immune response. In active AA patients, clonal TCRBs were observed. However, clonality varied in response for the newly diagnosed PNH patients. As discussed in **Chapter 6**, delays in initial diagnosis and difficulties in tracking when PNH began could be the limitations for finding clonality in the newly diagnosed patients. If a patient is responding to another infection at a given time this will have a considerable effect on the representation of the TCRB repertoire in the sample. Measures based on the "normal" TCRB repertoires allowed estimates to be put in place for response levels of clonally expanded TCRBs in this project but assessing the trajectory of a TCRB clone is only possible with multiple time points.

7.2.1.2. TCRB clonal expansions were not exclusively linked to diagnosis

Studies into TCR repertoires in PNH are limited and tend to be in animal models [**341**]. A study analysing GPI+ and -ve populations of T-cells focused on CD40 -dependent pathways that are thought to be important in the control of autoreactive T-cell clones [**342**].

However, in this thesis, 34 unique TCRB clonal expansions were identified in 77 AA and PNH patient repertoires. Twenty-six of these were novel and unique to AA or PNH patients at the time of analysis. Only two of the TCRB clonal expansions were found when literature searching or searching TCR databases. 'CASKGGNQPQHF' was found in three other patients in a study so not unique to AA. 'CSVGSGGTNEKLFF' was found with multiple hits linking to EBV and Influenza A. This CDR3 was also one of four TCRB clonal expansions found to be persistent (clonal at multiple time-points for one patient). It appeared in five PNH/AA patients along with 5 normals spanning a range of the patient categories.

Two clonal clonotypes 'CSVETPGTSGRYEQYF' and 'CSVDKAGAGELFF', whose origin is currently unidentified by research, were found in 23 and 24 of the 30 normals respectively. They were however, only found in one AA patient with an increasing PNH clone (n=6) and one AA no PNH patient (n=6) respectively. Monoclonal and polyclonal repertoires spanned the different diagnosis categories and were not restricted to a specific group. However, in the AA increasing PNH clone group, all patients with clonal repertories were male and aged above 65 years old. This most likely means that clonality was attributed to age rather than the disease. Interestingly, the majority of the clonal repertoire patients who had large stable PNH clones had had PNH for longer periods of time (10+ years) than some of the non-clonal (2+ years). Increasing multiple time-point studies in future combined with the advancing technologies and disease associated TCRB specific data available, will mean that more of these TCRB clones will be able to be identified to show whether they are truly AA or PNH specific.

7.2.1.3. Is the definition of TCRB clonality diluting TCRB clonality identification?

Defining the TCRB clonotype in this project

Another possible reason for the lack of TCRB clonality that needs to be considered is how TCRB clonality was defined. This is not exclusive to this project but TCR repertoire research as a whole. It is an extremely fine balance defining TCRB clonality so as not to overrepresent clonality or underrepresent it. However, this project went above the arbitrary values adopted by many TCRB studies of 1%, and defined TCRB clonality based on the proportion of top TCRB clones that were clonal in a normal population. On the assumption that the majority of normals would be non-clonal, but at time of sampling, current infection or other health conditions not noted in metadata, could lead to some TCRB clonality.

The definition of a TCRB clone being above 2.42% of the TCRB repertoire in this study, accounted for clonal variation observed naturally in TCRBs, artificial clones due to amplification TCRB PCRs and detected clonal populations related to factors such as EBV which are known to account for clonal populations in many people's TCRB repertoires [**343**].

It could be a case of some of the TCRB clonality being missed by the calculation, or due to the definition including a CDR3 amino acid sequence, defining clonality in an antigen specific manner. This is why analysis was also carried out irrespective of CDR3 (circos plots) and similarities in CDR3 amino acid properties were assessed rather than simply searching for an identical CDR3 amino acid sequence. The calculation makes the best of current methodologies and approaches to clonality currently available.

Does IMGT® have an exclusive list of TCR genes?

Another possibility for the lack of clonality could be successfully annotating the TCRB reads using IMGT®. Notably, a number of reads were discarded as they had no TCRBJ gene in IMGT®. One reason for this could be due to sequencing errors, this was a consistent percentage irrespective of normals, disease or sequencing run. Another possibility is that in a similar way to B cells, not all TCR genes or genetic variations in the sequences are detailed in IMGT®. Unlike B cells, TCRs do not undergo somatic mutation and variation is limited. However, genetic variation in the genes could lead to non-alignment of the TCR reads to IMGT®. Diversity sections in this study were also not included in the definition of TCRB clonality which is in line with many TCRB studies and should not have a great effect on results.

Another reason for not using diversity, was that in many clones, it was not annotated from IMGT® and the reads might be discarded unnecessarily if used in the definition. Again, this highlights that not all data for TCRs perhaps is in IMGT®, the gold standard for TCR repertoire analysis. This point was also highlighted during discussions at the AIRR 2019 conference [**344**] amongst experts in the field. Therefore, annotating the TCR genes when assembling TCRB clones could be a limitation for detecting some degree of TCRB clonality, the majority were identified using these methods.

7.2.1.4 TCRB CDR3 amino acid properties in PNH patients differed from normals

As the project progressed it became apparent a single, identical TCRB clone or series of shared TCRB were not present in and exclusive to all PNH patients. The finding that actually, TCRB clones present at non-clonal levels, but more prevalent in the repertoire (occurring 31-100 times) in PNH than normals, suggests there could be TCR specific responses, but dependent on clinical stage, may be memory T-cells rather than activated T-cells. PNH patients could be responding to a superantigen or similar form of an antigen to one another. However, the TCR responses have evolved to be different between individuals, attributed to genetics such as HLA [**345**]. These mechanisms also help prevent one pathogen killing an entire population.

Diversity in TCRB responses are advantageous in populations [**346**]. This process in some respect could be supported with PNH. It is an extremely rare disease, and if it is linked to an antigen, most likely the antigen infects healthy controls too, but the immune responses do not go on to result in PNH progression. To aid the hypothesis that PNH patients' TCRs could still be responding to a similar antigen but with varied TCRB responses, CDR3 properties were looked at. Irrespective of whether the repertoire was clonal PNH patients had significant differences to normals and the AA and AA no PNH groups.

As discussed in more detail in **Chapter 5**, PNH repertoires had CDR3s with more acidic residues, more negative residues and higher aliphatic indexes than normals and were shorter than normals and AA patients. Potentially indicative of self-reactive T-cells and increases in CD8+ populations. Interestingly, AA and AA no PNH could be distinguished from PNH based on some CDR3 characteristics. CDR3 populations linked to PNH in the AA patients could be attributed to the higher acidic, more negative CDR3s identified than in AA no PNH patients. These were characteristics that significantly differed from AA with no PNH patients, indicative of PNH specific responses. Interestingly the converse was found in the newly diagnosed PNH patients who had more basic residues than the PNH group as a whole, in line with CDR3s in AA. This suggested that different TCRBs could be involved or present at the start of PNH and during the progression or as the disease stabilisers or recovers. CDR3 properties were found to change with PNH activity, and response to disease. For example, recovering patients/patients with falling PNH clones had the shortest CDR3s and more negative CDR3s all indicative of antigen skewing, autoimmunity mechanisms and CD8+ populations.

Perhaps suggesting a PNH patient's recovery involves other autoimmune mechanisms, it would be interesting to see if these patients had other comorbidities, some could be age related as recovering PNH/falling PNH clones were older than average. Active PNH CDR3s tended to have fewer hydrophobic residues, which highlights how CDR3 property characteristics could in future, indicate PNH stage and severity. These results were discussed in more depth in **Chapter 6**. Machine learning techniques to analyse these trends would be beneficial in designing biomarkers.

7.2.1.5. CDR3s with more acidic and more negative residues in PNH could provide insight into antigen response

With limitations in this study for identifying a possible antigen due to factors such as only analysing TCRB chains and a lack of publicly available disease associated TCRBs, CDR3 characteristics could be used provide insight. TCRB CDR3s account for the majority of variation observed in a TCRB repertoire and are the portion that interact with the antigen. Therefore, any similarities in CDR3 properties in PNH patients not found in normals and AAs could indicate properties of the antigen they bind. The CDR3s in PNH tended to have more acidic and more negative residues. Acidic residues are aspartic acid and glutamic acids. CDR3s and peptides bind through forces such as hydrogen bonding, electrostatic interactions and Van der Waal forces [**296**]. Acidic, negative residues will tend to interact with more basic, positive peptides. Studies into the TCR CDR3s in relation to amino acid properties at present, are not extensive. However, one study provided some insight into potential immune responses. TCRs interact with APCs to either become activated in the case of CD8+ or to activate B cells for antibody production in the case of follicular helper T-cells (Tfh) by contributing to co-stimulatory signals [**347**]. Increase in Tfh have been noted in some autoimmune disease. This subset was not mentioned as they are relatively newly discovered, and studies are limited. But perhaps, the T-cell populations in PNH are less pathogenic and more regulatory, trying to re-balance the immune response.

The study found that antibodies produced by activated B cells with CDRs that were net positively charged, had a higher risk of low specificity for antigens, meaning that rather than binding to a specific antigen, it could bind to many off-targets, potentially reacting to self-peptides. The presence of arginine in the CDRs were implicated as a risk factor in non-specific interaction as it is larger and can form electrostatic interactions and H bonding with the peptides. The B-cell receptor tends to be almost identical in structure to the secreted antibodies to combat the antigen that activated the B-cell [**348**].

In PNH the TCRs, appear to have more negative residues, which would mean the antigen could have more positive residues. In turn the antibody produced in response to the antigen, would be more negative like the TCRs and more specific in relation to this study. There could be an autoimmune element related to the B-cells. In multiple sclerosis one event that can lead to pathogenesis is the escape of pathogenic B-cells from T-cell mediated control. If T-cells are altered by lack of GPI-AP, this could affect B-cell:T-cell interactions which could lead to immune dysfunction via B-cells [**349**].

Arginine would still act the same way in a TCR CDR3. Therefore, perhaps PNH having TCRB CDR3s with more negatively charged residues relates to increased specificity. It could be, combined with findings of re-population and increased diversity in recovering PNH patients/patients with falling PNH clones, that this is a mechanism in PNH designed to try and re-balance the immune system. It would be interesting to isolate clonal populations into T-cell subsets such as Tregs to assess the mechanisms and whether these populations change with recovery, to more balancing immune responses.

In conclusion, although no specific TCRBs were found exclusively in all PNH patients at clonal levels, the project provided sufficient evidence that there were changes in the TCRB repertoire occurring in response to PNH and its clinical status. Further work described in the next section will help improve the research findings.

7.3. Factors and immune mechanisms potentially linked to TCRB changes in PNH

In this section, multiple important findings from the previous chapter will be discussed assessing potential factors and links with immune mechanisms that could be involved in PNH.

7.3.1. PNH showed no sex bias

PNH as discussed in **Chapter 1** is partly caused by the somatic PIG-A mutation in HSCs, which is X-linked. PIG-A undergoes X inactivation in female somatic cells, which means both males and females only need one mutation in *PIG-A* to deplete its function [**350**].

Therefore, any variances in sex ratios between patient categories is most likely down to smaller subsets of PNH patients attributed to the rarity of the disease. However, in terms of autoimmune responses and the potential links with PNH, it was important to take into account sex when looking at TCRB responses and TCRB clonality. Research has suggested that females are more prone to autoimmune attack than males, due to evolutionary advantages for women having strong immune systems especially when pregnant [**351**]. Delving deeper into the differences, some papers have suggested that females have more CD8+ dominant repertoires.

AA involves an autoimmune mediated destruction of the bone marrow, however, if anything, male: female ratios were relatively balanced in the cohort. In contrast, one group, AA patients with no PNH clone, albeit a small dataset (n=6) had a ratio skewed towards males at 1:5. Like in the AA cohort, PNH does not see a skew towards females in its datasets. However, this does not rule out that the immune system is involved in the pathogenesis or progression of PNH.

7.3.2. A case for immune-ageing in PNH

Although no significant differences were observed between factors of TCRBs such as clonality, diversity, TCR numbers in regard to age, differences in age were still a point of interest in PNH patients.

7.3.2.1. Recovering PNH patients were older than the average PNH patient

Patients recovering from PNH/decreasing PNH clones were on average almost ten years older than the average PNH patient. This was linked with the fact that the average length the PNH patients had had PNH before recovering/seeing a decrease in PNH clone levels was 13.5 years so was not surprising. Immuno-ageing is associated with numerous negative processes, such as inflammation as cells enter immunosenescence. It is not just a response that affects TCRs. On the whole the immune system response deteriorates, becomes non-specific and activated by mediators such as stress [**352-353**].

CD4+CD25++FoxP3+ Treg cells are essential for mediating immune responses that occur between self and non-self-peptides [**354**]. With immune-ageing, if these subsets become “exhausted” over time and are less specific in response, they may not be able to keep the immune system homeostasis in balance. This could lead to autoimmunity and pathology. This could be helping PNH progress, or perhaps the disruption lessens the molecular mechanisms in PNH leading to recovery from PNH, but development of Treg dysfunctional, autoimmune disease. Sorting the T-cells according to their subsets by cell markers would be interesting. Assessing how many comorbidities the recovering PNH patients/those with falling PNH clones had, such as type II diabetes, that are linked to the immune ageing process, or whether in the case of PNH patients, immune ageing benefits recovery could also be beneficial. For example, changes in the immune system due to ageing may help reduce negative immune responses associated with PNH.

7.3.2.2. Recovering PNH patients' ages correspond with ages where thymic involution rates increase

All of the recovering PNH patients/those with falling PNH clones were 39 years old or above including both spontaneous remission patients who had had the disease for over 18 years each. As discussed previously, studies have shown that with age the human thymus, responsible for developing new TCRBs and for educating naïve T-cells, involutes. This process increases significantly beyond the age of 40 [**355**], the naïve T-cell pool shrinks considerably, diversity decreases and clonality in general increases with ongoing infections of CMV and EBV for example.

Another contributor to the shrinking of the TCRB repertoire is HSCs. They deviate from the lymphoid to the myeloid lineage with the ageing process, leading to a decrease in B and T-cells. Interestingly, CD4+ cells are maintained at a stable level with homeostatic proliferation (**Section 1.4.5**) whereas CD8+ T-cells shrink with age. Larger increases of homeostatic proliferation are required to maintain the CD8+ repertoire. This may relate to their differences in functions with CD8+ T-cells fighting infection and killing infected cells, whereas CD4+ T-cells elicit helper functions (**Section 1.2**). Both however, have been implicated in autoimmune diseases. Many autoimmune diseases have an MHC class II restriction such as in Rheumatoid Arthritis which would implicate CD4+ T-cells [**356**]. Potentially, the shrinking of the CD8+ pool may have been linked to the increased age of the recovering PNH patients.

7.3.3. TCRB repertoires likely to be affected by medical treatments, diagnosis, stage of disease, HLA and *PIG-A* mutations (PNH patients)

7.3.3.1. AA TCRB clonal responses differed according to treatment and stage of disease

To reiterate, in this work AA was used as a positive control in this study as it was known to be T-cell mediated. Autoreactive, cytotoxic, CD8+ T cell clonal populations have been identified in AA [**357**].

One study found that effector memory CD8+CD57+T-cell, TCRBV oligo clonal expansions in AA were a frequent occurrence in peripheral blood [**358**]. Oligoclonality in this paper was defined differently to the project based on spectratyping and deep sequencing. Samples were oligoclonal with 1-3 immunodominant TCRBs, whereas in this project, if there was only one clonal TCRB (greater than 2.42% of repertoire) it was defined as monoclonal if there was more than one clonal TCRB it was described as polyclonal.

To note, in that study, at the time of sampling, none of the patients had received treatment, therefore, these clonal populations may have been present due to this. This agrees with this project's findings detailed below. Treatment could reduce the appearance of TCRB clonal expansions in AA.

AA patients would be on a number of immunosuppressant treatments, such as cyclosporin, lowering immune responses, unless the sample was taken when newly diagnosed, before treatment. Depending on the dosage level, high levels inhibit T-cell activation, whereas low can lead to the production of pro-inflammatory cytokines, autoimmunity and hyperreactive immune responses [**359**]. It would be interesting to analyse the AA repertoires in the context of dosage to assess whether there are differences in TCR dynamics. Higher dosages could mean T-cell populations would be supressed to much lower levels or prevent T-cell activation and further TCR clonal expansions, which could change TCR dynamics observed in a repertoire when looking for clonality. Low levels could lead to unbalanced immune responses, heightened T-cell activation and perhaps the decrease of Tregs that would help re-balance the immune system.

T-cell clonal expansions attribute to AA progression, but effective treatments would seek to minimise this and therefore, some treated patients may not have the clonal TCRBs anymore. An example in this project was an AA patient who had had a bone marrow transplant prior to sampling and now had no clonal TCRBs in the PB on sampling. A lack of AA pathogenic clonal TCRBs associated with AA progression would be expected in the bone marrow, as the autoimmune TCRs, for the time being, would be eradicated or at very low levels due to the bone marrow treatment, but it was interesting to see this effect in the PB too.It would be interesting to perform a longitudinal study on this patient, with PB and BM matched samples, to assess whether pathogenic TCRBs re-populate the BM and whether these are observed in the periphery. However, an AA patient with progressive, active disease at time of sampling, where T-cells would be expected to be attacking the bone marrow, had a TCRB repertoire that was monoclonal with a moderate 14% TCRB clone specific to EBV, V29-1/J1-4, 'CSVGSGGTNEKLFF'. The PNH granulocyte was less than 1%.

7.3.3.2. Changes in treatment could affect PNH patient TCRB repertoires

Acute infection and vaccination can affect TCRB repertoire clonality both in the short term in initial response and long term. Naïve T-cells are activated by an antigen, for instance in the vaccine, which leads to the differentiation of these naïve T-cells to effector T-cells over the course of 1-2 weeks [**359**].

The majority of these cells will die off once the infection has cleared apart from a small subset that will become memory T-cells.

A suspected example of this was patient 004VH whose repertoire significantly differed over a sampling period of just two weeks. The patient had a large stable PNH clone. The first sample was non-clonal for TCRBs but two weeks later there was a clonal population of V15/J2-3. PNH patients are not thought to be more prone to infection. Metadata showed that the patient was transitioning on to the Eculizumab stage of a clinical trial.

