University College London

Investigating the Role of the T-Cell Receptor Using Targeted Capture and High-Throughput Sequencing

Lisa Louise Carter

Submitted for the degree of Doctor of Philosophy

Department of Immunology, Inflammation and Rheumatology

Institute of Child Health, University College London

Declaration

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

Acknowledgements

Firstly, I would like to express my appreciation for Robin Callard and Nigel Klein for their supervisory support during my time at Great Ormond Street Institute of Child Health. In addition, I am grateful for the support I received from Kenth Gustafsson, Mona Bajaj-Elliott, who very kindly stepped up when I needed them. I also thank the team at Great Ormond Street Children's Hospital, including Graham Davies, Paul Veys and Stuart Adams, for their assistance with clinical work, and to the UCL Genomics team, in particular Tony Brooks and Mark Kristiansen. Thanks also to John Anderson and Jonathan Fisher, who welcomed me into their group.

Secondly, office 646 where I sat for three years deserves a huge thank you for all of the support and good times. In particular, I would never have kept my sanity without regular tea breaks with Ronan Doyle, Garth Dixon, Hannah Poulsom, Rollo Hoare, Ariane Standing, Claire Walker and Phil Dusart. The wonderful Vania de Toledo deserves huge appreciation for her efforts at keeping us all in line.

The past few years have brought a lot of changes, so finally, I would like to thank the people that were there for it all and kept me grounded. Thank you to my parents, Dorothy and David Carter, to my sister Carolyn Bennett, and to Robert Jaques, Daniel Higgins, Evelyne Chaignat, Luke Meredith and Jan Baumann.

Abstract

Maintaining a diverse immune repertoire is crucial for protection against a wide range of pathogens. Until recently, it has been difficult to quantify this diversity and define the range of a repertoire in a healthy individual. The rise of massively parallel high-throughput sequencing has enabled researchers to gather more information than ever before, but many published works concentrate on describing individual T-cell and immunoglobulin chains. This report introduces a novel method for sequencing all α, β, γ and δ chains of the T-cell receptor and all immunoglobulin chains simultaneously, using high-throughput sequencing and targeted capture. Data was obtained through this method using two sequencing platforms, the Illumina MiSeq and the Ion Torrent Personal Genome Machine, and the analysis of data focused on the T-cell receptor using diversity measures borrowed from other scientific fields.

This work demonstrated the successes and limitations of the capture technique and suggests that immune repertoire sequencing could have dramatic impact on the understanding of the immune system across a range of disease states. Therefore, a preliminary investigation was carried out into the reconstitution of the immune repertoires following paediatric haematopoeitic stem cell transplants in the treatment of acute myeloid leukaemia.

Acronyms & Abbreviations

ADCC	Antibody dependent cell mediated cytotoxicity
Ag	Antigen
AML	Acute myeloid leukaemia
aAPC	Artificial antigen presenting cell
APC	Antigen presenting cell
ART	Antiretroviral therapy
ATG	Anti-thymocyte globulin
bp	Base pair
BWA	Burrows-Wheeler alignment
BWT	Burrows-Wheeler transform
CBT	Cord blood transplant
CCR	Chemokine receptor
CD	Cluster of differentiation
cDNA	Complementary DNA
CDR	Complementarity determining region
CMV	Cytomegalovirus
DMAPP	Dimethylallyl pyrophosphate
DMSO	Dimethyl sulfoxide
DN	Double negative
DNA	Deoxyribonucleic acid
DP	Double positive
ds	Double stranded
EBV	Epstein Barr virus

ENST	Ensembl gene spliced transcript number
FACS	Fluorescence-activated cell sorting
FCS	Foetal calf serum
FLASH	Fast length adjustment of short reads
G-CSF	Granulocyte colony-stimulating factor
GM-CSF	Granulocyte -macrophage colony stimulating factor
GOSH	Great Ormond Street Hospital
GVHD	Graft versus host disease
HIV	Human immunodeficiency virus
HMB-PP	(E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate
HS	High sensitivity
HSC	Haematopoeitic stem cell
HSCT	Haematopoeitic stem cell transplant
HTS	High-throughput sequencing
IFN	Interferon
Ig	Immunoglobulin
IGV	Integrated Genome Viewer
IL	Interleukin
IMGT	International Immunogenetics Information System
IPP	Isopentenyl pyrophosphate
ISFET	Ion-sensitive field-effect transistor
ISP	Ion Sphere Particles
MAMP	Microbe associated molecular pattern
MHC	Major histocompatibility complex

MHC-Ag	Major histocompatibility complex - antigen complex
MiTCR	Mi T-cell receptor
MRD	Minimal residual disease
mRNA	Messenger RNA
MUD	Matched unrelated donor
ng	Nanogram
nt	Nucleotide
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PAMP	Pathogen associated molecular pattern
PBMC	Peripheral blood mononucleocytes
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PGM	Personal Genome Machine
PID	Primary immunodeficiencies
PRR	Pattern recognition receptor
RACE	Rapid amplification of cDNA ends
RAG	Recombination activating gene
RIN	RNA Integrity number
RNA	Ribonucleic acid
RPMI	Roswell Park Memorial Institute
rRNA	Ribosomal RNA
RTE	Recent thymic emigrant
SAM	Sequence alignment map

SCF	Stem cell factor
TBI	Total body irradiation
TCD	T-cell depletion
TCR	T-cell receptor
TdT	Terminal deoxynucleotidyl transferase
TLR	Toll-like receptor
TRAV	T-cell receptor alpha variable
TRBV	T-cell receptor beta variable
TRDV	T-cell receptor delta variable
TRECs	T-cell receptor excision circles
TRGV	T-cell receptor gamma variable
tRNA	Transfer RNA
V(D)J	Variable (Diversity) Joining segments

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1 General Introduction

The immune system provides protection from pathogens including bacteria, viruses and parasites. It is a complex system made up of biological processes that can be divided into innate and adaptive arms. The innate immune system is a feature of both vertebrates and invertebrates alike and is the first line of host defense against microbial invasions (1). Innate immunity relies on the ability of germ line encoded receptors to recognise microbes. These receptors include pattern recognition receptors (PRRs), such as toll-like receptors (TLRs). These recognise pathogen associated molecular patterns (PAMPs) and microbe associated molecular patterns (MAMPs) which can be the first signals to the host that a response is necessary (2). The innate inflammatory response is mediated through cells such as monocytes, macrophages and polymorphonuclear leukocytes, and the release of cytokines and chemokines.

In addition to the innate immune system, vertebrates have also evolved an adaptive immune response that can be generally characterised by its capacity to generate antigen specific responses through the recombination of genes encoding their recognition receptor and the feature of immunological memory. This recombination results in a vast immune repertoire with the capacity to respond to a wide range of antigenic challenges. The cellular components to the adaptive immune response are the T and B-lymphocytes. Both kinds of lymphocytes derive their names from the specialised organs in which they mature, for the T-lymphocyte this is the thymus and for the B-lymphocyte this is named for the Bursa of Fabricus found in birds, although in humans, they mature in the bone marrow (3). T lymphocytes, commonly referred to as T-cells, underpin cell-mediated immunity whereas B-lymphocytes, or B-cells, underpin humoral immunity through antibody responses. Both T-cells and B-cells have germ line encoded receptors that undergo a process of recombination. These recombinatorial systems are thought to have appeared in both jawed and non-jawed vertebrates early on in their evolution, during the Cambrian period of the Paleozoic era, which was around 500 million years ago (4, 5).

The quantity of different T-cell and B-cell receptor combinations present in the peripheral blood stream and the frequency of each rearrangement are what begin to define the diversity of an immune repertoire of an individual. The success of the adaptive response depends on this diversity. The analysis of the peripheral repertoire may introduce a bias in the understanding of the entire repertoire of an individual, as the populations of cells present varies between different organs. In particular, there are sub populations of tissue resident memory cells that are maintained independently to circulating cells (6, 7). However, this is the most practical method of investigating repertoire due to the limitations of sampling human organs.

A highly diverse repertoire is required in order to appropriate an immune response to a wide range of potential disease-causing agents. It has been hypothesised that a decrease in diversity in certain disease states may affect the long-term health of an individual, for example, during the aging process, infection with human immunodeficiency virus (HIV), or following bone marrow transplantation (8, 9). The focus of this work in particular will be on investigating this dynamic diversity of the T-cell receptor (TCR) repertoire.

1.1 T-Lymphocytes and Major Histocompatibility Complex Interactions

T-cells are a crucial part of the adaptive immune system. They form a coordinated response to antigen exposure, producing cytokines to induce downstream effector functions or exhibiting direct cytotoxic activity. Each T-cell has a receptor on its cell surface, called the TCR, which interacts with the major histocompatibility complex (MHC). The MHC is a highly polymorphic family of molecules with great inter and intra individual variation. These molecules fall into two categories relating to antigen presentation including class I, which is expressed on the surfaces of all nucleated cells and class II, which is usually only expressed on the surfaces of professional antigen presenting cells (APCs), particularly dendritic cells, but also B-cells and macrophages (1). There is also an MHC class III, the genes of which encode cytokines and proteins involved in the complement cascade (10). The primary functions of both class I and class II MHC molecules are to secure fragments of antigens and expose them on the cell surface for presentation to T-cells.

APCs phagocytose potential pathogens and internally process them before presenting fragments of proteins derived from them on their cell surfaces via the MHC, for the adaptive immune system. When the APCs encounter a T-cell, its TCR will bind to these MHC-antigen (MHC-Ag) complexes and respond accordingly.

1.2 T-Lymphocyte Subsets

T-cells can be distinguished from other cell types by their expression of the CD3 marker. The CD3 molecule comprise four subunits, CDγ, CDδ and two CDε molecules, which associate with the ζ chain and the TCR, both of which are also present on the surface of T-cells, to form the TCR complex. The TCR complex, when activated, generates an intracellular signal to perform T-cell effector functions (11, 12). Cell surface expression of molecules other than CD3 further defines many subsets of T-cells. The first subtype differentiation to consider is αβ+ T-cells and γδ+ T-cells, as defined by their expression of a TCR made up of α and β subunits or γ and δ subunits. The most abundant of these subsets that can be found in peripheral blood are the αβ+ T-cells, which represent over 90% of all circulating T-cells in healthy individuals (13). However, the relative proportions of αβ+ and γδ+ T-cells in other organs varies (14).

T-cells express different set of molecules required for differentiation and downstream effector functions. The expression of these molecules is dynamic and therefore can shift following activation or during different stages of development. As a consequence, some of those molecules can be used as markers of differentiation and function in order to phenotypically describe cell subtypes.

1.2.1 αβ+ T-Lymphocytes: Helper and Cytotoxic

When in the periphery, and once a T-cell has bound an MHC-antigen complex complementary to its TCR, it becomes activated, rapidly undergoing clonal expansion and carrying out effector functions. The effector functions depend upon the subtype of T-cell. αβ+ T-cells can be sub categorised further depending on whether they express CD4 or CD8 glycoproteins on their surfaces. Expression of CD4 or CD8 is determined during T-cell maturation in the thymus.

CD4+ T-cells, often referred to as helper T-cells, provide assistance to other cells of the immune system during an immune response. This can include antigen specific signaling prompts to activate B-cell antibody production and the activation of cytotoxic T-cells (1). Activation of CD4+ T-cells happens through the interaction between its TCR and a specific formation of antigen-MHC complex found on the surface of an antigen-presenting cell. CD4+ T-cells comprise further subtypes of CD4+ regulatory cells, T_H1, T_H2 and T_H17 which mediate other immune responses. T_H1 cells secrete cytokines such as interferon-γ (IFNγ) that activate macrophages and stimulate B-cell production of antibody, T_H2 cells drive immunoglobulin E (IgE) by secretion of IL-4, and T_H17 cells mediate the immune response of macrophages and neutrophils (15-17).

CD8+ T-cells are also referred to as cytotoxic T-cells. They play a much more direct role in the killing of cells that are infected or cancerous through the release of cytotoxic substances including perforin, granzymes and granulysin (18). Cell death is induced through the action of perforin, which creates holes in the cell membrane, allowing entry of the cytotoxic substances. Once in the cytoplasm, these substances trigger apoptosis through the activation of the caspase cascade. Apoptosis can also be triggered by cytotoxic T-cells through the binding of Fas ligands between activated CD8+ T-cells and their target (19, 20).

However, the dividing lines between T-cell subsets can be blurred and the functions of cells may overlap, suggesting that rigid definition of T-cell subsets may represent an over simplistic view of T-cell dynamics.

1.2.2 αβ+ T-Lymphocytes: Naïve and Memory

Once matured T-cells exit the thymus as naïve T-cells, initially as CD31+ recent thymic emigrants (RTEs), but this marker is rapidly lost upon proliferation in the periphery.

Naïve T-cells circulate in the periphery and are constantly exposed to MHC-antigen complexes on the surfaces of APCs but have yet to encounter their cognate antigen.

Following antigen exposure, the T-cell will receive stimulatory signals from the APC in the draining lymph node of the affected organ, the strength of which affects the cell's subsequent differentiation. Strong stimulatory signals will result in terminal differentiation to an effector T-cell (T_{EFF}). However, signals of lesser strength may result in differentiation to a memory stem cell (T_{SCM}), central memory cell (T_{CM}) or effector memory cell (T_{EM}) (21). T-cell function depends on the stage of differentiation the cell has reached, as well as the identity of the coreceptor.

In humans, naïve T-cells can be phenotypically distinguished from memory cells by the expression of an isoform of the surface marker CD45. CD45, also known as lymphocyte common antigen, is a receptor linked tyrosine phosphatase (22). CD45 is expressed by T-cells in different isoforms; naïve T-cells express the CD45RA isoform whereas memory T-cells express the CD45RO isoform. However, there is growing evidence that some antigen-experienced T-cells may upregulate the CD45RA isoform (23). Therefore, further markers are required in order to improve characterisation of naïve and memory T-cells. There is a gradient of expression of surface markers, and there are still some exceptions to the rules. Therefore, in order to phenotype naïve and memory cells, it is necessary to include further surface markers, such as CCR7, CD62L, CD27 and CD28 which are primarily expressed on naïve T-cells help distinguish the cell-types (21).

1.2.3 γδ+ T-Lymphocytes

γδ+ T-cells represent the remaining 5-10% of circulating T-cells and are present in higher proportions in different organs. γδ+ T-cells exhibit both adaptive and innate properties (13). Data concerning the analysis of the repertoire of γδ+ TCRs is limited, and therefore this work may help to illuminate more information upon this subject.

γδ+ T-cells form a complex bridge between different cells of the immune system. They express a TCR generated through gene segment recombination that can produce antigen-specific responses and can potentially exhibit a memory phenotype, as studies of mycobacterial challenge have shown (24-26). γδ+ T-cells are unique in that they are able to recognise intact proteins and non-peptide compounds, independent of MHC, whereas αβ+ T-cells strictly recognise MHC-bound antigen (27-29).

Gene segment expression in γδ+ T-cells is tissue specific. For example in humans, Vγ9/Vδ2+ cells are the dominant subset amongst the γδ+ T-cell populations found in peripheral blood (29). This subset respond to an intermediate molecule of the non-mevalonate pathway, called (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate (HMB-PP) (30). This pathway is involved in the isoprenoid biosynthesis and occurs in many organisms. It is of particular importance in bacteria, such as the pathogen *Mycobacterium tuberculosis* (31). This pathway results in the production of isopentenyl pyrophosphate (IPP) which is structurally similar but less potent than HMB-PP. HMB-PP and IPP are phosphoantigens. Mammalian cells may also produce IPP through the mevalonate pathway, usually in cells under metabolic stress, for example tumor cells (32, 33). IPP is routinely used to induce the expansion of Vγ9/Vδ2 T cells *ex-vivo* (34).

Effector functions carried out by γδ+ T-cells include production of pro-inflammatory cytokines and the direct killing of infected cells *via* the release of perforin and granulysin. γδ+ T-cell cytotoxicity can also extend to tumor killing properties. Vγ9-Vδ2+ T-cells are the most well-characterised of the subsets, other subsets may have more potent antitumor properties and therefore require attention (35-37).

1.3 αβ+ T-Cell Development in the Thymus

Lymphocyte development occurs in specialised lymphoid organs. Both T and B cells share a common precursor cell, the haematopoeitic stem cell (HSC) , which is derived from red bone marrow, and defined through its surface expression of CD34 (38). Whereas B-cell progenitors remain in the bone marrow and undergo the maturation process *in-situ*, T-cell progenitors are mobilised to the periphery and migrate to the thymus, where the maturation process of T-cells continues (39-41). The environment within the thymus contains specialised stromal and epithelial cells upon which T-cells are dependent for maturation (42-44). Progenitor T-cells receive specific cell contact-dependent signals that drives their differentiation into different subtypes, and it is during this time that the T-cell repertoire is initially formed.

The thymus provides an environment that allows interactions that promote the development of T-cells, through the secretion of signaling molecules and binding between receptors. The thymus structure is made up of many lobules, these include an inner medulla and outer cortex, which each provide different microenvironments for immature T-cells, or thymocytes, during their development. Thymocytes migrate through these microenvironments, each of which is important for a different developmental stage. Each stage of T-cell differentiation in the thymus can be described by the expression of CD4 and CD8. The initial immature thymocytes are often referred to as double negative (DN) thymocytes, due to their lack of expression of either coreceptor (45). DN cells can be further subdivided into four groups according to their expression of CD25 and CD44 cells, which have been shown to be functionally distinct (46). These groups include CD25-CD44+, followed by CD25+CD44+, then CD25+CD44- and finally the thymocytes lose expression of both receptors and become CD25-CD44-. Differences in functionality include their responsiveness to interleukin-7 (IL-7) and stem cell factor (SCF). Due to regulation of the expression of c-kit, the SCF receptor, cells that are CD25+CD44+ are more responsive to these stimulants than cells that are CD44- (47).

At this point during the DN stage, one of the first lineage determining events occurs. Immature thymocytes undergo somatic recombination of variable, diversity and joining gene segments (V(D)J) at the TCR loci, at which point they diverge into αβ+ T-cells or γδ+ T-cells. As this is one of the main focal points of the present study, V(D)J recombination will be described in more detail in section 1.5.

Following the successful expression of a TCR, the immature T-cells continue to develop. Many thymocytes that generate a TCR will still be non-functional, so the next downstream event is a checkpoint that demonstrates their ability to bind MHC, a process referred as ‘positive’ selection. The coreceptor CD4 is required for a T-cell to bind MHC-II and CD8 is required to bind MHC-I, therefore at this stage, the cells upregulate both coreceptors. This is referred to as the double positive (DP) stage. If a positive signal is not received through TCR activation upon an attempt to bind MHC (no affinity for the receptor), the cell will undergo apoptosis. If a signal is received, the thymocyte will commit to a CD4+ or CD8+ lineage, becoming single positive for either one (48, 49).

CD4 and CD8 single positive thymocytes upregulate the CCR7 chemokine receptor, initiating a migration from the cortex to the medulla. During this transit, negative selection is initiated, occurring both in the cortex and medulla (50). It is possible that through the process of recombination, TCRs are generated that are auto reactive and negative selection in the healthy, should eliminate these cells. During negative selection, the T-cells that are activated following exposure to a variety of self-antigens are selected against. The fate of the immature T-cell activated at this stage can be affected by the affinity with which it binds self-antigen, as those that bind with low affinity may differentiate into regulatory T-cells. However, those that bind self-peptides bound to MHC with a high affinity will undergo apoptosis. If a thymocyte is able to pass all of these checkpoints, final maturation occurs before it is released from the thymus into the periphery. It exists as a naïve T-cell, constantly sampling MHC-Antigen complexes of APCs, until it encounters an antigen and is able to produce an effector response.

In humans, thymic released T-cells enter the periphery as naïve CD45RA+ T-cells. The rate at which naïve T-cells leave the thymus is known as thymic output, which is high in neonates and during childhood, but drops rapidly after adolescence and continues to decline with age, this output correlates with a decrease in mass of the thymus itself (51).

This process, known as thymic involution has a profound effect on the evolution of the TCR repertoire over the course of an individual's lifetime (52).

Following infection and T-cell activation, a small proportion of T-cells of the expanded clone remain in the periphery as memory T-cells. These memory cells provide long-term immune protection and can rapidly induce an effective immune response upon a subsequent encounter of the same antigen (53).

The development of γδ+ T-cells in the thymus differs and is partially determined by a family of transmembrane proteins called notch proteins, that are involved in various developmental and cell fate progressions. Notch signaling plays an important role during development of lymphocytes and is a factor in determining cell lineage (54). The strength of the notch signal may determine whether a thymocyte follows the αβ or γδ lineage (55, 56). High notch signals favor the development of γδ+ T-cells over αβ+ T-cells (57). However, γδ+ T-cells development is not as well characterised as that for αβ+ T-cells.

1.4 Structure and Function of the T-Cell Receptor

The TCR is a transmembrane protein, found on the surfaces of all subsets of T-cell. It is a heterodimeric structure, which may be composed of an α and a β chain, which accounts for 90-95% of T-cells in the blood, or a γ and a δ chain, accounting for the remaining 5-10% of T-cells (13). Each chain consists of an intracellular region, which is a short cytoplasmic tail, a transmembrane region, a constant region on the cell surface, which anchors the structure to the cell surface, and a variable region. The two chains that form the receptor are bound together through a disulphide bridge that forms between conserved cysteine residues found in both chains. The crystal structure of the human αβ TCR is represented in Figure 1.

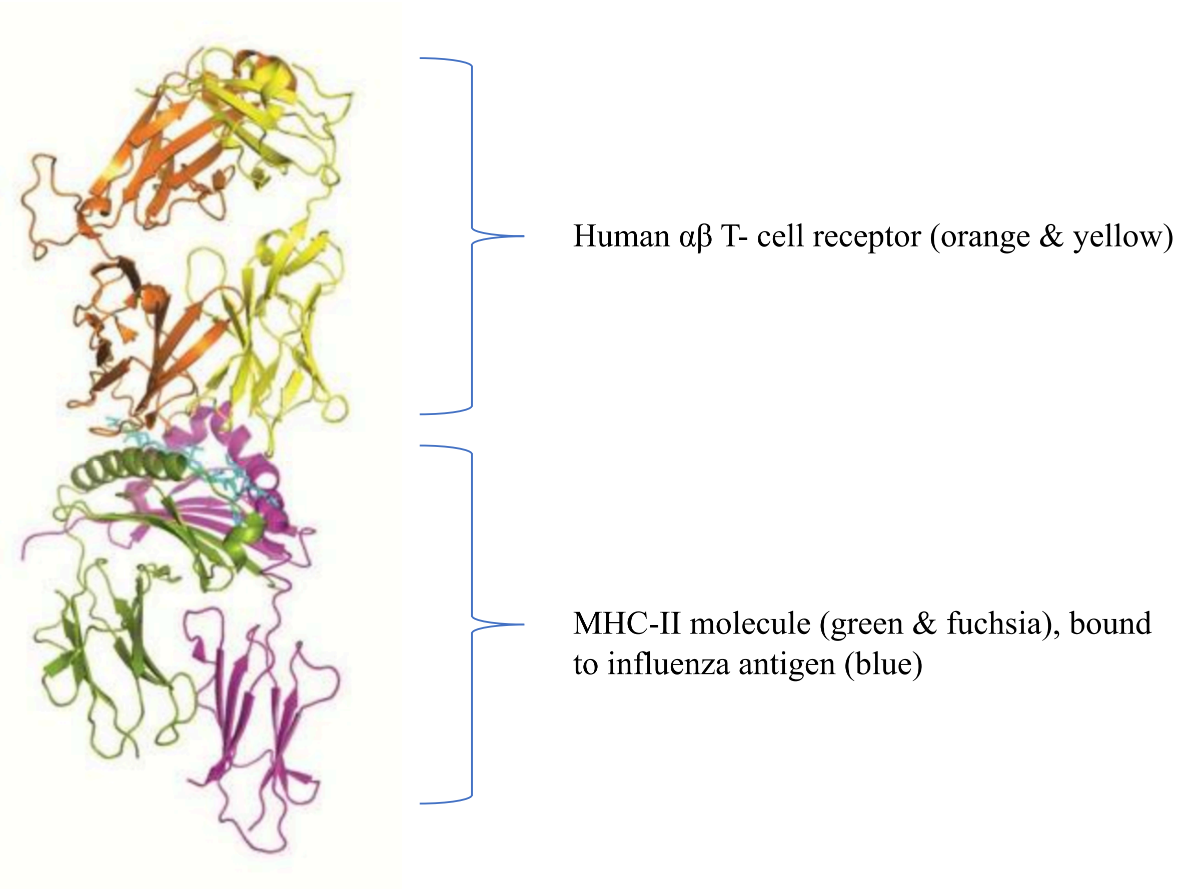

Human αβ T- cell receptor (orange & yellow)
MHC-II molecule (green & fuchsia), bound to influenza antigen (blue)

Figure 1. Crystal structure and arrangement of the T-cell receptor. The crystal structure of human αβ TCR is shown in association with the haemagglutinin antigen of an influenza virus, presented through MHC-II (58). The hypervariable region of the T-cell receptor binds antigen.

The variable region protrudes farthest from the cell membrane. This is the region responsible for antigen and MHC recognition and is therefore where the most sequence diversity occurs. It is generally divided into three complementarity determining regions (CDR), CDR1, CDR2 and CDR3. CDR1 interacts with the N-terminus of the antigenic peptide, CDR2 interacts with the MHC and CDR3 binds the bulk portion of the processed peptide. The CDR1 and CDR2 loops are less variable than the CDR3, which is hypervariable as it interacts with the most diverse sequence of the antigenic peptide. The structure of the γδ+ TCR is similar to that of the αβ+ TCR (59). However, there are a few distinct differences as a result of smaller angles between intra-structural domains in the γδ+ TCR.

In summary, activation of the TCR is dependent on concurrent activation of a coreceptor, which may be CD4 or CD8 depending on cell lineage. CD4 binds MHC-II, whereas CD8 binds MHC-I (60). The TCR associates with CD3 which is comprised of four distinct chains, the CD3γ, the CD3δ and two CD3ε chains, and also the TCRζ chain. These together form the TCR complex, which is responsible for cell activation and effector function in response to antigen recognition.

1.5 V(D)J Recombination

Immature thymocytes undergo somatic V(D)J recombination at the TCR loci, where they diverge into αβ+ T-cells or γδ+ T-cells. The germline TCR locus of each chain is organised into variable (V), joining (J) and constant (C) regions, with an additional diversity (D) segment between V and J in the β and δ loci. Each V, D and J region is made up of a series of discrete segments, defined by distinct open reading frames (ORFs), the numbers of which are summarised in Table 1 and illustrated in Figure 2. Any combination of these may be selected to produce the final sequence of the TCR. The nomenclature for TCR genes includes the TCR alpha variable (TRAV), TCR beta variable (TRBV), TCR gamma variable (TRGV) and TCR delta variable (TRDV). As some TRDV genes lie within the TRAV locus and are shared, they may be referred to as TCR alpha delta variable (TRADV).

Chain	V Segments	D Segments	J Segments	C Segments
α	54	0	61	1
β	64 to 67	2	14	2
γ	12 to 15	0	5	2
δ	8*	3	4	1

Table 1. Numbers of V, D and J gene segments of each TCR chain. Each chain can be found at a specific gene locus, with multiple ORFs, which are rearranged to create the final receptor sequence. For each rearrangement one each of V, D and J are included for each β and δ out of the total number of ORFs listed above. For each α and γ rearrangement only the V and J segments are included. The number of genes for the Vδ includes the 5 segments that are TRAV/TRADV (61).