However, patients about to be given Eculizumab, are sometimes given a meningococcal vaccine, dependent on risk factors, prior to this as they are more susceptible to meningitis once taking Eculizumab. This is because Eculizumab prevents the formation of terminal complement. Terminal complement is needed to prevent *Neisserial* infection, but not most other infections [**360**]. Potentially the rapid change in TCRB repertoire is as a result of this. It could also be that the patient was more prone to infection in general since starting Eculizumab as an anomaly finding rather than the normal response in PNH patients. It would be interesting to see if this TCRB clone has subsequently contracted again in line with predicted T-cell dynamics. The TCRB clone could be in response to the flu, a flu vaccine or the common cold. It highlights the dynamic nature of TCRB repertoires and how quickly they can change. Interestingly, the clonal TCRB found in the 2017 sample was not present two weeks before but was present in the patient sample in 2013 but at very low levels. In reality, the TCRB would be present in the repertoire two weeks before if present in 2013, but at memory T-cell levels (below 1.2% which would mean it was not always represented in the sequenced sample).

However, the rapid increase does suggest infection or reaction to a vaccine. For future work, the samples could be sorted prior to sequencing into effector/memory/naïve T-cell subsets to help decipher the types of mechanisms occurring. For instance, if the TCR clone is a Treg it may highlight trying to re-balance immune responses. CD8+ responses would be indicative of infection. It would be interesting to take samples before and after the vaccine had been administered, to assess the effect on the TCR repertoire as B-cells responding to the vaccine become activated. Changes could be analysed in the context of being above or below 40 years old, relating to the thymic involution process. It could be interesting to assess whether the changes are in the GPI+ or GPI- fraction of T-cells.

Although Eculizumab is not involved in the TCRB process and currently not thought to affect the adaptive immune response, for future reference, the majority of patients were on Eculizumab at time of sampling apart from the two spontaneous remission from PNH patients who had been taken off prior to sampling. Their repertoires were not clonal for TCRBs (all TCRBs <2.42%), supporting the hypothesis of T-cell involvement in the pathogenesis and/or progression of PNH as PNH is now absent in these patients. However, interestingly, they shared a higher than average number of TCRB clonotypes between individuals in this study. The number was actually the highest comparison in the study, whether the immune system was responding in similar way to recover from PNH would need further investigation. Perhaps remission occurred in this sub-set of patients as they could be defined by TCRs and potentially a specific antigen because of the high number of shared TCRs. Tracking these patients after recovery would help assess the effect of coming off Eculizumab, if any, on TCRB repertoires. Hopefully, more patients in future will recover from PNH, allowing larger numbers of them to be assessed and trends in the repertoires identified.

7.3.3.3. Delayed diagnosis could attribute to lack of clonal TCRBs detected

PNH itself is easily diagnosed by GPI- neutrophils by flow cytometry. However, one of the challenges with PNH is realising the symptoms are as a result of the disease. PNH is rare and often not the first thought when a patient presents symptoms such as fatigue. By the time the patient has been diagnosed and samples taken, the initial pathogenesis will have set in which could be another reason for large clonal T-cell populations not being observed in all PNH new or increasing patients.

Tissue resident T-cell subsets and PNH

BM matched PB samples showed that TCRB clones present in the BM may not be present in the blood at the same time or below the limits of detection by this method. PB samples were taken 7 months later than these matched samples detecting TCRB clones previously found in the BM but not the matched PB. This does not mean that the TCRB clones are not present until 7 months later, but multiple short-term time points are essential to ensure that changes in TCRBs in the BM are captured in the PB blood. This could also be a case of tissue resident T-cells in the BM attributing to PNH pathogenesis or progression and not those TCRBs circulating in the PB as PB are only thought to account for 1-2% of the T-cells in the human body, the majority are found in the gut which is not relevant to this study. As PNH affects the bone marrow, isolating tissue resident T-cells found in the BM could be interesting for future work.

Matched BM and PB did show degrees of similarity increasing with more time points (**Section 6.5**) so even if responses are mainly in tissue resident subsets, some degree of TCR repertoire change would be expected to be picked up in the PB.

A study suggested that tissue resident T-cells needed GPI anchored proteins in order to home to tissues. The example studied was a peptidase inhibitor 16 protein that was preferentially expressed by skin homing CD8+ T-cells specifically [**361**]. Some T-cell subsets in PNH would not have these GPI anchored homing molecules, therefore, would be in the blood, rather than homing to tissues. This would mean GPI- T-cells would be sequenced from blood samples as an assessment of PNH progression. GPI +ve T-cells involved in PNH pathogenesis would still express the molecules and could home to tissues and so may not be present in peripheral blood. In relation to the BM, it would be interesting to investigate as to whether tissue resident memory T-cells had the GPI- or + phenotype.

Perhaps the ratios of GPI+/- are different between the periphery and the bone marrow. Equally assessing how these GPI+/- subsets in PB and BM changed in response to vaccine, for instance, would help understand how new immune responses are affected in the context of PNH and whether PNH affects the differentiation of naïve T-cells to memory cells.

7.3.3.4. Different mutations in *PIG-A* genes could result in different TCR responses

Investigations into variations in *PIG* mutations could explain differences in TCR responses

In PNH, different types of *PIG* mutation can have an effect on PNH pathogenesis. For instance, the degree of sensitivity that PNH red blood cells have to complement mediated lysis causing PNH depends on the mutation. Type III cells have complete GPI-AP deficiency and are extremely sensitive to complement mediated attack whereas type II cells only have partial deficiency and are less sensitive to attack [**362**]. Partial deficiency causing some TCRs to attack GPI is a current hypothesis in the field. However, it is not known whether it is against GPI itself or a differentially expressed surface protein or a configuration change that is enforced by the absence of GPI. This could aid understanding as to why some T-cells responses differ between patients. Clonal expansions of HSCs may be polyclonal due to multiple types of acquired mutations in the HSCs, for example frameshift. PNH granulocyte clones and TCRB clones were correlated in patient 004V3 in this project. Perhaps if there are multiple PNH clones, there are multiple clonal TCRBs in the repertoire. This was one reason for looking at any shared characteristics within the individual's TCRB repertoire regarding CDR3 for example, irrespective of clonality and comparing these to "normal" TCRB repertories.

It would be interesting to assess differences in GPI anchored proteins in different T-cell subsets in PNH repertoires to see whether differences in TCR repertoires follow this trend.

Differences in *PIG* mutations affect pathogenesis of PNH

Other rarer GPI biosynthetic pathway mutations have been found in PNH patients; however, *PIG-A* remains the most common and others are extremely rare. Pathological germline mutations have never been detected in *PIG-A* mutated PNH patients as they are thought to be lethal [**363**]. The mutation is acquired. Whereas mutations in *PIG-M*, *PIG-V* and *PIG-N* have been detected in germline. One patient has had *PIG-T* mutations identified [**364**]. *PIG-A* mutations affect the beginning of GPI anchor biosynthesis and mutations are therefore extremely disruptive, whereas *PIG-T* is involved at the end of the pathway, therefore biosynthesis is not disrupted entirely [**365**].

Mutations occurring in other *PIG* genes rather than *PIG-A* generate different phenotypes in patients, for instance, a *PIG-T* mutated patient developed other immune related disorders such as irritable bowel syndrome at a similar time, which is not unique. *PIG-M* mutations have been linked to symptoms such as thrombosis but not haemolytic anaemia [**366**]. This suggests different *PIG* mutations affect PNH symptoms and PNH severity, attributed to the stage at which the mutation affects the GPI synthesis pathway. In future, PNH patients' *PIG* mutations could be included in the metadata although extremely rare. This would allow inferences to be made between differences in TCRB responses potentially in response to differences in PNH severity caused by differences in *PIG* mutations and how this could vary TCRB repertoire responses. In reality, there are fewer mutations in the other *PIG* genes, but it may be worth investigating the heterogeneity.

7.3.4. A case for autoimmunity in PNH

7.3.4.1. Shorter CDR3s associated with autoimmune disease found in PNH and AA TCRB repertoires

Shorter TCRB CDR3s contain fewer insertions and have been found to be highly enriched in thymic selection. Skewed TCRB repertoires towards shorter lengths have shown to be linked with autoimmune disease (self-reactive) and can also be involved in antigen driven selection [**91**].

AA is known to have an autoimmune mechanism whereas in PNH it is hypothesised. By analysing the CDR3 lengths it allowed for comparisons between other autoimmune TCRB repertoires to be made.

Interestingly, AA and PNH patients had significantly shorter CDR3s in their repertoires than normals. PNH CDR3s were even shorter than in AA. Firstly, this indicates that different T-cells could be involved in PNH than AA, responding to a different antigen. Secondly, this characteristic indicates that the CDR3s could be self-reactive despite clonal populations to a specific antigen not being identified or shared between PNH patients.

7.3.4.2. Inverse CD4:CD8 ratios in GPI- T-cell subsets

The other finding from this work that might support a case for autoimmunity was the flow cytometry data. Two of the patients who were younger than average at 26 and 35 showed interesting flow profiles with inverse CD4:CD8 ratios in the GPI- T-cell subsets. Both had more CD8+ T-cells than expected and were selected for either having had PNH a long time or a large PNH clone. These PNH patients were not recovering and were younger than the average age of recovery, potentially when their immune systems were stronger and more potent. In general, most PNH patients will not fully recover [**194**].

7.3.4.3. EBV specific TCRB clones in AA and PNH patients could link with autoimmunity

In the progressive AA patient, the monoclonal TCRB response was EBV specific. Other EBV specific TCRBs may have been present but were not currently in any publicly available datasets. In the long-term AA patient with a rising PNH clone, the monoclonal response was EBV, persistently above 40% over 6 years. This TCRB clone has previously been identified as CD8+ specific to the antigen BMLF-1 which has been linked to multiple sclerosis [**292**]. EBV has previously been shown to be linked with AA [**334**]. Patient 00551 and 005D7 were both PNH large clone stable with EBV specific clonal expansions. 005D7 was highlighted as an interesting case because the cells did not haemolyse like expected. This EBV specific TCRB clone was only found in five of the normals, which was surprising as 90% of the population are thought to have been infected by EBV. It could be that the TCRB clone is in a memory state and not detectable by the current methods. As a person ages, the repertoire will become more clonal and potentially dominated by the likes of EBV responding TCRBs. As these clones are detected in normals too, it is hard to assess whether the EBV infection is linked to pathogenesis or progression of AA or PNH.

Many studies have looked into EBV and T-cell responses, which could also be why databases contain EBV specific/linked TCRBs and a limitation of these databases would sometimes be a lack of comparison with normal datasets.

However, in this case the EBV specific TCRB seemed sufficiently characterised by experimental data. The other clonal TCRBs identified in this study had no known origins, they may all be linked to a specific comorbidity, but until more research is performed it cannot be assessed.

Chronic, persistent infection with EBV could also lead to a process known as "T-cell exhaustion" which leads to loss of effector functions [**335**]. It would be interesting to assess changes in transcriptional and epigenetic activity in the patients with and without EBV infection to assess factors such as "T-cell exhaustion" on the TCR repertoires. RNA-seq and flow cytometry carried out by Hosokowa *et al.* has looked into T-cell transcriptomes in PNH patients [**367**]. Interestingly, they found that expression of CD69, TNFSF8 and PD-1 were higher than in normals. PD-1 was particularly highly expressed on effector T-cells indicative of persistent chronic infection. There were two caveats to the study. The first being the expression levels were not defined in activated T-cells so not indicative of TCRB clones and no GPI- T-cells subsets were sequenced due to low frequencies. Considerably more research has gone into the role of T-cells in AA than in PNH. A recent study using single cell and transcriptomic analysis suggested Hsp70 proteins, were important in AA pathogenesis [**368**]. Future work into T-cell Transcriptomes in PNH patients could reveal novel immune pathways that cannot be determined by TCR repertoire analysis. GPI+ versus GPI- T-cells may differentially express molecules compared to each other. Antibodies could be used to surface label cells prior to sequencing. This could highlight whether PNH or AA patients may be suffering from processes such as "T-cell exhaustion" either causing or as a product of AA and PNH.

Another potential link was the finding of the EBV responsive CD8+ TCRBs at clonal response levels in AA patients with no PNH, AA patients with small PNH clones and PNH patients with large stable clones. It was also found in 5 of the normals but not necessarily at clonal levels.

A strong, sometimes persistent, chronic response to EBV (one of the AA patients had the substantial clone at 46% being persistent over 6-year time samples) could suggest a hyperactive CD8+ population leading to an increase PNH clone and delayed recovery in some PNH patients. However, as EBV infection is common with age and appeared in patients without PNH and normals, the process in not specific to PNH.

7.3.5. GPI and T-cell signalling

GPI negative HSCs allow for the progression of PNH. Subsets of T-cells in PNH patients will also be GPI negative as PNH progresses highlighted in the flow cytometry data. These GPI- populations amongst the 8 patients ranged from 1.44 -26.7%. Understanding how GPI and GPI linked proteins interact or are involved with T-cell processes and combining this with the project results may aid understanding of the role of T-cells in PNH. T-cells that are not present at the start of the disease may be involved in its progression. The mechanisms as to whether T-cells are directly, indirectly involved or are clonal as a result of PNH pathogenesis or progression are unknown. Some ideas of mechanisms will be discussed here in the context of future project work.

GPI anchored proteins have been found to activate T-cells [**369**]. When GPI anchored proteins cross link, it promotes kinase activity. GPI anchored proteins such as Thy-1 [**370**] are associated with tyrosine kinases in the src family mentioned in **Section 1.1.** essential for T-cell activity. Potentially, loss of some of these proteins on the surface of some of the T-cells in a patient can lead to variances of T-cell activation. This results in a lack of new clonal TCRBs. Newly generated T-cells may not be able to activate to the same degree as the older GPI+ T-cells.

As GPI- T -cells only made up around a quarter of the T-cell population in the 8 patients identified by Dr S. Richards, the immune system would still have some older, GPI+ activated T-cells. However, if subsets of naïve T-cells that are GPI- are unable to be activated once PNH pathogenesis sets in, this could mean that PNH patients have lower capabilities of generating activated and consequently differentiated memory T-cells. These processes could affect the balance of subsets such as Tregs and Th17 which have been linked to AA as mentioned in **Chapter 1.** Decreased Tregs and increases in Th17 could indicate autoimmune responses. Phenotyping the subsets would help identify if there are Tregs or Th17s present and their ratios, potentially indicative of immune stability [**228-229**].

GPI associated proteins such as CD59 located in cell membranes, have a number of other important roles in membrane trafficking, endocytosis, immune signalling and T-cell activation [**371**]. GPI-anchor protein cross linking of GPI associated proteins is important for formation of actin patches where proteins that are phosphorylated by tyrosine kinases are located. Formation of these actin patches aids the formation of immunological synapse.

In order for T-cells to become activated by a peptide via an MHC molecule, the APC and T-cell need to have a stable contact in which signalling can occur, this is the role of the immunological synapse. Potentially, if some vital GPI associated proteins are missing, some T-cells may fail to form immunological synapses and thus fail to be activated [**372**]. Alternatively, populations of acidic TCRBs identified in PNH patient repertoires, could interact with part of an antigen abundant in basic residues. If parts of the GPI anchor structure on the surface of normal HSCs are abundant in basic residues, they may be targeted by these acidic TCRB subsets. Taking parts of the GPI anchor structures and assessing whether they complement the CDR3s of the acidic TCR structures could help investigate this using 3D modelling techniques. Further work could also involve investigating these mechanisms in LTBMC and implementing TCR phenotyping in the GPI +/- subsets and in the clonal TCRBs identified in this project to understand more about GPI associated proteins and the lack of them in TCR interactions with HSCs.

7.4. Limitations of TCRB repertoire studies

Many of the limitations of this project are also limitations in the TCR research field as a whole and are detailed in **Table 28**. In this section, the project limitations and measures taken to address them will be discussed with references to improve future project work.

Table 28. Comparison of topics that need developing to further research in this project along with areas that need expansion in the TCR repertoire research field as a whole.

Topic	Future work	
	Project work	TCR repertoire research field
BM studies	Single cell sequence populations in bone marrow.	Studies carrying out single cell BM
CD4/CD8 responses	Use FACS to split T-cell subsets or algorithms	Algorithms available for this
Disease vs normals	Develop machine learning techniques to find trends in TCRB repertoires	Machine learning techniques under development
HLA	HLA modelling to help epitope studies	Algorithms and modelling developed
Infection detection	Be able to identify TCRB related to infections versus PNH or AA	Standardised databases for TCRB identified in infections and disease, e.g McPAS-TCR, VDJdb
Long term studies	Multiple time point studies for RTB, assess spontaneous remission patients	Increase in long term studies with ageing, same patients, clonal dynamic data
Metadata	Generate basic metadata for normals	Linking medical records with TCRB repertoire samples
	Link TCRB changes with changes in treatment for patients	
Modelling TCRBs	Collaborate on modelling clonal TCRBs to understand antigen specificity	Topic is developing e.g TCRex
Normal repertoires	RTB collection of normals above 57 years old	Central database for data generated from studies using normals
	RTB collection of more normals who are male	Standardised and normalised normals data
Paired TCRA/TCRB	Move towards paired TCRB sequencing	Provide standard for paired TCR
Single cell	Single cell sequence clonal populations idnetified	Provide standard for single cell sequence
UMIs	Develop TCRB working with UMIs	Acquire UMI gold standard e.g Euroclonality

7.4.1. Capturing TCR diversity and clonality representative of the entire TCRB repertoire

TCRB studies usually prioritise searching for clonality or diversity in the TCRB repertoire. Computational modelling and mathematical algorithms [**106**] are gaining pace in the immune repertoire research field to estimate diversity as it is not possible to detect an individual's TCRB repertoire from a single blood sample, even when carrying out exhaustive TCRB sequencing.

Clonality was the priority when designing the methods as it was hypothesised that a single TCRB clone or series of TCRB clones would be involved in PNH. However, it was important to be able to reliably and accurately detect clonal TCRBs and to some extent overall diversity in the blood sample itself, as a representation of a portion of the patient's TCRB repertoire. Factors that had the potential to skew the diversity of the sample repertoire were biases attributed to PCR amplification, sequence errors, sample types and concentration of genetic material in samples.

7.4.1.1. TCRB clones present at percentages below 1.2% varied between biological replicates

An interesting finding from using replicate samples used to test PCR and sequencing biases was that when TCRB clonotypes were present at percentages below 1.2%, they could vary from replicate to replicate, with variations in diversity. They were not always picked up in replicate samples. As the rarefaction plots in normals showed, in general the TCRB sequencing methods identified the majority of the TCRBs in that particular sample. In a repertoire, the majority of TCRBs would be non-clonal as shown in the normal dataset in **Chapter 4** and would, perhaps, compete for available TCRB amplifying primers in PCR reactions. Clonal TCRBs were always detected at similar levels in replicates and, generally, samples split at the index stage to investigate sequencing bias shared the top non-clonal TCRB clonotype too and had similar values for diversity and unique TCRB number. Again, this supported the hypothesis that it is at the PCR amplification stage, rather than sequencing methods, that biases in diversity could occur.