In αβ+ T-cells, the β chain rearranges first, initiated by a wave of recombination activating gene (RAG) 1 and RAG2 expression (62, 63). A Vβ segment will rearrange to a Dβ segment, which then rearranges to a Jβ segment, followed by the transcription and splicing of the VDJβ exon to the Cβ gene segment (64). The rearranged β chain then associates with a pre-α to form the pre-TCR complex. There is then a second wave of RAG gene expression, which triggers the rearrangement of the α chain, which is similar to the β chain, but lacks a diversity (D) section so it produces only a VJC rearrangement. After synthesis of the proteins encoded in the α and β messenger ribonucleic acid (mRNA) transcripts, the two TCR chains associate, joined by disulphide bridges (65).

This rearrangement will occur at either the α and the β loci, or at the γ and the δ loci. The δ locus contains a D segment, but the γ locus does not (66). Not all of these rearrangements are successful, as some will produce a TCR that is out of frame and therefore not produce a functional protein; these are termed nonproductive rearrangements. If a productive rearrangement is not made on this first attempt, or a subsequent attempt at the equivalent loci on the other chromosome, the cell will die. Cell surface expression of a functional TCR at this stage in T-cell development is referred to as the β selection checkpoint (67).

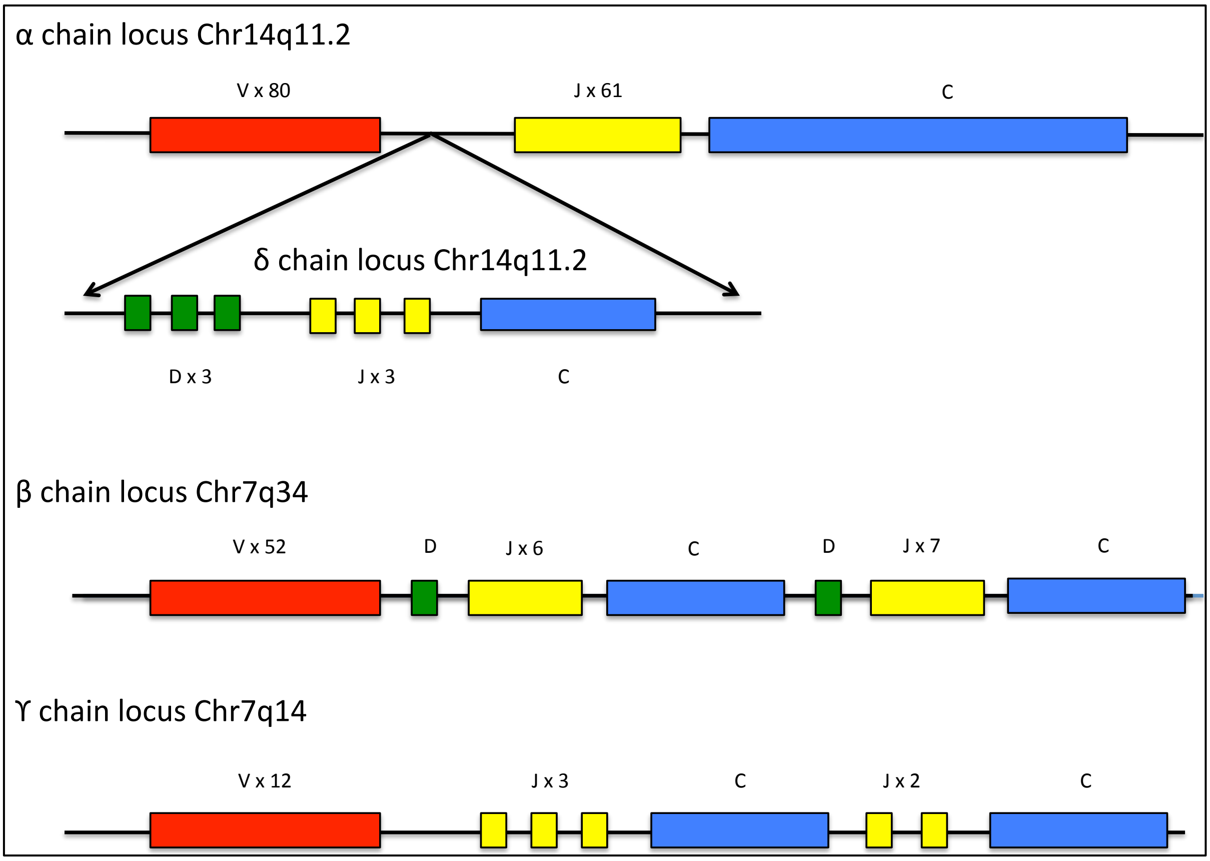


Figure 2. Organisation and Location of Human TCR α, β, γ and δ Genes. Not all gene segments are illustrated here for practical reasons, but their numbers are. Adapted from Janeway 2005 (68).

Further diversity is introduced into the CDR3 region through the addition and deletion of nucleotides at the joining regions between the V and J sections of the α and γ chains and the V, D and J sections of the β and δ chains. The process of introducing this junctional diversity begins with RAG1, RAG2, and artemis, a deoxyribonucleic acid (DNA) repair protein (64, 69). These enzymes are responsible for the removal of hairpin-like loops of nucleotide sequences that are left behind following V(D)J recombination, and for the addition of palindromic sequences, known as "P" nucleotides. Terminal deoxynucleotidyl transferase (TdT), then adds random bases at these junctions, known as "N" nucleotides. This process often leaves mismatched bases, which are removed and repaired by other enzymes as the two DNA strands anneal. It is as yet unclear how some bases are removed in another diversity-generating step. Unlike in B-cell immunoglobulin (Ig) generation, TCRs do not undergo somatic hypermutation. This limits the diversity that can be generated in the CDR1 and CDR2 loops, which may be beneficial in binding interactions with the MHC molecule. Whereas Ig molecules are able to bind free antigen, TCRs must retain their ability to bind less variable MHC complexes. It may also reduce the risk of producing self-reactive TCRs, which is more likely to induce pathogenesis than the generation of autoreactive Ig, since B-cells usually require help from T-cells in order to secrete antibody.

1.6 The T-Cell Receptor Repertoire

Maintaining a diverse repertoire of TCRs is crucial for sufficient protection against a wide range of pathogens. Therefore, it is important to be able to characterise the TCR repertoire and to define ways of measuring its diversity, to be able to compare a healthy repertoire with changes in certain disease states. Understanding and standardising these measures presents a further challenge.

V(D)J recombination results in a great number of possible TCR combinations and further diversity is generated through N and P diversity. At present it is not clear how many of these potential rearrangements leave the thymus and what the actual diversity found within an individual is. Current paradigm suggests diversity to be much lower than theoretically possible. For example, in mice there are 10^{15} potential TCR combinations, whereas an experimental estimate found that there were 2 x 10^6 distinct αβ TCRs in a healthy mouse (70, 71).

Element	α-Chain	β-Chain
V Segments	70	52
D Segments	0	2
J Segments	61	13
Joining Regions with N and P Diversity	1	2
Number of V Gene Pairs	5.8 x 10^6	
Junctional Diversity	~ 2 x 10^{11}	
Total Diversity	~ 10^{18}	

Table 2. Estimated potential diversity of the human T-cell receptor repertoire. This calculation is based on the number of theoretically possible rearrangements, rather than the number of TCRs that have been identified in test subjects (68).

Similarly, in humans, when all combinations of variable and joining genes, as well as nucleotide additions and deletions are taken into consideration, V(D)J recombination could generate a potential repertoire of up to 10^{18} beta chain clonotypes, which is shown in Table 2 (70). However, only a fraction of this potential repertoire actually exists in the human periphery, with experimental estimates of true numbers of unique clonotypes being approximately 10^6 to 10^7 different clonotypes (71-73). The TCR repertoire is dynamic and therefore there are many external factors that may influence the numbers of clonotypes throughout an individual's lifetime. For example the number of unique clonotypes decreases in old age due to thymic involution, which leads to a decrease in thymic output (74). Chronic infections, such as Epstein Barr virus (EBV) or cytomegalovirus (CMV) also play a role in the contraction of the repertoire as antigen specific clonotypes will expand over time (75, 76). Note that clonotype refers to an individual T-cell species defined by its TCR, and clone size refers to the number of individuals within that species.

Part of the reason why there are fewer circulating distinct TCRs than theoretically possible is related to the actual number of cells; an adult human has ~10^{12} circulating T-cells, which is far fewer than the potential 10^{18} different rearrangements, and some of these cells will be derived from clonal expansions (72). Additionally, many of these possible rearrangements are likely to be non-functional or may not pass positive or negative selection in the thymus, due to their particular binding affinities. Furthermore, although there are so many potential rearrangements, many CDR3 sequences may actually be shared between individuals, known as public T-cell sequences (77, 78). The likelihood of this happening by chance is very low; suggesting that there are genetic and environmental factors that contribute to the convergence seen in the diversity of the repertoire, possibly as a result of exposure to common antigen.

1.7 T-Cell Receptor Repertoire Analysis

There have been previous attempts to investigate the TCR repertoire, using a variety of methods. Spectratyping was one of the first methods to be developed (79, 80). This involves polymerase chain reaction (PCR) amplification of the CDR3 regions through the use of different primers for each V and J gene segment. CDR3 amplicons are then analysed by polyacrylamide gel electrophoresis (PAGE) to obtain the relative frequency of different length CDR3s organised by their V or J usage.

Spectratyping has been used in parallel with Sanger sequencing of CDR3 amplicons, leading to an estimate of the size of the TCR β repertoire in the peripheral blood of adult humans generating $\sim 10^6$ different clonotypes (72). The study, by Arstila et al., isolated a single amplicon through amplification of the TRBV18-TRBJ1-4 gene segments, which were then exhaustively sequenced. The resulting diversity was then extrapolated to take into account all TCR β families, and their relative expression.

Similar methods include Amplicot, which has been used to investigate the change in diversity of T-cell subsets in HIV infection and found that the overall repertoire decreases in diversity, but that diversity is maintained within subsets, suggesting that this decrease is due to a loss of total cell number, rather than a depletion of particular clonotypes (81). More recently, attempts to quantify the TCR repertoire have been made utilising high-throughput sequencing, following advancement in technologies (82, 83).

1.8 Sequencing Technologies

Sequencing technologies have improved dramatically in the past decade, and large data sets can be obtained from one sequencing experiment. Next generation sequencing refers to methods that have been developed as a higher-throughput alternative to capillary sequencing, also known as Sanger sequencing. These modern high-throughput sequencing technologies perform massively parallel sequencing of millions of short fragments of DNA.

The advent of high-throughput sequencing technologies has allowed characterisation of the TCR repertoire in greater resolution than ever before. Millions of sequencing reads can now be obtained in a few days, which would have been unfeasible using traditional Sanger sequencing due to financial and time constraints. Different sequencing technologies have been applied to the analysis of immune repertoires, for example 454 pyrosequencing (84, 85), Illumina (86-89) and Ion Torrent (90). Studies have also attempted to exhaustively sequence the entire TCR β chain repertoire of an individual (88).

These new technologies however do generate errors and therefore progress must be made to understand these errors and the consequent limitations of high-throughput antigen sequencing. For example, Illumina sequencing demonstrates a non-random error distribution and has been documented to be dependent on physical factors such as read direction, lane position and GC bias, all of which lead to errors (91, 92). This is highly relevant when sequencing the TCR, particularly the CDR3 region, as it is highly variable and therefore it is difficult to distinguish between a true variant and a sequencing error. Therefore error correcting algorithms specifically developed for use with antigen receptor sequencing are necessary to avoid overestimation of repertoire diversity (90).

Some error limiting methods have utilised the quality score assigned to each base by sequencing software. This quality score, Q, is defined as follows;

$$Q = - 10 \log_{10} (P)$$

Here, P is the probability that base call is incorrect and is related to the Phred score, which was developed during implementation of the Human Genome Project (93). Therefore, a Q score of 30 corresponds to the inferred base call accuracy of 99.9%.

Additionally, acquiring millions of sequencing reads in one run results in so much data that new ways of analysis are required to process the data. There are multiple computational methods available to do this, but each method has its own strengths and limitations, depending on which kind of datasets it was designed to contend with (94, 95).

The next section describes the processes involved in the two high-throughput sequencing platforms that were selected for use in this project in depth; the Illumina MiSeq and the Themo Fisher Ion Torrent Personal Genome Machine (PGM). These two platforms were chosen based on their suitability and practicality for this project.

1.8.1 Illumina MiSeq Platform

The Illumina platform uses sequencing by synthesis technology, which means that a template sequence is used to build a complementary strand that is used to determine the sequence of the original strand. Libraries of samples prepared for sequencing are ligated to Illumina adapters, before being injected into a flow cell and run on the sequencer. The flow cells are covered with a dense “lawn” of sequences complementary to the adapters on the sample sequences. When samples are washed through the flow cell, fragments of library hybridise to its surface (96). Library fragments are clonally amplified through a series of extensions and isothermal bridge amplification, forming millions of clusters of identical sequences. Reverse strands are cleaved and removed by washing. Following cluster amplification, the libraries are ready for sequencing. Primers bind to fragments in these clusters, and a mix of four fluorescently labeled reversibly terminated nucleotides is washed over the clusters. Each base, adenosine, cytosine, guanine and thymine, is labeled with a different fluorochrome. All four bases are washed across the template at the same time and compete with each other for binding sites. The complementary labeled base will bind to the next base in the sequence. Following each round of synthesis, a laser is used to excite the fluorochromes, which emit a specific color depending on the base to identify the locations at which different bases have bound. The fluorochrome and blocker are then removed, and the process repeated to determine the next base in the sequence.

After each addition of labeled bases, an image is taken of the flow cell, which captures information on where fluorochromes are bound, and therefore which bases have been incorporated in the chain. The clusters of sequences amplify the signal, as the optical hardware requires a certain intensity to recognise the colour. Because each individual base is read after each cycle, there is greater accuracy across high polymer frequency regions or repetitive sequences.

1.8.2 Ion Torrent Personal Genome Machine Platform

The Ion Torrent PGM is another sequencing platform utilising synthesis technology. It does not rely on optics; instead it is based on semiconductor sequencing, which detects the decrease in pH following the binding of a base in a sequence of DNA (97). This decrease occurs when a covalent bond is formed between nucleotides releasing a pyrophosphate, which is also the basis of Roche 454 sequencing, and a hydrogen ion (98). In this case, template sequences bind to microscopic beads and amplification of library occurs in emulsion, amplifying the sequence across the bead.

Instead of a flow cell, Ion Torrent technology relies on a chip containing micro wells. A single bead bound to an individual template sequence occupies a single micro well. The chip is flooded with each (unmodified) base in turn, and if that particular base is complementary to the next base in the sequence, it will be incorporated into the chain and an ion-sensitive field-effect transistor (ISFET) detects the corresponding decrease in pH (99). If more than one base has been incorporated in the sequence, there will be proportionally a greater signal, and so the system is able to estimate the length of homopolymer regions. However, as this is estimation, there is a higher rate of sequencing error associated with reads containing many homopolymers (100).

1.9 Use of High-Throughput Sequencing to Investigate the Repertoire

The iteration of high-throughput sequencing technologies has enabled the acquisition of more data, which is beginning to illuminate the dynamics of both T-cell and B-cell repertoires (84, 86-88, 101, 102). One study produced 1.7bn paired end reads, from which 1,062,522 distinct TCR β CDR3 were obtained, highlighting the magnitude of TCR β clonotypes that exist in peripheral blood (88). A method utilising 5' rapid amplification of cDNA ends (5'RACE), a technique based on PCR amplification of known gene segment sequences, was used to amplify TCR β transcripts prior to sequencing. This study, by Warren et al., also compared sequences between individuals and found that up to 1.1% of CDR3 nucleotide sequences may be shared between people, but that this rises to up to 14.2% of translated amino acid sequences, suggesting that there is some sequence redundancy and that there are certain public T-cell sequences found across individuals.

1.10 Application of Diversity Estimates to Sequencing Data

Methods of calculating diversity of the repertoire from high-throughput sequencing data can be adapted from other fields. For example, comparisons can be drawn between clonotype diversity of the TCR repertoire and species diversity within an ecological population. Similarities include the need to determine the presence of different species; in this case species would refer to a particular T-cell clonotype, and the density of that species in the environment, a parallel to clone size. The use of diversity indices may include species richness estimators, similarity indices such as the Jaccard and Morisita-Horn, or dispersion metrics, such as the Simpson Index, and Shannon and Renyi entropies. Previous studies have used the Simpson, Shannon and Gini indices to monitor the TCR repertoire of HIV specific CD8 T-cells during antiretroviral therapy (ART) and after haematopoeitic stem cell transplant (HSCT) (9, 103, 104).

1.11 Use of Targeted Capture to Sequence TCR Repertoires

Immune repertoires have the potential to provide a vast amount of clinically useful information but, until recently, have been difficult to investigate in detail. Most attempts to describe the TCR repertoire have relied on amplifying from the variable and constant segments, using a different primer for each variable segment, such as in spectratyping and sequencing experiments. However, due to the nature of the amplification process, published methods very often focus on a specific subset of the repertoire. The most-well characterised of these in existing literature is the TRBV chain.

This thesis describes a novel method of simultaneously acquiring data on all TCR and Ig chains, using Agilent’s SureSelect Target Enrichment system, and a bespoke set of RNA probes. These RNA probes are specifically designed to be complementary to all Variable, Joining and Constant gene segments, and can therefore capture these sequences from a larger sample library for sequencing. This method can be adapted in multiple ways to increase throughput and cost-effectiveness, including the use of pre-capture pooling of samples. In the future, it may be possible to automate for potential clinical purposes.

1.12 Aims of this Study

The primary aim of this study was to develop a novel targeted capture method to sequence the TCR repertoire in peripheral blood. Following this, attempts were made to quantitatively describe the diversity of the repertoire and to follow the TCR repertoire during recovery from HSCTs. As a long-term objective, stored patient samples may be used for future sequencing, in order to investigate the diversity of the TCR repertoire across immune reconstitution following treatments for immune disorders, including leukemia and Di George syndrome.

Chapter 2

Materials & Methods

2 Materials & Methods

This chapter details the reagents and general experimental methods used to obtain the data described in this work. The list of reagents and materials below applies throughout this work, unless otherwise specified.

All patients were seen at Great Ormond Institute of Child Health (GOSH) National Health Service Foundation Trust, United Kingdom. All data and samples were anonymised prior to analysis.

2.1 Reagents and Materials

This section lists all reagents and materials that were used to generate data included in this work.

2.1.1 General Reagents

Reagent	Supplier	Catalogue Number
Chloroform	Sigma-Aldrich	C2432
FCS	Thermo Fisher Scientific	10-082-147
Isopropanol	Sigma-Aldrich	I9516
Kimwipes	Sigma-Aldrich	Z188956
Lithium heparin	Sigma-Aldrich	9045-22-1
Lymphoprep	Stemcell	07851
Nuclease-free water	Ambien	AM9930
P5 Illumina reverse complement oligo	Sigma-Aldrich	Custom
P7 Illumina oligo	Sigma-Aldrich	Custom
PBS	Thermo Fisher Scientific	10010023
RNase Zap	Sigma-Aldrich	R2020
RPMI 1640 Media	Thermo Fisher Scientific	11875093
Sodium acetate	Sigma-Aldrich	W302406
Sodium hydroxide	Sigma-Aldrich	1091371000
TRIzol	Invitrogen	15596018
Tryphan blue	Thermo Fisher Scientific	15250061

Table 3. General reagents used throughout this work. All general reagents not otherwise specified are included here.

2.1.2 General Kits

Kit Name	Supplier	Catalogue Number
500 Cycle MiSeq Reagent Kit v2	Illumina	MS-102-2003
Agencourt AMPure XP Kit	Beckman Coulter Genomics	A63880
Agilent SureSelect Target Enrichment	Agilent	Custom
Bioanalyzer DNA 1000	Agilent	5067-1504
Bioanalyzer DNA High Sensitivity	Agilent	5067-4626
Bioanalyzer Nano RNA 6000	Agilent	5067-1511
Dynabeads mRNA Purification Kit	Thermo Fisher Scientific	61006
Dynabeads MyOne Streptavidin C1 Beads	Thermo Fisher Scientific	65001
Dynabeads MyOne Streptavidin T1 Magnetic Beads	Life Technologies	65601
E-Gel SizeSelect 2% Agarose Gels	Invitrogen	G661012
Ion 318 Chip kits	Life Technologies	4488150
Ion OneTouch 200 Ion Sphere Particles	Life Technologies	4478525
Ion PGM 200 Sequencing Kit	Life Technologies	4474004
KAPA HiFi Mastermix	Thermo Fisher Scientific	KK2601
NEBNext First Strand cDNA Synthesis Module	NEBNext	E7771
NEBNext Multiplex Oligos Illumina Index Set 1	NEBNext	E7335
NEBNext PolyA mRNA Magnetic Isolation Module	NEBNext	E7490

Kit Name (continued)	Supplier	Catalogue Number
NEBNext Second Strand cDNA Synthesis Module	NEBNext	E6111
NEBNext Ultra RNA Library Prep for Illumina	NEBNext	E7530
NEBNext® Fast DNA Library Prep Set for Ion Torrent	NEBNext	E6270
PhiX Control	Illumina	FC-110-3001
Qubit dsDNA High Sensitivity Kit	Thermo Fisher Scientific	Q32854
The Ion OneTouch 200 Template Kit v2	Life Technologies	4478316
SureSelect Reagent Kit	Agilent	G9611A
SureSelect Custom RNA Target Enrichment Probes	Agilent	N/A

Table 4. General kits used throughout this work. Kits were used according to the manufacturer's instructions, unless otherwise stated. The SureSelect Custom RNA Target Enrichment Probes were part of a custom order, although the reagents used for hybridisation, described later, were part of the kit.

2.1.3 Sample Preparation

Most samples collected for this work were taken from healthy donor volunteers. However, some samples were obtained from clinical samples that had been stored previously. This section will describe the details of both sets of samples.

2.1.4 Sample Taking

Peripheral blood mononuclear cells (PBMCs) were obtained from fresh blood draws with informed consent from multiple individuals between the ages of 24 and 43. 5ml of blood was taken per blood draw per individual, using a 21G butterfly needle and syringe. For each blood draw, a 15ml Falcon tube was prepared with 50µl of lithium heparin. 10µl of lithium heparin per milliliter of blood was used to prevent the blood from clotting whilst avoiding potential inhibition of PCR reactions downstream due to a high concentration of the anticoagulant. The amount of anticoagulant used was conservative but effective, so as not to approach the inhibitory concentration (105).

2.1.5 Isolation of Peripheral Blood Mononuclear Cells

Total PBMCs were isolated using a density gradient centrifugation technique. Prior to cell isolation, Roswell Park Memorial Institute (RPMI) media was placed in a water bath at 37°C. RPMI supplemented with 10% fetal calf serum (FCS) was also placed in the water bath and was used downstream to provide nutrients for the cells. Blood samples were diluted 1 in 2 with the pre-warmed RPMI without FCS. The media was added to the blood using a stripette and mixed thoroughly by pipetting up and down before proceeding to the next step.

After dilution, the blood was layered on top of a density gradient media. The media used in this case was Lymphoprep, at a ratio of 2:1 diluted blood to Lymphoprep. Since 5ml of blood was taken and diluted with 5ml of RPMI supplemented with 10% FCS, the mixture was layered on a volume of 5ml of Lymphoprep, according to the manufacturer's recommendation.

The layered tubes were then centrifuged at 7000g for 25 minutes at ambient temperature. To prevent the layers from being disrupted from abrupt slowing of the centrifuge rotor, the brakes were turned off and the spinning rotor allowed to come to a stop smoothly.

Following centrifugation, the layer of PBMCs was removed using a Pasteur pipette and placed in a clean tube. An excess amount of pre-warmed RPMI media supplemented with 10% FCS was added to the tube to wash the cells once. The tube was then centrifuged at 3500g for 10 minutes at ambient temperature and the pellet was resuspended in 1ml RPMI without FCS for cell counting.

2.1.6 Cell Counting

Each cell sample obtained was counted both before and after freezing. To do this, an aliquot of 10μl of resuspended cells was diluted 1 in 2 with a diazonium salt stain, Trypan blue. This is a dye exclusion method, where the stain enters dead or damaged cells (due to membrane permeability), staining cells blue. Stained cells were excluded from counting, whilst ensuring the intact live cells were counted under a light microscope.

Diluted and stained samples were counted using a haemocytometer. Cells were allotted onto the haemocytometer in the recommended area and covered with a glass slide. Cells were visualised under a light microscope at 100x magnification. For each sample, the same four squares in the counting chamber were counted (using a clicker), and the following formula applied to calculate the final concentration of cells;

$$\frac{(t_1 + \ t_2 + t_3 + t_4)}{2} \times 10{,}000 = Total\ cells\ per\ ml$$

In this equation, t_1, t_2, t_3 and t_4 represent the total cell counts from each square on the counting chamber respectively. The total of this is divided by 2 to calculate the average cell count across the four squares whilst taking into account the dilution factor. The multiplication by 10,000 accounts for the scaling of the counting chamber. To calculate the total cells in a sample, the result of this equation, the total cells per ml, can be multiplied by the volume of the sample in milliliters.

2.1.7 Freezing & Thawing of Peripheral Blood Mononuclear Cells

In some cases, it was necessary to freeze cells for further downstream processing. Cells suspended in extraction medium were centrifuged at 3500g for 10 minutes at ambient temperature and resuspended in freezing media, comprising 90% FCS and 10% dimethyl sulfoxide (DMSO), that had been stored at 4°C. Approximately 1ml of freezing media was used per two million cells and cells were frozen in 1ml aliquots in cryovials. Aliquots were frozen in a Mr. Frosty Freezing container (Nalgene, H × diam. 86 mm × 117 mm) at -80°C. For long-term storage (greater than one month), cells were transferred to liquid nitrogen. To thaw cells, aliquots were placed in a water bath at 37°C and washed with an excess of FCS and RPMI, this procedure was followed by centrifugation (3500g for 10 minutes) at ambient temperature. Next the supernatant was discarded, and the cells resuspended in 100µl of RPMI media without FCS or phosphate-buffered saline (PBS) for RNA extraction.

2.2 Isolation of Total RNA and Quality Control

This section describes the process of RNA isolation and the ways in which quality control measures were applied.

2.2.1 RNA Isolation

Surfaces were cleaned prior to extraction using RNase Zap to prevent sample contamination and degradation by nucleases. Total RNA was isolated from cell samples using TRIzol, supplied by Invitrogen. Cell samples were centrifuged at 3500g for 10 minutes at room temperature. The supernatant was discarded, and cells resuspended in the residual liquid by pipetting up and down using a Pasteur pipette. For each sample, 1ml of TRIzol was added to the resuspended cells. The mix was vortexed for 2 minutes. 200µl of chloroform was then added and the samples were vortexed for a further 3 minutes. Samples were centrifuged at 12,000g for 15 minutes at 4°C. This results in an aqueous layer forming at the top of the tube, which contains the RNA. This layer is removed and added to a new tube with 500µl of isopropanol. Samples were then incubated for 10 minutes at room temperature, before being centrifuged at 12,000g for 10 minutes at 4°C to pellet the RNA. To wash, the supernatant was removed and 1ml of 70% ethanol added, before centrifugation at 12,000g for 10 minutes at 4°C. The wash step was repeated once, for a total of two washes, before samples were resuspended in 50µl of nuclease-free water.

2.2.2 RNA Quality Assessment

The quality and quantity of RNA was assessed initially using the Nanodrop 2000 Spectrophotometer (Thermo Scientific). The Nanodrop measures the absorbance of UV-visible light by DNA and RNA and can measure nucleic acid concentrations between 2 and 15,000 ng/µl. It provides purity ratios that can indicate the level of impurity in a sample. 1µl of sample was pipetted on to the sample well for measurement. The well was cleaned using Kimtech Kimwipes and ethanol between samples. The measurement produces two ratios, the 260/280 and the 260/230 ratio. The maximum absorbance of light by nucleic acids occurs at a wavelength of 260nm. For RNA, a 260/280 ratio of 2.0 and a 260/230 ratio of 2.2 are considered pure, so samples that were out of a ± 0.3 range were discarded or cleaned up using ethanol precipitation to remove contamination of protein and/or phenol.

The use of the Nanodrop was treated as a screening tool, which was followed by more accurate determination of quality and concentration by the Agilent 2100 Bioanalyzer system. The RNA Nano 6000 chip is able to quantify RNA within the range 25 – 500 ng/µl. To begin, 550 µl of RNA gel matrix was added to a spin filter and centrifuged at 1500g for 10 minutes. For each chip, 65 µl of this filtered gel was added to a 0.5ml RNase-free micro-centrifuge tube. 1 µl of RNA dye concentrate was added to the 65 µl of gel and the mixture vortexed for 1 minute, then centrifuged at 13,000g for 10 minutes at ambient temperature. 9 µl of this gel-dye mix was loaded into the RNA chip, while positioned in the chip-priming station. The plunger of the chip-priming station was used to evenly disperse the gel across the chip matrix. 5 µl of RNA marker was added to each well, including the ladder well, then 1 µl of sample or ladder was added to the appropriate wells. The chip was then vortexed for 1 minute and run in the Bioanalyzer within 5 minutes. Up to 12 samples could be analysed in one run. All reagents were supplied in the kit.

RNA with an RNA integrity number (RIN) of less than 8 was discarded or cleaned up further using an ethanol precipitation step (section 2.2.3). Following this, library preparation for the two (Ion Torrent and Illumina) sequencing platforms diverges, as described in the section below (see section 2.4).

2.2.3 Ethanol Precipitation

Ethanol based precipitation of RNA is a routine procedure used to purify RNA samples with high concentration of salt, phenol or other impurities. RNA volume was made up to 50µl using nuclease free water, to which 5µl of sodium acetate at a concentration of 3M at pH5.2 was added and vortexed thoroughly. 450µl of cold (4°C) 100% ethanol was added and mixed by inversion of the tube. Samples were then placed on ice for a minimum of 30 minutes and kept for a maximum overnight at 4°C, and spun at 14,000g for 20 minutes at 4°C. The supernatant was aspirated, leaving the pellet undisturbed. The pellets were washed using 500µl of 4°C 70% ethanol, with minimal disturbance to the pellets, before being spun again at 14,000g for 5 minutes in a cold (4°C) centrifuge. The supernatant was then aspirated; to remove as much ethanol as possible and the tubes were then left to air dry at ambient temperature for a minimum of 10 minutes to allow the remaining ethanol to evaporate. RNA was then resuspended in 50µl of nuclease free water and quantified (as described above, section 2.2.2).

2.3 Preparation of Patient Samples and Expanded γδ+ T-cell Populations

Clinical samples were obtained from leftover material following routine blood draws for clinical testing. Samples obtained were frozen RNA aliquots. Samples that were used in the experiments involving expansion of γδ+ T-cells were obtained as frozen RNA aliquots. Dr Stuart Adams and Dr Jonathan Fisher kindly provided these stored samples.

2.4 Library Preparation for Sequencing

The convenience of high-throughput sequencing technology has improved drastically over the past few years. However, it is still necessary to use only high-quality samples and to follow a strict protocol of library preparation before sequencing. Here, the process of library preparation for the two different sequencing platforms, Ion Torrent

PGM and Illumina MiSeq is described. If a library preparation protocol needed to be paused, cleaned up samples were kept in a -20°C freezer for up to 3 nights before the protocol was continued.

For all procedures in sections 2.4.1 to 2.5.2, samples and reagents were manipulated using RNase and DNase free barrier filter tips for 1000 µl (Thermo Scientific catalogue no. AM12660), 200 µl (Thermo Scientific catalogue no. AM12650) and 10 µl (Thermo Scientific catalogue no. AM12635). Samples were stored in 0.2ml (Thermo Scientific catalogue no. AM12230) or 1.5ml (Thermo Scientific catalogue no. AM12400) microfuge tubes unless otherwise specified.

2.4.1 mRNA Isolation

The mRNA was isolated from total RNA, using a polyA selection kit, either the NEBNext Poly(A) mRNA Magnetic Isolation Module, or the Dynabeads mRNA Purification Kit for mRNA Purification from Total RNA. Both kits were chosen as they were recommended by experts in the field at the time and were easily available on the market.

For the NEBNext kit, the total RNA was diluted in nuclease-free water to a final volume of 50µl in a 0.2 ml PCR tube. 20 µl of magnetic oligo d(T)$_{25}$ beads were resuspended and added to another 0.2 ml PCR tube, then washed with 100 µl of RNA binding buffer. The tube with the beads was placed on the magnetic rack and the supernatant removed using a pipette. This wash step was repeated so that there were 2 washes in total and the beads were resuspended in 50 µl of RNA binding buffer, to which the 50 µl of sample RNA was added.

The sample-bead mixtures were incubated at 65°C for 5 minutes and then put on ice to allow for the RNA binding, before being incubated at room temperature for a further 5 minutes. The tubes were put on the magnetic rack for 2 minutes and the supernatant was removed using a pipette, without disturbing the beads that were at that point bound to the mRNA. The beads were washed twice by adding 200 µl of wash buffer, placing on the magnetic rack and removing the liquid.

After washing, 50 µl of Tris buffer at pH 7.5 was added to the beads. The tubes were incubated at 80°C for 2 minutes, and then held at 25°C. 50 µl of RNA binding buffer was added to the beads and this process was repeated once. After the wash steps, 17 µl of Tris buffer at pH 7.5 was added to the beads to elute the mRNA and incubated at 80°C for 2 minutes, and then held at 25°C as described above. The samples were placed on the magnetic rack and the purified mRNA in the eluate was collected and transferred to a new RNase-free 0.2ml PCR tube. All required reagents were included in the kit.

For the Dynabeads kit, the volume of total RNA was adjusted to 100 µl using nuclease-free water in 0.5 ml microcentrifuge tube and samples were incubated at 65°C for 2 minutes. 200µl of Dynabeads were added to a separate microcentrifuge tube and placed on the magnetic rack for 30 seconds. The supernatant was removed using a pipette and 100 µl of binding buffer was added, before the tube was placed back on the rack again for 30 seconds. The supernatant was removed, and the beads were resuspended in 100 µl of binding buffer. The RNA was added to the bead suspension and samples were left on a roller for 5 minutes. The tubes were then placed on the magnetic rack for 30 seconds and the supernatant was removed with a pipette. The mRNA-bead complex was washed twice with washing buffer using the magnetic rack and both times the supernatant was completely removed. After the wash steps, the beads were resuspended in 10 µl of Tris buffer at pH 7.5 and incubated at 80°C for 2 minutes to elute the mRNA from the beads and placed immediately on the magnetic rack. The eluted mRNA was transferred to a new nuclease-free microcentrifuge tube. All reagents needed were provided in the Invitrogen kit.

2.4.2 Library Preparation for Ion Torrent Sequencing

Isolated mRNA was used as a template to produce first and second strand cDNA. First strand cDNA was synthesized and fragmented using the NEBNext RNA First Strand Synthesis Module. To do this, the first strand buffer was prepared by adding 5 µl of mRNA to 4 µl of first strand synthesis reaction buffer and 1 µl of random primers, both of which are provided in the kit, and this mixture was incubated for 15 minutes at 94°C. To this mixture, 0.5 µl of murine RNase inhibitor, 1 µl of ProtoScript II reverse transcriptase and 8.5 µl of nuclease-free water was added. The mix was the incubated in a thermocycler for the conditions described in Table 5, which resulted in the complete first strand cDNA.

Stage	**Temperature/ºC**	**Time/Minutes**
1	25	10
2	42	50
3	70	15
4	4	Hold

Table 5. First strand cDNA synthesis conditions. For use with the NEBNext RNA First Strand Synthesis Module. The 4 stages described above are the cycling conditions that were programmed into the thermocycler, also known as a PCR machine. The thermocycler held samples at each stage, and they

were removed once the thermocycler had reached stage 4 and cooled the samples down to 4ºC. Synthesis of the second strand of cDNA was done immediately after.

The second strand cDNA synthesis followed immediately from the first strand, using the NEBNext Second Strand cDNA Synthesis Module. 48 µl of nuclease-free water was added to the microfuge tube containing the first strand, followed by 8 µl of 10 x second strand synthesis reaction buffer and 4 µl of second strand synthesis reaction mix. The mixture was then incubated for 2.5 hours at 16°C in a thermocycler. The double stranded cDNA was then cleaned up using Agencourt AMPure XP Beads, described below in section 2.4.4.

A sequencing library preparation kit from NEBNext, the Fast DNA Library Prep Set for Ion Torrent kit, was used for the remainder of the process. This kit includes the materials required for end repair, adaptor ligation, size selection and PCR amplification. All clean up steps were done using Agencourt AMPure XP Beads, described in section 2.4.4. Additionally, E-Gel SizeSelect Agarose Gels were used for size selection in place of the module in the NEBNext kit, for improved accuracy, as described in section 2.4.6.

For end repair, the fragmented cDNA, eluted in 51 µl of nuclease-free water, was added to 6 µl of end repair reaction buffer and 3 µl of enzyme mix for a total volume of 60 µl. This mix was incubated in a thermocycler for 20 minutes at 25°C, 10 minutes at 70°C and then held at 4°C. Adapter ligation immediately followed, and the reagents listed in Table 6 were added to the end repair reaction tube and incubated in a thermocycler for 15 minutes at 25°C, 5 minutes at 65°C and then held at 4°C. The samples were then cleaned up as described in section 2.4.4.

Reagent	**Volume/µl**
Nuclease-free water	18
T4 DNA ligase buffer*	10
Ion Torrent adapters*	5
Warm start DNA polymerase*	1
T4 DNA ligase*	6

Table 6. Reagents used for Ion Torrent adapter ligation. The items labelled with an asterisk (*) were provided in the Ion PGM 200 Sequencing Kit (see Table 4), those not labelled with asterisk were part of the general reagents used (see Table 3).

The final step in the Ion Torrent library preparation was to amplify sample prior to the capture step. 4 µl of primers and 50 µl of Q5 hot start hifi PCR master mix was added to 40 µl of adapted ligated sample. The reaction was carried out in a thermocycler using the conditions described in Table 7.

PCR Stage		**Temperature/ ºC**	**Time/s**
Initial Denaturation		98	30
Denaturing	12 Cycles	98	10
Annealing		58	30
Extending		65	30
Final Extension		65	300
Hold		4	Hold

Table 7. PCR cycling conditions for pre-capture amplification of Ion Torrent libraries. The conditions above were programmed into the thermocycler. Samples were placed in the thermocycler and each stage was carried out by the machine. For denaturing, annealing and extending, the conditions were repeated 11 times, for a total of 12 cycles, in order to amplify the material.

Samples were then cleaned up and quantified before being sequenced individually on the Ion Torrent Personal Genome Machine, as described in section 2.5.1.

2.4.3 Library Preparation for Illumina Sequencing

The NEBNext Ultra RNA Library Prep Kit for Illumina was used to create the library, including the cDNA synthesis so there was no need for separate modules, except for the adaptors. This kit also included the materials required for end repair, adaptor ligation, size selection and PCR amplification. Purified mRNA was used as a template to produce first and second strand cDNA. First strand cDNA was generated by adding 5 µl of mRNA to 1 µl of random primers and this mixture was heated at 65°C for 5 minutes, with a heated lid at 105°C to prevent evaporation and subsequent loss of product, then held at 4°C. To this primed mixture, 4 µl of first strand synthesis reaction buffer, 0.5 µl of murine RNase inhibitor, 1 µl of ProtoScript reverse transcriptase and 8.5 µl of nuclease-free water was added. This was incubated in a thermocycler as described in Table 8.

Stage	Temperature/ºC	Time/Minutes
1	25	10
2	42	15
3	70	15
4	4	Hold

Table 8. First strand cDNA synthesis conditions for Illumina library preparation. The above conditions were programmed into the thermocycler. Samples were placed in the thermocycler and the machine carried out each stage 1 through 4. Samples were removed once the thermocycler had reached stage 4 and cooled the samples down to 4ºC.

The second strand cDNA synthesis followed immediately from the first strand. 48 µl of nuclease-free water, 8 µl of 10x second strand synthesis reaction buffer and 4 µl of second strand synthesis reaction mix was added to the first strand reaction tube. The mixture was then incubated for 1 hour at 16°C, with a heated lid at 40°C to prevent sample evaporation. The double stranded cDNA was then cleaned up using Agencourt AMPure XP Beads, described below in section 2.4.4. At this step, samples were fragmented through sonication as described in 2.4.5.

For end repair, the cDNA, eluted in 55.5 µl of nuclease-free water, was added to 6.5 µl of 10x end repair reaction buffer and 3 µl of end prep enzyme mix so the total volume was 65 µl. This was incubated in a thermocycler for 30 minutes at 20°C, 30 minutes at 65°C and then held at 4°C. Adapter ligation and sample indexing followed post-end repair, and the reagents listed in Table 9 were added to the end repair reaction tube and incubated in a thermocycler for 15 minutes at 20°C. After 15 minutes, 3 µl of USER enzyme was added to the ligation mixture and incubated for a further 15 minutes at 37°C.

Reagent	**Volume/µl**
Nuclease-free water	2.5
T4 DNA ligase buffer	10
Illumina adaptor (Index 1 to 12)	5
Warm start DNA polymerase	1
T4 DNA ligase	6

Table 9. List of reagents for Illumina adapter ligation. The Illumina adapters were provided in the NEBNext Multiplex Oligos for Illumina (Index Primers Set 1). 12 indexes were provided in this kit so it was possible to multiplex up to 12 samples at once.

Adaptor ligation and indexing was followed by size selection of fragments using E-Gel SizeSelect Agarose Gels, as described in section 2.4.6. Size selected fragments were then amplified in the pre-capture PCR reaction. 2.5 µl of the index primer, 2.5 µl of the universal primer and 25 µl of the Q5 hot start hifi PCR master mix was added to 20 µl of sample. The reaction was carried out using the conditions described in Table 10.

PCR Stage		**Temperature/ºC**	**Time/Seconds**
Initial Denaturation		98	30
Denaturing	12 Cycles	98	10
Annealing & Extension		65	75
Final Extension		65	300
Hold		4	Hold

Table 10. PCR cycling conditions for pre-capture amplification of Indexed Illumina libraries. The conditions above were programmed into the thermocycler. Samples were placed in the thermocycler and each stage was carried out by the machine. For denaturing, annealing and extending, the conditions were repeated 11 times, for a total of 12 cycles, in order to amplify the material.

Samples were then cleaned up and quantified before being pooled for the capture reaction. Between 40 and 100ng of indexed cDNA library material was pooled into one tube. All clean up steps were done using Agencourt AMPure XP Beads, described in section 2.4.4 and quantification was done according to section 2.4.7.

PCR cycling conditions varied between the Ion Torrent PGM and the Illumina MiSeq platforms as different enzymes were provided in each kit. The different polymerase(s) work optimally at different temperatures due to structural differences. In addition to this, each kit provided primers with different DNA sequences which therefore had different annealing temperatures.

2.4.4 Sample Clean Up

For each high-throughput sequencing protocol used, multiple clean up steps are required, which were carried out using Agencourt AMPure XP Beads and a 95 well magnetic separation plate. A 1:1 ratio of beads to sample volume was employed unless otherwise stated in the protocol. The mixture was incubated for 5 minutes at room temperature, before tubes were placed on a magnetic rack. Once the beads had migrated to the magnet, the solution was removed and discarded. To wash, 200 µl of 70% ethanol was added to the tube, whilst still on the magnet, and removed after 30 seconds. This was repeated once for a total of two washes. Beads were air dried for ten minutes before being resuspended in the volume of nuclease-free water required for the next step and replaced on the magnet. The clear solution containing the cleaned-up DNA was transferred to a new tube.

2.4.5 Fragmentation Using Sonication

The Covaris S220 Focused Sonicator (Applied Biosystems) was used to shear double stranded (ds) cDNA into fragments ranging between 200 and 300bp in length. cDNA samples (in 50 µl of nuclease-free water) were transferred to Covaris microtubes. The machine settings were adjusted as described in Table 11. Samples were sonicated for 300 seconds (detailed in Chapter 3).

Covaris Settings	
Duty cycle	10%
Intensity	5
Cycles per burst	200

Table 11. Covaris sonication settings for 200-300bp fragments. The settings described above were programmed into the Covaris machine. Samples were then placed inside and the machine switched on. These sonification settings resulted in DNA fragments approximately 200 and 300bp in length. Sonication works through a series of high frequency signals, or oscillations, delivered to the samples in bursts. The duty cycle is the percent of time with active bursts, the intensity is the power of the bursts and the cycles per burst are the number of oscillations per burst.

2.4.6 Library Size Selection

For both Ion Torrent and Illumina protocols, appropriately sized fragments were selected using the E-Gel SizeSelect Agarose Gel system (version 2.0) to maximise sample recovery. The iBase was placed on top of the Safe Imager and the SizeSelect 2% program was selected. 25 µl of each sample was loaded into wells on the E-Gel and 10 µl of DNA ladder was added to the middle well. 25µl of deionized water was used to fill all empty sample wells and the collection wells at the bottom of the E-Gel. The program was switched on and left to run for approximately 15 minutes. From 10 minutes onwards, the gel was monitored in order to select the correct size fragments. Samples were collected from the collection wells and cleaned up using AMPure XP beads.

2.4.7 DNA Quantification

Complementary DNA obtained was quantified using both the Life Technologies Qubit system and the Agilent 2100 Bioanalyzer system. The Qubit dsDNA High Sensitivity

(HS) kit was used for accurate determination of concentration; whereas the Bioanalyzer was used to determine the size distribution of fragments.

For the Qubit, the working solution was made up by diluting the dsDNA HS reagent 1 in 200 with the dsDNA HS buffer. 190 µl of the working solution was added to each sample tube and two additional tubes for the standards. 10 µl of sample or standard was then added to each tube and vortexed. Tubes were incubated at room temperature for 2 minutes and then measured using the Qubit 1.0 Fluorometer.

The DNA 1000 chip and High Sensitivity DNA Kit were used to measure concentration, within the ranges 0.5 – 50 ng/µl and 5 – 500 pg/µl respectively. For the DNA 1000 kit, 9 µl of this gel-dye mix was loaded into the DNA chip, while positioned in the chip-priming station. The plunger of the chip-priming station was used to evenly disperse the gel across the chip matrix. A further 9 µl of gel-dye mix was added to two more wells of the chip. 5 µl of RNA marker was added to each well, including the ladder well, then 1 µl of sample or ladder was added to the appropriate wells. The chip was then vortexed for 1 minute and run in the Bioanalyzer within 5 minutes. Up to 12 samples could be analysed in one run. All reagents were supplied in the kit.

The protocol for the DNA High Sensitivity kit was very similar. 9 µl of this gel-dye mix was loaded into the DNA chip, while positioned in the chip-priming station. The plunger of the chip-priming station was used to evenly disperse the gel across the chip matrix. A further 9 µl of gel-dye mix was added to three more wells. 5 µl of RNA marker was added to each well, including the ladder well, then 1 µl of sample or ladder was added to the appropriate wells. The chip was then vortexed for 1 minute and run in the Bioanalyzer within 5 minutes. Up to 11 samples could be analysed in one run. All reagents were supplied in the kit.

2.4.8 Hybridisation and Capture

Hybridisation between the amplified libraries and the custom capture library containing the bespoke set of baits was carried out according to the protocol for Agilent's SureSelect Target Enrichment System. Samples were vacuum centrifuged to concentrate them at 221 ng/µl. The hybridisation buffer was prepared by mixing the reagents in Table 12.

Reagent	Volume/µl
Hybridisation buffer 1	6.63
Hybridisation buffer 2	0.27
Hybridisation buffer 3	2.65
Hybridisation buffer 4	3.45

Table 12. Preparation of hybridisation buffer. Volumes are listed as required per sample. All reagents are included in the SureSelect Target Enrichment kit.

A block mix was then prepared by mixing the reagents in Table 13. This mix of DNA sequences blocks non-specific binding and binding of adaptor sequences to each other during the hybridisation reaction.

Reagent	Volume/µl
Indexing block 1	2.5
Block 2	2.5
ILM indexing block 3	0.6

Table 13. Preparation of block mix. Volumes are listed as required per sample. All reagents are included in the SureSelect Target Enrichment kit.

The hybridisation buffer was kept at ambient temperature and the block mix was kept on ice until required. 5.6 µl of block mix was added to 3.4 µl of prepared library (100 ng of sample material in total). This mixture was then incubated in a thermocycler at 95°C for 5 minutes, followed by 65°C for at least another 5 minutes. The RNase block was diluted 1 in 3 with nuclease-free water and 2 µl of this was added to the sample and block mix. Finally, 13 µl of the hybridisation buffer was added, while the samples were already placed in the block of the thermocycler. The tubes or wells were thoroughly sealed to prevent evaporation during hybridisation and left to undergo capture for 24 hours at 65°C with a heated lid at 105°C, to prevent loss through evaporation.

After this time, the bound biotinylated baits (to the targeted sequences) were isolated using streptavidin coated magnetic beads. Wash buffer 2 was warmed to 65°C in a water bath. 50 µl of resuspended Dynabeads (MyOne Streptavidin T1 magnetic beads) was added to a nuclease-free microfuge tube and washed using 200 µl of binding buffer. The tube was placed on the magnetic rack, the supernatant was removed, and the beads resuspended in 200 µl of binding buffer. This was repeated twice for a total of 3 washes.

The hybridisation sample was then added straight from the thermocycler to the binding buffer and bead mix and left on a mixer for 30 minutes at room temperature, after which the tube was placed on the magnetic rack, supernatant removed and the beads were resuspended in 200 µl of wash buffer 1 and incubated at room temperature for 15 minutes followed by washing with wash buffer 2 (pre-warmed to 65°C), and incubated at 65°C for 10 minutes. The tube was placed on the magnetic rack and the beads resuspended in wash buffer 2. This process was repeated twice for a total of 3 washes. After the final wash the beads were resuspended in 50 µl of elution buffer and incubated at room temperature for 10 minutes. For a final time, the tube was placed on the magnetic rack, and the supernatant transferred to a clean nuclease-free tube. Samples were further purified using the AMPure XP beads before undergoing the post-capture PCR amplification.

2.4.9 Post-Capture PCR Amplification

After capture, samples were amplified once more. As this step was modified throughout the development of the capture technique, more details are described in Chapter 3. The primers used were custom ordered oligos from Sigma Aldrich. The forward primer was the reverse complement of the Illumina P5 adaptor sequence and the reverse primer was the Illumina P7 adaptor sequence. To set up the 50 µl reaction, 25 µl of 2X KAPA HiFi Hot Start Mastermix was added to 2.5 µl of each 25µM primer and 22 µl of the post-capture reaction material. The PCR conditions are described in Table 14. PCR amplification was followed by a clean-up step and quantification, as described in 2.4.4 and 2.4.7 before samples were sequenced.

PCR Stage		**Temperature/°C**	**Time/s**
Initial Denaturation		98	45
Denaturing	12 Cycles	98	15
Annealing		60	30
Extending		72	30
Final Extension		72	60

Table 14. PCR cycling conditions for the post-capture PCR amplification. These cycling conditions were programmed into the thermocycler. The samples were placed in the thermocycler as it cycled through each stage. The denaturing, annealing and extending stages were repeated 11 times for a total of 12 cycles, in order to appropriately amplify the material.

2.5 Sequencing

Samples were sequenced using either the Ion Torrent PGM or the Illumina MiSeq machines. The preparation for each platform differs according to the manufacturer's protocols, as described below.

2.5.1 Preparation for Ion Torrent Sequencing

The Ion OneTouch 200 Template Kit v2, Ion OneTouch 200 Ion Sphere Particles, Dynabeads MyOne Streptavidin C1 Beads, Ion PGM 200 Sequencing Kit and Ion 318 Chip kits were used for sequencing. The OneTouch machine was set up according to manufacturer instructions. The amplified, captured library material was diluted to 25nM using nuclease-free water and used to make up the amplification solution, shown in Table 15.

Reagent	**Volume/µl**
Nuclease-free water	280
Ion OneTouch 2X Reagent Mix	500
Ion OneTouch Enzyme Mix	100
Diluted library	20

Table 15. Ion OneTouch amplification solution. The diluted library here refers to the amplified, captured material from the hybridisation reaction at 25nM.

100 µl of Ion Sphere Particles (ISPs) were added to the 900 µl of amplification solution. This entire mix was then pipetted into the sample port on the Plus Reaction Filter, followed by 1.5 ml of Reaction Oil, before the Reaction Plus Filter was placed on the Ion OneTouch. Following the OneTouch reaction, the then template-positive ISPs were recovered and washed. The tube containing the ISPs was centrifuged and the

supernatant was removed with a pipette. The ISPs were then resuspended in Recovery Solution and 1ml of Wash Solution was added. This mixture was centrifuged and all but 100 µl of the supernatant was removed, in which the ISPs were resuspended.

The template-positive ISPs were then enriched using the OneTouch ES. The sample was transferred to an 8-well strip with the relevant reagents, as shown in Table 16. The strip was then loaded on to the Ion OneTouch ES machine. This enrichment process took approximately 35 minutes.

Well	Volume/µl	Reagent
1	100	ISPs
2	130	Dynabeads
3	300	Wash Solution
4	300	Wash Solution
5	300	Wash Solution
6	0	Empty
7	300	Melt Off Solution
8	0	Empty

Table 16. Reagents and volumes to fill the 8-well strip prior to loading on the Ion OneTouch ES.

The enriched, template-positive ISPs were washed in 200 µl of Wash Solution once, centrifuged at 15,000g for 90 seconds, resuspended in 100 µl of Wash Solution at ambient temperature, before being ready for sequencing.

Samples were sequenced using Ion 318 Chips. 100 µl of Annealing Buffer and 5 µl of Control Ion Sphere Particles were added to the template ISPs. The mixture was centrifuged at 15,000g for 2 minutes at ambient temperature. All but 15 µl of the supernatant was removed, and the ISPs were resuspended in the residual volume and

12 µl of Sequencing Primer. Tubes were then incubated at 95°C for 2 minutes then 37°C for 2 minutes with a heated lid at 105°C. Once the sequencing primers were annealed, 3 µl of Sequencing Polymerase was added to the ISPs and incubated at room temperature for 5 minutes. This mixture was then loaded on to the chip and placed in the PGM for sequencing. The sequencing took approximately 4 hours to complete.

2.5.2 Preparation for Illumina MiSeq Sequencing

The 500 Cycle MiSeq Reagent kit v2 and PhiX Control kits were used for sequencing. All reagents were included in the kits, excluding the sodium hydroxide (NaOH). The first step in setting up a run on the MiSeq was to thaw the reagent cartridge. Cartridges were removed from the -20°C freezer and left overnight at 4 to 8°C. Immediately before use, cartridges were inverted 10 times to mix the reagents.

The libraries that had been prepared for sequencing were diluted to a concentration of 2nM with nuclease free water, having been quantified previously as described in 2.4.7. They were then denatured by adding 10 µl of 0.2M NaOH to 10 µl of 2nM library and the mix incubated for 5 minutes at room temperature. 980 µl of HT1 buffer was then added to the denatured libraries, giving final library concentration of 20pM. 400 µl of this 20pM library was added to 600 µl of HT1 buffer, making a library with a concentration of 8pM. The diluting and denaturing process was repeated with PhiX, which was then used to spike the library at a concentration of 1% for quality control. The final volume of library was 1ml, all of which was loaded into the reagent cartridge.

A run was created on the MiSeq using the BaseSpace software, and on-screen instructions were followed. The flow cell was cleaned carefully by rinsing in deionized water to remove excess salts and using a lint-free lens tissue to dry. The flow cell was then loaded into the MiSeq (software version 2.4). The reagent cartridge was loaded into the machine, as prompted by the software and the run was started. Each run would take approximately 27 hours.

2.6 Bioinformatics and Data Processing

High-throughput sequencing generates large quantities of data, requiring the use of sophisticated bioinformatics software. This section provides a brief description of how sequencing data has been handled for TCR repertoire analysis. All processing was done using command line tools.