7.4.1.2. Genetic input of 200ng per TCRB PCR provided enough material to accurately sequence TCRB repertoires

Using 200ng, approximately 30,000 T-cells, of genetic input into the TCRB amplifying PCR based on the results from **Chapter 4**, showed that 200ng was enough gDNA to be able to pick up clonality in a TCRB repertoire, but also a realistic amount to be extracted from patient samples that have very little DNA.

It was important to keep all sample inputs as close to 200ng as possible across normals and patients to reduce biases associated with genetic material concentration. Genomic DNA was chosen over common cDNA methods as only one TCRB per cell would be productive and would therefore capture accurate TCRB diversity. TCRBs from complementary DNA made from mRNA would vary per cell, depending on the expression of the TCRB at the given time and would therefore vary according to activation state, potentially skewing the repertoire representation.

7.4.1.3. Buffy coat samples did not capture diversity as well as Lymphoprep®

Buffy coat samples were only used when no other sample type was available. This was because comparisons from the same blood sample, buffy coat versus Lymphoprep ® preparation, highlighted discrepancies in diversity attributed to the sample preparation method used. Buffy coat samples contained around half the number of T-cells than the standard method.

This was because in the Lymphoprep® method, granulocytes are removed with density centrifugation allowing more T-cells to be inputted into the PCR. This was taken into account when calculating T-cell inputs. It would be expected that due to fewer TCRBs being present in the sample before amplification, repertoires might appear more clonal than expected. However, this was the opposite. In matched samples, buffy coat clonal TCRBs were generally lower in percentage than in standard preparations. This could be because fewer TCRBs are captured at the sampling stage prior to PCR. The TCRB pool is much smaller and, therefore, diversity and clonality captured is less than in the standard method. In future, buffy coat samples should perhaps be avoided completely. Ideally T-cells would have been purified for maximum counts but unfortunately this is not a standard process in clinical laboratories. Another factor for not capturing as high diversity in a sample is unproductive TCRB reads that were amplified by the PCR primers. Around 75% of the reads were productive, with 25% of the reads being unproductive and removed during the bioinformatics analysis. Currently, there is no way to reduce this.

7.4.1.4. Clonal TCRB repertoires are the best indicators for assessing potential sequencing biases between sequence runs

The final aspect of the project that was used to ensure accurate diversity and clonality measures, was having a sample that was sequenced in every sequencing library acting as a comparison and an indicator if something was wrong in the TCRB sequencing library reaction. Both a mixed donor gDNA sample and a clonal patient TCRB repertoire were used.

Interestingly, the clonal patient TCRB repertoire was the best marker for variation, due to the stability of amplifying clonal TCRBs. The Promega® commercial mixed donor gDNA sample was not clonal and all TCRBs were present at levels below 1.2%, therefore, TCRBs that came up as top clonotype would differ each time regardless of sequencing run. Fortunately, no stark contrasts were found in the replicates of samples run across the sequencing runs used in this project to serve as controls in **Chapter 4**.

In future, where applicable, as this is difficult and potentially wasteful with rare patient samples, running technical and biological replicates alongside each sample would allow accurate conclusions to be drawn about capturing the diversity and clonality of a sample.

7.4.1.5. Problems associated with developing a UMI TCRB sequencing method

Extensive work throughout this project went into developing a better TCRB sequencing method. The idea of UMIs is that each biological TCRB amplicon has the addition of a unique UMI by, ideally, a single round of PCR. These sequences undergo subsequent PCR amplification rounds to generate large enough concentrations for sequencing. After sequencing, these reads can in theory be collapsed according to their UMI. Reads with the same UMI, V/J genes and CDR3 amino acid sequences are annotated as technical replicates as they are duplicated and collapsed to give the true biological representation of the repertoire rather than one amplified by PCR. Even though diversity was not the primary question here, developing a method using UMIs would have been the ultimate aim to ensure sequencing and other PCR associated biases were kept to a minimum. Other methods put in place in this project have helped reduce this as much as possible and hopefully, PCR amplifies all the reads at similar rates, resulting in proportional amplification, so a top clone would still appear at the highest frequency. However, in future, a UMI method should be incorporated. In this section problems that may have contributed to the failure of the UMI sequencing reads passing quality control filtering will be discussed.

UMI TCRB library samples were unstable compared to non-UMI counterparts

The main areas of concern discovered whilst working with UMIs was their stability, tag swapping, their properties, such as GC content and potential secondary structure formation. The UMI sample (**Section 6.7.**) that reached the TCRB sequencing stage had a significant number of reads produced by the sample, but the majority did not pass quality control checks. A number of observations were made that may have contributed to this happening. When working with UMIs the final TCRB library was much more unstable than non-UMI libraries. Within a day in the fridge the library would degrade substantially. This created a problem. Adding a 6 random base pair sequence into each TCRBV/J primer – adapter sequence changed the overall stability drastically. Tape station images of the TCRB sequencing product were taken before putting in the fridge and then a day later, and the concentrations had dropped by about 10-fold (data not shown in results).

Biases in the UMI tags incorporated into the TCRB amplified reads

As discussed in **Section 6.7.**, the GC contents of the overall TCRB sequencing reads saw similar ranges to those without UMIs. However, this is averaged across the read as a whole and 12bp is small enough not to skew the result for a 250bp read. It would be interesting in future to analyse the UMI tags from successful TCRB sequencing runs and those from TCRB populations that are clonal to assess that there are no biases or differences in GC content that could have an effect on the result. Although the methods were developed to reduce GC bias in the samples, for example no heat was applied when extracting TCRB amplicons from the agarose gel as this can create GC bias, biases in the UMIs were likely to occur. Interestingly, sequences such as the UMIs that are not neutral, are much harder to incorporate into TCRB amplicons. The UMIs are randomly generated in order to create enough combinations to cover sample diversity. In each TCRBV or J primer aliquot, all random iterations of 6bp were present. Until the reads were sequenced, UMI combinations could not be identified. It would be interesting to assess the overall charge of these tags and identify whether other UMIs were not incorporated due to charge. Although all TCRBV/J primer aliquots should have contained the same number of the random UMI tags, properties such as charge could affect incorporation into the commercially produced primer aliquot. In turn this could introduce biases into the PCR unintentionally and without knowing. A targeted UMI approach may be more successful.

Using randomly generated tags although having benefits made the bioinformatic analysis challenging. Parameters had to be decided such as the 'hamming distance' [**373**] which was decided as 1, meaning across the 12 bp UMI for each TCRB read, only 1 base was allowed to be different per UMI group. In reality it is extremely difficult to assess whether the original UMI sequence incorporated pre-sequencing was sequenced correctly. Illumina® error rates are a serious consideration for this work. TCRB clones with the same UMI (technical replicate) allowing for one mismatch were collapsed to form an accurate TCRB repertoire not artificially amplified by PCR. Mathematical and computational methods would have to be applied to assess likelihood of sequencing errors occurring, especially as the UMIs are at the start of a sequence read. Illumina® sequencing tends to encounter sequence errors towards the start of the sequencing read [**374**]. The hamming distance could have been increased to allow for more groupings between UMIs that were the same but differed due to sequencing errors, but again, it would be difficult to assess between sequence errors causing variation, PCR variation and actual biological variation. The general diversity and TCRB numbers for the TCR repertoires was on average 18,514 for the patient samples in this study.

A number capable of capturing sample diversity, for example, thirty thousand, known UMI tags (to account for sequences that did not pass the strict quality control filters) could be generated and used in the experiments, rather than completely random ones. Perhaps knowing the input beforehand could help with analysis and calculations although would encounter other caveats.

The final issue attributed to UMI biases was the phenomenon known as 'tag swapping'. The phenomenon has been mainly studied in Illumina® technology that uses patterned flow cells, such as the HiSeq ®. However, the mechanisms potentially still apply when using MiSeq®. Research is in its infancy and not much is known about the consequences of this to sequencing libraries. In general, tag swapping occurs when reads are mislabelled during sequencing and appear in different multiplexed sequencing samples. For instance, one TCRB clone may occur in patient 1 but due to tag swapping, the clone appears in patient 2 in the same sequence library. This could be a problem for multiplexed libraries as in this study where many samples are collected together for sequencing. MiSeq® can sequence up to 96 different samples at time. Recent estimates suggest that TCRB read swapping could occur in around 2.5% of reads [**375**]. This phenomenon could occur in non-UMI libraries as it is generally associated with the Illumina® adapters. However, it adds additional complexity to UMI analysis trying to differentiate between biological and technical TCRBs along with discriminating between different patients. UMI analysis makes assumptions for UMI occurrences and groupings based on an individual's repertoire. If samples between patients could overlap due to sequencing methods, this would cause inaccuracies.

Potential formation of secondary structures in UMI sequences

Although when testing the UMI libraries, sequence lengths were as expected, potentially changes in TCRB sequence structure, such as secondary loops, formed because of the UMI tags added. The majority of the reads in the UMI samples that did not generate successful TCRB reads contained mainly Ns. This did not occur at noticeable levels in non-UMI TCRB reads. Ns are produced when the sequencer is unable to call the nucleotide base as the signal is ambiguous. It can also be caused if the expected spacing between nucleotides is not as expected for example if one nucleotide is inserted too close to the next and could be indicative of odd structures forming. Differences in TCRB structures could alter interactions with adapters, enzymes and potentially binding to flow cells reducing sequencing potential. To combat this, in future, spacer regions could be incorporated and tested in the UMI -primer – adapter TCRB sequences to prevent secondary structure formations. As the UMIs are random, until the TCRB is sequenced, it is not known which of the possible combinations is present. Each UMI, potentially can form 2 amino acids.

Certain UMI combinations may, therefore, have certain properties that decrease their likelihood of being incorporated into the sequence, for instance if they are not neutral as discussed. This can also relate back to secondary structure formation. Creating a method to assess this in future would be beneficial.

UMIs on both ends of TCRB reads are challenging to analyse

One of the issues developing the method were that UMIs needed to be on the ends of each TCRBV and J gene, one at each end of the sequence in order to assess biases in both primer sets. The majority of current methodologies use RNA-seq which only needs the UMI on one end [**376**]. Many methods also use a smaller pool of known UMIs for barcoding. Since performing the experiments in this project, a number of companies and research collaborations are developing methods for UMIs for use with gDNA and double ended primers, therefore, the field is moving in the right direction to combat issues outlined above.

UMI technologies being developed in the TCRB sequencing field

The majority of UMI technologies are tailored for RNA-seq type data. As discovered through working on a collaborative project looking at T-cell rearrangements in RNA-seq (data not shown) the T-cell output is not large enough to be able to perform an accurate TCR repertoire analysis due to low abundance of TCR mRNA. The use of RNA and cDNA adds to the complexity of expression levels of certain TCRs linked to activation status rather than clonal expansions. This is why the need for UMIs is much greater for RNA-seq data than gDNA. However, it would be worth developing an effective UMI method for use with gDNA building on the work in this project, to help reduce bias introduced through PCR and sequencing, especially if the project wanted to focus on diversity rather than clonality in future.

Whilst carrying out the research for this project, the number of technologies using UMIs and the capabilities of these technologies have rapidly expanded. Using the technologies in future work would enable a number of the fundamental questions discussed above to be researched further.

A number of commercial kits have recently come into production which could be incorporated. The first is the NEBNext Direct® kit [**377**] which uses hairpin loops to incorporate a 12bp UMI to the 5’ end of the target gene. However, this method has a couple of caveats.

It incorporates the UMI into only one end of the TCRB gene which would not infer PCR bias for one set of either TCRBV or J primers in the TCRB amplification reactions and another possible issue is the processes used to incorporate the UMI. Various temperatures and buffers need to be used which could interfere with the TCR structures, cause degradation of some samples, or incorporate GC biases associated with heating steps. However, it does highlight how technology is improving in regard to UMI usage. Chung *et al.* (2019) [**378**] constructed an evaluation of a number of commercial UMI kits which showed variability between products and set out interesting benchmark data to be considered when using UMIs in research. Euroclonality [**379**] the consortium behind the development of BIOMED-2, as of 2020, have developed a UMI method which they recommend using. This project would look to build on this work in future.

7.5. Ideas for further work

Some ideas for future work have already been discussed in the context of project results. In this section a number of additional ideas will be introduced.

7.5.1. Improved methods for assessing trends between TCRB repertoires in PNH

One of the difficulties with this project was being able to display the analysed data clearly. There is such a breadth of analysis that can be performed on TCR repertoires that sometimes it can be difficult to decide what methods to use. Additional analysis methods in this project could investigate diversity regions of the CDR3. This project mainly focussed on CDR3 amino acid sequences to reduce convergent recombination biases but perhaps this approach was too antigen specific when assessing clonality. Consequently, it was important to look at TCRBV/J pairings irrespective of the CDR3. This allowed findings of extremely high usages of particular V and J genes in non-clonal repertoires for instance the PNH patient with LGL and showed the need for variations in analysis in order to assess trends in the TCRB repertoires. Running alongside this, the more samples available, the better the gathering of accurate information, but the more challenging and complex comparison analysis can become. Averaging repertoires results per category, for instance, although helpful with making comparisons and inferring conclusions about PNH repertoires compared to normals, could nullify significant findings on a more individual basis. As stated previously, TCRB repertoires are diverse, dynamic and even if two people are exposed to the same antigens, their responses and TCRB profiles can be very different. The ideal would be to analyse all repertoires on an individual basis. Potentially, this is where machine learning developments could be implemented.

Models could be trained to look for subtler differences and agreements between a large volume of repertoires, taking into account ages, sex and potentially previous treatments, where current bioinformatics analysis may fall short. There has also been an increase in mathematical studies and modelling on TCRB data. Collaboration with these research groups could provide interesting insight into PNH progression.

7.5.2. Extracting cell populations, new technologies and algorithms

7.5.2.1. Paired chain and single cell sequencing to assess potential antigen targets

In future, it would be interesting to select the clonal TCRBs identified in this study from their cell populations and use advancing technologies such as the paired TCR sequencing (alpha and beta paired chains) or single – cell sequencing [**380,381**], to gain knowledge of the alpha chain that pairs with these clonal populations. Paired TCRB sequencing would generate data that would allow 3D protein modelling to be used to assess possible antigens that these clonal expansions are responding to. Prediction tools such as TCRex [**382**] could also be used to predict sequence epitope specificity. These antigens may have similarities between PNH patients, or it would help identify clonal expansions in response to infections such as CMV or the common cold rather than a chronic disease like PNH.

7.5.2.2. HLA typing, identifying CD4+ or CD8+ specific TCRB responses, T-cell subset sorting and single cell sequencing

Identifying CD4+ or CD8+ specific TCRB responses

Although some TCRB repertoire measurements, for instance net charge can provide some insight into whether TCRB responses are CD4+ or CD8+ dominant, splitting the original samples into CD4+ and CD8+ T-cells before sequencing would decipher whether responses in PNH are CD8+ dominated. Splitting into subgroups and identifying any CD1d restricted T-cells would be further project work. One benefit to splitting and phenotyping cells into T-cells subsets is that it allows changes in these populations to be identified over time and in response to disease state. This is something that TCRB sequencing results in this project cannot identify. For example, Kordasti *et al.* [**229**] found that regulatory T-cells (Tregs) were reduced in function and number in AA. PNH could have similar mechanisms and this provides a stronger case for phenotyping T-cells in the PNH samples where possible**.** A possible technicality with this process would be that for some samples such as the BM samples, there are very few T-cells present. Splitting them further, could lead to a loss of gDNA and inaccurate results for diversity representation.

As the use of algorithms and machine learning increases perhaps they could be used in combination with experimental data. Different HLAs present peptides in different manners which could also be interesting to investigate in the context of TCR repertoires[**383**].

HLA-typing to better understand TCRB repertoires of PNH patients with LOH

HLA typing of patients in the RTB tissue bank would also help to aid this level of research as discussed. The interesting patient with a loss of heterozygosity on the short arm of chromosome 6 from 6p25.3 to p21.32 is a good example as to why HLA typing would be beneficial. Recent studies have identified acquired copy number neutral loss of heterozygosity on the short arm of chromosome 6 (6p CN-LOH). Previously, LOH in chromosome 6 was only linked with AA . This LOH results in a loss of allele from one of the parent HLA haplotypes and was thought to help HSC clones escape auto immune responses and toxic conditions caused by AA progression. The PNH patient with LOH in this study, in 2010, was diagnosed with AA but at the time of sampling it was just PNH. The TCRB response was monoclonal, low responding and V19-1/J2-1. The overall CDR3 repertoire was shorter than normals. It had CDR3s with fewer basic residues than PNH new or increasing clones which was its diagnosis category, but values were in line with normals. The repertoire had generally more negative CDR3s indicative of CD8+ responses. The repertoire appeared quite normal in terms of TCRBV and J gene usage. To see whether the TCRB clonal expansion was affected by the LOH, it would be interesting to see whether it was GPI+ or –ve and which HLA type. 6pCN-LOH clones have been found at lower levels in peripheral blood than bone marrow which may be true if a BM matched sample was available. This case highlights why HLA typing would be interesting to test in PNH patients especially those with similar responses to see whether HLA may be causing variance in responses [**384-385**].

T-cell subset sorting

T-cell subsets can be distinguished from one another based on the expression of cell surface markers. For example, central memory T-cells express CD45RO, CCR7 and CD62L. Whereas effector memory T-cells express CD45RO, but not CCR7. Sorting T-cells into subsets prior to sequencing using biomarkers could provide insight into whether populations such as Tregs, or tissue resident memory T-cells are particular involved in PNH. It could also highlight as to whether these populations change over time. The caveat is a decrease in T-cell numbers being inputted into each PCR after separation. Some patients may have very low populations of a particular T-cell subset that may not amplify but could be important in PNH.

Single cell TCR sequencing

Issues with single cell sequencing currently, involve resolution and depth. The technology can only sequence around 6,000 cells at a time when using a technology such as 10x genomics [**386**] amounting to around 1000 reads per sample. When assessing TCRB repertoire diversity, this technology does not sequence deeply enough to be able to capture diversity in the system, this also makes the technology expensive. However, selecting out interesting populations of cells, such as V6-5/J1-4 identified in **Chapter 5** or a specific TCRB clonal expansion in patients, such as for patient 004V3 where it is potentially linked to PNH diagnosis, and then performing single cell or paired sequencing, would alleviate the issue of depth and could help improve the study. Once TCRB clones linked to PNH pathogenesis are identified, these could be included in future mouse models. Recreating PNH pathogenesis in mouse models has so far failed due to *PIG-A* knockdown in HSCs not being enough to create PNH in a mouse model (**Chapter 1.**). TCRB clones that cause a destructive bone marrow environment can be identified and used in these mouse models, it could significantly improve research. The specific T-cells could also be used in the LTBMC experiments to assess whether those populations are pathogenic and give *PIG-A* mutated HSCs a proliferative advantage over normal HSCs.