2.6.1 Burrows Wheeler Alignment

Once sequencing data was obtained using the targeted capture method described, it was necessary to determine the proportions of reads that aligned to the target regions. In this case, the target refers to TCR and Ig transcript sequences. This was done using Burrows-Wheeler Alignment tool (BWA), an open source package that uses the Burrows-Wheeler Transform (BWT) to align sequencing reads to a reference file (found at http://bio-bwa.sourceforge.net/), version 0.7. The reference file used in this case was the human transcriptome (UCSC reference transcriptome version hg19). BWA generates a sequence alignment map (SAM) file as output, which contains all information on read alignment for that file. BWA was utilised through the command line using the following command;

```
./bwa aln -t 8 index.fa read1.fa > read1.sai
```

Followed by;

```
./bwa aln -t 8 index.fa read2.fa > read2.sai
```

This results in two output files per sample, read1.sai and read2.sai, one for each read. These are then combined into one SAM file using the following;

```
./bwa sampe index.fa read1.sai read2.sai > sample.sam
```

SAM files can be analysed and visualised using various programs, for example the Integrated Genome Viewer (IGV) and alignment information can be exported for downstream analysis. IGV version 2.3.26 was used for this purpose.

IGV can be found at https://software.broadinstitute.org/software/igv/home.

2.6.2 FASTQ Data Processing

Sequencing data was processed to improve quality and enable paired ends to be aligned. The workflow can be seen below (in Figure 3) and is described in further detail in sections 2.6.3, 2.6.4, 2.6.5 and 2.6.6.

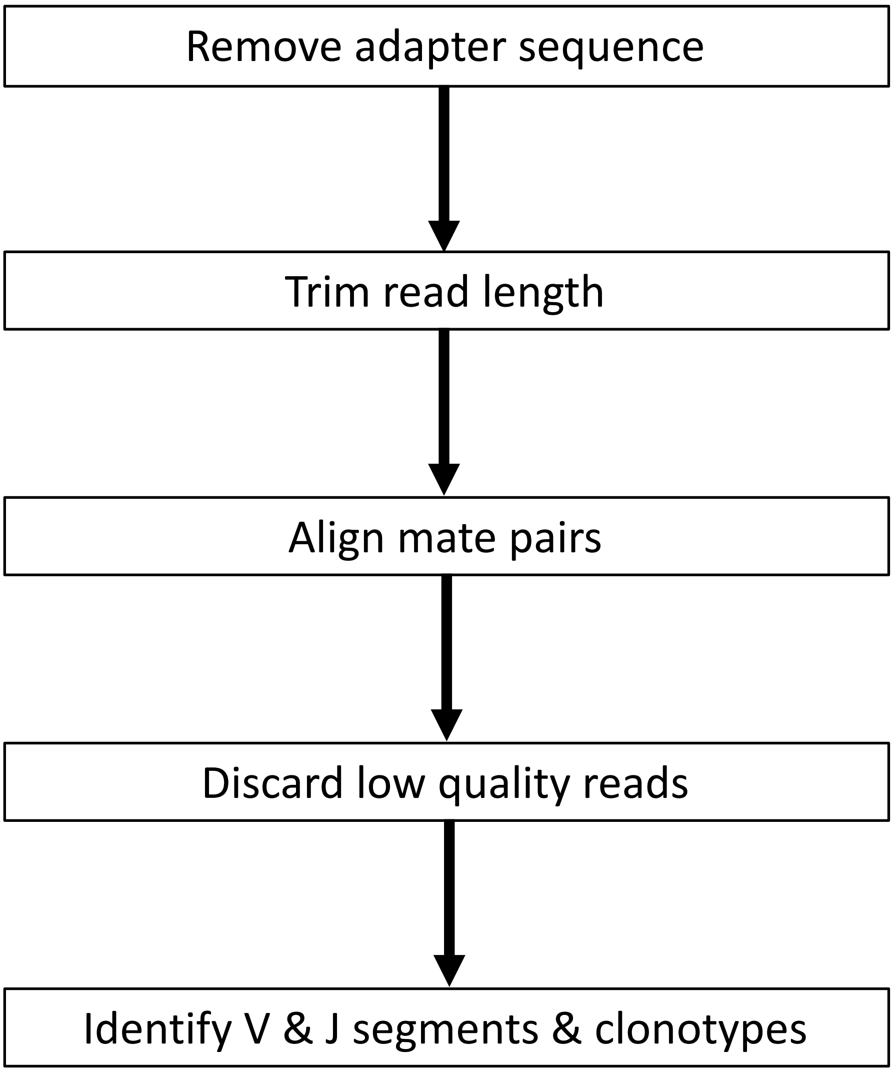

Remove adapter sequence
Trim read length
Align mate pairs
Discard low quality reads
Identify V & J segments & clonotypes

Figure 3. The workflow for bioinformatic processing of FASTQ files.

2.6.3 Adaptor Trimming

For all MiSeq runs carried out, 500 cycles of sequencing were done. As this was paired end, this means 250 cycles in each direction, forward and reverse. In some cases, the lengths of the sample insert sequences were shorter than 250 and some of the adapter was also sequenced. Therefore, the first step in data processing is to remove the Illumina adapter sequences from the data. This is done using Cutadapt (found at https://cutadapt.readthedocs.io/en/stable/), version 1.2, and the following commands:

```
cutadapt -a adaptorsequence input.fastq > output.fastq
```

The input.fastq and output.fastq refer to the full path to the input file and for the designated output. The option –a (adaptor) indicates where the adaptor sequence should be inserted in the command. The adaptor sequences are as follows;

For read 1 (forward):

5' AAT GAT ACG GCG ACC ACC GA 3'

For read 2 (reverse):

5’ ATC TCG TAT GCC GTC TTC TGC TTG 3’

The forward read cutting sequence is the Illumina P5 sequencing adaptor sequence, and the reverse read cutting sequence is the reverse complement of the Illumina P7 adaptor sequence. This was not necessary for Ion Torrent sequencing as the adaptor sequences did not affect the data.

2.6.4 Trimming the Read Length

For some data files it is then necessary to trim the length of the reads, depending on the size of the insert sequence and the number of sequencing cycles carried out. This was done using an open source program called Fastx-Toolkit (found at http://hannonlab.cshl.edu/fastx_toolkit/), version 0.0.13, and the following command;

```
fastx_trimmer -l 200 -Q33 -i input.fastq -o output.fastq
```

The input.fastq and output.fastq refer to the full path to the input file and for the designated output. The –Q33 flag tells it that the quality scores are Illumina encoded.

The option –l (length) determines the length to be trimmed to. The -i and -o refer to the input and output files.

This was done for only Illumina sequences in order to facilitate the joining of mate pairs. The sequences for both read 1 and read 2 were trimmed to equal lengths each time, so they could be paired downstream.

2.6.5 Discarding the Poor-Quality Reads

Next, reads were retained or discarded based on their quality scores. Poor quality reads should be removed from analysis of data sets as they are unreliable and may introduce bias. This was done for both Ion Torrent and Illumina sequences using Fastx-Toolkit;

```
fastq_quality_filter -Q33 -q 30 –p  75 -i input.fastq -o output.fastq –v
```

The input.fastq and output.fastq refer to the full path to the input files and for the designated output file that will be created by this process. Here, the –Q33 flag tells it that the quality scores are Illumina encoded. The options –q and –p give the desired cut off quality score, Q30, and the percentage of bases within a read that must be above this score. Reads that do not fall within this description are discarded. The option –v stands for verbose and gives a printout of the number of reads discarded.

2.6.6 Aligning Mate Pairs from Illumina Paired End Sequencing

The final step in processing data prior to analysis was the alignment of paired end reads. This was only done for Illumina as the Ion Torrent data was single end. Pairing was done using the open source program FLASH (fast length adjustment of short reads, found at https://ccb.jhu.edu/software/FLASH/), version 1.2.7, using the following command;

```
./ flash read1.fastq read2.fastq
```

This aligns reads 1 and 2 and outputs a file containing all paired sequences. The data was then ready to undergo analysis for TCR sequences.

2.6.7 Computational Analysis of the TCR Repertoire

Following data processing, FASTQ files were used to identify TCR clonotypes and V and J gene segment analysis. There are multiple programs designed specifically to analyse immune repertoires. Programs utilised in this work include the Decombinator (found at https://github.com/innate2adaptive/Decombinator) and MiTCR (found at https://github.com/milaboratory/mitcr). For both programs are run through the command line and request an input file, a FASTQ, file for analysis. The output files list the clonotypes identified and the V and J segments used for each. Further details are described in Chapter 3. Here, Niclas Thomas is acknowledged for the adaptation of the Decombinator for use with γ and δ TCR chains, as well as α and β chains.

2.7 Statistical Analysis

An unpaired t-test was carried out where relevant to compare variables, as described in the text. A p value of <0.05 was considered to be statistically significant. All instances where statistical analysis was undertaken in this work are highlighted and described in the text. For all other instances, statistical analysis was not undertaken according to experimental setup and statistical reliably. For these experiments further explanations are provided in the respective sections. Where there were repeated data points, the mean and standard deviation were calculated and included in the figures. All calculations were done in Excel.

Chapter 3

Development of the Capture Technique

3 Development of the Capture Technique

3.1 Introduction

T-cell and B-cell receptor sequencing traditionally depends on PCR techniques and amplification of each different V and J gene segment. Targeted capture had not previously been used for this purpose and it presented an opportunity to look at all TCR and Ig chains simultaneously. As this was a novel technique, a molecular assay that was compatible with targeted capture needed to be developed, requiring optimisation of library preparation for sequencing and iteration of a data analysis workflow, both of which form the focus of this chapter. The data analysis can be split into two workflows; low level and high-level processing (106). The low-level processing includes the quality control steps, V(D)J gene assignment and CDR3 identification. The high-level processing involves a more detailed analysis of repertoire diversity to facilitate interpretation of the data, which will be addressed in Chapter 4.

High throughput sequencing requires highly specialised procedures for effective library preparation. The stages vary depending on the purpose of the library and from user to user. The first consideration is the starting template, as either DNA or RNA may be used to create sequencing libraries, depending on the research question.

The TCR rearrangements occur at the mRNA level, therefore isolating total RNA was the appropriate step. The main stages that were involved in generating libraries for RNA sequencing are illustrated in Figure 4, which illustrates differences in workflow between the two different sequencing platforms, the Ion Torrent PGM and the Illumina MiSeq.

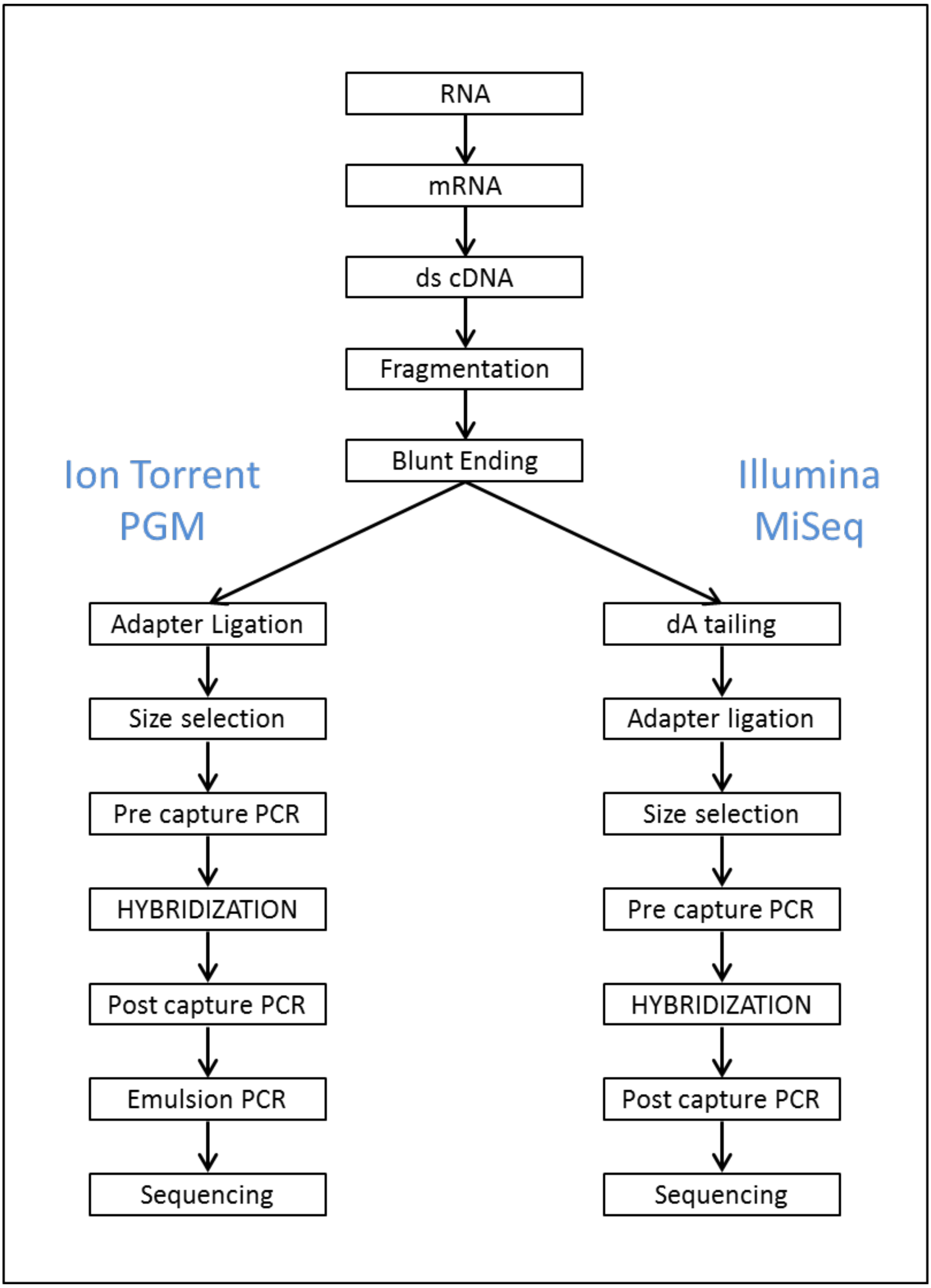


Figure 4. Library preparation procedures vary between sequencing platforms. Both Ion Torrent PGM and Illumina MiSeq platforms were used to generate data, but each have platform specific techniques and kits that should be used for optimal results.

3.1.1 Design of the Capture Kit

Agilent's SureSelect Target Enrichment kits allow for targeted sequencing of a specific region of a genome or transcriptome (107). The process involved the hybridisation of a set of biotinylated RNA baits, also known as probes, of 120bp in length with sequences complementary to the area of interest. The hybridised sequences were then isolated using magnetic bead separation. For this work, Agilent's online application, eArray, was used to generate the baits. The sequences of the baits were designed by Michael Epstein, as part of his MSc project. These sequences were complementary to the mRNA product of the T-cell and B-cell receptor germline rearrangements, as found in the ImMunoGeneTics (IMGT) database (61).

The aim was to indirectly capture sequence information of the highly variable CDR3 region. The CDR3 region was defined as beginning from a conserved cysteine motif in the V segment and ending at a corresponding conserved phenylalanine (or tryptophan for Ig segments) in the J region. Baits were computationally designed to be complementary to all V and J segments of each TCR α, β, γ and δ chains, and Ig heavy, κ and λ chains. As the J segments are typically only 50 to 70bp long, baits were designed to include all combinations of J and C segments. Further baits continued downstream into the C region. Baits are staggered across all regions, with their starting points separated by 60bp, as shown in Figure 5.

The concept was that by designing 120bp baits that reach the boundaries of the CDR3 region and assuming an average CDR3 length of 39nt with sequencing libraries randomly fragmented to 200bp, there would be some fragments that extend across the entire hypervariable region. Therefore, capturing these sequences could help to identify the rearrangements, shown in Figure 5. Each bait was matched by its complementary sequence, which increased sensitivity of the kit and captured both the forward and the reverse cDNA strand. Sequence information for the V, J and C regions were downloaded in FASTA format from the IMGT database. A computational approach ensured that the final library comprising 8,830 different bait sequences covered all T-cell and B-cell receptor sequences. Although the Ig heavy, κ and λ chains were also included in the capture kit, the focus of this thesis is the TCR. The sequences of the baits can be made available upon request.

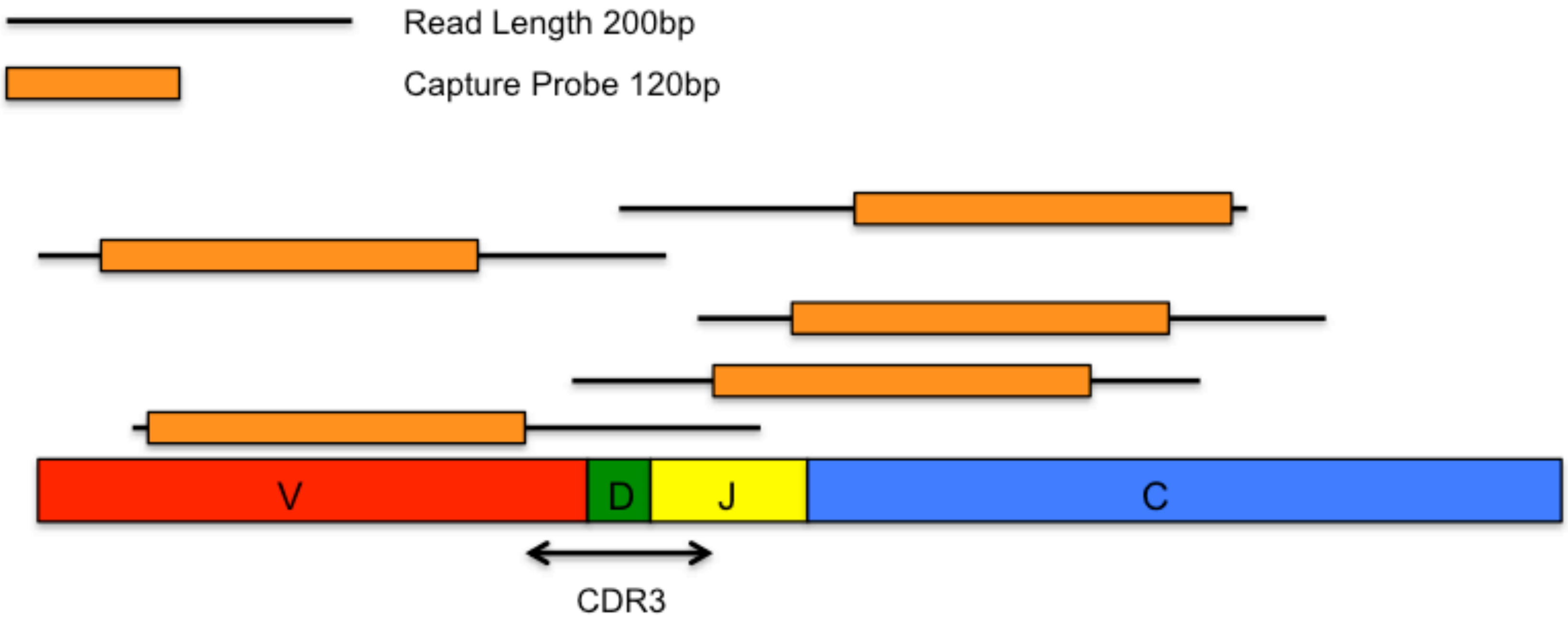


Figure 5. Capture bait design and coverage of the CDR3 region. Each capture bait comprised a 120bp nucleotide sequence complementary to a part of a V, J or C segment of the TCR α, β, γ or δ chains, or Ig heavy, κ or λ. To cover all V and J segments, 4415 baits were generated, and the reverse complements included. Therefore, the final bait library had 8,830 unique sequence.

3.2 Methods

Data presented in Chapters 3, 4 and 5 were obtained at the UCL Genomics facility at Great Ormond Street Hospital, United Kingdom. Support for data analysis and bioinformatics between 2011 and 2015 was received from Tony Brooks, Mike Hubank, James Heather and Benny Chain.

3.2.1 Summary of Data

The workflow was initially tested using samples from healthy individuals. The data presented here was generated from samples taken from 19 consenting volunteer donors with no current history of illness, aged between 23 and 43. From here on, these samples and datasets are referred to as “healthy control”, labelled with “H”, numbers 1 to 19. As shown in Table 17, two of these individuals were sequenced using the Ion Torrent PGM, and the remaining individuals using the Illumina MiSeq.

Other samples were also included to economically demonstrate the use of the capture kit, including CD14+ cell depleted samples, as these cells were used by another group for another set of experiments. I acknowledge Alastair Copland for the cell separation component of all T-cell samples and CD14+ cell depleted samples (H5, H6, H10, H11, H12, H13 and H15), using magnetic activated cell sorting (MACS). In total, 14 separate sequencing runs were carried out, including Ion Torrent PGM and Illumina MiSeq platforms, as shown in Table 18.

Sample Reference	Starting Material	Sequencing Platform
H1	PBMC	PGM
H2	PBMC	PGM
H3	PBMC	MiSeq
H4	PBMC	MiSeq
H5	T-cells	MiSeq
H6	T-cells	MiSeq
H7	PBMC	MiSeq
H8	PBMC	MiSeq
H9	PBMC	MiSeq
H10	CD14-Depleted	MiSeq
H11	CD14-Depleted	MiSeq
H12	CD14-Depleted	MiSeq
H13	CD14-Depleted	MiSeq
H14	PBMC	MiSeq
H15	CD14-Depleted	MiSeq
H16	PBMC	MiSeq
H17	PBMC	MiSeq
H18	PBMC	MiSeq
H19	PBMC	MiSeq

Table 17. A summary of the samples taken from healthy individuals. This work comprises samples of peripheral blood mononucleocytes (PBMCs), T-cells and CD14+ depleted PBMCs. The difference in starting material is due to ethical and efficiciency considerations in the use of samples for experimental work.

Run Number	**Platform**	**Samples Included**
TCR1	Ion Torrent PGM	H1a*
TCR2	Illumina MiSeq	H1b
TCR3	Ion Torrent PGM	H2*
TCR4	Ion Torrent PGM	G1
TCR5	Illumina MiSeq	H3/G1*
		G2
TCR6	Illumina MiSeq	H4/G3*
		G4
TCR7	Illumina MiSeq	H5a*
		H5b
		H5c
TCR8	Illumina MiSeq	H6a*
		H6b
		H6c
TCR9	Illumina MiSeq	H5a
		H5b
		H5c
TCR10	Illumina MiSeq	H8*
		H9*
		H10*
		H11*
		H12*
		H13*
		H14*
		H15*
TCR11	Illumina MiSeq	G5
		G6
		G7
		G8
		G9
		G10
		G11
		G12
		G13
		H7*
		H16*

Run Number	Platform	Samples Included
TCR12	Illumina MiSeq	H17a*
		H17b
TCR13	Illumina MiSeq	H18a*
		H18b
		H19a*
		H19b
TCR14	MiSeq	T1a
		T1b
		T1c
		T1d
		T2a
		T2b
		T2c
		T3a
		T3b
		T3c
		T4a
		T4b

Table 18. Summary of all 14 sequencing runs carried out. Two different sequencing platforms were used, including the Ion Torrent PGM and Illumina MiSeq. Samples highlighted above with an asterisk (*) have been part of specific experiments, which will be described in this and later chapters. Samples labelled with "H" refer to healthy controls, "G" refers to samples sequenced for the γδ information, and "T" refers to clinical samples from the cord blood transplantations.

3.2.2 Development of the Library Preparation Workflow

The standard protocols were for library preparation, hybridisation and sequencing were combined and altered for the purposes described in this work. In the sections below, results of the optimisation process are reported upon. Please note that all standard methods are described more comprehensively in Chapter 2.

3.2.3 Comparison of mRNA Isolation Techniques

Total RNA was extracted from PBMCs from healthy controls using TRIzol. mRNA can be isolated from the total RNA through a variety of methods, including poly-A selection, or ribosomal RNA depletion. Here, poly-A selection was selected to minimise the likelihood of sequencing non-expressed RNA molecules. High throughput sequencing experiments can be sensitive so the addition of a carrier molecule, such as glycogen, was omitted to reduce inhibition of downstream PCR reactions.

In the early stages of the methodological development, Agilent, the company producing the custom capture kit, recommended the use of Dynabeads to isolate messenger RNA from the total RNA. However, the same company that produces the library preparation kits, NEBNext also began producing a kit to isolate mRNA. Therefore, a comparison was done between these two kits, which both use a poly-A selection technique to ensure consistency between kits.

Sample	Kit	Total RNA/ng	mRNA Yield/ng	Mean mRNA Yield/ng
1	Dynabeads	1000	14.4	8.93 (±4.49)$^{n.s.}$
2	Dynabeads	1000	9.0	
3	Dynabeads	1000	3.4	
4	NEBNext	1000	10.1	8.93 (±3.25)$^{n.s.}$
5	NEBNext	1000	12.2	
6	NEBNext	1000	4.5	

Table 19. Comparison of mRNA yield using two different poly-A selection kits. Two different poly-A selection kits were used to isolate mRNA, Dynabeads and NEBNext. Three samples were extracted using each kit and compared. An unpaired t-test was performed for comparison of both methods (n.s.: not significant).

The results demonstrated no significant difference in yield between the two kits, as both kits yielded 3-15ng mRNA/1000ng total RNA (Table 19), mRNA generally represents <5% of total RNA. The yields in the current study ranged between 0.34-1.4%. The results demonstrated high variability in eluted mRNA yield. As both protocols gave similar yields, time efficiency was taken into account for continuous use with this library preparation. Magnetic bead isolation protocols like Dynabeads can be performed within twenty minutes, whereas the NEBNext kit requires an additional hour and thirty minutes.

3.2.4 Comparison of Nucleic Acid Fragmentation Techniques

Once mRNA has been isolated, it is then fragmented. Different library preparation kits suggest different techniques. The NEBNext RNA Library Preparation kits that were used for most of the methodological development included an enzymatic fragmentation module designed to fragment mRNA strands to an appropriate length. However, it was found that this technique frequently failed, and material was lost completely.

The protocol defined by NEBNext, as included for this kit and described in the methods sections in 2.4.2 and 2.4.3, was therefore altered and enzymatic fragmentation was abandoned in favor of sonication. Rather than proceeding to nucleic acid fragmentation immediately following mRNA isolation, samples were first taken through the stages of first and second strand cDNA synthesis. Once double stranded cDNA had been synthesised, the material was then fragmented to the desired size using the Covaris S220 Focused Ultrasonicator (Applied Biosystems).

A time course was used to determine optimal settings to achieve the desired fragment sizes. For experiments described in this thesis, dsDNA was sheared to fragments that were 200bp in length, to complement the Illumina MiSeq library preparation, but here dsDNA was also sheared to 400bp for comparison. The desired fragment sizes were produced by adjusting the length and intensity of sonication. The Agilent Bioanalyzer was used to measure the size and yield of fragments produced following sonication.

Covaris Settings 200-300bp	
Duty cycle	10%
Intensity	5
Cycles per burst	200

Table 20 Settings used for the Covaris to produce fragments between 200 and 300bp.

Covaris Settings 400-500bp	
Duty cycle	10%
Intensity	4
Cycles per burst	200

Table 21. Settings used for the Covaris to produce fragments between 400 and 500bp.

This method is imprecise, and a distribution of different sized fragments is produced, therefore a time course was compared to determine the most appropriate length of fragmentation required to achieve the desired fragment size. The size of fragments at the peak of the distribution, measured using the Bioanalyzer, was taken as the length of the fragments in the sample.