Further work for the LTBMC could involve sampling the T-cell populations at the start and then every 5 days or so for the length of the experiment, assessing the TCRB repertoires and seeing whether they change as the PNH model progresses.

7.5.3. Unproductive TCRB analysis detecting thymic selection process in PNH

Another project that would be interesting to look at would be unproductive reads. These were ignored in this project, but differences may occur between normal TCRB repertoires and those with PNH and/or AA. Unproductive reads are sometimes used to look at thymic selection events as they did not successfully rearrange and pass the selection process [**387**], and this may be linked with the immune-ageing processes and perhaps with PNH progression. Linked to this, comorbidity studies of recovering PNH patients would be interesting to give an indication of the current immune state of the patient.

It would be interesting to compare GPI+/- subsets here as well to see whether these populations change with age, as it can be tricky to assess specifically when PNH starts.

7.5.4. Data mining medical records to assess immunological history of TCRB repertoires

TCR repertoires contain in depth information about a person's immune response over decades. No doubt, the effect of the TCRB repertoire would change as a result of treatments that may or may not be related to PNH or AA. As technologies progress it would be interesting to be able to track TCRB repertoire changes with treatments for instance. Many of the AA patients would have been on immunosuppressants. Antithymocyte globulin (ATG) and cyclosporine (CsA) are commonly used to treat AA and are thought to suppress CD8+ T-cells [**388**] . AA was used as a "positive control", as it is known to be T-cell mediated and repertoires would expect to show T-cell clonal expansions. However, this would be dependent on the patient's treatment. Those who were clonal, may have paused treatment allowing T-cells to expand again or may no longer be responsive to treatment. It would, therefore, be interesting to retrospectively look back at the clinical data of these AA and PNH patients, to discover any major factors that could be affecting TCRB repertoire profiles.

In parallel to this, many patients with PNH are treated with Eculizumab which targets the complement system. Although not originally thought to have any effect on T-cells, a study found changes in expression of molecules on T-cells with Eculizumab, for example CXCR4 expression [**389**]. This is most likely because the complement system is involved in immunity and there may be some downstream effects on the immune system that might be specific to an individual or common across treated patients. Whether a PNH patient has just started treatment, has had treatment for a long time, or is no longer responsive to treatment would be really interesting for comparison with T-cell receptor beta responses. Patients' samples taken before treatment and during would again aid this research. Patients are often given meningococcal vaccine before taking Eculizumab, as the drug makes patients more prone to bacterial infections.

Investigating TCRB repertoire changes would be interesting from the point of increased susceptibility to infection by Eculizumab targeting the complement system, although currently, patients are not thought to be more prone to infections. This could investigate whether T-cells may be responding the increased rate of infection linked with taking Eculizumab rather than being linked to PNH progression, for instance. Linking medical records would also be useful for assessing the "normal" repertoires. Only the age and sex of the normals were noted by the RTB. Perhaps, within ethical grounds, asking some additional questions about previous use of immunosuppressants, known diagnosis or viral infections in the month prior to sampling would help provide insight into potential TCRB clonal expansions observed in normals. As mentioned previously, vaccines can also cause clonal TCRB expansions.

Potentially, if normals have had vaccinations for example, before travelling, a couple of weeks prior to sampling this may be indicated in the TCRB repertoire results.

7.5.5. Tracking normal and patient TCRB repertoires over long periods of time

The most interesting and informative results were when long term points were used. This allowed for the identification of persistent clones, such as the large EBV specific TCRB clone and would allow tracking of diagnosis with TCRB repertoire over time to allow inferences. For instance, patient 004V3 saw a falling PNH clone over 6 years. The repertoire remained monoclonal with the clonal TCRB of V15/1-4, with a CDR3 'CATSSQAGEKLFF' falling from 18.3% to 4.15% over the six years, suggesting a link between persistent TCRB clone size and PNH clone size. This data is more useful than one sample at a given time point, which is static and does not capture the diversity and dynamic nature of TCRB repertoires. Again, if TCRB clones are in the middle of contracting following infection, or expanding, this is impossible to infer from a single time point but multiple would be easier.

Some RTB patients may come in multiple times a year, or before starting a trial or change in treatment. If a blood sample is taken each time, patients could be tracked over the short and long term, for changes in the repertoire. This would help assess changes that occur when PNH is progressing and active, or stable in the long term. It will also mean that any patients who spontaneously remit will have TCRB repertoire samples for when they had PNH as well to better understand roles of T-cells. Some patients become resistant to Eculizumab over time, detecting changes in the repertoire could aid further understanding of the mechanisms behind this happening. Assessing normals over time periods would also help provide more informative and accurate background values for comparison and alleviate the need for known TCRB clonal expansions, for instance linked to acute viral infections.

7.5.6. Expansion of normal TCRB repertoire datasets

TCRB repertoire research does not currently have any curated TCRB databases specialising in "normal" TCRB repertoires, to be used as background when comparing repertoires with disease. One reason for this is how to define a "normal" repertoire. This is not an easy task. An individual may not have any diagnosed illnesses but at point of sampling could be suffering from an infection such as a CMV. Variation between individuals as discussed previously again, adds to the complexity of a "normal" immune repertoire. Large research institutions like the "Allen Institute for Immunology" are devoted to researching the diverse cell types and networks that make up healthy immune systems in humans.

Another problem is that depending on the type of TCR sequencing method used or nomenclature for TCRs, the results can differ. For example, the paper used originally in **Section 2.5.** [**241**] for testing common TCRBs, identified V20-1 as the most common. However, this method used 5' RACE and cDNA which could be why this differed from V29-1 in this project's data. Consequently, it was important to sequence "normals" in this project using the same methods as for the PNH and AA patients to allow as accurate a comparison as possible. To further this work, the range of normals could be increased and trends in "normal" repertoires could be identified further. More male normals and older normals are needed for this study. PNH patients tend to be older and many are in their 70s and 80s whereas the normals only went up to around the age of 60 years old. A selection of normals more representative in age and sex to PNH are essential for comparing variations between groups. Tracking normal TCRB repertoire variation over time would also be essential for comparisons with patient samples over time enabling clear identification of changes due to PNH progression, for example, rather than, those attributed with age.

7.5.7. Standardisation of methods and databases in the immune repertoire research field

Since the start of this project there have been advances in technologies and publicly available datasets. However, comparisons between project datasets and these data-sets produced by these different TCRB methods stored in publicly available databases can differ technically. To improve this issue, a gold standard should be adopted in regard to nomenclature, methodologies and data curation and storage. IMGT® has been adopted by most as the standard for TCR nomenclature. However, for instance, the flow cytometry TCRBV beta antibodies results were originally based on Wei *et al.* nomenclature and converted to IMGT® for comparisons. The Adaptive Immune Receptor Repertoire (AIRR) Community of The Antibody Society [**390**] have made significant contribution and progress in the standardisation of methods, data curation and comparison in the immune repertoire field. Euroclonality/BIOMED-2 [**379**] are another consortium aiming to standardise the field. Generating a 'minimum information protocol' style approach to immune repertoire studies will really benefit the field as a whole. Currently, curated databases such as McPAS-TCR and VDJdb used in this [**254-255**] study are providing important insight into disease associated TCRs. With standardisation of methods and an increase in TCR studies these databases will hopefully be expanded. More disease specific TCRBs will be identified and able to be compared with TCRB clonal expansions discovered in this study. In this project only two of the clonal AA/PNH TCRBs were linked to the data in databases. At present, due to limited data in these databases, accurately assessing TCRB clonotypes involved in disease are mainly determined by internally generated normals but the wider origin remains unknown.

7.6. Final summary

In summary, this project has highlighted that there are differences in TCRB repertoires between patients with PNH, AA and normals. However, identifying potential TCRBs linked with PNH will require further study. Multiple time point samples for patients to assess TCRB dynamics over time and comparison with their changes in diagnosis, for example PNH recovery, helped suggest links between TCR dynamics with PNH clonal expansions. However, increasing these types of samples will benefit this research further. Analysing TCR samples in the context of previous treatments such as immunosuppressants and bone marrow transplants will also help identify changes linked specifically to PNH versus previous infections or treatment history. As the findings potentially identified that in PNH, the TCRB clonal space is in the more moderately expanded clones' size, whereas the hyperexpanded clones appear to be more in response to infections like EBV, it will be interesting, as more disease specific TCRBs are identified to see if this still holds true.

Moving towards paired alpha beta chain and single cell methods on clonal populations identified in this project will allow TCR receptor 3D protein modelling to help evaluate potential peptides that the TCRs are responding to and whether these are linked to PNH. Deciphering whether these TCRBs are GPI+ or –ve could help to identify TCRB clones that are linked with either the original PNH pathogenesis or subsequent progression of the disease. Adoption of future machine learning techniques will enable subtle trends between category groups and patients to be identified and increase the number of patient repertoires analysed. Finally, the standardisation of TCRB methods, and TCR nomenclature for instance through the use of 'Minimum information protocols', will allow the expansion of public databases and effective comparison with experimental sequencing data to link TCRBs to specific diseases.

List of References

1. Nikolich-Žugich, J., Slifka, M. K., & Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. In *Nature Reviews Immunology*.
2. Yang, Q., Jeremiah Bell, J., & Bhandoola, A. (2010). T-cell lineage determination. *Immunological Reviews*.
3. Bhandoola, A., & Sambandam, A. (2006). From stem cell to T cell: One route or many? In *Nature Reviews Immunology*.
4. Shortman, K., & Wu, L. (1996). Early T lymphocyte progenitors. In *Annual Review of Immunology*.
5. Roth, D. B., & Craig, N. L. (1998). VDJ Recombination. *Cell*.
6. Liu, H., Rhodes, M., Wiest, D. L., & Vignali, D. A. A. (2000). On the dynamics of TCR:CD3 complex cell surface expression and downmodulation. *Immunity*.
7. Morris, G. P., & Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. In *Nature Immunology*.
8. Klein, L., Kyewski, B., Allen, P. M., & Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). In *Nature Reviews Immunology*.
9. Takaba, H., & Takayanagi, H. (2017). The Mechanisms of T Cell Selection in the Thymus. In *Trends in Immunology*.
10. Mosaad, Y. M. (2015). Clinical Role of Human Leukocyte Antigen in Health and Disease. In *Scandinavian Journal of Immunology*.
11. Von Boehmer, H., Aifantis, I., Gounari, F., Azogui, O., Haughn, L., Apostolou, I., Jaeckel, E., Grassi, F., & Klein, L. (2003). Thymic selection revisited: How essential is it? In *Immunological Reviews*.
12. Abramson, J., & Anderson, G. (2017). Thymic epithelial cells. In *Annual Review of Immunology*.
13. Hu, Z., Lancaster, J. N., & Ehrlich, L. I. R. (2015). The contribution of chemokines and migration to the induction of central tolerance in the thymus. In *Frontiers in Immunology*.
14. Bortnick, A., & Murre, C. (2018). mTECs Aire on the side of caution. *Nature Immunology*.
15. Moran, A. E., & Hogquist, K. A. (2012). T-cell receptor affinity in thymic development. In *Immunology*.
16. Morris, G. P., & Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. In *Nature Immunology*.
17. Carl, J. W., Liu, J.-Q., Joshi, P. S., El-Omrani, H. Y., Yin, L., Zheng, X., Whitacre, C. C., Liu, Y., & Bai, X.-F. (2008). Autoreactive T Cells Escape Clonal Deletion in the Thymus by a CD24-Dependent Pathway. *The Journal of Immunology*.
18. Kurd, N., & Robey, E. A. (2016). T-cell selection in the thymus: A spatial and temporal perspective. *Immunological Reviews*.
19. Kannan, A., Huang, W., Huang, F., & August, A. (2012). Signal transduction via the T cell antigen receptor in naïve and effector/memory T cells. In *International Journal of Biochemistry and Cell Biology*.
20. Rane, S., Hogan, T., Seddon, B., & Yates, A. J. (2018). Age is not just a number: Naive T cells increase their ability to persist in the circulation over time. *PLoS Biology*.
21. Hubo, M., Trinschek, B., Kryczanowsky, F., Tuettenberg, A., Steinbrink, K., & Jonuleit, H. (2013). Costimulatory molecules on immunogenic versus tolerogenic human dendritic cells. *Frontiers in Immunology*.
22. Smith-Garvin, J. E., Koretzky, G. A., & Jordan, M. S. (2009). T cell activation. In*Annual Review of Immunology*.
23. Underhill, D. M., & Goodridge, H. S. (2007). The many faces of ITAMs. In *Trends in Immunology*.

24. Ballek, O., Valecka, J., Dobešová, M., Broucková, A., Manning, J., Rehulka, P., Stulík, J., & Filipp, D. (2016). TCR triggering induces the formation of Lck-RACK1-actinin-1 multiprotein network affecting Lck redistribution. *Frontiers in Immunology*.

25. Hernández-Hoyos, G., Sohn, S. J., Rothenberg, E. V., & Alberola-Ila, J. (2000). Lck activity controls CD4/CD8 T cell lineage commitment. *Immunity*.

26. Lin, K., Longo, N. S., Wang, X., Hewitt, J. A., & Abraham, K. M. (2000). Lck domains differentially contribute to pre-T cell receptor (TCR)- and TCR-α/β-regulated developmental transitions. *Journal of Experimental Medicine*.

27. Pennock, N. D., White, J. T., Cross, E. W., Cheney, E. E., Tamburini, B. A., & Kedl, R. M. (2013). T cell responses: Naïve to memory and everything in between. *American Journal of Physiology - Advances in Physiology Education*.

28. Wan, Y. Y., & Flavell, R. A. (2009). How diverse-CD4 effector T cells and their functions. In *Journal of Molecular Cell Biology*.

29. Omilusik, K. D., & Goldrath, A. W. (2017). The origins of memory T cells. In *Nature*.

30. Wong, M. T., Ong, D. E. H., Lim, F. S. H., Teng, K. W. W., McGovern, N., Narayanan, S., Ho, W. Q., Cerny, D., Tan, H. K. K., Anicete, R., Tan, B. K., Lim, T. K. H., Chan, C. Y., Cheow, P. C., Lee, S. Y., Takano, A., Tan, E. H., Tam, J. K. C., Tan, E. Y., … Newell, E. W. (2016). A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity*.

31. Evans, G. A., Lewis, K. A., & Lawless, G. M. (1988). Molecular organization of the human CD3 gene family on chromosome 11q23. *Immunogenetics*.

32. Yang, Q., Jeremiah Bell, J., & Bhandoola, A. (2010). T-cell lineage determination. *Immunological Reviews*.

33. Van der Wel, N. N., Sugita, M., Fluitsma, D. M., Cao, X., Schreibelt, G., Brenner, M. B., & Peters, P. J. (2003). CD1 and major histocompatibility complex II molecules follow a different course during dendritic cell maturation. *Molecular Biology of the Cell*.

34. Borgulya, P., Kishi, H., Muller, U., Kirberg, J., & Von Boehmer, H. (1991). Development of the CD4 and CD8 lineage of T cells: Instruction versus selection. *EMBO Journal*.

35. Doyle, C., & Strominger, J. L. (1987). Interaction between CD4 and class II MHC molecules mediates cell adhesion. *Nature*.

36. Luckheeram, R. V., Zhou, R., Verma, A. D., & Xia, B. (2012). CD4 +T cells: Differentiation and functions. In *Clinical and Developmental Immunology*.

37. Sakaguchi, S., Wing, K., & Miyara, M. (2007). Regulatory T cells - A brief history and perspective. *European Journal of Immunology*.

38. Shevyrev, D., & Tereshchenko, V. (2020). Treg Heterogeneity, Function, and Homeostasis. In *Frontiers in Immunology*.

39. Okeke, E. B., & Uzonna, J. E. (2019). The pivotal role of regulatory T cells in the regulation of innate immune cells. In *Frontiers in Immunology*.

40. Sakaguchi, S., Yamaguchi, T., Nomura, T., & Ono, M. (2008). Regulatory T Cells and Immune Tolerance. In *Cell*.

41. Arellano, B., Graber, D. J., & Sentman, C. L. (2016). Regulatory T cell-basedtherapies for autoimmunity. *Discovery Medicine*.

42. Zhang, N., & Bevan, M. J. (2011). CD8+ T Cells: Foot Soldiers of the Immune System. In *Immunity*.

43. Golubovskaya, V., & Wu, L. (2016). Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy. In *Cancers*.

44. Laidlaw, B. J., Craft, J. E., & Kaech, S. M. (2016). The multifaceted role of CD4+ T cells in CD8+ T cell memory. In *Nature Reviews Immunology*.

45. Joshi, N. S., & Kaech, S. M. (2008). Effector CD8 T Cell Development: A Balancing Act between Memory Cell Potential and Terminal Differentiation. *The Journal of Immunology*.

46. Jameson, S. C., & Masopust, D. (2018). Understanding Subset Diversity in T Cell Memory. In *Immunity*.

47. Mami-Chouaib, F., & Tartour, E. (2019). Editorial: Tissue resident memory T cells. In *Frontiers in Immunology*.

48. Schenkel, J. M., & Masopust, D. (2014). Tissue-resident memory T cells. In *Immunity*.

49. Wang, Y., & Zhang, C. (2019). The Roles of Liver-Resident Lymphocytes in Liver Diseases. In *Frontiers in immunology*.

50. Wu, L., & Kaer, L. Van. (2011). Natural killer T cells in health and disease. *Frontiers in Bioscience - Scholar*.

51. Van Kaer, L., Parekh, V. V., & Wu, L. (2011). Invariant natural killer T cells: Bridging innate and adaptive immunity. In *Cell and Tissue Research*.

52. Salio, M., & Cerundolo, V. (2009). Linking inflammation to natural killer T cell activation. In *PLoS Biology*.

53. Van Kaer, L., Parekh, V. V., & Wu, L. (2015). The response of CD1d-restricted invariant NKT cells to microbial pathogens and their products. In *Frontiers in Immunology*.

54. Girardi, E., & Zajonc, D. M. (2012). Molecular basis of lipid antigen presentation by CD1d and recognition by natural killer T cells. *Immunological Reviews*.