Time Course 200-300bp					
Sample	**Time /s**	**cDNA Starting Material/ng**	**Fragmented cDNA yield/ng**	**Fragmented cDNA Yield/%**	**Peak Fragment Size/bp**
1	240	8.76	3.2	36.53	330
2	260	8.76	2.0	22.83	224
3	280	8.76	3.4	38.81	280
4	300	8.76	6.0	68.49	229
5	320	8.76	3.6	41.10	222
Mean	**N/A**	**N/A**	**N/A**	**41.55**	**425**

Table 22. Time course for 200 to 300bp. The settings detailed in Table 20 were used to generate the fragmented cDNA showed in the table above, as recommended by previous protocols. Above 320s sample yield was deemed too low.

Time Course 400-500bp					
Sample	**Time /s**	**cDNA Starting Material/ng**	**Fragmented cDNA yield/ng**	**Fragmented cDNA Yield/%**	**Peak Fragment Size/bp**
1	35	14.38	5.7	39.64	951
2	45	14.38	10.5	73.02	628
3	55	14.38	5.7	39.64	480
4	65	14.38	7.8	54.24	462
5	75	14.38	6.0	41.72	425
Mean	**N/A**	**N/A**	**N/A**	**49.65**	**425**

Table 23. Time course for 400 to 500bp. The settings detailed in Table 21 were used to generate the fragmented cDNA showed in the table above, as recommended by previous protocols. Above 75s sample yield was deemed too low.

Between 260 and 320 seconds for 200 to 300bp, or 55 and 75 seconds for 400 to 500bp, appears to produce the most appropriately sized fragments, shown in Table 22 and Table 23. However, much of the material was lost through sample manipulation and yield was low at 41.55% and 49.65% for 200 to 300bp and 400 to 500bp respectively. Statistical analysis was not performed due to sample size (n = 1).

3.2.5 Pre-Capture PCR Optimisation

Following size selection and adapter ligation, cDNA is amplified by PCR to ensure there is enough material for a successful hybirdisation. The number of PCR cycles was adjusted and minimised for each sample to reduce potential introduction of PCR errors and subsequent bias.

Sample	Adapter-Ligated Material/ng	PCR 1 Material/ng	PCR 2 Material/ng
1	5.40	10.02	288.90
2	3.60	6.39	145.80
3	< Range	11.37	576.00
4	20.25	139.80	N/A
5	22.56	110.10	N/A
6	4.80	40.50	312.00

Table 24. Optimisation of pre-capture PCR amplification. Samples were amplified through different numbers of PCR cycles. Note that sample 3 was out of the limit of detection of the device used to quantify the DNA.

Material was amplified through the pre-capture PCR reaction, described in Table 10. The first round PCR reaction went through 8 cycles of denaturation, annealing and extension. However, for three of the six samples shown for example in Table 24, there was not sufficient material for hybridisation, which stipulates 40ng material minimum per sample. Therefore, a second round of pre-capture PCR was done for samples 1, 2, 3 and 6, with 8 PCR cycles. Sample 6 had enough material, but additional PCR was carried out to ensure there would be leftover sample for storage. This demonstrates issues with pre-capture PCR.

3.2.6 Pre-Capture Pooling and Hybridisation

High-throughput sequencing experiments can be prohibitively expensive. It is cost effect effective to pool samples when sequencing, so multiple samples can be sequenced in one run. In this work, samples were pooled following the pre-capture PCR and before the hybridisation, which had not been previously attempted, as the cost of the baits was high.

Both Illumina MiSeq and Ion Torrent PGM sequencing platforms can generate upwards of ten million reads in a single run. This produces an excessive depth of sequencing for a single sample for most applications. For the application here, it was determined approximately one million reads per sample would be sufficient depth to extract the information desired, as discussed in previous literature (87, 108). Note that samples sequenced on the PGM were not pooled, only those sequenced on the MiSeq.

For the Illumina MiSeq platform, the index sequences themselves were identifying barcodes 6 nucleotides in length, and there were constraints that must be considered as to which index sequence combinations may be used with the MiSeq, due to the optical system of Illumina technology. Index sequences were contained within the adapters, therefore were simultaneously ligated to samples during adapter ligation, as shown in Figure 6 below.

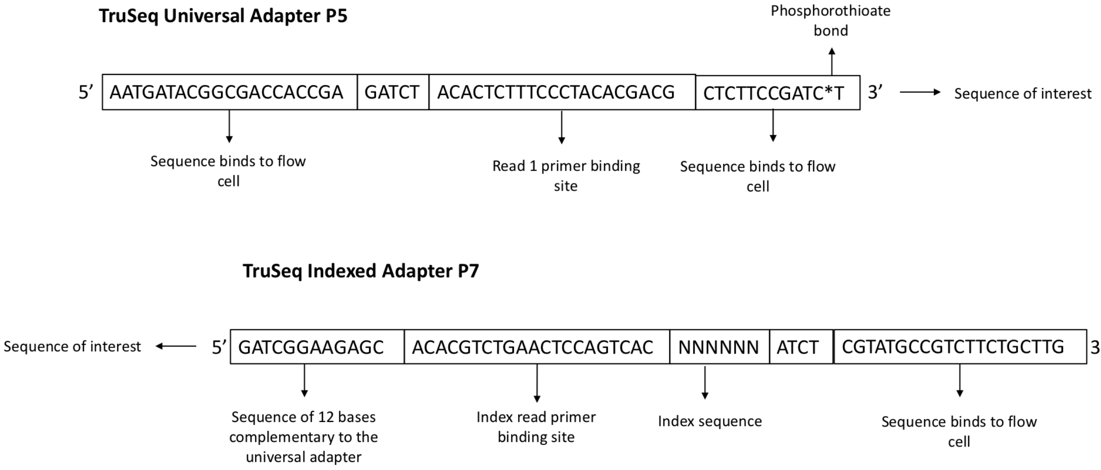

TruSeq Universal Adapter P5
Phosphorothioate bond
5′
AATGATACGGCGACCACCGA
GATCT
ACACTCTTTCCCTACACGACG
CTCTTCCGATC*T
3′
Sequence of interest
Sequence binds to flow cell
Read 1 primer binding site
Sequence binds to flow cell
TruSeq Indexed Adapter P7
Sequence of interest
5′
GATCGGAAGAGC
ACACGTCTGAACTCCAGTCAC
NNNNNN
ATCT
CGTATGCCGTCTTCTGCTTG
3′
Sequence of 12 bases complementary to the universal adapter
Index read primer binding site
Index sequence
Sequence binds to flow cell

Figure 6. Illumina Adapter Sequences. Two adapters are ligated to cDNA fragments during library preparation for the Illumina platform, one to the 5’ end (P5) and one to the 3’ end (P7). Adapters allow the sample to bind to the flow cell, and for the sequencing primers to bind during sequencing. Additionally, the P7 adapter also included the index sequence. The index sequence was 6 nucleotides in length and allowed samples to be multiplexed for sequencing and demultiplexed using specialised software downstream.

Adapter	**Adapter Nucleotide Sequence**
TruSeq Adapter Index 1	ATCACG
TruSeq Adapter Index 2	CGATGT
TruSeq Adapter Index 3	TTAGGC
TruSeq Adapter Index 4	TGACCA
TruSeq Adapter Index 5	ACAGTG
TruSeq Adapter Index 6	GCCAAT
TruSeq Adapter Index 7	CAGATC
TruSeq Adapter Index 8	ACTTGA
TruSeq Adapter Index 9	GATCAG
TruSeq Adapter Index 10	TAGCTT
TruSeq Adapter Index 11	GGCTAC
TruSeq Adapter Index 12	CTTGTA

Table 25. List of Illumina indexes used during this workflow. Indexes are located within the P7 adapter sequence. Samples were pooled together for sequencing on the MiSeq, up to 12 at a time, using these index sequences.

The SureSelect kit from Agilent does not support multiplexing and therefore the workflow required alteration to accommodate sample pooling. Following library preparation, index sequences were added during a pre-capture PCR amplification. Following amplification, the quality and yield of each sample was quantified using a combination of the Qubit (Life Technologies) and Bioanalyzer (Agilent). 40ng of product from each sample was pooled into a single tube. The total volume was adjusted to the correct volume for capture, 3.6µl, through vacuum centrifugation. Each sample to be indexed had one of the index sequences shown in Table 25 ligated to it so the software could identify the different samples. Ultimately, in each sequencing run, up to 12 samples were pooled together. Pooling 12 samples in one run also meant that the amount of material needed for each sample to undergo hybridisation was lower, thus reducing the number of PCR cycles necessary.

3.2.7 Post-Capture PCR Amplification

The protocol provided by the company for use with the SureSelect kit was modified to include the pre-capture sample pooling step described above. Because of this modification, the post-capture PCR reaction had to also be optimised. The Agilent PCR protocol was altered so that the amplification was primed from Illumina adapter sequences, which are used as a start point for base calling during sequencing.

Primer	**Primer Sequence**
P5 reverse complement oligo	TCGGTGGTCGCCGTATCATT
P7 oligo	ATCTCGTATGCCGTCTTCTGCTTG

Table 26. Sequences of the primers used for post-capture PCR amplification.

To amplify the pooled post-capture material, oligos of the reverse complement P5 sequence and the forward direction P7 sequence were used as PCR primers. Table 26 shows the sequences of the oligos used, which correspond to the parts of the Illumina adapters that bind to the flow cell. Like the pre-capture PCR amplification, it was important to limit the number of PCR cycles to preserve sequence integrity. The amount of material present in the captured samples was consistently too low to be detected by either of Qubit or Bioanalyzer. Therefore, the only quantification that could be done was following the post-capture PCR.

Reagent	**Concentration/µM**	**Volume/µl**	**PCR Stage**		**Temperature/ºC**	**Time/s**
			Initial Denaturation		98	45
Captured Sample	Unknown	23	Denaturing	x12 Cycles	98	15
KAPA mastermix	Unknown	25	Annealing		65	30
P5 Oligo	25	1	Extending		72	30
P7 Oligo	25	1	Final Extension		72	60

Table 27. Initial protocol for post-capture PCR amplification.

However, following multiple attempts with these conditions, no material was amplified. This could be due to either issues with the capture, or issues with the PCR protocol. Therefore, both pre-capture and post-capture samples were amplified again, with a slight modification to the protocol, shown in Table 28. The annealing temperature used was adjusted to 60°C, to optimise the efficiency of the polymerase enzyme, rather than 65°C, which was the optimal annealing temperature for the primers.

PCR Stage		**Temperature/ºC**	**Time/s**
Initial Denaturation		98	45
Denaturing	x12 Cycles	98	15
Annealing		60	30
Extending		72	30
Final Extension		72	60

Table 28. Modified PCR cycling conditions for the post-capture PCR.

Once the annealing temperature was adjusted, both sets of samples from pre-capture and post-capture amplified, as shown in Table 29. Therefore, both the PCR and the capture reactions were successful and pooled samples can be sequenced using this protocol. Once the successful PCR cycling conditions had been determined, this protocol was used for all samples multiplexed and sequenced on the MiSeq. Sequencing of captured, amplified samples was then performed as detailed in the chapter on General Methods.

Sample		**PCR Amplified Material/ng**
Pre-capture	1	253.00
	2	127.00
	3	504.00
Post-capture	1	1.16
	2	1.04
	3	2.15

Table 29. Yield of post-capture PCR amplification.

3.2.8 Sequencing Data Metrics and Quality Control

It was important to monitor and track sequencing data metrics for quality control. Therefore, FASTQ generated from sequencing captured T-cell receptor sequences using both Illumina MiSeq and Ion Torrent PGM platforms were downloaded, demultiplexed and quality assessed using FastQC (109). This program gives an overview of various run metrics including basis statistics, per base sequence quality and content, GC and N content. These metrics enable rapid identification of issues with the sequencing run and if these results cause concern, it may be necessary to repeat the run.

Figure 7 demonstrates the run metrics for quality scores obtained for an example sample. The quality scores, Q scores, demonstrate the software's confidence in the data, according to an internal algorithm. If a sample were to have unusually low sequence quality across all sequences, it is likely that the run would be aborted, and the sequence data would have been discarded. Although the quality scores varied across reads, there were no issues with low sequence quality following TCR capture. It was anticipated that this might be an issue due to sequencing many cDNA fragments of highly similar sequence. That there were no issues indicated that the capture method produces significant sequence diversity for both Illumina MiSeq and Ion Torrent PGM platforms. The number of sequences obtained per sample varied widely, shown in Table 30.

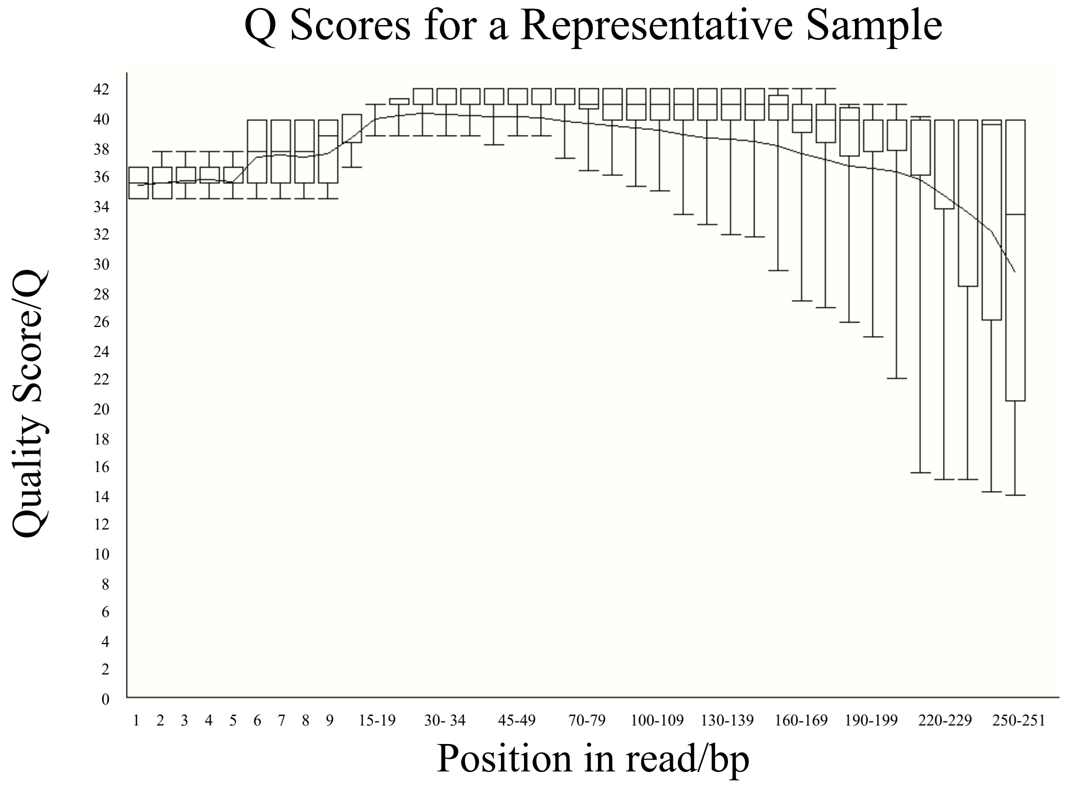


Figure 7. Quality scores across all bases (Sanger/Illumina 1.9 encoding). Quality scores are high, with a Q score >30 for each base called during sequencing (figure created using FASTQC software). As sequencing progresses the quality score tends to decrease, therefore a cutoff of Q30 was applied to all samples as described in 2.6.5.

Sample	Platform	Multiplexing	Total Sequences
H1	Ion Torrent PGM	1	2,127,004
H2	Ion Torrent PGM	1	2,265,226
H3	Illumina MiSeq	2	8,852,295
H4	Illumina MiSeq	2	6,606,392
H5	Illumina MiSeq	3	2,675,141
H6	Illumina MiSeq	3	7,063,767
H7	Illumina MiSeq	12	2,394,924
H8	Illumina MiSeq	8	222,079
H9	Illumina MiSeq	8	766,509
H10	Illumina MiSeq	8	431,497
H11	Illumina MiSeq	8	585,580
H12	Illumina MiSeq	8	532,042
H13	Illumina MiSeq	8	316,579
H14	Illumina MiSeq	8	321,645
H15	Illumina MiSeq	8	1,070,866
H16	Illumina MiSeq	12	675519
H17	Illumina MiSeq	2	7457921
H18	Illumina MiSeq	4	5937409
H19	Illumina MiSeq	4	4476809

Table 30. Sequencing reads acquired per sample. The number of reads acquired for each sample was not proportional when multiplexing was considered. Only the samples from the healthy individuals were included here to demonstrate.

3.3 Results

3.3.1 Alignment of Data to Target Transcripts

It was hypothesised that most sequencing reads obtained from both platforms would align to the TCR and Ig target sequences. To investigate this, FASTQ files were analysed using the Burrows Wheeler Alignment (BWA) tool. BWA aligns reads according to a reference library, in this case the transcriptome, and outputs results in a SAM file. Data presented below was acquired from the SAM files of 11 representative samples from the healthy individuals that were sequenced.

To simplify the analysis of alignment data, a script was created in Python that read the SAM files and extracted the relevant information. This script labelled "Dickey3", an abbreviation of "dictionary key version 3", can be found in the appendix (Appendix 1). This script generated a ".txt" file as output, which could be exported to Excel for analysis allowing for a rapid comparison of transcriptome alignment.

The design of the capture kit included probes for both Ig and TCR sequences and it was hypothesized that most sequences obtained should align to one of these regions. Therefore, both Ig and TCR are included in the analysis of alignment of data to target transcripts.

Sample	Total TCR/%	Total Ig/%	Total TCR & Ig/%	Off-Target/%
H8	25.09	33.86	58.95	41.05
H9	13.07	34.99	48.06	51.94
H10	15.00	32.72	47.72	52.28
H11	10.10	38.38	48.48	51.52
H12	10.75	36.97	47.72	52.28
H13	13.42	25.54	38.96	61.04
H14	20.31	20.29	40.59	59.41
H15	8.75	37.88	46.63	53.37
H17	12.60	37.41	50.01	49.99
H18	24.01	33.49	57.50	42.50
Mean	15.31	33.15	48.46	51.54

Table 31. Proportions of reads that were aligned to the identified target region. Analysis was carried out using BWT, as described in 2.6.1.

Table 31 shows the results for 10 representative samples taken from the healthy individuals that were sequenced and the starting material, which was either total PBMCs or CD14 depleted samples. Therefore, all samples shown here contained the total T-cells and B-cells. This shows the proportions of reads that were aligned to either TCR or Ig sequences, and the average proportion of on-target reads.

A mean of 48.46% of reads across all samples tested were aligned to target sequences. In this case, this was a library produced using RNA extracted from total PBMCs. A substantial difference between alignments to T-cell and B-cell sequences was observed, 15.31% and 33.15% of alignments respectively, as shown in Figure 8.

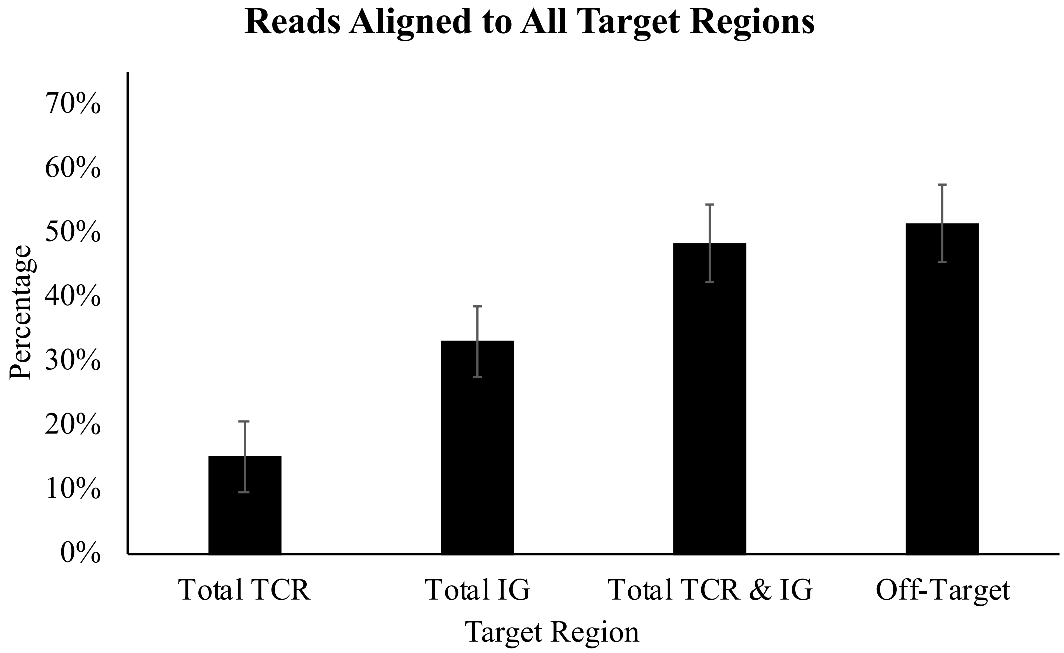


Figure 8 Proportions of reads aligned to the target regions. The mean values and respective standard deviations were calculated from Table 31 and are represented here.

This data also demonstrates a variation across individuals. For example, sample H15 appeared to have a lower proportion of T-cell alignments than the other samples, but the proportion of Ig aligned sequences is towards the higher end of the range.

Most reads that were successfully aligned to the TCR were aligned to the β chain, as shown in Figure 9. Similarly, most reads that were successfully aligned to the Ig were aligned to the Ig heavy chain, as shown in Figure 10.

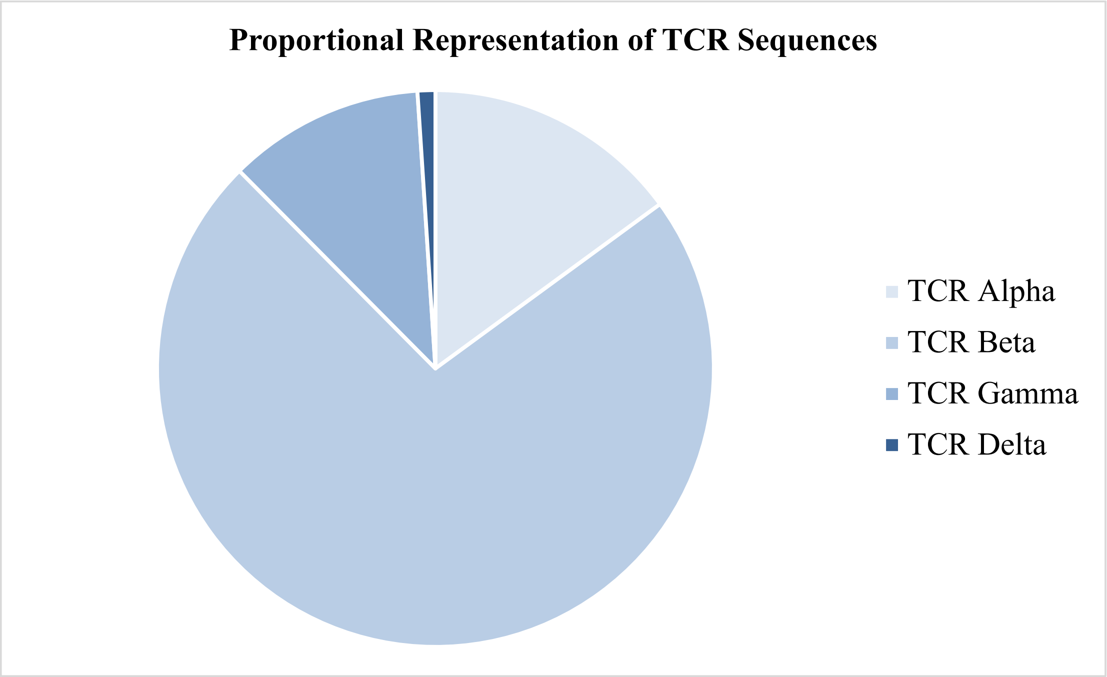


Figure 9. Proportional Representation of TCR Sequences. Proportions of reads that were aligned to TCR sequences, for each of the four TCR chains, for the ten representative samples.

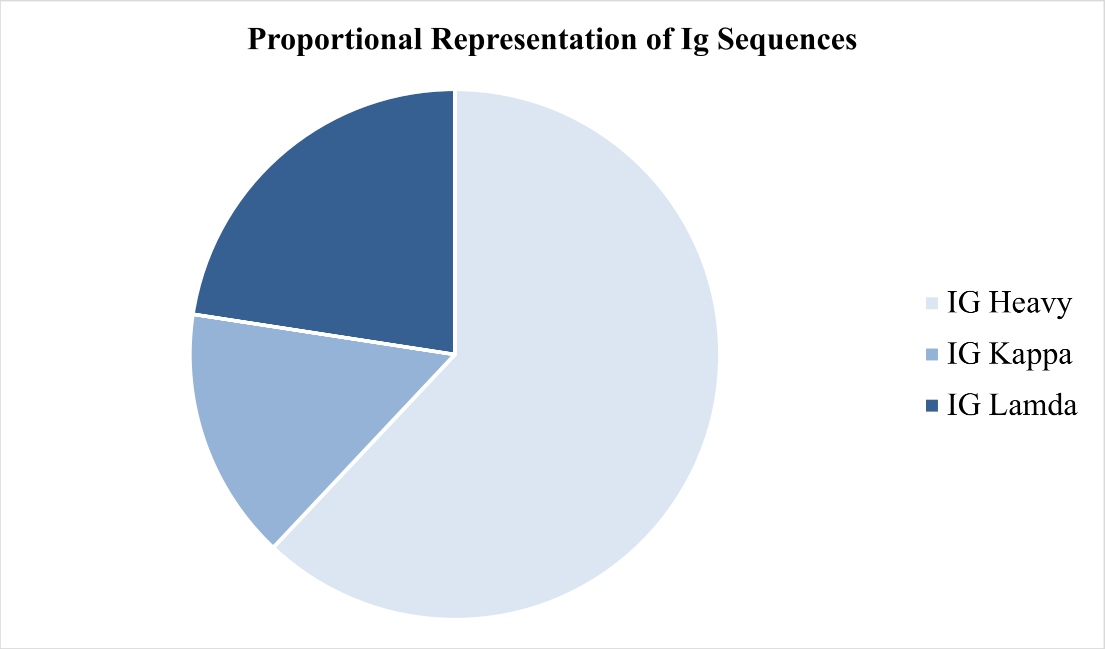


Figure 10. Proportional representation of Ig sequences. Proportions of reads that were aligned to TCR sequences, for each of the three Ig chains, for the ten representative samples.

Figure 11 and Figure 12 show the proportions of reads aligned to the TCR chains and Ig chains respectively. For the TCR, alignments to the β were higher than to the α, and alignments to γ and δ are lower than to α and β.

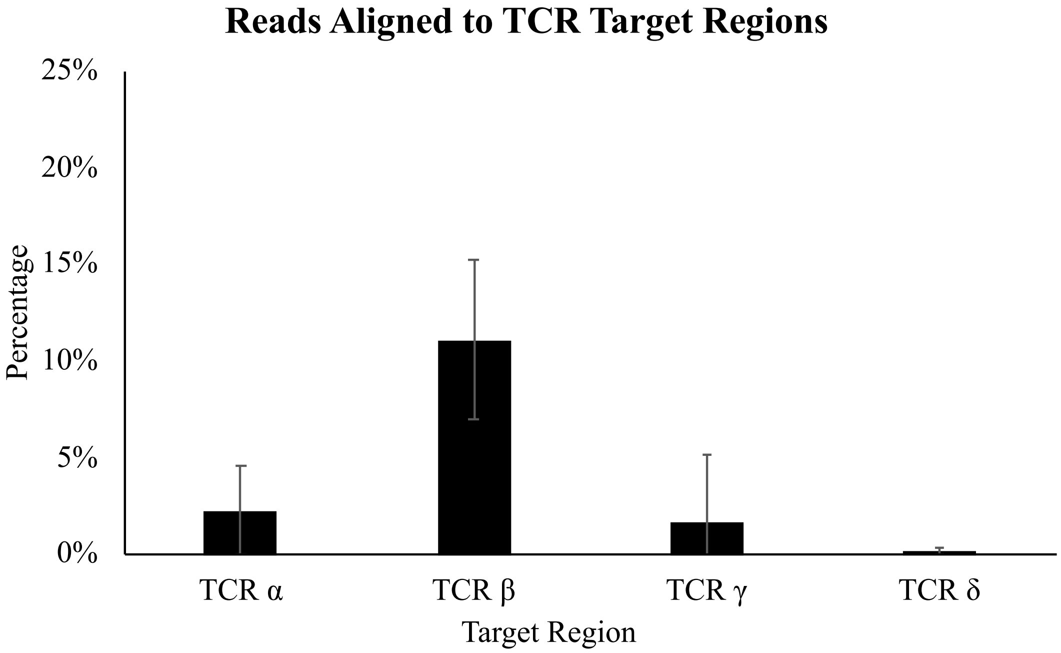


Figure 11 Proportion of total aligned reads that were aligned to TCR chains. The highest proportion of alignments were to the β chain, followed by the α chain. γ and δ chain alignments were much lower. The mean values and respective standard deviations of samples H8 to H18 were calculated.