55. Harner, S., Noessner, E., Nadas, K., Leumann-Runge, A., Schiemann, M., Faber, F. L., Heinrich, J., & Krauss-Etschmann, S. (2011). Cord blood Vα24-Vβ11+ natural killer T cells display a Th2-chemokine receptor profile and cytokine responses. *PLoS ONE*.

56. Pereira, C. S., & Macedo, M. F. (2016). CD1-restricted T cells at the crossroad of innate and adaptive immunity. In *Journal of Immunology Research*.

57. Brigl, M., & Brenner, M. B. (2004). CD1: Antigen presentation and T cell function. In *Annual Review of Immunology*.

58. Zhao, J., Weng, X., Bagchi, S., & Wang, C. R. (2014). Polyclonal type II natural killer T cells require PLZF and SAP for their development and contribute to CpG-mediated antitumor response. *Proceedings of the National Academy of Sciences of the United States of America*.

59. Ryu, S., Park, J. S., Kim, H. Y., & Kim, J. H. (2018). Lipid-Reactive T Cells in Immunological Disorders of the Lung. In *Frontiers in immunology*.

60. de Jong, A. (2015). Activation of human T cells by CD1 and self-lipids. *Immunological Reviews*.

61. Krovi, S. H., & Gapin, L. (2018). Invariant natural killer T cell subsets-more than just developmental intermediates. In *Frontiers in Immunology*.

62. Greenaway, H. Y., Ng, B., Price, D. A., Douek, D. C., Davenport, M. P., & Venturi, V. (2013). NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination. *Immunobiology*.

63. Jerud, E. S., Bricard, G., & Porcelli, S. A. (2006). CD1d-restricted natural killer T cells: Roles in tumor immunosurveillance and tolerance. In *Transfusion Medicine and Hemotherapy*.

64. Sullivan, B. A., Nagarajan, N. A., & Kronenberg, M. (2005). CD1 and MHC II find different means to the same end. In *Trends in Immunology*.

65. Kreslavsky, T., Gleimer, M., Garbe, A. I., & von Boehmer, H. (2010). αβ versus γδ fate choice: Counting the T-cell lineages at the branch point. *Immunological Reviews*.

66. Legut, M., Cole, D. K., & Sewell, A. K. (2015). The promise of γδT cells and the γδT cell receptor for cancer immunotherapy. *Cellular and Molecular Immunology*.

67. Hayes, S. M., Laird, R. M., & Love, P. E. (2010). Beyond αβ/γδ lineage commitment: TCR signal strength regulates γδ T cell maturation and effector fate. In *Seminars in Immunology*.

68. Janeway CA Jr, Travers P, Walport M, et al. (2001). Immunobiology: The Immune System in Health and Disease. 5th edition. T-cell receptor gene rearrangement. *NewYork: Garland Science;*

69. Clark, S. P., Arden, B., Kabelitz, D., & Mak, T. W. (1995). Comparison of human and mouse T-cell receptor variable gene segment subfamilies. In *Immunogenetics*.

70. Chaplin, D. D. (2010). Overview of the immune response. *Journal of Allergy and Clinical Immunology*.

71. Roth, D. B. (2014). V(D)J Recombination: Mechanism, Errors, and Fidelity Generation of antigen receptor diversity: a double-edged sword HHS Public Access. *Microbiol Spectr.*

72. Little, A. J., Matthews, A., Oettinger, M., Roth, D. B., & Schatz, D. G. (2015). The Mechanism of V(D)J Recombination. In *Molecular Biology of B Cells: Second Edition*.

73. Olaru, A., Petrie, H. T., & Livák, F. (2005). Beyond the 12/23 Rule of VDJ Recombination Independent of the Rag Proteins. *The Journal of Immunology*, *174*(10), 6220 LP – 6226.

74. Meier, J. T., & Lewis, S. M. (1993). P nucleotides in V(D)J recombination: a fine-structure analysis. *Molecular and Cellular Biology*.

75. Murphy, K., & Weaver, C. (2017). Janeway ' S 9 Th Edition. In *America*.

76. Conformational changes and flexibility in T-cell receptor recognition of peptide-MHC complexes.Armstrong KM, Piepenbrink KH, Baker BM. *Biochem. J.* 415, 183-96, (2008)

77. Hockett, R. D., De Villartay, J. P., Pollock, K., Poplack, D. G., Cohen, D. I., & Korsmeyer, S. J. (1988). Human T-cell antigen receptor (TCR) δ-chain locus and elements responsible for its deletion are within the TCR α-chain locus. *Proceedings of the National Academy of Sciences of the United States of America*.

78. Naik, A. K., Hawwari, A., & Krangel, M. S. (2015). Specification of Vδ and Vα Usage by Tcra/Tcrd Locus V Gene Segment Promoters . *The Journal of Immunology*.

79. Fang, H., Yamaguchi, R., Liu, X., Daigo, Y., Yew, P. Y., Tanikawa, C., Matsuda, K., Imoto, S., Miyano, S., & Nakamura, Y. (2014). Quantitative T cell repertoire analysis by deep cDNA sequencing of T cell receptor α and β chains using next-generation sequencing (NGS). *OncoImmunology*.

80. Sleckman, B. P., Bassing, C. H., Hughes, M. M., Okada, A., D'Auteuil, M., Wehrly, T. D., Woodman, B. B., Davidson, L., Chen, J., & Alt, F. W. (2000). Mechanisms that direct ordered assembly of T cell receptor β locus V, D, and J gene segments. *Proceedings of the National Academy of Sciences of the United States of America*.

81. Brooks, E. G., Balk, S. P., Aupeix, K., Colonna, M., Strominger, J. L., & Groh-Spies, V. (1993). Human T-cell receptor (TCR) alpha/beta + CD4-CD8- T cells express oligoclonal TCRs, share junctional motifs across TCR V beta-gene families, and phenotypically resemble memory T cells. *Proceedings of the National Academy of Sciences*, *90*(24), 11787 LP – 11791.

82. Moss, P. A. H., & Bell, J. I. (1996). Comparative sequence analysis of the human T cell receptor TCRA and TCRB CDR3 regions. *Human Immunology*.

83. Danska, J. S., Livingstone, A. M., Paragas, V., Ishihara, T., & Garrison Fathman, C. (1990). The presumptive CDR3 regions of both T cell receptor α and β chains determine t cell specificity for myoglobin peptides. *Journal of Experimental Medicine*.

84. Heather, J. M., Ismail, M., Oakes, T., & Chain, B. (2018). High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*.

85. Lefranc, M. P. (2003). IMGT, the international ImMunoGeneTics database®. In *Nucleic Acids Research*.

86. Warren, E. H., Matsen IV, F. A., & Chou, J. (2013). High-throughput sequencing ofB- And T-lymphocyte antigen receptors in hematology. *Blood*.

87. Ma, L., Yang, L., Shi, B., He, X., Peng, A., Li, Y., Zhang, T., Sun, S., Ma, R., & Yao, X. (2016). Analyzing the CDR3 Repertoire with respect to TCR—Beta Chain V-D-J and V-J Rearrangements in Peripheral T Cells using HTS. *Scientific Reports*, *6*(1), 29544.

88. Stadinski, B. D., Shekhar, K., Gómez-Touriño, I., Jung, J., Sasaki, K., Sewell, A. K., Peakman, M., Chakraborty, A. K., & Huseby, E. S. (2016). Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nature Immunology*.

89. Rock, E. P., Sibbald, P. R., Davis, M. M., & Chien, Y. H. (1994). CDR3 length in antigen-specific immune receptors. *Journal of Experimental Medicine*.

90. Flaherty, D. K. B. T.-I. for P. (Ed.). (2012). *Chapter 10 - Antibody Diversity* (pp. 79–86). Mosby.

91. Hou, X., Zeng, P., Zhang, X., Chen, J., Liang, Y., Yang, J., Yang, Y., Liu, X., & Diao, H. (2019). Shorter TCR β-chains are highly enriched during thymic selection and antigen driven slection. *Frontiers in Immunology*.

92. Chapman, C. G., Yamaguchi, R., Tamura, K., Weidner, J., Imoto, S., Kwon, J., Fang, H., Yew, P. Y., Marino, S. R., Miyano, S., Nakamura, Y., & Kiyotani, K. (2016). Characterization of T-cell receptor repertoire in inflamed tissues of patients with Crohn's disease through deep sequencing. *Inflammatory Bowel Diseases*.

93. Fischer, D. S., Wu, Y., Schubert, B., & Theis, F. J. (2020). Predicting antigen specificity of single T cells based on TCR CDR 3 regions . *Molecular Systems Biology*.

94. Rudd, B. D., Venturi, V., Davenport, M. P., & Nikolich-Žugich, J. (2011). Evolution of the Antigen-Specific CD8 + TCR Repertoire across the Life Span: Evidence for Clonal Homogenization of the Old TCR Repertoire . *The Journal of Immunology*.

95. Marlin, R., Pappalardo, A., Kaminski, H., Willcox, C. R., Pitard, V., Netzer, S., Khairallah, C., Lomenech, A. M., Harly, C., Bonnevill, M., Moreau, J. F., Scotet, E., Willcox, B. E., Faustin, B., & Déchanet-Mervillea, J. (2017). Sensing of cell stress by human γδ TCR-dependent recognition of annexin A2. *Proceedings of the National Academy of Sciences of the United States of America*.

96. McDonnell, W. J., Koethe, J. R., Mallal, S. A., Pilkinton, M. A., Kirabo, A., Ameka, M. K., Cottam, M. A., Hasty, A. H., & Kennedy, A. J. (2018). High CD8 T-cell receptor clonality and altered CDR3 properties are associated with elevated isolevuglandins in adipose tissue during diet-induced obesity. *Diabetes*.

97. Krishna, C., Chowell, D., Gönen, M., Elhanati, Y., & Chan, T. A. (2020). Genetic and environmental determinants of human TCR repertoire diversity. *Immunity & Ageing*

98. Schneider-Hohendorf, T., Görlich, D., Savola, P., Kelkka, T., Mustjoki, S., Gross, C. C., Owens, G. C., Klotz, L., Dornmair, K., Wiendl, H., & Schwab, N. (2018). Sex bias in MHC I-associated shaping of the adaptive immune system. *Proceedings of the National Academy of Sciences of the United States of America*.

99. Naylor, K., Li, G., Vallejo, A. N., Lee, W.-W., Koetz, K., Bryl, E., Witkowski, J., Fulbright, J., Weyand, C. M., & Goronzy, J. J. (2005). The Influence of Age on T Cell Generation and TCR Diversity. *The Journal of Immunology*.

100. Attaf, M., & Sewell, A. K. (2016). Disease etiology and diagnosis by TCR repertoire analysis goes viral. In *European Journal of Immunology*.Woodsworth, D. J.,

101. Castellarin, M., & Holt, R. A. (2013). Sequence analysis of T-cell repertoires in health and disease. In *Genome Medicine*.

102. Duffy, D. (2020). Understanding immune variation for improved translationalmedicine. In *Current Opinion in Immunology*.

103. Laydon, D. J., Bangham, C. R. M., & Asquith, B. (2015). Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach.*Philosophical Transactions of the Royal Society B: Biological Sciences*.

104. Kleiveland, C., & Kleiveland, C. (2015). Peripheral blood mononuclear cells.In *The Impact of Food Bioactives on Health: In Vitro and Ex Vivo Models*.

105. Goldrath, A. W., & Bevan, M. J. (1999). Selecting and maintaining a diverseT-cell repertoire. In *Nature*.

106. Lythe, G., Callard, R. E., Hoare, R. L., & Molina-París, C. (2016). How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*.

107. Ganusov, V. V., & De Boer, R. J. (2007). Do most lymphocytes in humans really reside in the gut? *Trends in Immunology*.

108. Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O. I., Verma-Gaur, J., Loguercio, S., Su, A. I., Feeney, A. J., Artyomov, M. N., & Oltz, E. M. (2013). Unifying model for molecular determinants of the preselection Vβ repertoire. *Proceedings of the National Academy of Sciences of the United States of America*.

109. Harsha Krovi, S., Kappler, J. W., Marrack, P., & Gapin, L. (2019). Inherent reactivity of unselected TCR repertoires to peptide-MHC molecules. *Proceedings of the National Academy of Sciences of the United States of America*.

110. Attaf, M., Huseby, E., & Sewell, A. K. (2015). αβ T cell receptors aspredictors of health and disease. *Cellular and Molecular Immunology*.

111. Pogorelyy, M. V., Minervina, A. A., Touzel, M. P., Sycheva, A. L., Komech, E. A., Kovalenko, E. I., Karganova, G. G., Egorov, E. S., Komkov, A. Y., Chudakov, D. M., Mamedov, I. Z., Mora, T., Walczak, A. M., & Lebedev, Y. B. (2018). Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences of the United States of America*.

112. Burrows, S. R., Silins, S. L., Moss, D. J., Khanna, R., Misko, I. S., & Argaet, V. P. (1995). T cell receptor repertoire for a viral epitope in humans is diversified by tolerance to a background major histocompatibility complex antigen. *Journal of Experimental Medicine*.

113. Venturi, V., Price, D. A., Douek, D. C., & Davenport, M. P. (2008). Themolecular basis for public T-cell responses? In *Nature Reviews Immunology*.

114. Carey, A. J., Hope, J. L., Mueller, Y. M., Fike, A. J., Kumova, O. K., van Zessen, D. B. H., Steegers, E. A. P., van der Burg, M., & Katsikis, P. D. (2017). Public clonotypes and convergent recombination characterize the Naïve CD8+ T-cellreceptor repertoire of extremely preterm neonates. *Frontiers in Immunology*.

115. Venturi, V., Kedzierska, K., Price, D. A., Doherty, P. C., Douek, D. C., Turner, S. J., & Davenport, M. P. (2006). Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences of the United States of America*.

116. Quigley, M. F., Greenaway, H. Y., Venturi, V., Lindsay, R., Quinn, K. M., Seder, R. A., Douek, D. C., Davenport, M. P., & Price, D. A. (2010). Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proceedings of the National Academy of Sciences of theUnited States of America*.

117. Huth, A., Liang, X., Krebs, S., Blum, H., & Moosmann, A. (2019). Antigen Specific TCR Signatures of Cytomegalovirus Infection. *The Journal of Immunology*.

118. Sharon, E., Sibener, L. V., Battle, A., Fraser, H. B., Christopher Garcia, K., & Pritchard, J. K. (2016). Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature Genetics*.

119. Alves Sousa, A. de P., Johnson, K. R., Ohayon, J., Zhu, J., Muraro, P. A., & Jacobson, S. (2019). Comprehensive Analysis of TCR-β Repertoire in Patients with Neurological Immune-mediated Disorders. *Scientific Reports*, *9*(1), 344.

120. Zvyagin, I. V., Pogorelyy, M. V., Ivanova, M. E., Komech, E. A., Shugay, M., Bolotin, D. A., Shelenkov, A. A., Kurnosov, A. A., Staroverov, D. B., Chudakov, D. M., Lebedev, Y. B., & Mamedov, I. Z. (2014). Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of theNational Academy of Sciences of the United States of America*.

121. Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R., & Holt, R. A. (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*.

122. Sebzda, E., Mariathasan, S., Ohteki, T., Jones, R., Bachmann, M. F., & Ohashi, P. S. (1999). Selection of the T cell repertoire. In *Annual Review of Immunology*.

123. Kedzierska, K., Day, E. B., Pi, J., Heard, S. B., Doherty, P. C., Turner, S. J., & Perlman, S. (2006). Quantification of Repertoire Diversity of Influenza-Specific Epitopes with Predominant Public or Private TCR Usage. *The Journal of Immunology*.

124. Zinkernagel, R. M. (1996). Immunology taught by viruses. *Science*.

125. Wolf, K., Maybruck, J., & DiPaolo, R. J. (2018). Diagnostic Assessment of Immunological History by High-throughput TCR sequence Analyses. *The Journal ofImmunology*, *200*(1 Supplement), 120.4 LP-120.4.

126. DeWitt, W. S., Smith, A., Schoch, G., Hansen, J. A., Matsen, F. A., & Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *ELife*.

127. Merkenschlager, J., & Kassiotis, G. (2015). Narrowing the gap: Preserving repertoire diversity despite clonal selection during the CD4 T cell response. In *Frontiers in Immunology*.

128. Chu, N. D., Bi, H. S., Emerson, R. O., Sherwood, A. M., Birnbaum, M. E., Robins, H. S., & Alm, E. J. (2019). Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunology*.

129. Kamimura, D., Atsumi, T., Stofkova, A., Nishikawa, N., Ohki, T., Suzuki, H., Katsunuma, K., Jiang, J., Bando, H., Meng, J., Sabharwal, L., Ogura, H., Hirano, T., Arima, Y., & Murakami, M. (2015). Naïve T Cell Homeostasis Regulated by Stress Responses and TCR Signaling. *Frontiers in Immunology*.

130. Thome, J. J. C., Bickham, K. L., Ohmura, Y., Kubota, M., Matsuoka, N., Gordon, C., Granot, T., Griesemer, A., Lerner, H., Kato, T., & Farber, D. L. (2016). Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. *Nature Medicine*.

131. Boyman, O., Létourneau, S., Krieg, C., & Sprent, J. (2009). Homeostatic proliferation and survival of naïve and memory T cells. In *European Journal of*

132. Min, B. (2018). Spontaneous T cell proliferation: A physiologic process to create and maintain homeostatic balance and diversity of the immune system. In *Frontiers in Immunology*.

133. Vallejo, A. N. (2006). Age-dependent alterations of the T cell repertoire and functional diversity of T cells of the aged. In *Immunologic Research*.

134. Mancebo, E., Clemente, J., Sanchez, J., Ruiz-Contreras, J., De Pablos, P., Cortezon, S., Romo, E., Paz-Artal, E., & Allende, L. M. (2008). Longitudinal analysis of immune function in the first 3 years of life in thymectomized neonates during cardiac surgery. *Clinical and Experimental Immunology*.

135. Palmer, S., Albergante, L., Blackburn, C. C., & Newman, T. J. (2018). Thymic involution and rising disease incidence with age. *Proceedings of the National Academy of Sciences of the United States of America*.

136. Thomas, R., Wang, W., & Su, D. M. (2020). Contributions of Age-Related Thymic Involution to Immunosenescence and Inflammaging. In *Immunity and*

137. Khan, N., Shariff, N., Cobbold, M., Bruton, R., Ainsworth, J. A., Sinclair, A. J., Nayak, L., & Moss, P. A. H. (2002). Cytomegalovirus Seropositivity Drives the CD8 T Cell Repertoire Toward Greater Clonality in Healthy Elderly Individuals. *The Journal of Immunology*.