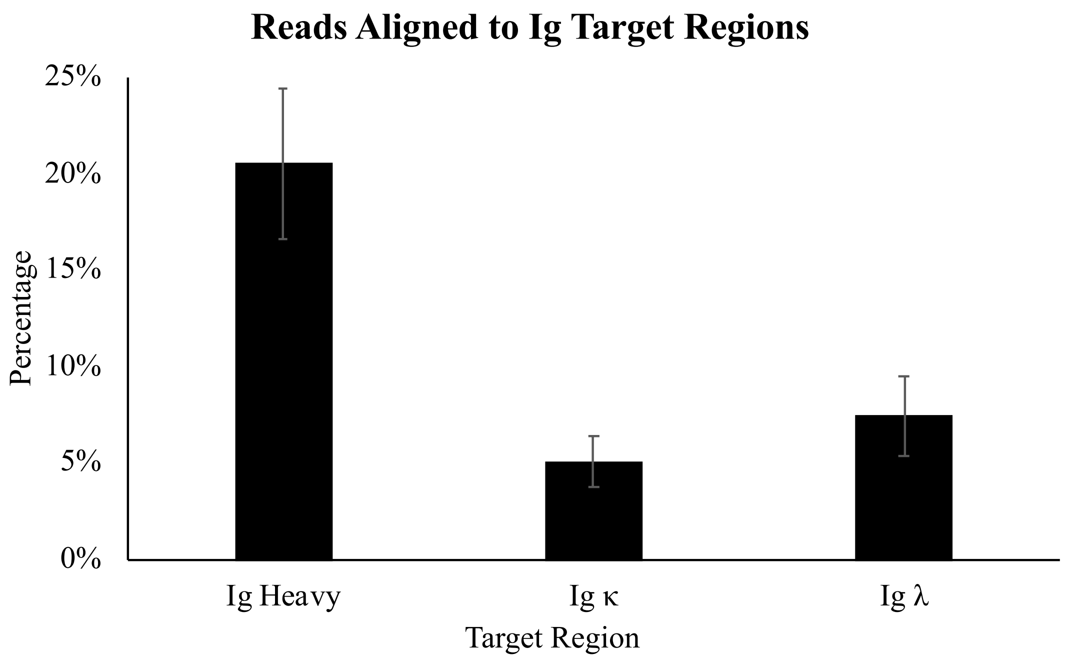


Figure 12 Proportion of total reads aligned that were aligned to Ig chains. Quantified proportions of reads that were aligned. The mean values and respective standard deviations of samples H8 to H18 were calculated.

The custom capture baits were ordered in two separate batches from Agilent and there was a concern that the quality of these two batches may differ, as the first two attempts at the capture reaction were unsuccessful. Therefore, the two batches were run in parallel using the same sample and the alignment compared. Direct comparison was achieved through splitting of a pre-capture library (H17) and performing separate capture reactions on each in parallel using a different batch of the capture baits, as shown in Table 32.

Chain	**H17a**		**H17b**	
	Number	**Proportion/%**	**Number**	**Proportion/%**
Total	15404497	100.00	4698812	100.00
TCR α	446334	2.90	117964	2.51
TCR β	1385118	8.99	438121	9.32
TCR γ	88620	0.58	12147	0.26
TCR δ	24001	0.16	8918	0.19
Ig Heavy	3051972	19.81	939325	19.99
Ig κ	974852	6.33	335854	7.15
Ig λ	1736645	11.27	598040	12.73
Total TCR	1944073	12.62	577150	12.28
Total Ig	5763469	37.41	1873219	39.87
Total TCR/Ig	7707542	50.03	2450369	52.15
Other	7696955	49.97	2248443	47.85

Table 32. Alignment of H17 to the target regions. Sample H19 was split into two libraries before capture, and each library underwent capture with a different batch of capture kit.

So far, only the alignment of data generated using total PBMCs as starting material has been described. Additional libraries were captured and sequenced from samples of sorted T-cells. It was observed that for experiments in which only the T-cells were sequenced, there was an increase in the number of reads that aligned to the target regions, without having additional Ig reads (Table 33). Very few Ig sequences were detected (1.39%) and the proportion of reads that aligned to TCR sequences increased from 15.31% (Table 31), to 44.33%. However, there was still a large proportion of reads that aligned outside of the target areas, 51.89% on average. Statistical analysis was not undertaken on this set of results due to low sample size (n = 2).

Sample	**H5a**	**H6a**
TCR α	11.54%	18.52%
TCR β	21.70%	29.77%
TCR γ	1.89%	4.20%
TCR δ	0.18%	0.86%
Ig Heavy	2.41%	0.71%
Ig κ	1.39%	0.27%
Ig λ	2.18%	0.60%
Total TCR	35.31%	53.35%
Total Ig	5.99%	1.57%
Total TCR/Ig	41.29%	54.92%
Other	58.71%	45.08%

Table 33. Alignment of reads generated from sorted T-cell libraries. Samples H5a and H6a, both originated from PBMC samples but were sorted for T-cells only, prior to RNA extraction.

3.3.2 Off-Target Alignment of Data

As mentioned previously, the results were showing significant numbers of reads of off-target sequences. An alignment of sample H9 was produced using BWA, and the number of on- and off-target reads plotted (Figure 13). The off-target reads were analysed to determine the most frequent occurrences within the sample. Many of these alignments mapped to genes that are transcribed at a high frequency. High frequency off-target alignments include mitochondrial, MHC, beta actin and ubiquitin transcripts. It is unclear whether a high abundance of these in the original RNA samples has meant they have been retained through overwhelming the capture baits, or due to the enrichment of these transcripts by the baits themselves. This indicated that the purity of the sample had a significant effect on the quality and quantity of usable data obtained and also that the capture may not have been specific enough for true enrichment.

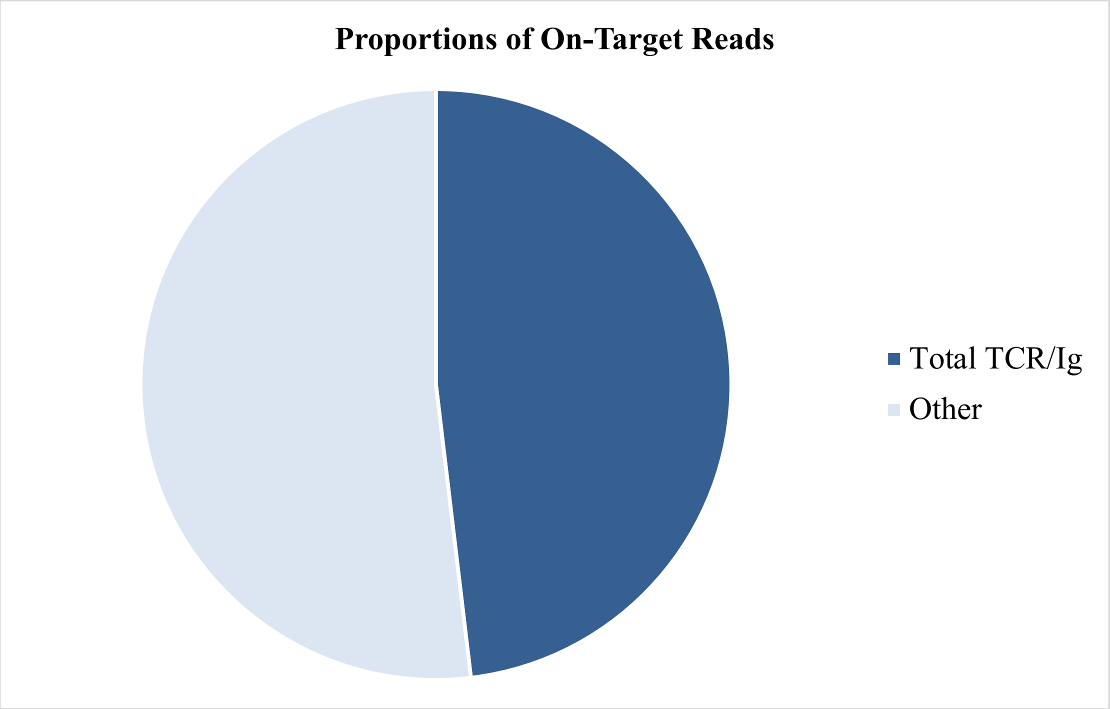


Figure 13. Proportional representation of on target and off target reads. Over half of the sequencing reads on average from were off target and aligned to other sequences.

A high proportion of sequencing reads aligned off-target, on average 51.54%. Files were analysed using BWA and the output files were exported to Excel. Table 34 lists the top 20 off-target alignments from the Excel file, for a represeentative sample. Two of the off target reads were identifed only through their Ensembl gene spliced transcript (ENST) number. The majority of the remaining alignments originated from mitochondrial RNA sequences. This suggests that the off target alignments may be due to non-specific binding of transcripts that are high in concentration.

ENST code	Frequency	Gene Name	Gene Info
ENST00000430694	179225	AC096579.7	
ENST00000361624	17665	MT-CO1	mitochondrially encoded cytochrome c oxidase I [Source:HGNC Symbol;Acc:7419]
ENST00000361381	10452	MT-ND4	mitochondrially encoded NADH dehydrogenase 4 [Source:HGNC Symbol;Acc:7459]
ENST00000361899	9601	MT-ATP6	mitochondrially encoded ATP synthase 6 [Source:HGNC Symbol;Acc:7414]
ENST00000316292	7661	EEF1A1	eukaryotic translation elongation factor 1 alpha 1 [Source:HGNC Symbol;Acc:3189]
ENST00000361739	6799	MT-CO2	mitochondrially encoded cytochrome c oxidase II [Source:HGNC Symbol;Acc:7421]
ENST00000309311	6374	EEF2	eukaryotic translation elongation factor 2 [Source:HGNC Symbol;Acc:3214]
ENST00000361789	5977	MT-CYB	mitochondrially encoded cytochrome b [Source:HGNC Symbol;Acc:7427]
ENST00000302754	4901	JUNB	jun B proto-oncogene [Source:HGNC Symbol;Acc:6205]
ENST00000361453	4890	MT-ND2	mitochondrially encoded NADH dehydrogenase 2 [Source:HGNC Symbol;Acc:7456]
ENST00000361567	4533	MT-ND5	mitochondrially encoded NADH dehydrogenase 5 [Source:HGNC Symbol;Acc:7461]

ENST code (cont.)	Frequency	Gene Name	Gene Info
ENST00000362079	4001	MT-CO3	mitochondrially encoded cytochrome c oxidase III [Source:HGNC Symbol;Acc:7422]
ENST00000361390	3585	MT-ND1	mitochondrially encoded NADH dehydrogenase 1 [Source:HGNC Symbol;Acc:7455]
ENST00000414273	3493	hsa-mir-6723	hsa-mir-6723 [Source:miRBase;Acc:MI0022558]
ENST00000303562	3320	FOS	FBJ murine osteosarcoma viral oncogene homolog [Source:HGNC Symbol;Acc:3796]
ENST00000301740	2884	SRRM2	serine/arginine repetitive matrix 2 [Source:HGNC Symbol;Acc:16639]
ENST00000426066	2724	LL22NC03-80A10.6	
ENST00000361681	2688	MT-ND6	mitochondrially encoded NADH dehydrogenase 6 [Source:HGNC Symbol;Acc:7462]
ENST00000318052	2611	EEF1A1P6	eukaryotic translation elongation factor 1 alpha 1 pseudogene 6 [Source:HGNC Symbol;Acc:3201]
ENST00000282388	2433	ZFP36L2	ZFP36 ring finger protein-like 2 [Source:HGNC Symbol;Acc:1108]

Table 34. The 20 most frequent off-target alignments identified for a representative sample (H10).

3.3.3 Processing of Illumina MiSeq and Ion Torrent PGM TCR Reads

High-throughput sequencing can generate millions of reads in a single run. However, across all platforms, quality control steps must be taken to remove low quality reads from datasets to ensure accurate results. Although true for all sequencing platforms to various degrees, analysis must be tailored to each method. As discussed, two different sequencing platforms were used, the Illumina MiSeq and the Ion Torrent PGM, each of which requires an individually designed pipeline. After sequencing on the MiSeq, the mate-pairs produced from Illumina paired-end sequencing had to be aligned.

Due to base mismatches, not all paired-end reads could be aligned to their mate. It was initially found that very few reads were paired together from the raw, unprocessed FASTQ data files. After the Illumina adapter sequences were trimmed from both read 1 and read 2, still a low proportion of reads were paired. It was hypothesised that many mate pairs did not align due to the distribution of cDNA fragment sizes. Therefore, the program Cutadapt was used to trim read lengths and identify the optimum sized fragment for the identification of TCR sequences, using H9 as an example. Table 35 shows that the most data was retained when 125bp length fragments were combined, so this was applied to any subsequent FASTQ file that required trimming. This is represented in Figure 14. Table 36 demonstrates the importance of read trimming.

	Trimmed Length/bp				
	100	**125**	**150**	**175**	**200**
Uncombined Reads/%	43.65	2.08	6.80	21.19	78.95
Combined Reads/%	56.35	97.92	93.20	78.81	21.05

Table 35. Optimisation of read trimming. Percentages of reads that were paired together once they were trimmed to different lengths for sample H9. Statistical analysis was not performed due to sample size (n = 1).

	125bp	**Adapter Trimmed**	**Raw FASTQ**
Uncombined Reads/%	2.08	81.99	99.80
Combined Reads/%	97.92	18.01	0.20

Table 36 Necessity of read trimming. Percentages of reads that were paired together once they were trimmed to the optimal 125bp, compared to the reads that were not trimmed. Of the reads that were adapter trimmed, 18.01% were combined, and only 0.20% of the raw FASTQ reads were combined, highlighting the need for trimming.

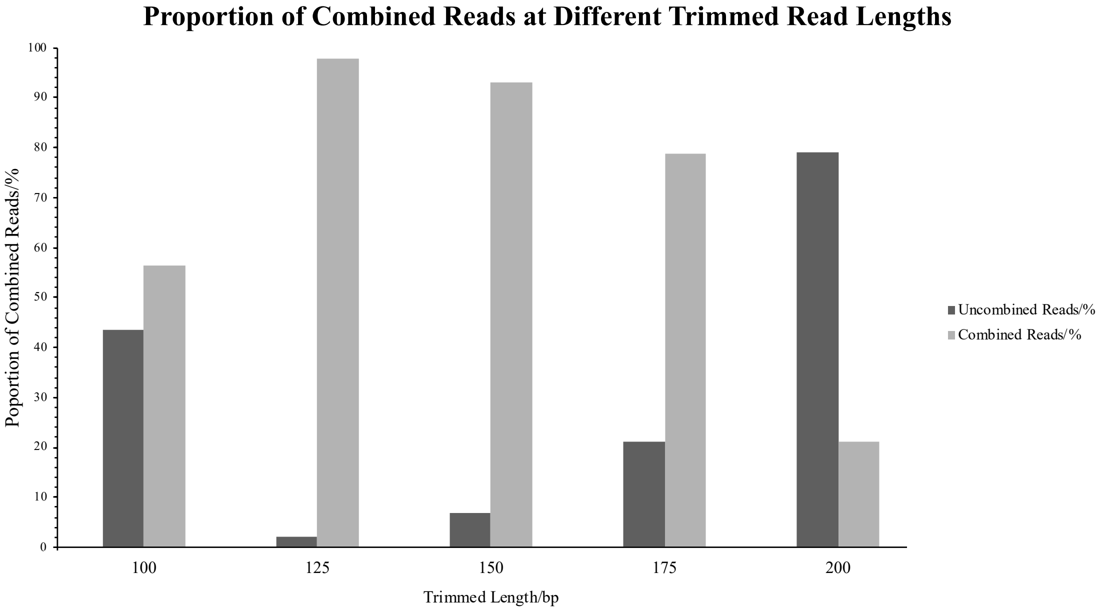


Figure 14. Proportion of combined reads at different trimmed read lengths. Sample H9 was used to demonstrate this information. Statistical analysis was not performed due to sample size (n = 1).

After pairing reads together, an additional program, Fastx Toolkit, was used for the rejection of poor-quality sequences and applying a quality score threshold. This reduced the impact of sequencing errors on the identification of unique TCR sequences downstream. The total number of reads obtained that passed the MiSeq filter per sample varied from 675,519 to 7,457,921, with an average of 3.22 million. On average, 16.3% of these reads were lost following the application of a stringent quality score of Q30, as seen in Figure 16. The data obtained using the Ion Torrent PGM platform was single end, and therefore the only step required was to discard reads based on quality. Quality scores gave an indication of the accuracy and reliability of sequences. Sequencing TCR repertoires requires the ability to accurately identify unique clonotypes and therefore sequencing errors that could introduce unique species through nucleotide substitutions, additions or deletions. The majority of reads were of high quality, with a Q score of at least 30, meaning that there is in a 1 in 1000 likelihood that the software has called an incorrect base, as shown in Figure 15 and Figure 16. Therefore, a cutoff of Q30 was applied to all samples before analysis.

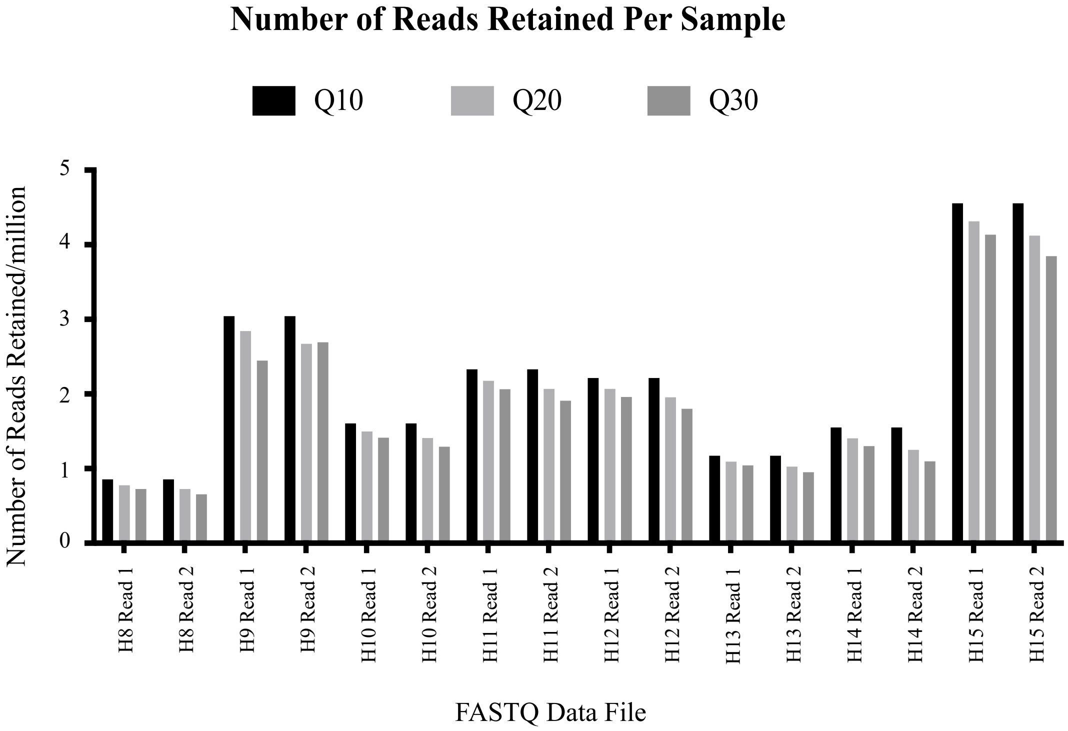


Figure 15 Number of Reads Retained at different Q score filters. The number of reads with a quality score of Q10, Q20 and Q30 is shown for each sample is shown.

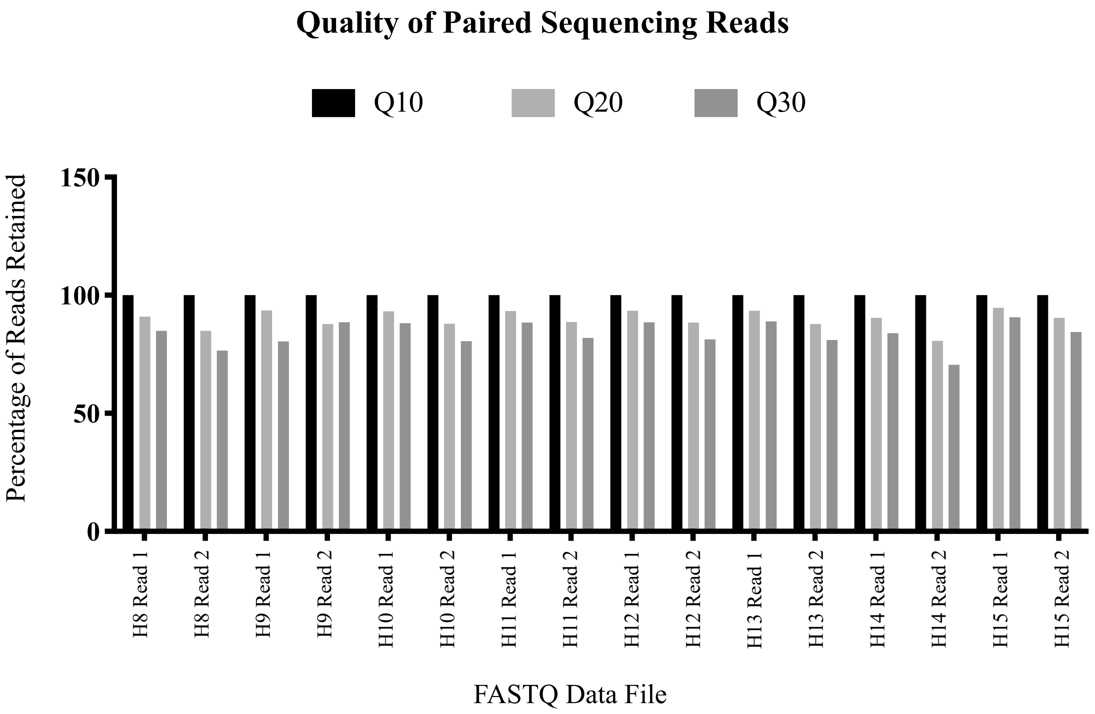


Figure 16. Quality of sequencing reads for 8 samples. The number of reads retained at quality score cutoffs of Q10, Q20 and Q30 (bottom) is shown. Samples H8 to H15 were used to represent.

3.3.4 Investigating VJ Usage of TCR α β γ and δ Chains Simultaneously

In addition to obtaining sequence information for the β chain, which has been well documented previously, this capture technique also allowed the gathering of data on the α, γ and δ chains as well as Immunoglobulin heavy, κ and λ chains, all of which are sequenced at the same time. The V and J gene usage for each TCR chain is shown for a single healthy individual below in Figure 17, Figure 18, Figure 19 and Figure 20. Data represented here was generated using the Decombinator.

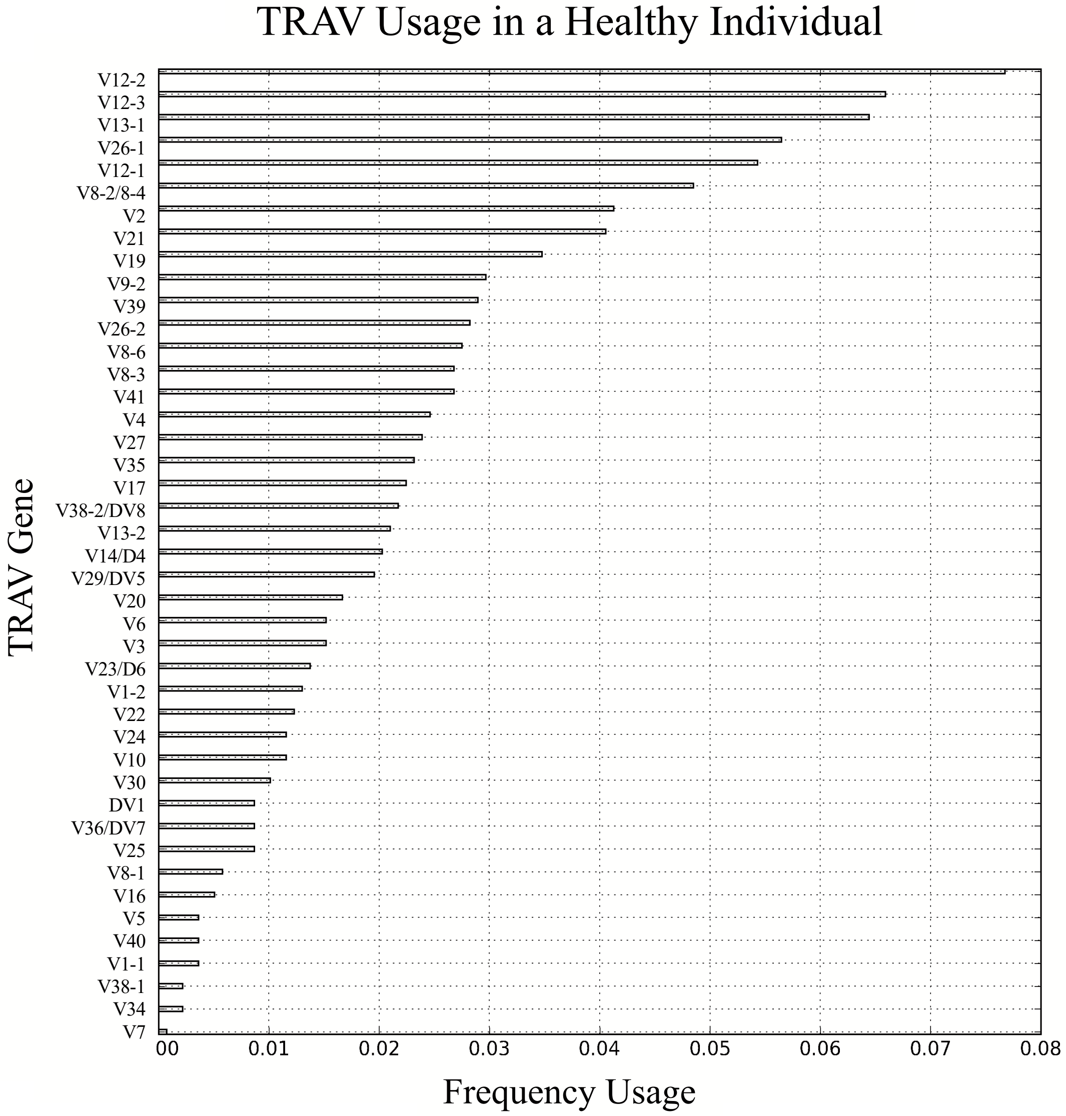


Figure 17. TCR α chain gene usage in a representative healthy individual. A single example has been selected to demonstrate the gene usage of the TCR α chain (H9). Statistical analysis was not performed due to sample size (n = 1).