138. Yanes, R. E., Gustafson, C. E., Weyand, C. M., & Goronzy, J. J. (2017). Lymphocyte generation and population homeostasis throughout life. In *Seminars in Hematology*.

139. de Greef, P. C., Oakes, T., Gerritsen, B., Ismail, M., Heather, J. M., Hermsen, R., Chain, B., & de Boer, R. J. (2020). The naive t-cell receptor repertoire has an extremely broad distribution of clone sizes. *ELife*.

140. Gil, A., Mishra, R., Song, I., Aslan, N., Luzuriaga, K., & Selin, L. K. (2017). Influenza A virus (IAV) infection in humans leads to expansion of highly diverse CD8 T cell repertoires cross-reactive with Epstein Barr virus (EBV). *The Journal of Immunology*.

141. Zhao, Y., Nguyen, P., Ma, J., Wu, T., Jones, L. L., Pei, D., Cheng, C., & Geiger, T. L. (2016). Preferential Use of Public TCR during Autoimmune Encephalomyelitis. *The Journal of Immunology*.

142. Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. In *Current Opinion in Immunology*.

143. Mahe, E., Pugh, T., & Kamel-Reid, S. (2018). T cell clonality assessment: Past, present and future. In *Journal of Clinical Pathology*.

144. Cui, J. H., Lin, K. R., Yuan, S. H., Jin, Y. Bin, Chen, X. P., Su, X. K., Jiang, J., Pan, Y. M., Mao, S. L., Mao, X. F., & Luo, W. (2018). TCR repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervicalcancer. *Frontiers in Immunology*.

145. Li, Y., & Xu, L. (2015). Evaluation of TCR repertoire diversity in patientsafter hematopoietic stem cell transplantation. *Stem Cell Investigation*.

146. Shomuradova, A. S., Vagida, M. S., Sheetikov, S. A., Zornikova, K. V, Kiryukhin, D. D., Titov, A., Peshkova, I. O., Khmelevskaya, A., Dianov, D. V, Malasheva, M., Shmelev, A., Serdyuk, Y., Bagaev, D. V, Pivnyuk, A., Shcherbinin, D. S., Maleeva, A. V, Shakirova, N. T., Pilunov, A., Malko, D. B., … Efimov, G. A. (2020). SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T-cell receptors. *Medrxiv*.

147. Tzifi, F., Kanariou, M., Tzanoudaki, M., Mihas, C., Paschali, E., Chrousos, G., & Kanaka-Gantenbein, C. (2013). Flow cytometric analysis of the CD4+ TCR Vβ repertoire in the peripheral blood of children with type 1 diabetes mellitus, systemic lupus erythematosus and age-matched healthy controls. *BMC Immunology*.

148. Fozza, C., Barraqueddu, F., Corda, G., Contini, S., Virdis, P., Dore, F., Bonfigli, S., & Longinotti, M. (2017). Study of the T-cell receptor repertoire by CDR3 spectratyping. In *Journal of Immunological Methods*.

149. Xu, L., You, X., Zheng, P. P., Zhang, B. M., Gupta, P. K., Lavori, P., Meyer, E., & Zehnder, J. L. (2017). Methodologic Considerations in the Application of Next-Generation Sequencing of Human TRB Repertoires for Clinical Use. *Journal of Molecular Diagnostics*.

150. Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. In *Nature Biotechnology*.

151. Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE*.

152. Rosati, E., Dowds, C. M., Liaskou, E., Henriksen, E. K. K., Karlsen, T. H., & Franke, A. (2017). Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology*.

153. Six, A., Mariotti-Ferrandiz, M. E., Chaara, W., Magadan, S., Pham, H. P., Lefranc, M. P., Mora, T., Thomas-Vaslin, V., Walczak, A. M., & Boudinot, P. (2013). The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Frontiers in Immunology*.

154. van Dongen, J. J. M., Langerak, A. W., Brüggemann, M., Evans, P. A. S., Hummel, M., Lavender, F. L., Delabesse, E., Davi, F., Schuuring, E., García-Sanz, R., van Krieken, J. H. J. M., Droese, J., González, D., Bastard, C., White, H. E., Spaargaren, M., González, M., Parreira, A., Smith, J. L., … Macintyre, E. A. (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 concerted action BMH4-CT98-3936. In *Leukemia*.

155. Mold, J. E., Réu, P., Olin, A., Bernard, S., Michaëlsson, J., Rane, S., Yates, A., Khosravi, A., Salehpour, M., Possnert, G., Brodin, P., & Frisén, J. (2019). Cell generation dynamics underlying naive T-cell homeostasis in adult humans. *PLoS*

156. Zschaler, J., Schlorke, D., & Arnhold, J. (2014). Differences in innate immune response between man and mouse. *Critical Reviews in Immunology*.

157. Sufficool, K. E., Lockwood, C. M., Abel, H. J., Hagemann, I. S., Schumacher, J. A., Kelley, T. W., & Duncavage, E. J. (2015). T-cell clonality assessment by next-generation sequencing improves detection sensitivity in mycosis fungoides. *Journal of the American Academy of Dermatology*.

158. Kitaura, K., Shini, T., Matsutani, T., & Suzuki, R. (2016). A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) α and β repertoires and identifying potential new invariant TCR α chains. *BMC Immunology*.

159. Thomas, P. G., Handel, A., Doherty, P. C., & La Gruta, N. L. (2013). Ecological analysis of antigen-specific CTL repertoires defines the relationship between naïve and immune T-cell populations. *Proceedings of the National Academy of Sciences of the United States of America*.

160. Aversa, I., Malanga, D., Fiume, G., & Palmieri, C. (2020). Molecular T-cell repertoire analysis as source of prognostic and predictive biomarkers for checkpoint blockade immunotherapy. In *International Journal of Molecular Sciences*.

161. Bousso, P., Wahn, V., Douagi, I., Horneff, G., Pannetier, C., Le Deist, F., Zepp, F., Niehues, T., Kourilsky, P., Fischer, A., & De Saint Basile, G. (2000). Diversity, functionality, and stability of the t cell repertoire derived in vivo from a single human T cell precursor. *Proceedings of the National Academy of Sciences of the United States of America*.

162. Chao, A. (1984). Non-parametric estimation of the classes in a population. *Scandinavian Journal of Statistics*.

163. Kaplinsky, J., & Arnaout, R. (2016). Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nature Communications*.

164. Charles, J., Mouret, S., Challende, I., Leccia, M. T., De Fraipont, F., Perez, S., Plantier, N., Plumas, J., Manuel, M., Chaperot, L., & Aspord, C. (2020). T-cell receptor diversity as a prognostic biomarker in melanoma patients. *Pigment Cell and Melanoma Research*.

165. Han, J.-F., Wang, Z., Bai, H., Chen, S., Wang, Y., Duan, J., & Wang, J. (2019). A novel noninvasive biomarker based on peripheral PD-1 posi CD8 T-cell receptor repertoire correlated with clinical outcomes to immunotherapy in non-small cell lung cancer. *Journal of Clinical Oncology*

166. Sena, J. A., Galotto, G., Devitt, N. P., Connick, M. C., Jacobi, J. L., Umale, P. E., Vidali, L., & Bell, C. J. (2018). Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Scientific Reports*.

167. Roberts, M. (2011). Phred quality score. In *Genome*.

168. Jost, L. (2006). Entropy and diversity. In *Oikos*.

169. Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. In *Ecology Letters*.

170. Vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N. H., O'connor, K. C., Hafler, D. A., Vigneault, F., & Kleinstein, S. H. (2014). PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires.

171. Shugay, M., Britanova, O. V., Merzlyak, E. M., Turchaninova, M. A., Mamedov, I. Z., Tuganbaev, T. R., Bolotin, D. A., Staroverov, D. B., Putintseva, E. V., Plevova, K., Linnemann, C., Shagin, D., Pospisilova, S., Lukyanov, S., Schumacher, T. N., & Chudakov, D. M. (2014). Towards error-free profiling of immune repertoires. *Nature Methods*.

172. https://github.com/alastair-droop/clipumi

173. Mangul, S., Driesche, S. Van, Martin, L., Martin, K., & Eskin, E. (2017). UMI-Reducer: Collapsing duplicate sequencing reads via Unique Molecular

174. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., & Chain, B. (2013). Decombinator: A tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*.

175. Andrews, S. (2010). FastQC. *Babraham Bioinformatics*.

176. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*.

177. https://github.com/FelixKrueger/TrimGalore

178. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*.

179. Gordon, A., Hannon, G. J., & Gordon. (2014). FASTX-Toolkit. In *[Online] http://hannonlab. cshl. edu/fastx_ toolkit http://hannonlab. cshl. edu/fastx_ toolkit*.

180. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics*.

181. Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*.

182. Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*.

183. Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. In *Nature Methods*.

184. Brochet, X., Lefranc, M. P., & Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Research*.

185. ImmunoMind Team. (2019). immunarch: An R Package for Painless Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires. Zenodo.

186. Shugay, M., Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., Putintseva, E. V., Pogorelyy, M. V., Nazarov, V. I., Zvyagin, I. V., Kirgizova, V. I., Kirgizov, K. I., Skorobogatova, E. V., & Chudakov, D. M. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Computational Biology*.

187. Redmond, D., Poran, A., & Elemento, O. (2016). Single-cell TCRseq: Paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Medicine*.

188. Bolotin, D. A., Shugay, M., Mamedov, I. Z., Putintseva, E. V., Turchaninova, M. A., Zvyagin, I. V., Britanova, O. V., & Chudakov, D. M. (2013). MiTCR: Software for T-cell receptor sequencing data analysis. In *Nature Methods*.

189. Hanaoka, N., Kawaguchi, T., Horikawa, K., Nagakura, S., Mitsuya, H., & Nakakuma, H. (2006). Immunoselection by natural killer cells of PIGA mutant cells missing stress-inducible ULBP. *Blood*.

190. Brodsky, R. A. (2014). Paroxysmal nocturnal hemoglobinuria. In *Blood*.

191. Hillmen, P., Lewis, S. M., Bessler, M., Luzzatto, L., & Dacie, J. V. (1995). Natural history of paroxysmal nocturnal hemoglobinuria. *New England Journal of*

192. Hill, A., Dezern, A. E., Kinoshita, T., & Brodsky, R. A. (2017). Paroxysmal nocturnal haemoglobinuria. In *Nature Reviews Disease Primers*.

193. Hill, A., Platts, P. J., Smith, A., Richards, S. J., Cullen, M. J., Hill, Q. A., Roman, E., & Hillmen, P. (2006). The Incidence and Prevalence of Paroxysmal Nocturnal Hemoglobinuria (PNH) and Survival of Patients in Yorkshire. *Blood*.

194. Richards, S. J., Dickinson, A. J., Cullen, M. J., Griffin, M., Munir, T., McKinley, C., Mitchell, L. D., Newton, D. J., Arnold, L., Hill, A., & Hillmen, P. (2020). Presentation clinical, haematological and immunophenotypic features of 1081 patients with GPI-deficient (paroxysmal nocturnal haemoglobinuria) cells detected by flow cytometry.

195. John Richards, S., Kelly, R., Hill, A., Dickinson, A., Cullen, F., Shingles, J., Cullen, M., & Hillmen, P. (2013). Insights Into The Natural History Of Paroxysmal Nocturnal Hemoglobinuria (PNH): Analysis Of The Presenting Clinical, Haematological and Flow Cytometric Features Of 705 Patients Leads To Improved Classification and Prediction Of Clinical Course. *Blood*.

196. Maciejewski, J. P., Sloand, E. M., Sato, T., Anderson, S., & Young, N. S. (1997). Impaired hematopoiesis in paroxysmal nocturnal hemoglobinuria/aplastic anemia is not associated with a selective proliferative defect in theglycosylphosphatidylinositol-anchored protein-deficient clone. *Blood*.

197. Boccuni, P., Del Vecchio, L., Di Noto, R., & Rotoli, B. (2000). Glycosyl phosphatidylinositol (GPI)-anchored molecules and the pathogenesis of paroxysmal nocturnal hemoglobinuria. In *Critical Reviews in Oncology/Hematology*.

198. Kinoshita, T., Fujita, M., & Maeda, Y. (2008). Biosynthesis, remodelling and functions of mammalian GPI-anchored proteins: Recent progress. In *Journal of*

199. Ware, R. E., Rosse, W. F., & Howard, T. A. (1994). Mutations within the Piga gene in patients with paroxysmal nocturnal hemoglobinuria. *Blood*.

200. Mon Père, N., Lenaerts, T., Pacheco, J. M., & Dingli, D. (2018). Evolutionary dynamics of paroxysmal nocturnal hemoglobinuria. *PLoS Computational Biology*.

201. Sarma, J. V., & Ward, P. A. (2011). The complement system. In *Cell and Tissue Research*.

202. Markiewski, M. M., & Lambris, J. D. (2007). The role of complement in inflammatory diseases from behind the scenes into the spotlight. *American Journal ofPathology*.

203. Risitano, A. M., & Rotoli, B. (2008). Paroxysmal nocturnal hemoglobinuria: Pathophysiology, natural history and treatment options in the era of biological agents. In *Biologics: Targets and Therapy*.

204. Ruiz-Argüelles, A., & Llorente, L. (2007). The role of complement regulatory proteins (CD55 and CD59) in the pathogenesis of autoimmune hemocytopenias. In *Autoimmunity Reviews*.

205. Harder, M. J., Kuhn, N., Schrezenmeier, H., Höchsmann, B., Von Zabern, I., Weinstock, C., Simmet, T., Ricklin, D., Lambris, J. D., Skerra, A., Anliker, M., & Schmidt, C. Q. (2017). Incomplete inhibition by eculizumab: Mechanistic evidence for residual C5 activity during strong complement activation. *Blood*.

206. Brodsky, R. A. (2009). How I treat paroxysmal nocturnal hemoglobinuria. In *Blood*.

207. Peerschke, E. I., Yin, W., & Ghebrehiwet, B. (2010). Complement activation on platelets: Implications for vascular inflammation and thrombosis. In *Molecular Immunology*.

208. Miyata, T., Takeda, J., Iida, Y., Yamada, N., Inoue, N., Takahashi, M., Maeda, K., Kitani, T., & Kinoshita, T. (1993). The cloning of PIG-A, a component in the early step of GPI-anchor biosynthesis. *Science*.

209. Araten, D. J., Nafa, K., Pakdeesuwan, K., & Luzzatto, L. (1999). Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. *Proceedings of the National Academy of Sciences of the United States of America*.

210. Wang, H., Chuhjo, T., Yamazaki, H., Shiobara, S., Teramura, M., Mizoguchi, H., & Nakao, S. (2001). Relative increase of granulocytes with a paroxysmal nocturnal haemoglobinuria phenotype in aplastic anaemia patients: The high prevalence at diagnosis. *European Journal of Haematology*.

211. Hu, R., Mukhina, G. L., Piantadosi, S., Barber, J. P., Jones, R. J., & Brodsky, R. A. (2005). PIG-A mutations in normal hematopoiesis. *Blood*.

212. Rosti, V. (2002). Murine models of paroxysmal nocturnal hemoglobinuria. *Annals of the New York Academy of Sciences*.

213. Endo, M., Ware, R. E., Vreeke, T. M., Singh, S. P., Howard, T. A., Tomita, A., Holguin, M. H., & Parker, C. J. (1996). Molecular basis of the heterogeneity of expression of glycosyl phosphatidylinositol anchored proteins in paroxysmal nocturnal hemoglobinuria. *Blood*.

214. Inoue, N., Izui-Sarumaru, T., Murakami, Y., Endo, Y., Nishimura, J. I., Kurokawa, K., Kuwayama, M., Shime, H., Machii, T., Kanakura, Y., Meyers, G., Wittwer, C., Chen, Z., Babcock, W., Frei-Lahr, D., Parker, C. J., & Kinoshita, T. (2006). Molecular basis of clonal expansion of hematopoiesis in 2 patients with paroxysmal nocturnal hemoglobinuria (PNH). *Blood*.

215. Sugimori, C., Padron, E., Caceres, G., Shain, K., Sokol, L., Zhang, L., Tiu, R., O'Keefe, C. L., Afable, M., Clemente, M., Lee, J. M., MacIejewski, J. P., List, A. F., Epling-Burnette, P. K., &

Araten, D. J. (2012). Paroxysmal nocturnal hemoglobinuria and concurrent JAK2 V617F mutation. In *Blood Cancer Journal*.

216. Araten, D. J., Bains, A., Lobry, C., Aifantis, I., & Ibrahim, S. (2012). A Role for TET2 Mutations in Paroxysmal Nocturnal Hemoglobinuria (PNH). *Blood*.

217. Lobry, C., Bains, A., Zamechek, L. B., Ibrahim, S., Aifantis, I., & Araten, D. J. (2019). Analysis of TET2 mutations in paroxysmal nocturnal hemoglobinuria (PNH). *Experimental Hematology and Oncology*.

218. Santagostino, A., Lombardi, L., Dine, G., Hirsch, P., & Misra, S. C. (2019). Paroxysmal Nocturnal Hemoglobinuria with a Distinct Molecular Signature Diagnosed Ten Years after Allogenic Bone Marrow Transplantation for Acute Myeloid Leukemia. *Case Reports in Hematology*.

219. Conrad ME, Barton JC. The aplastic anemia-paroxysmal nocturnal hemoglobinuria syndrome. Am J Hematol. 1979;7(1):61-7. doi: 10.1002/ajh.2830070108. PMID: 507047.

220. Risitano, A. M., Kook, H., Zeng, W., Chen, G., Young, N. S., & Maciejewski, J. P. (2002). Oligoclonal and polyclonal CD4 and CD8 lymphocytes in aplastic anemia and paroxysmal nocturnal hemoglobinuria measured by Vβ CDR3 spectratyping and flow cytometry. *Blood*.

221. Pu, J. J., Mukhina, G., Wang, H., Savage, W. J., & Brodsky, R. A. (2011). Natural history of paroxysmal nocturnal hemoglobinuria clones in patients presentingas aplastic anemia. *European Journal of Haematology*.

222. Takahashi, T., Maruyama, Y., Saitoh, M., Itoh, H., Yoshimoto, M., & Tsujisaki, M. (2015). Fatal Epstein-Barr Virus Reactivation in an Acquired Aplastic Anemia Patient Treated with Rabbit Antithymocyte Globulin and Cyclosporine A.*Case Reports in Hematology*.

223. Sugimori, C., Chuhjo, T., Feng, X., Yamazaki, H., Takami, A., Teramura, M., Mizoguchi, H., Omine, M., & Nakao, S. (2006). Minor population of CD55-CD59- blood cells predicts response to immunosuppressive therapy and prognosis in patientswith aplastic anemia. *Blood*, *107*(4), 1308–1314.

224. Brodsky, R. A. (2008). Paroxysmal Nocturnal Hemoglobinuria: Stem Cells and Clonality. *Hematology*, *2008*(1), 111–115.

225. Young, N. S., Maciejewski, J. P., Sloand, E., Chen, G., Zeng, W., Risitano, A., & Miyazato, A. (2002). The relationship of aplastic anemia and PNH. In *International journal of hematology*.

226. Kinoshita, T., & Inoue, N. (2002). Relationship between aplastic anemia and paroxysmal nocturnal hemoglobinuria. In *International Journal of Hematology*.

227. Manivannan, P., Ahuja, A., & Pati, H. P. (2017). Diagnosis of Paroxysmal Nocturnal Hemoglobinuria: Recent Advances. *Indian Journal of Hematology & Blood Transfusion : An Official Journal of Indian Society of Hematology and BloodTransfusion*, *33*(4), 453–462.

228. Platanias, L. C. (2010). Abnormalities in Th17 T cells in aplastic anemia. In *Blood*. h

229. Kordasti, S. Y., AlKhan, S. M., Lim, Z., Abellan, P. P., Marsh, J. C. W., & Mufti, G. J. (2009). Contrasting Roles of Th1 and Th17 Cells in Aplastic Anaemia(AA) and Myelodysplastic Syndrome (MDS). *Blood*.

230. Johnson, R. J., & Hillmen, P. (2002). Paroxysmal nocturnal haemoglobinuria: Nature's gene therapy? In *Journal of Clinical Pathology - Molecular Pathology*.

231. Kelly, R. J., Hill, A., Arnold, L. M., Brooksbank, G. L., Richards, S. J., Cullen, M., Mitchell, L. D., Cohen, D. R., Gregory, W. M., & Hillmen, P. (2011). Long-term treatment with eculizumab in paroxysmal nocturnal hemoglobinuria: Sustained efficacy and improved survival. *Blood*.

232. Malaspina H, Childs BH, Boulad F, Karadimitris A, Notaro R, Luzzatto L. Dynamics of hematopoiesis in paroxysmal nocturnal hemoglobinuria (PNH): no evidence for intrinsic growth advantage of PNH clones. Leukemia. 2002 Nov;16(11):2243-8. doi: 10.1038/sj.leu.2402694. PMID: 12399968.

233. Maciejewski, J. P., Follmann, D., Nakamura, R., Saunthararajah, Y., Rivera, C. E., Simonis, T., Brown, K. E., Barrett, J. A., & Young, N. S. (2001). Increased frequency of HLA-DR2 in patients with paroxysmal nocturnal hemoglobinuria and the PNH/aplastic anemia syndrome. *Blood*.

234. Kelly, R., et al., 2011. ASH Abstract

235. Karadimitris, A., Manavalan, J. S., Thaler, H. T., Notaro, R., Araten, D. J., Nafa, K., Roberts, I. A. G., Weksler, M. E., & Luzzatto, L. (2000). Abnormal T-cell repertoire is consistent with immune process underlying the pathogenesis of paroxysmal nocturnal hemoglobinuria. *Blood*.

236. Gargiulo, L., Papaioannou, M., Sica, M., Talini, G., Chaidos, A., Richichi, B., Nikolaev, A. V., Nativi, C., Layton, M., De La Fuente, J., Roberts, I., Luzzatto, L., Notaro, R., & Karadimitris, A. (2013). Glycosylphosphatidylinositol-specific, CD1d restricted T cells in paroxysmal nocturnal hemoglobinuria. *Blood*.

237. Richards, S. J., & Barnett, D. (2007). The Role of Flow Cytometry in the Diagnosis of Paroxysmal Nocturnal Hemoglobinuria in the Clinical Laboratory. In*Clinics in Laboratory Medicine*.

238. Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., & Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood*.

239. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 06 22]. Available from: https://www.ncbi.nlm.nih.gov/gene/

240. https://genome.ucsc.edu/cgi-bin/hgPcr

241. Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., & Holt, R. A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*.

242. Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*.

243. Holm, S. (1979). A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*.

244. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*.

245. Gupta and Vander Heiden, et al (2017) <doi:10.1093/bioinformatics/btv359>, Stern, Yaari and Vander Heiden, et al (2014) <doi:10.1126/scitranslmed.3008879>.

246. Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*.

247. W. S. Cleveland, E. Grosse and W. M. Shyu (1992) Local regression models. Chapter 8 of *Statistical Models in S* eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

248. Hou, X. L., Wang, L., Ding, Y. L., Xie, Q., & Diao, H. Y. (2016). Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes and Immunity*.

249. Simpson, E., Measurement of diversity, *Nature*, 1949, 163p.688.

250. https://immunarch.com/articles/web_only/v4_overlap.html

251. M.K, V., & K, K. (2016). A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications: An International Journal*.

252. PubMed [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 06 22]. Available from: https://pubmed.ncbi.nlm.nih.gov

253. https://scholar.google.com - last used for update analysis - 8.8.2020

254. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., & Friedman, N. (2017). McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*.

255. Bagaev, D. V., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E. S., Zvyagin, I. V., Babel, N., Cole, D. K., Godkin, A. J., Sewell, A. K., Kesmir, C., Chudakov, D. M., Luciani, F., & Shugay, M. (2020). VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*.

256. Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*.

257. Gini, C. (1912). Variabilità e mutabilità. *Memorie Di Metodologica Statistica*.

258. Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B., & Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*.

259. Illumina. (2014). MiSeq ® System User Guide. *Illumina Application Note*.

260. Ngs, I. (2009). Technology Spotlight: Illumina Sequencing Technology. *Manual*.

261. Billi, A. C., Kahlenberg, J. M., & Gudjonsson, J. E. (2019). Sex bias in autoimmunity. In *Current opinion in rheumatology*.

262. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*.

263. https://www.adaptivebiotech.com/products-services/immunoseq/

264. Scott-Browne, J. P., White, J., Kappler, J. W., Gapin, L., & Marrack, P. (2009). Germline-encoded amino acids in the αB T-cell receptor control thymic selection. *Nature*.

265. Kandulski, A., Malfertheiner, P., & Wex, T. (2010). Role of regulatory T-cells in H. pylori-induced gastritis and gastric cancer. In *Anticancer Research*.

266. Nikolich-Žugich, J., & van Lier, R. A. W. (2017). Cytomegalovirus (CMV) research in immune senescence comes of age: overview of the 6th International Workshop on CMV and Immunosenescence. In *GeroScience*.

267. Carter, J. A., Preall, J. B., Grigaityte, K., Goldfless, S. J., Jeffery, E., Briggs, A. W., Vigneault, F., & Atwal, G. S. (2019). Single T Cell Sequencing Demonstrates the Functional Role of αβ TCR Pairing in Cell Lineage and Antigen Specificity. *Frontiers in Immunology*.

268. https://www.cdc.gov/ibd/#epidIBD

269. Kelly, A., & Trowsdale, J. (2019). Genetics of antigen processing and presentation. In *Immunogenetics*.

270. Jaccard, Paul (1912), "The Distribution of the flora in the alpine zone", *New Phytologist*, **11** (2): 37–50,

271. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157, 105-32 (1982).

272. suchiya, Y., Namiuchi, Y., Wako, H., & Tsurui, H. (2018). A study of CDR3 loop dynamics reveals distinct mechanisms of peptide recognition by T-cell receptors exhibiting different levels of cross-reactivity. *Immunology*.

273. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A., & Peakman, M. (2017). T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nature Communications*.

274. Carlson, C. S., Emerson, R. O., Sherwood, A. M., Desmarais, C., Chung, M. W., Parsons, J. M., Steen, M. S., LaMadrid-Herrmannsfeldt, M. A., Williamson, D. W., Livingston, R. J., Wu, D., Wood, B. L., Rieder, M. J., & Robins, H. (2013). Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Communications*.

275. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C., & Klein, S. L. (2020). Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nature Reviews Immunology*.

276. Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J. Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., & Goronzy, J. J. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences ofthe United States of America*.

277. Risitano, A. M. (2017). Immune insights into AA. *Blood*.

278. Parker, C. J. (2008). Paroxysmal nocturnal hemoglobinuria: an historical overview. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*.

279. Rizzetto, S., Eltahla, A. A., Lin, P., Bull, R., Lloyd, A. R., Ho, J. W. K., Venturi, V., & Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Scientific Reports*.

280. Greiff, V., Miho, E., Menzel, U., & Reddy, S. T. (2015). Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. In *Trends in Immunology*.

281. Tuomisto, H. (2010). A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamm diversity. *Ecography*.

282. Chang, C. M., Hsu, Y. W., Wong, H. S. C., Wei, J. C. C., Liu, X., Liao, H. T., & Chang, W. C. (2019). Characterization of T-Cell Receptor Repertoire in Patients with Rheumatoid Arthritis Receiving Biologic Therapies. *Disease Markers*.

283. Root-Bernstein, R. S. (1982). Amino acid pairing. *Journal of Theoretical Biology*.

284. Wolf, K., Hether, T., Gilchuk, P., Kumar, A., Rajeh, A., Schiebout, C., Maybruck, J., Buller, R. M., Ahn, T. H., Joyce, S., & DiPaolo, R. J. (2018). Identifying and Tracking Low-Frequency Virus-Specific TCR Clonotypes Using High-Throughput Sequencing. *Cell Reports*.

285. Hsu, S. C., Chang, C. P., Tsai, C. Y., Hsieh, S. H., Wu-Hsieh, B. A., Lo, Y. S., & Yang, J. M. (2012). Steric recognition of T-cell receptor contact residues is required to map mutant epitopes by immunoinformatical programmes. *Immunology*.

286. Roth, C. M., Neal, B. L., & Lenhoff, A. M. (1996). Van der Waals interactions involving proteins.

287. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21, 170-201 (1968).

288. Ikai AJ. Thermostability and aliphatic index of globular proteins. J Biochem 88, 1895-1898 (1980).

289. Grantham R. Amino acid difference formula to help explain protein evolution. Science 185, 862-864 (1974).

290. Moore DS. Amino acid and peptide net charges: A simple calculational procedure. Biochem Educ 13, 10-11 (1985).

291. Reynolds, C., Chong, D., Raynsford, E., Quigley, K., Kelly, D., Llewellyn-Hughes, J., Altmann, D., & Boyton, R. (2014). Elongated TCR alpha chain CDR3 favors an altered CD4 cytokine profile. *BMC Biology*.

292. Rozenblum, G. T., Kaufman, T., & Vitullo, A. D. (2014). Myelin basic protein and a multiple sclerosis-related MBP-peptide bind to oligonucleotides. *Molecular Therapy - Nucleic Acids*.

293. Gimsa, J., & Haberland, L. (2005). Electric and Magnetic Fields in Cells and Tissues. In *Encyclopedia of Condensed Matter Physics*.

294. Razzaq, T. M., Ozegbe, P., Jury, E. C., Sembi, P., Blackwell, N. M., & Kabouridis, P. S. (2004). Regulation of T-cell receptor signaling by membrane microdomains. In *Immunology*.

295. Lu, J., Van Laethem, F., Bhattacharya, A. *et al.* Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nat Commun* **10,** 1019 (2019).

296. Liu YC, Miles JJ, Neller MA, et al. Highly divergent T-cell receptor binding modes underlie specific recognition of a bulged viral peptide bound to a human leukocyte antigen class I molecule. *J Biol Chem*. 2013;288(22):15442-15454.

297. Hsu SC, Chang CP, Tsai CY, et al. Steric recognition of T-cell receptor contact residues is required to map mutant epitopes by immunoinformatical programmes [published correction appears in Immunology. 2012 Aug;136(4):459]. *Immunology*. 2012;136(2):139-152.

298. Daley, S. R., Koay, H. F., Dobbs, K., Bosticardo, M., Wirasinha, R. C., Pala, F., Castagnoli, R., Rowe, J. H., Ott de Bruin, L. M., Keles, S., Lee, Y. N., Somech, R., Holland, S. M., Delmonte, O. M., Draper, D., Maxwell, S., Niemela, J., Stoddard, J., Rosenzweig, S. D., … Notarangelo, L. D. (2019). Cysteine and hydrophobic residues in CDR3 serve as distinct T-cell self-reactivity indices. *Journal of Allergy and Clinical Immunology*.

299. Donald, J. E., Kulp, D. W., & DeGrado, W. F. (2011). Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function and Bioinformatics*.

300. Liu X, Nguyen P, Liu W, Cheng C, Steeves M, Obenauer JC, Ma J, Geiger TL. T cell receptor CDR3 sequence but not recognition characteristics distinguish autoreactive effector and Foxp3(+) regulatory T cells. Immunity. 2009 Dec 18;31(6):909-20.

301. Li HM, Hiroi T, Zhang Y, et al. TCRβ repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *J Leukoc Biol*. 2016;99(3):505-

302. Deleidi, M., Jäggle, M., & Rubino, G. (2015). Immune ageing, dysmetabolism and inflammation in neurological diseases. *Frontiers in Neuroscience*.

303. Maskew, M., Brennan, A. T., Westreich, D., McNamara, L., MacPhail, A. P., & Fox, M. P. (2013). Gender differences in mortality and CD4 count response among virally suppressed HIV-positive patients. *Journal of Women's Health*.

304. Lefranc, M.-P., The Immunologist, 7, 132-136 (1999).

305. Brunner, S., Herndler-Brandstetter, D., Weinberger, B., & Grubeck-Loebenstein, B. (2011). Persistent viral infections and immune aging. In *Ageing Research Reviews*.

306. Dingli, D., Luzzatto, L., & Pacheco, J. M. (2008). Neutral evolution in paroxysmal nocturnal hemoglobinuria. *Proceedings of the National Academy of Sciences of the United States of America*.

307. Cura Daball, P., Ventura Ferreira, M. S., Ammann, S., Klemann, C., Lorenz, M. R., Warthorst, U., Leahy, T. R., Conlon, N., Roche, J., Soler-Palacín, P., Garcia-Prat, M., Fuchs, I., Fuchs, S., Beier, F., Brümmendorf, T. H., Speckmann, C., Olbrich, P., Neth, O., Schwarz, K., … Rensing-Ehl, A. (2018). CD57 identifies T cells with functional senescence before terminal differentiation and relative telomere shortening in patients with activated PI3 kinase delta syndrome. *Immunology and Cell Biology*.

308. Xing LM, Liu CY, Fu R, Wang HQ, Wang J, Liu X, et al. . CD8(+)HLA-DR+ T cells are increased in patients with severe aplastic anemia. Mol Med Rep. (2014) 10:1252–58.

309. Ohara, T., Koyama, K., Kusunoki, Y., Hayashi, T., Tsuyama, N., Kubo, Y., & Kyoizumi, S. (2002). Memory Functions and Death Proneness in Three CD4 + CD45RO + Human T Cell Subsets . *The Journal of Immunology*.

310. Zhang, H., Chen, Y., Li, G., Zeng, H., & Chen, H. (2017). Research on the diversity of T cell receptor repertoire in the process of malignant transformation of HBV chronic infection. *Journal of Clinical Oncology*.

311. Richert-Spuhler, L. E., & Lund, J. M. (2015). The Immune Fulcrum: Regulatory T Cells Tip the Balance between Pro- and Anti-inflammatory Outcomes upon Infection. *Progress in Molecular Biology and Translational Science*.

312. Li, H. M., Hiroi, T., Zhang, Y., Shi, A., Chen, G., De, S., Metter, E. J., Wood, W. H., Sharov, A., Milner, J. D., Becker, K. G., Zhan, M., & Weng, N. -p. (2016). TCR repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *Journal of Leukocyte Biology*.

313. Liu, L., Zhang, X., & Feng, S. (2018). Epstein-Barr Virus-Related Post-Transplantation Lymphoproliferative Disorders After Allogeneic Hematopoietic Stem Cell Transplantation. *Biology of Blood and Marrow Transplantation*, *24*(7), 1341–1349.

314. Rho, H., Wells A Game of Clones: The Complex Interplay of Aplastic Anaemia, Myelodysplastic Syndrome, and Paroxysmal Nocturnal Haemoglobinuria, 2018, EMJ. 2018;3[3]:108-115.

315. Krummel, M. F., Bartumeus, F., & Gérard, A. (2016). T cell migration, search strategies and mechanisms. In *Nature Reviews Immunology*.

316. Yura, Y., Sano, S., & Walsh, K. (2020). Clonal Hematopoiesis: A New Step Linking Inflammation to Heart Failure. In *JACC: Basic to Translational Science*.

317. Guan, Y., Jakimovski, D., Ramanathan, M., Weinstock-Guttman, B., & Zivadinov, R. (2019). The role of Epstein-Barr virus in multiple sclerosis: From molecular pathophysiology to in vivo imaging. In *Neural Regeneration Research*.

318. Zhang, T., Liu, C., Liu, H., Li, L., Wang, T., & Fu, R. (2018). Epstein Barr Virus Infection Affects Function of Cytotoxic T Lymphocytes in Patients with SevereAplastic Anemia. *BioMed Research International*.

319. McEvoy, C. R. E., Morley, A. A., & Firgaira, F. A. (2003). Evidence for whole chromosome 6 loss and duplication of the remaining chromosome in acute lymphoblastic leukemia. *Genes Chromosomes and Cancer*.

320. Choo, S. Y. (2007). The HLA system: Genetics, immunology, clinical testing, and clinical implications. *Yonsei Medical Journal*.

321. Hirsch, C. M., Clemente, M., Chomczynski, P., Przychodzen, B. P., Nagata, Y., Adema, V., Williams, L., Visconte, V., Lichtin, MD, A., Mustjoki, S., Sekeres, M. A., & Maciejewski, J. P. (2018). Polyclonal Immune Response in T-LGL Leads to Clonal Expansions Preceding Occurrence of STAT3 Mutations Further Solidifying Clonal Dominance. *Blood*.

322. Zens, K., & Münz, C. (2019). Tissue resident T cell memory or how the magnificent seven are chilling in the bone. In *European Journal of Immunology*.

323. Bonomo, A., Monteiro, A. C., Gonçalves-Silva, T., Cordeiro-Spinetti, E., Galvani, R. G., & Balduino, A. (2016). A T cell view of the bone marrow. *Frontiers in Immunology*.

324. Cusick, M. F., Libbey, J. E., & Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clinical Reviews in Allergy and Immunology*.