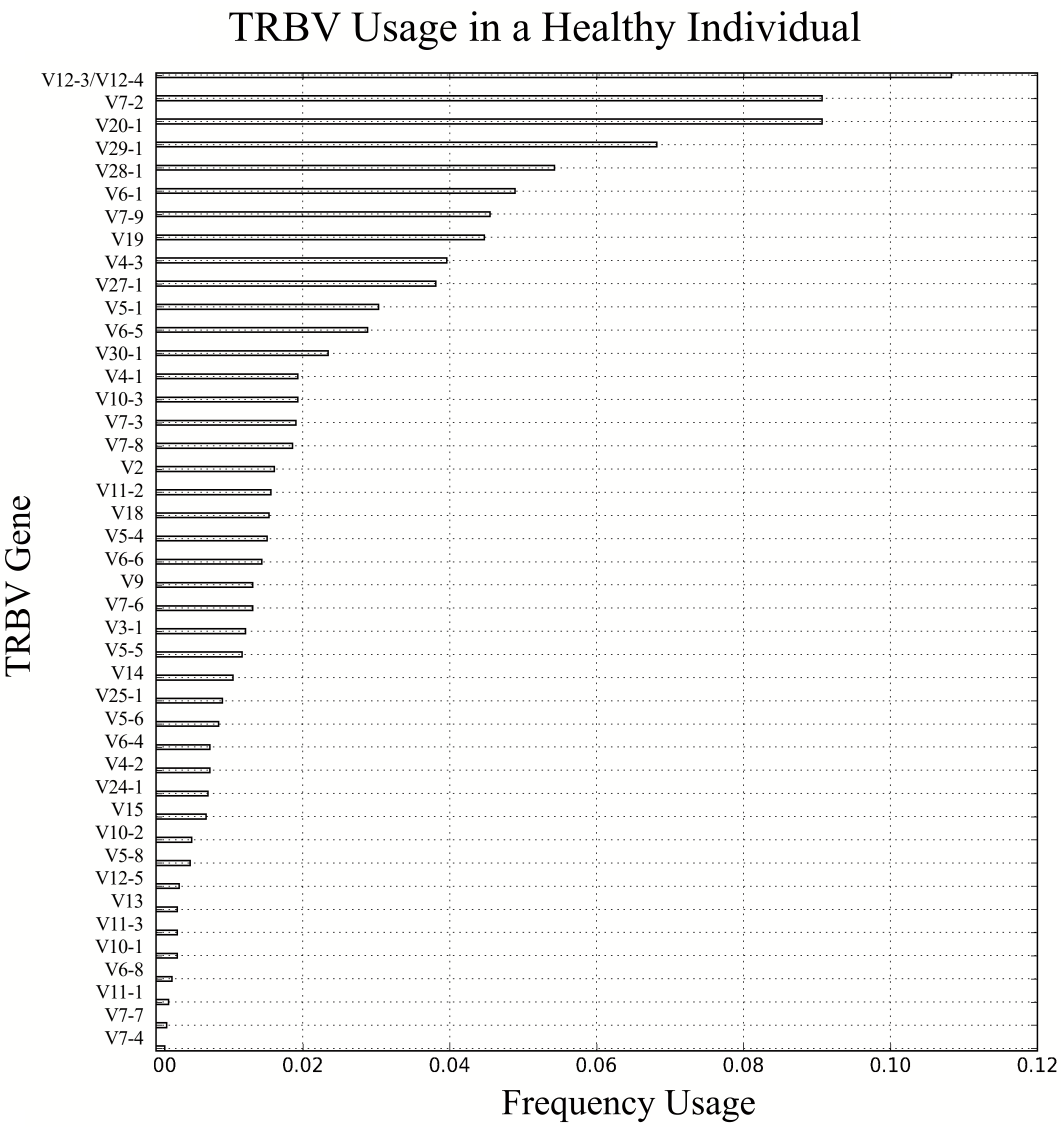


Figure 18. TCR β chain gene usage in a representative healthy individual. A single example has been selected to demonstrate the gene usage of the TCR β chain (H9). Statistical analysis was not performed due to sample size (n = 1).

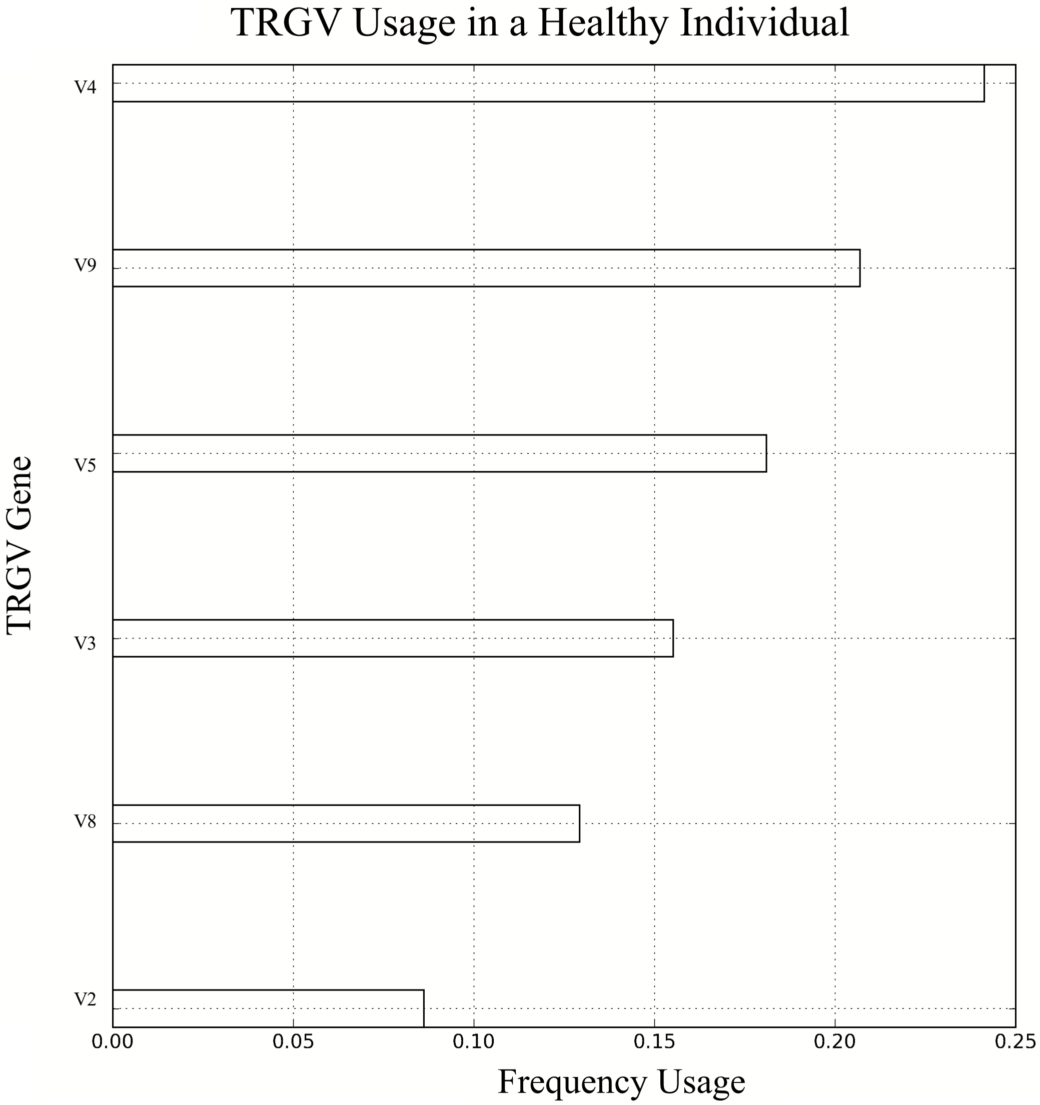


Figure 19. TCR γ chain gene usage in a representative healthy individual. A single example has been selected to demonstrate the gene usage of the TCR γ chain (H9). Statistical analysis was not performed due to sample size (n = 1).

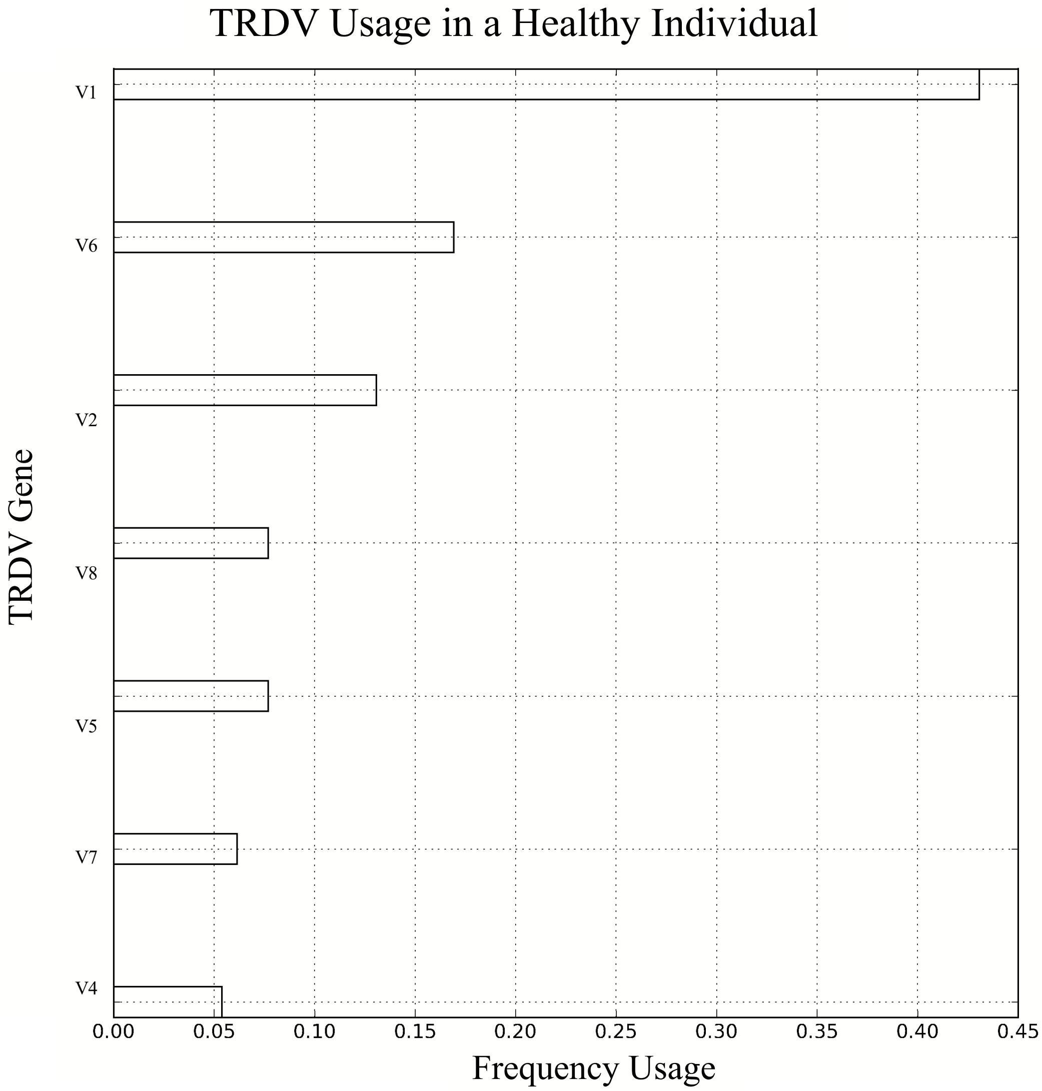


Figure 20. TCR δ chain gene usage in a representative healthy individual. A single example has been selected to demonstrate the gene usage of the TCR δ chain (H9). Statistical analysis was not performed due to sample size (n = 1).

Each TCR chain demonstrated a bias in gene expression, shown in Figure 17, Figure 18, Figure 19 and Figure 20. Some V and J segments are utilised more frequently than others, which is most likely a result of bias during recombination. As well as the representation of each of the V and J being non-uniform, there was also a bias in the pairings of these two gene segments during recombination. This is shown in the heat map in Figure 21. It could be due to clonal proliferation or a bias in the recombination process.

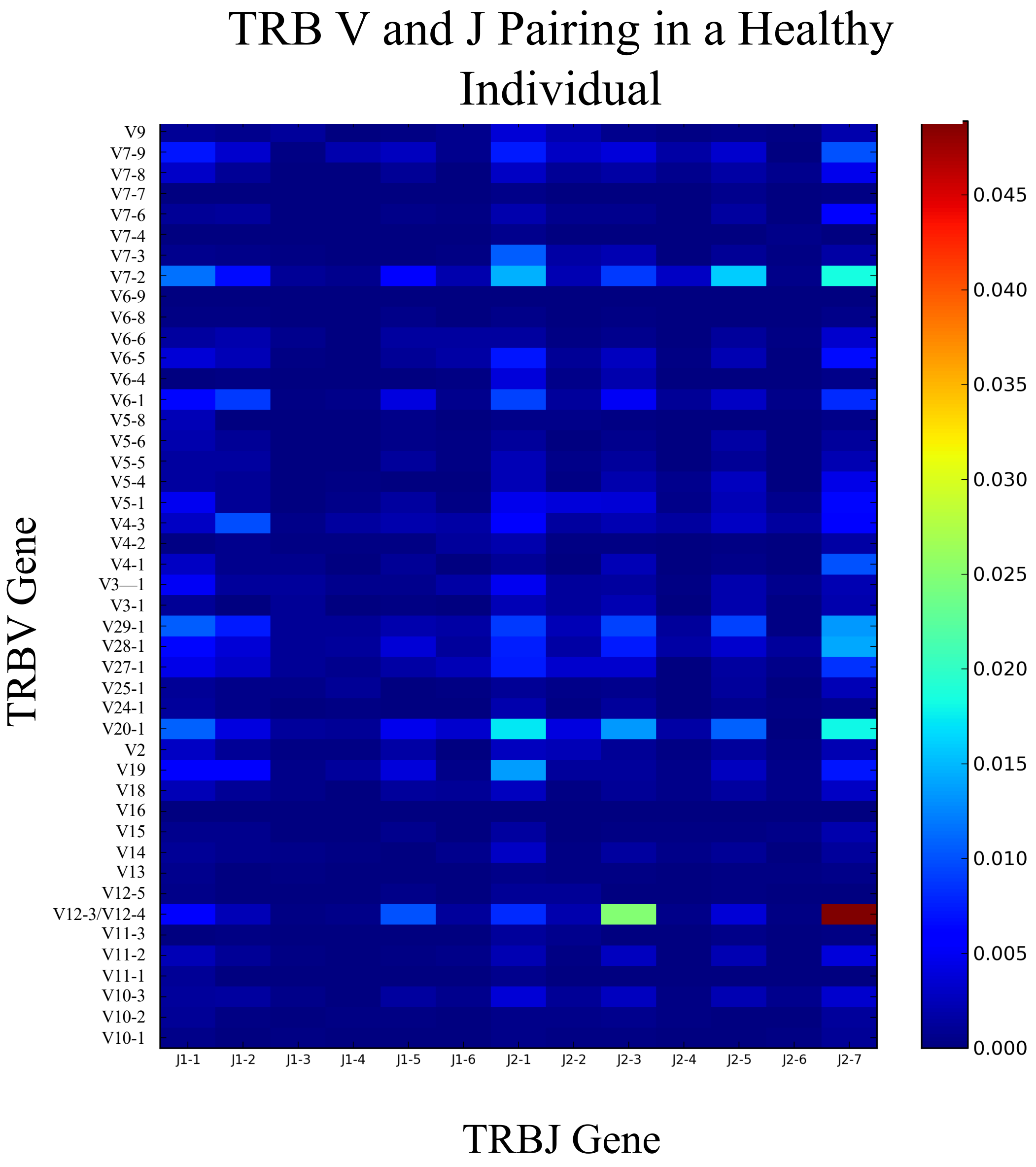


Figure 21. VJ usage heat map for the TCR β chain. V segments genes are on the y axis and J segment genes are on the x axis. The legend on the right displays the scale; the highest frequency of pairing was 0.05 (5.0%), which demonstrates non-random pairing of V and J segments. The pairing of V12-3/V12-4 with J2-7 was the most frequent. A single example has been selected to demonstrate the VJ pairing of the β chain (H9).

3.3.5 Subsampling of FASTQ Data Files

As previously described, high-throughput sequencing produces large amounts of data, which may not be necessary for all experimental questions. Therefore, to optimise multiplexing, the number of reads required to produce meaningful information regarding the TCR repertoire was addressed, looking at different sized subsamples of data. This was done using a Python script that takes FASTQ files as input, with an option to determine how large the subsample will be and creates a new FASTQ file of randomly selected sequences, which is shown in the appendix (Appendix 2). Sequences were selected without replacement, to prevent repeated sequences. Subsamples of 10,000, 100,000, 500,000 1,000,000, 2,000,000, 3,000,000 and 4,000,000 sequences were taken from the largest set of processed data, sample number H3, which contained over 8,000,000 sequences. Subsamples larger than 4,000,000 were not taken due to time constraints, as the script requires a large amount of computing memory to run and becomes increasingly slower as the size of the subsample increases. Each subsampled output file was analysed using the Decombinator and the results are displayed in Figure 22 and Figure 23 below.

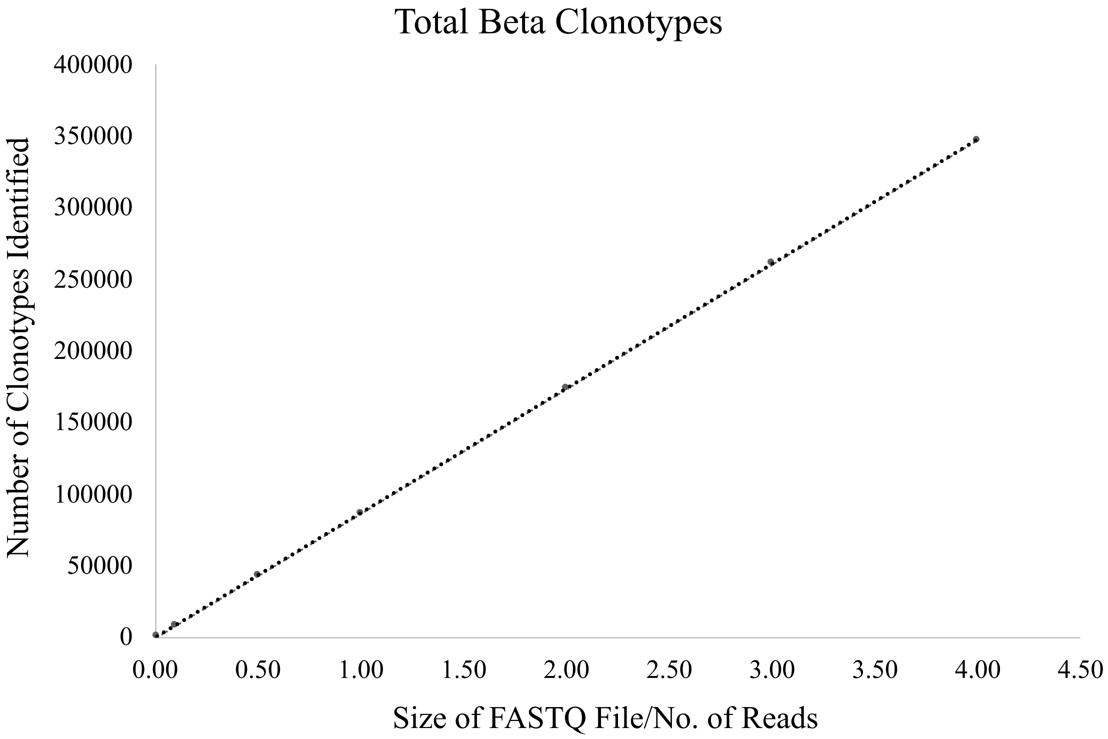


Figure 22. Total number of β chain sequences identified from subsampled data. A large FASTQ file was subsampled using a Python script, and smaller numbers of sequences randomly selected from it and compiled into a file of their own, which was analysed using the Decombinator. Statistical analysis was not performed due to sample size (n = 1).

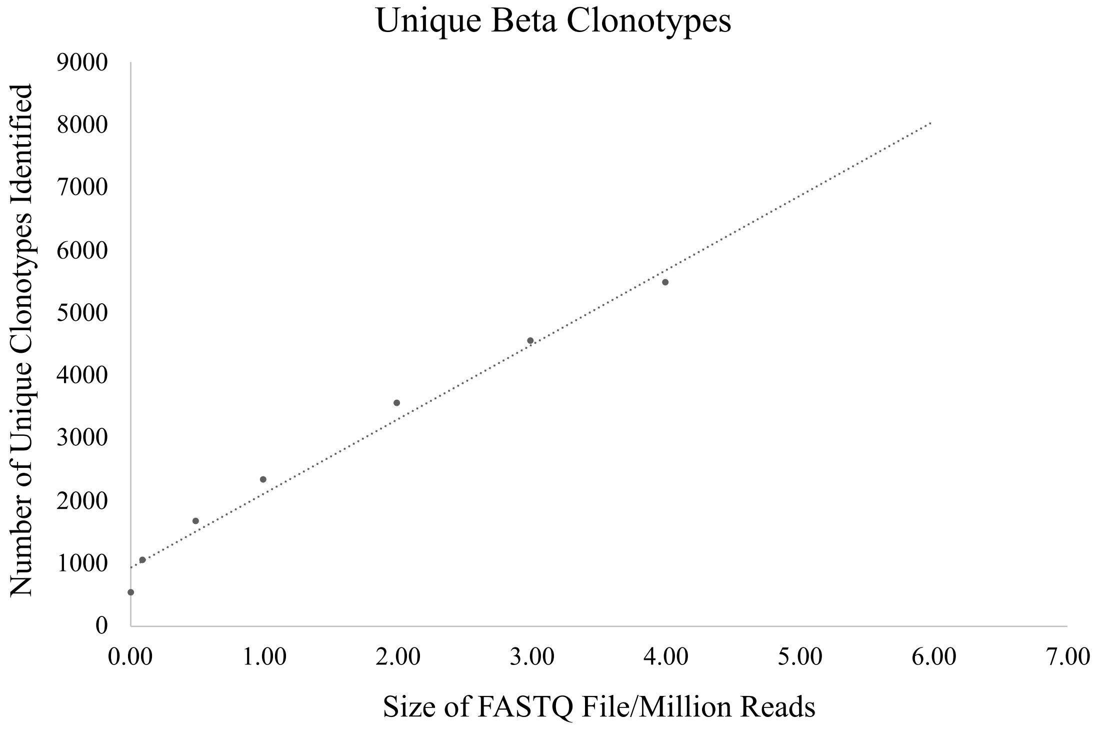


Figure 23. Number of Unique β chain sequences identified from subsampled data. The same subsampling method was used as above to identify all unique clonotypes. Statistical analysis was not performed due to sample size (n = 1).

The rate of identification of new clonotypes increases rapidly as sample size increases, Figure 23, suggesting that it would be difficult to estimate the total number of unique clonotypes within one sample without the use of exhaustive sequencing. For the purposes of this work, 1,000,000 sequences were the target for all sequencing experiments, to maximise multiplex capacity whilst retaining quality information on as many TCR species as possible. However, this may not be appropriate in all cases as some experiments may require greater sequencing depth, or exhaustive sequencing, in order to identify more unique TCR species.

Size of Sample	Highest Frequency Clone	Clone Size	Clone Size (%)	Mean CDR3 Length
10000	GTCCCGGACTAGCCA	12	0.0133	15.03
100000	AACAGGGGGCTC	94	0.0107	15.57
500000	AACAGGGGGCTC	468	0.0108	16.15
1000000	AACAGGGGGCTC	897	0.0104	16.41
2000000	AACAGGGGGCTC	1826	0.0105	16.73
3000000	AACAGGGGGCTC	2763	0.0106	16.93
4000000	AACAGGGGGCTC	3584	0.0103	16.98

Table 37. Representation of highest frequency clonotypes within subsampled data. As the size of the sample decreases, the representation of the highest frequency clonotype is consistent until a sample size of above 10,000 sequences.

Of note, it appears that the average length of the CDR3 region is steadily increasing as sample size increases, as shown in Table 37.

3.3.6 Reproducibility of Assay (Sample Splitting)

The stability and reproducibility of the high-throughput sequencing data obtained was established using the method described earlier, a series of sample splits and repeats were carried out. Blood was drawn from two healthy volunteers and cell samples were split into two aliquots. Total RNA was extracted from all samples, followed by mRNA isolation and first and second strand cDNA synthesis. Once double stranded (ds) cDNA had been synthesised, two samples were split further into duplicate samples, which would theoretically contain the same transcripts at the same frequencies (illustrated in Figure 24). Each of the final six samples, H5a, H5b, H5c, H6a, H6b and H6c, was prepared for capture and sequencing on the Illumina MiSeq, in the same manner as all other samples, as previously described in Chapter 2.

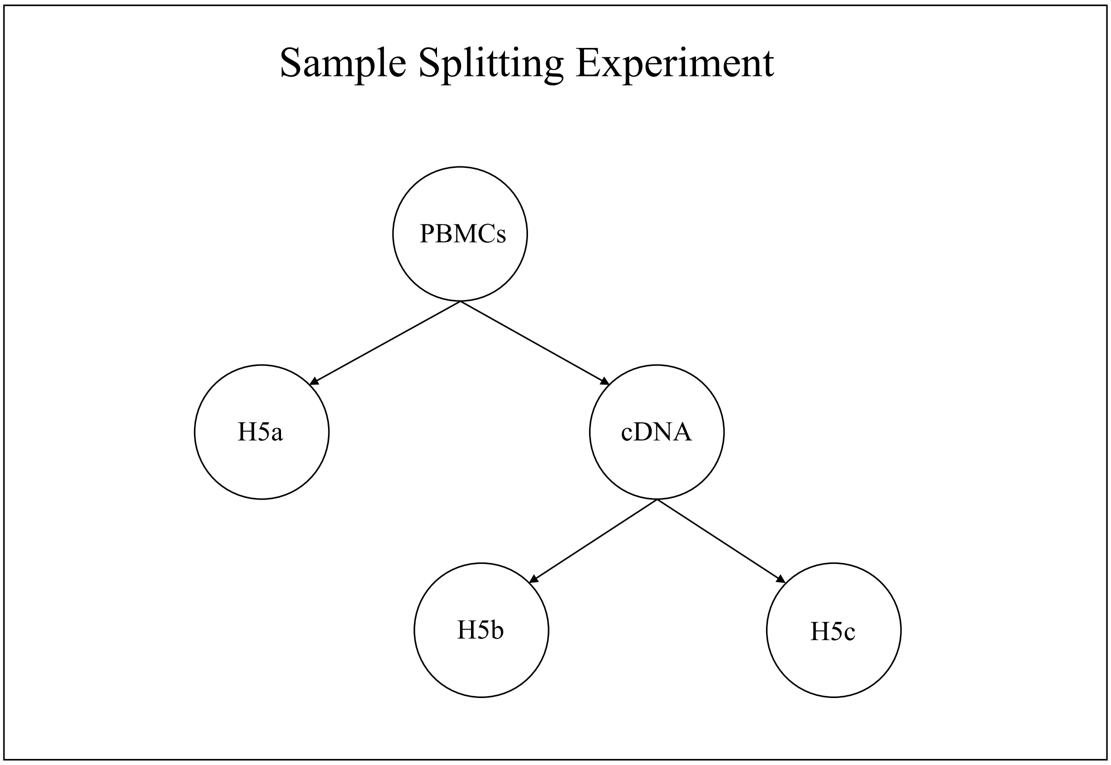


Figure 24. Visualisation of the sample splitting method. Cells were isolated, and split into two samples, “a” and “b”. Following library preparation, double stranded cDNA was synthesised, and the sample “b” was split further, into “b” and “c”. This was done for samples from two healthy individuals, H5 and H6. Therefore, for each of the two H samples, three samples were sequenced, all of which came from the same original sample, with two that came from the same transcript pool.