325. Dejaco, C., Duftner, C., Grubeck-Loebenstein, B., & Schirmer, M. (2006). Imbalance of regulatory T cells in human autoimmune diseases. In *Immunolog*

326. Barennes, P., Quiniou, V., Shugay, M., Egorov, E. S., Davydov, A. N., Chudakov, D. M., Uddin, I., Ismail, M., Oakes, T., Chain, B., Eugster, A., Kashofer, K., Rainer, P. P., Darko, S., Ransier, A., Douek, D. C., Klatzmann, D., & Mariotti-Ferrandiz, E. (2020). Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology*.

327. Zhang, B., Jia, Q., Bock, C., Chen, G., Yu, H., Ni, Q., Wan, Y., Li, Q., & Zhuang, Y. (2016). Glimpse of natural selection of long-lived T-cell clones in healthy life. *Proceedings of the National Academy of Sciences of the United States of America*.

328. Savola, P., Rajala, H., & Mustjoki, S. (2016). LGL leukemia and autoimmunity - the borderline between autoimmune disease and cancer is becoming blurred. In *Duodecim; laaketieteellinen aikakauskirja*.

329. Smatti, M. K., Al-Sadeq, D. W., Ali, N. H., Pintus, G., Abou-Saleh, H., & Nasrallah, G. K. (2018). Epstein-barr virus epidemiology, serology, and genetic variability of LMP-1 oncogene among healthy population: An update. In *Frontiers in Oncology*.

330. Memon, A., & Hutchinson, D. (2017). 19. A case of Large Granular Lymphocytosis (LGL) with neutropaenic sepsis associated with positive ANA, anti-Ro, anti-Ribosomal P and anti-Neutrophil antibodies responding to Methotrexate. *Rheumatology Advances in Practice*.

331. Bonarius HP, Baas F, Remmerswaal EB, van Lier RA, ten Berge I, Tak PP, et al. Monitoring the T-cell receptor repertoire at single-clone resolution. PLoS One (2006) 1:e55.10.1371/journal.pone.0000055

332. Bigouret V, Hoffmann T, Arlettaz L, et al. Monoclonal T-cell expansions in asymptomatic individuals and in patients with large granular leukemia consist of cytotoxic effector T cells expressing the activatingCD94:NKG2C/E and NKD2D killer cell receptors. *Blood.* 2003;101:3198-3204.

333. Weyand, C. M., Yang, Z., & Goronzy, J. J. (2014). T-cell aging in rheumatoid arthritis. In *Current Opinion in Rheumatology*.

334. Khan, I. (2012). EBV Infection Resulting in Aplastic Anemia: A Case Report and Literature Review. *Journal of Blood Disorders & Transfusion*.

335. Wherry, E. J., & Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. In *Nature Reviews Immunology*.

336. Wedderburn, L. R., Patel, A., Varsani, H., & Woo, P. (2001). The developing human immune system: T-cell receptor repertoire of children and young adults shows a wide discrepancy in the frequency of persistent oligoclonal T-cell expansions. *Immunology*.

337. Shi, M., Olteanu, H., Jevremovic, D., He, R., Viswanatha, D., Corley, H., & Horna, P. (2020). T-cell clones of uncertain significance are highly prevalent and show close resemblance to T-cell large granular lymphocytic leukemia. Implications or laboratory diagnostics. *Modern Pathology*.

338. Xiao, Y., Zhao, S., & Li, B. (2017). Aplastic anemia is related to alterations in T cell receptor signaling. In *Stem Cell Investigation*.

339. Coles, C. H., Mulvaney, R. M., Malla, S., Walker, A., Smith, K. J., Lloyd, A., Lowe, K. L., McCully, M. L., Martinez Hague, R., Aleksic, M., Harper, J., Paston, S. J., Donnellan, Z., Chester, F., Wiederhold, K., Robinson, R. A., Knox, A., Stacey, A. R., Dukes, J., … Harper, S. (2020). TCRs with Distinct Specificity Profiles Use Different Binding Modes to Engage an Identical Peptide–HLA Complex. *The Journal of Immunology*.

340. Kelly, R., Richards, S., Arnold, L., Valters, G., Cullen, M., Hill, A., & Hillmen, P. (2009). A Spontaneous Reduction of Clone Size in Paroxysmal Nocturnal Hemoglobinuria Patients Treated with Eculizumab for Greater Than 12 Months. *Blood*, *114*(22), 1992.

341. Visconte, V., Raghavachari, N., Liu, D., Keyvanfar, K., Desierto, M. J., Chen, J., & Young, N. S. (2010). Phenotypic and functional characterization of a mouse model of targeted Pig-a deletion in hematopoietic cells. *Haematologica*.

342. Terrazzano, G. (2005). T cells from paroxysmal nocturnal haemoglobinuria (PNH) patients show an altered CD40-dependent pathway. *Journal of Leukocyte Biology*.

343. Schaft, N., Lankiewicz, B., Drexhage, J., Berrevoets, C., Moss, D. J., Levitsky, V., Bonneville, M., Lee, S. P., McMichael, A. J., Gratama, J. W., Bolhuis, R. L. H., Willemsen, R., & Debets, R. (2006). T cell re-targeting to EBV antigens following TCR gene transfer: CD28-containing receptors mediate enhanced antigen-specific IFNγ production. *International Immunology*.

344. AIRR Community Meeting IV: “Bridging the Gaps”.May 11-15, 2019, University of Genoa, Italy

345. Johnson, S., Gittelman, R. M., Sanders, C., & Robins, H. (2020). Determining the impact of HLA type and chronic viral infection on peripheral T-cell receptor sharing between unrelated individuals. *The Journal of Immunology*, *204*(1 Supplement), 140.16 LP-140.16.

346. Chaara, W., Mariotti-Ferrandiz, E., Gonzalez-Tort, A., Florez, L., Six, A., & Klatzmann, D. (2018). Representativeness and robustness of TCR repertoire diversity assessment by high-throughput sequencing. *BioRxiv*.

347. Crotty, S. (2014). T Follicular Helper Cell Differentiation, Function, and Roles in Disease. In *Immunity*.

348. Rabia, L. A., Zhang, Y., Ludwig, S. D., Julian, M. C., & Tessier, P. M. (2018). Net charge of antibody complementarity-determining regions is a key predictor of specificity. *Protein Engineering, Design and Selection*.

349. van Langelaar, J., Rijvers, L., Smolders, J., & van Luijn, M. M. (2020). B and T Cells Driving Multiple Sclerosis: Identity, Mechanisms and Potential Triggers. In *Frontiers in Immunology*.

350. Bessler, M., & Hillmen, P. (1998). Somatic mutation and clonal selection in the pathogenesis and in the control of paroxysmal nocturnal hemoglobinuria. In *Seminars in Hematology*.

351. Aghaeepour, N., Ganio, E. A., Mcilwain, D., Tsai, A. S., Tingle, M., Van Gassen, S., Gaudilliere, D. K., Baca, Q., McNeil, L., Okada, R., Ghaemi, M. S., Furman, D., Wong, R. J., Winn, V. D., Druzin, M. L., El-Sayed, Y. Y., Quaintance, C., Gibbs, R., Darmstadt, G. L., … Gaudilliere, B. (2017). An immune clock of human pregnancy. *Science Immunology*.

352. Aiello, A., Farzaneh, F., Candore, G., Caruso, C., Davinelli, S., Gambino, C. M., Ligotti, M. E., Zareian, N., & Accardi, G. (2019). Immunosenescence and its hallmarks: How to oppose aging strategically? A review of potential options for therapeutic intervention. In *Frontiers in Immunology*.

353. Walford, R. L. (1969). THE IMMUNOLOGIC THEORY OF AGING. *Immunological Reviews*.

354. Sakaguchi, S., Wing, K., Onishi, Y., Prieto-Martin, P., & Yamaguchi, T. (2009). Regulatory T cells: How do they suppress immune responses? In *International Immunology*.

355. Britanova, O. V., Putintseva, E. V., Shugay, M., Merzlyak, E. M., Turchaninova, M. A., Staroverov, D. B., Bolotin, D. A., Lukyanov, S., Bogdanova, E. A., Mamedov, I. Z., Lebedev, Y. B., & Chudakov, D. M. (2014). Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *The Journal of Immunology*.

356. Kampstra, A. S. B., & Toes, R. E. M. (2017). HLA class II and rheumatoid arthritis: the bumpy road of revelation. In *Immunogenetics*.

357. Kordasti, S., Costantini, B., Seidl, T., Perez Abellan, P., Martinez Llordella, M., McLornan, D., Diggins, K. E., Kulasekararaj, A., Benfatto, C., Feng, X., Smith, A., Mian, S. A., Melchiotti, R., de Rinaldis, E., Ellis, R., Petrov, N., Povoleri, G. A., Chung, S. S., Thomas, N. S., Farzaneh, F., … Mufti, G. J. (2016). Deep phenotyping of Tregs identifies an immune signature for idiopathic aplastic anemia and predicts response to treatment. *Blood*, *128*(9), 1193–1205.

358. Giudice, V., Feng, X., Lin, Z., Hu, W., Zhang, F., Qiao, W., Ibanez, M. del P. F., Rios, O., & Young, N. S. (2018). Deep sequencing and flow cytometric characterization of expanded effector memory CD8 + CD57 + T cells frequently reveals T-cell receptor Vβ oligoclonality and CDR3 homology in acquired aplastic anemia. *Haematologica*.

359. Kaech, S. M., Wherry, E. J., & Ahmed, R. (2002). Effector and memory T-cell differentiation: Implications for vaccine development. In *Nature Reviews Immunology*.

360. McNamara, L. A., Topaz, N., Wang, X., Hariri, S., Fox, L., & MacNeil, J. R. (2017). High Risk for Invasive Meningococcal Disease Among Patients Receiving Eculizumab (Soliris) Despite Receipt of Meningococcal Vaccine. *MMWR. Morbidity and Mortality Weekly Report*.

361. Lupsa, N., Érsek, B., Horváth, A., Bencsik, A., Lajkó, E., Silló, P., Oszvald, Á., Wiener, Z., Reményi, P., Mikala, G., Masszi, T., Buzás, E. I., & Pós, Z. (2018). Skin-homing CD8 + T cells preferentially express GPI-anchored peptidase inhibitor 16, an inhibitor of cathepsin K. *European Journal of Immunology*.

362. Rosse, W. F. (1990). Phosphatidylinositol-linked proteins and paroxysmal nocturnal hemoglobinuria. *Blood*, *75*(8), 1595–1601.

363. Bessler, M., Mason, P. J., Hillmen, P., Miyata, T., Yamada, N., Takeda, J., Luzzatto, L., & Kinoshita, T. (1994). Paroxysmal nocturnal haemoglobinuria (PNH) is caused by somatic mutations in the PIG-A gene. *EMBO Journal*.

364. Krawitz, P. M., Höchsmann, B., Murakami, Y., Teubner, B., Krüger, U., Klopocki, E., Neitzel, H., Hoellein, A., Schneider, C., Parkhomchuk, D., Hecht, J., Robinson, P. N., Mundlos, S., Kinoshita, T., & Schrezenmeier, H. (2013). A case of paroxysmal nocturnal hemoglobinuria caused by a germline mutation and a somatic mutation in PIGT. *Blood*.

365. Kinoshita, T. (2014). Biosynthesis and deficiencies of glycosylphosphatidylinositol. In *Proceedings of the Japan Academy Series B:Physical and Biological Sciences*.

366. Almeida, A. M., Murakami, Y., Baker, A., Maeda, Y., Roberts, I. A. G., Kinoshita, T., Layton, D. M., & Karadimitris, A. (2007). Targeted therapy for inherited GPI deficiency. *New England Journal of Medicine*.

367. Hosokawa, K., Kajigaya, S., Keyvanfar, K., Qiao, W., Xie, Y., Townsley, D. M., Feng, X., & Young, N. S. (2017). T Cell Transcriptomes from Paroxysmal Nocturnal Hemoglobinuria Patients Reveal Novel Signaling Pathways. *The Journal of Immunology*.

368. LIU, Y. A. N., Wang, S., Zuo, X., & Peng, J. (2020). Landscape of bone marrow T cells in acquired aplastic anaemia revealed by single-cell sequencing. *The Journal of Immunology*, *204*(1 Supplement), 224.31 LP-224.31.

369. Takahama, Y., Ohishi, K., Tokoro, Y., Sugawara, T., Yoshimura, Y., Okabe, M., Kinoshita, T., & Takeda, J. (1998). Functional competence of T cells in the absence of glycosylphosphatidylinositol-anchored proteins caused by T cell-specific disruption of the Pig-a gene. *European Journal of Immunology*.

370. Chen, Y., Veracini, L., Benistant, C., & Jacobson, K. (2009). The transmembrane protein CBP plays a role in transiently anchoring small clusters of Thy-1, a GPI-anchored protein, to the cytoskeleton. *Journal of Cell Science*.

371. Wang, L. N., Gao, M. H., Wang, B., Cong, B. B., & Zhang, S. C. (2018). A role for GPI-CD59 in promoting T-cell signal transduction via LAT. *Oncology Letters*.

372. Lipid rafts, major histocompatibility complex molecules, and immune regulation. *Goebel J, Forrest K, Flynn D, Rao R, Roszman TL. Hum Immunol. 2002 Oct; 63(10):813-20.*

373. Waggener, Bill (1995). *Pulse Code Modulation Techniques*. Springer. p. 206. ISBN 9780442014360. Retrieved 13 June 2020.

374. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*.

375. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L., & Marioni, J. C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature Communications*.

376. De Simone, M., Rossetti, G., & Pagani, M. (2018). Single cell T cell receptor sequencing: Techniques and future challenges. In *Frontiers in Immunology*.

377. https://international.neb.com/tools-and-resources/usage-guidelines/using unique-molecular-ids-with-directed-genomics-datausage-guideline-page

378. Chung, J., Lee, K. W., Lee, C., Shin, S. H., Kyung, S., Jeon, H. J., Kim, S. Y., Cho, E., Yoo, C. E., Son, D. S., Park, W. Y., & Park, D. (2019). Performance evaluation of commercial library construction kits for PCR-based targeted sequencing using a unique molecular identifier. *BMC Genomics*.

379. Brüggemann, M., Kotrová, M., Knecht, H., Bartram, J., Boudjogrha, M., Bystry, V., Fazio, G., Froňková, E., Giraud, M., Grioni, A., Hancock, J., Herrmann, D., Jiménez, C., Krejci, A., Moppett, J., Reigl, T., Salson, M., Scheijen, B., Schwarz, M., … Langerak, A. W. (2019). Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia*.

380. Turchaninova, M. A., Britanova, O. V., Bolotin, D. A., Shugay, M., Putintseva, E. V., Staroverov, D. B., Sharonov, G., Shcherbo, D., Zvyagin, I. V., Mamedov, I. Z., Linnemann, C., Schumacher, T. N., & Chudakov, D. M. (2013). Pairing of T-cell receptor chains via emulsion PCR. *European Journal ofImmunology*.

381. Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J. M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Giang Phan, T., Junankar, S., Jackson, K., Goodnow, C. C., Smith, M. A., & Swarbrick, A. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes.*Nature Communications*.

382. Gielis, S., Moris, P., Neuter, N. De, Bittremieux, W., Ogunjimi, B., Laukens, K., & Meysman, P. (2018). TCRex: a webtool for the prediction of T-cell receptor sequence epitope specificity. *BioRxiv*.

383. Greiff, V., Weber, C. R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., & Reddy, S. T. (2017). Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *The Journal of Immunology*.

384. Betensky, M., Babushok, D., Roth, J. J., Mason, P. J., Biegel, J. A., Busse, T. M., Li, Y., Lind, C., Papazoglou, A., Monos, D., Podsakoff, G., Bessler, M., & Olson, T. S. (2016). Clonal evolution and clinical significance of copy number neutral loss of heterozygosity of chromosome arm 6p in acquired aplastic anemia. *Cancer Genetics*.

385. Stanley, N., Olson, T. S., & Babushok, D. V. (2017). Recent advances in understanding clonal haematopoiesis in aplastic anaemia. In *British Journal of Haematology*.

386. 10X Genomics. (2018). Single Cell 3 ' Reagent Kits v2 User Guide. *10X Genomics*.

387. Khosravi-Maharlooei, M., Obradovic, A., Misra, A., Motwani, K., Holzl, M., Seay, H. R., DeWolf, S., Nauman, G., Danzl, N., Li, H., Ho, S. hong, Winchester, R., Shen, Y., Brusko, T. M., & Sykes, M. (2019). Cross-reactive public TCR sequences undergo positive selection in the human thymic repertoire. *Journal of Clinical Investigation*.

388. Scheinberg, P., & Young, N. S. (2012). How I treat acquired aplastic anemia. *Blood*, *120*(6), 1185–1196.

389. Alfinito, F., Ruggiero, G., Sica, M., Udhayachandran, A., Rubino, V., Pepa, R. Della, Palatucci, A. T., Annunziatella, M., Notaro, R., Risitano, A. M., & Terrazzano, G. (2012). Eculizumab treatment modifies the immune profile of PNH patients. *Immunobiology*.

390. https://www.antibodysociety.org

391. Putintseva, E. V., Britanova, O. V., Staroverov, D. B., Merzlyak, E. M., Turchaninova, M. A., Shugay, M., Bolotin, D. A., Pogorelyy, M. V., Mamedov, I. Z., Bobrynina, V., Maschan, M., Lebedev, Y. B., & Chudakov, D. M. (2013). Mother and child T cell receptor repertoires: Deep profiling study. *Frontiers in Immunology*.

392. Figures 4,5,6,7,13 and 21 created in part of full with BioRender.com.

393. Spellerberg, Ian F., and Peter J. Fedor. (2003) A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. Global ecology and biogeography 12.3, 177-179.

394. Babushok, D. V., Stanley, N., Xie, H. M., Huang, H., Bagg, A., Olson, T. S., & Bessler, M. (2017). Clonal Replacement Underlies Spontaneous Remission in Paroxysmal Nocturnal Haemoglobinuria. In *British Journal of Haematology*.

395. Korkama, E. S., Armstrong, A. E., Jarva, H., & Meri, S. (2018). Spontaneous remission in paroxysmal nocturnal hemoglobinuria-Return to health or transition into malignancy? *Frontiers in Immunology*.

396. Killian, M. S., Matud, J., Detels, R., Giorgi, J. V., & Jamieson, B. D. (2002). MaGiK method of T-cell receptor repertoire analysis. *Clinical and Diagnostic Laboratory Immunology*.

397. Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*.