Variable gene segment usages of the split samples, H5a to H6c, are shown in Figure 25. Results showed that sample H6 displayed mirrored frequencies of gene usage across all samples. Furthermore, the most frequently represented CDR3 clonotypes are consistent in both sequence and proportions across split samples. However, gene usage across sample H5 does not appear uniform. In fact, sample H5 appears to be anomalous in comparison to both its split samples and to the rest of the healthy individuals, despite being treated in the same manner as the other samples.

The expression of TRAV gene segments is shown in Figure 25 and Figure 26 for H5 and H6 respectively. The expression of TRBV gene segments is shown in Figure 27 and Figure 28 for H5 and H6 respectively. There is substantial mirroring of gene expression for H6 but despite each of the three samples originating from the same blood draw, and H5b and H5c from the same transcript pool, mirroring of gene expression is not demonstrated for H5. This suggests that different TCRs have been amplified and sequenced across the samples. Each figure describes the mean of gene usage across the split samples, with the error bars representing the standard deviation.

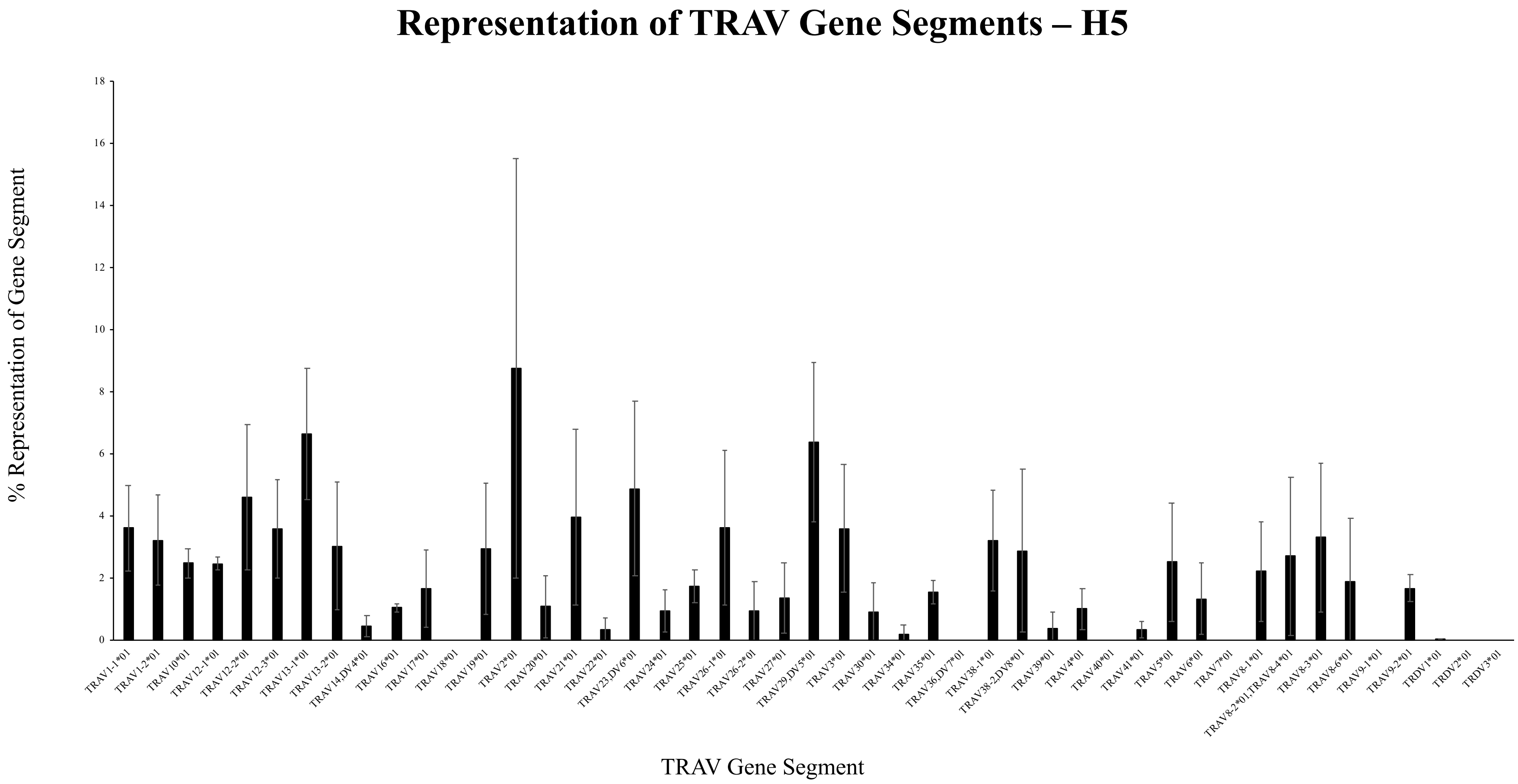


Figure 25. Mean representation of each TRAV gene segment for the split samples H5a, H5b and H5c. The mean values of individual TRAV genes and respective standard deviations from samples H5a, H5b and H5c were calculated.

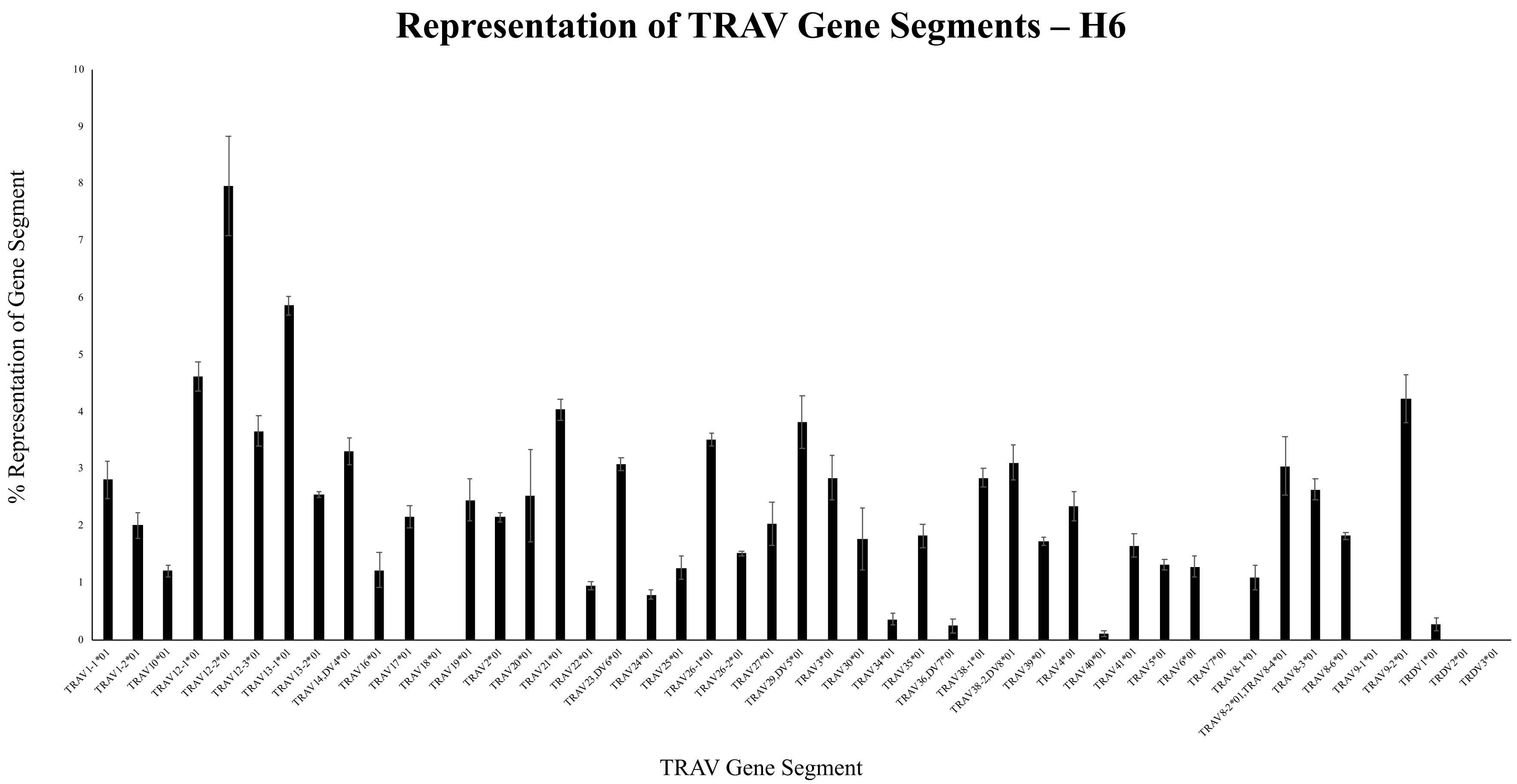


Figure 26. Mean representation of each TRAV gene segment for the split samples H6a, H6b and H6c. The mean values of individual TRAV genes and respective standard deviations from samples H6a, H6b and H6c were calculated.

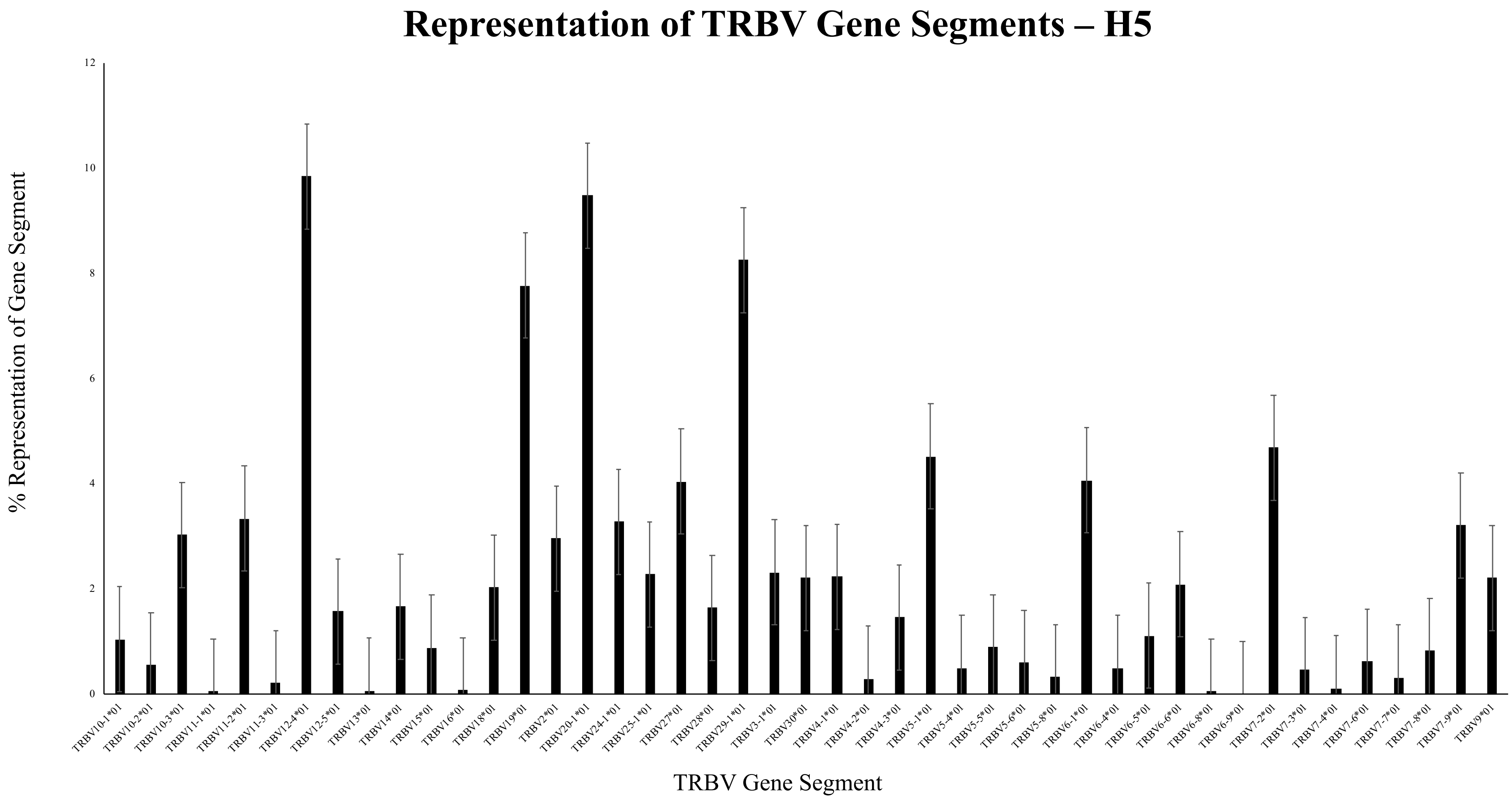


Figure 27. Mean representation of each TRBV gene segment for the split samples H5a, H5b and H5c. The mean values of individual TRBV genes and respective standard deviations from samples H5a, H5b and H5c were calculated.

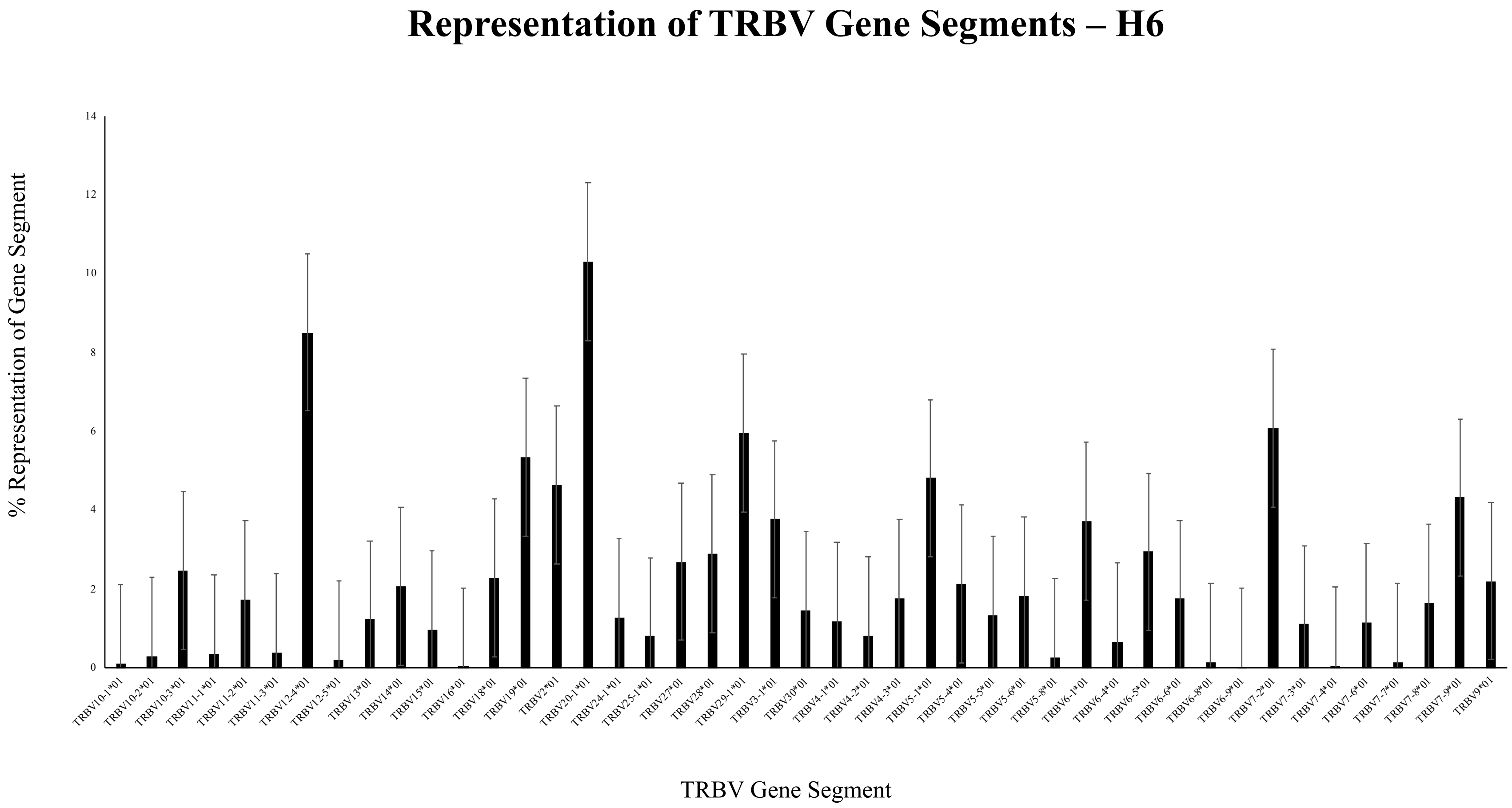


Figure 28. Mean representation of each TRBV gene segment for the split samples H6a, H6b and H6c. The mean values of individual TRBV genes and respective standard deviations from samples H6a, H6b and H6c were calculated.

Sample H6 demonstrated strong mirroring across all three split samples, which suggests that the same TCR transcripts were present, amplified and sequenced in each. H5 demonstrated mirroring of gene segment expression across the split samples to a lesser degree. To compare clonotypes identified, Table 38 and Table 39 show the top 5 most frequent clonotypes and their CDR3 sequences.

Sample	**Freq**	**Sequence**
H5a	1669	TCCCGGGGGGGCCAATAAG
	1124	TCCCCGGGACAGGGGG
	1039	ACAGCAGG
	985	AGGGCAGCGGCGGGGGGGGAA
	974	GATGACAGGGCGGGTCTTCAGGATCC
H5b	3048	AGGACTGAGGCA
	3028	CCTCTACAGACAGGGGGGACTG
	2710	TTATACAGGGAGCGG
	2455	ACCATTGGGG
	2291	CCGGGACTATGTGGG
H5c	1262	GACTAGCGGGAGCGG
	1236	GTAC
	1142	AGTGGGGGGCGCCCTGGG
	1049	TACGGG
	1000	TCCCCCGCCGGACTAGCGTCG

Table 38. The top 5 most frequent CDR3 sequences identified in H5a, H5b and H5c. Each of the sequences in the table above was different, suggesting that different TCR transcripts were amplified from the original sample.

Sample	Freq	Sequence
H6a	4627	TCGG
	3702	CCCGGGGCGGGTA
	3540	CCCGGGACAGAC
	3169	GCCTGGGGGAGA
	2485	AAGCGGACAGGT
H6b	3088	TCGG
	2922	CCCGGGGCGGGTA
	2801	CCCGGGACAGAC
	2531	GCCTGGGGGAGA
	2040	GTTCAGCGGTCAA
H6c	3151	CCCGGGACAGAC
	2049	TCGG
	1778	GCCTGGGGGAGA
	1225	GTTCAGCGGTCAA
	1000	GCCCCACAGGAACACAA

Table 39. The top 5 most frequent CDR3 sequences identified in H6a, H6b and H6c. This table contains sequences that were shared across the three split samples.

H6a, H6b and H6c demonstrated mirroring of both gene and clonotype expression but this was less visible across H5a, H5b and H5c.

3.4 Discussion

The landscape of available sequencing technology has been highly changeable in recent years since the advent of high throughput platforms. "Off-the-shelf" kits are being constantly refined as improvements and new developments are made. In addition, these experiments constituted the first attempts at using targeted capture to sequence immune repertoires. There was therefore no pre-determined workflow. This meant that sequencing the target region was an experiment in itself. Consequently, it was necessary to optimise many steps of the library preparation process prior to sequencing, a process which carried on throughout the entirety of this project. Data was acquired through use of this technique; therefore, this has been successful. However, there were issues with library preparation and standardisation of the assay.

Two different methods of mRNA isolation were used. The yield was 0.89% of total RNA on average for both methods, compared to between 2 to 5% of total RNA that is made up of mRNA, depending on cell type (3, 110). The remainder of the total RNA extracted comprises approximately 80% of ribosomal RNA (rRNA) and at least 15% of transfer RNA (tRNA). There may be issues with RNA degradation, as RNA is fragile and prone to degradation by nucleases (111). Quantification may have been inaccurate, as it more difficult to accurately quantify small amounts and low concentrations of nucleic acids (112). The low yield of starting material could distort the interpretation of the repertoire, although it is likely that this is also an issue for other methods of repertoire sequencing as the extraction techniques and quantification steps are not unique to this capture method (104, 113).

The failure of the enzymatic fragmentation could have been due to the quality of the starting material; if RNA is not in the best condition prior to fragmentation there may be more chance it would be degraded through application of the fragmentation enzyme. Loss of sample was also encountered during sonication and further material was lost following magnetic bead clean ups and quantification, both of which are necessary to produce a high-quality library, as this reduces contamination by protein or salt and allow for quality assurance (114). Each step involves substantial manipulation of the sample, which can lead to the loss of material as it binds to the magnetic beads or is washed away by ethanol. It is unclear whether there is a bias in the material that is lost, for example, the length of the fragment may affect the likelihood of loss, both during the mRNA isolation and subsequent library preparation. It is challenging to determine whether this may skew the repertoire by giving an advantage to the amplification and capture of some TCR sequences. This limitation should be taken into account when interpreting data, and the number of steps involved in library preparation should be minimised whenever possible.

Low yield during library preparation meant some samples underwent high numbers of PCR cycles, anywhere between 8 and 20. The greater the number of PCR cycles used during library preparation, the greater the number of fidelity errors that may be introduced (115). This is a major concern with this project as the target region to be sequenced is hypervariable and further introduction of variability through PCR error may create a false impression of repertoire diversity. Over-amplification may also intensify the effect of PCR bias, where some TCR sequences might be preferentially amplified. Both of these issues may skew the repertoire and affect sensitivity and specificity of the assay.

Across all samples, no reads were flagged as being low quality. Read quality decreases towards the end of the run at the higher cycle numbers, which was commonly seen across all runs. However, the quality scores for reads generated by the Ion Torrent PGM were lower, indicating that there may have been an issue with quality, although this program was designed for use with Illumina MiSeq and it may be that the analysis pipeline was incompatible. The Ion Torrent PGM was used to generate data in the early stages of this project, and this data was analysed further, as data generated from different sequencing methods has previously found to be correlated (108).

The success demonstrated by pre-capture pooling, drives down the cost of the method, making it potentially feasible for large scale applications, for example in a clinical setting. The sequencing depth must always be considered when pooling samples. Different depths may be appropriate for different questions. It should be noted that exhaustive sequencing of the TCR repertoire is highly challenging, as there are many rare clonotypes that may be missed unless the organism in its entirety is sequenced. In this work, 1,000,000 reads was used as a bench mark for sequencing depth (88). Although the number of sequences did not appear to affect downstream analysis, it could have an impact on the representation of T-cell receptor sequences. However, even within one run, the proportion of reads per sample was not equal, suggesting that some samples were favorably amplified, or bound preferentially to the flow cell, since the same amount of starting material was sequenced from each.

Much of the data obtained does not align to the target regions of the probes, which cover both the TCR and Ig transcripts. Therefore, this assay is currently not as efficient at sequencing TCR repertoires as other PCR based methods (65, 88, 104). Also, many of the sequences that align to TCR transcripts using BWA are lost when using specific

TCR repertoire analysis software. This is because the landscape of currently available TCR analysis software requires identifier sequences to be present on both sides of the CDR3. The method may be improved on in future by shearing the material to 500bp to improve the proportion of reads that span the CDR3 region, or the bait library could be altered to only include baits that capture directly adjacent to the CDR3.

The difference in alignments to TCR or Ig sequences may suggest a bias within the capture library or that there was a greater number of Ig transcripts present in these samples than TCR, due to a greater number of Ig sequences being transcribed compared to TCR sequences, independent of the cell number. This explanation could correlate with a greater number of B-cells than T-cells in peripheral blood samples or suggest that the abundance of Ig transcripts was greater than that of TCR transcripts. It may be that the increase in Ig transcripts, rather than TCR, was required to create secreted antibodies, rather than the T-cell receptor, which remains on the cell surface. The variation of proportions across individuals may reflect different proportions of T-cells and B-cells in the initial PBMC sample, and potentially different states of health of the individuals. However, this remains speculative at this time.

The alignment to T-cell receptor sequences did not change significantly when T-cells were isolated from PBMCs. It was considered that since there was no Ig material to amplify and capture that the concentration of TCR rearrangements would increase. However, it did not which suggests the TCR specific baits may have become saturated during hybridisation and unable to capture more sequences. This highlights a potential issue with the bait design.

In both TCR and Ig, there was an overexpression of some chains relative to others. For the TCR, the β chain was aligned at a higher frequency that the α chain. This may reflect a relative abundance of β-chain transcripts relative to the α-chain. It may be possible, as above, that there is an unevenness in transcription; there are many variable factors in transcription dynamics. Alternatively, it could be a result of bias within either the experimental procedure or the analysis pipeline preferentially amplifying or aligning β-chain transcripts. The alignments to γ and δ are lower than to α and β, but this is expected as there are far fewer circulating γδ+ T-cells than αβ+ T-cells.

Note that the differences between the alignments between the chain pairs may be due to differences in gene expression or bias in either the experimental assay or the analysis pipeline. However, the method does reflect similar biases in V and J usage that has been noted in other publications (116). Similarly, the Ig heavy chain was aligned to more frequently than either the κ or λ chains (Figure 9 and Figure 10, and **Error! Reference source not found.** and **Error! Reference source not found.**).

Further inconsistencies appeared in this data, particularly when the reproducibility of the assay was investigated. The mirroring of frequencies demonstrated in Figure 25 and Figure 26 suggested this method can produce reliable results that are representative of the T-cell receptor of an individual as it is at the point of blood draw. One of the split libraries, H6, demonstrated strong reproducibility and samples mirrored one another regarding gene expression. The other library, H5, did not appear consistent, but none of the results were statistically significantly different from one another. This may demonstrate reproducibility of the method, and that the sample of an individual's repertoire obtained during sequencing is a representative snapshot of this dynamic repertoire at that moment in time. Successful sample splitting, as shown by H6, has been demonstrated by other groups (116). However, sample variation highlights the requirement for suitable quality controls at set points across this assay and that refinement of the technique may be required in future experiments.

There is a compounding effect of sample loss which is not usually documented in publications but is not unique to this work. It does not nullify the results but should be considered in their interpretation as it is unclear whether material is lost at random or introduces bias. Optimisation of this assay requires further work but data generated using this technique has been investigated in the following chapters to demonstrate a pilot project for the use of targeted capture.

Chapter 4

The Healthy T-Cell Receptor Repertoire

4 The T-Cell Receptor Repertoire in Health

4.1 Introduction to the T-Cell Receptor Repertoire

Having established the high-throughput sequencing and targeted capture method, it was applied to investigate the TCR repertoire in greater detail. This included comparison of different sequencing technologies and analysis pipelines. It has been reported in some instances that the choice of sequencing has minimal impact on the repertoire analysis downstream so in this case the Ion Torrent PGM and Illumina MiSeq were compared (108).

There are of a number of computational methods that can be used to identify TCR reads from sequencing data. It was therefore of interest to compare them. Two open source programs were used throughout this work for this purpose, the Decombinator (94) and MiTCR (95). These programs were chosen at the time as they had the most validation data in support of their use. Each program works by assigning TCR V and J gene segments of α, β, γ and δ chains to sequencing reads using an Aho-Corasick algorithm (117). Here, the differences in identification of TCR clonotypes and the frequencies at which they appear were considered.

Understanding the diversity of the TCR repertoire is crucial to understand the ability of the immune system to respond to pathogens. A diverse repertoire is needed to respond to a wide range of pathogens. However, diversity can be difficult to quantify. Therefore, methods of statistical estimation, including the Simpson index, Shannon entropy and Gini index methodologies that are routinely utilised in other scientific fields are being increasingly employed in the TCR research. The term richness is often used to describe this diversity, referring to how many different entities a dataset contains. In the context of TCR repertoires the individuals refer to the individual TCR clonotypes.

4.1.1 Simpson Index

The Simpson index was introduced in 1949 and has since been used to measure the diversity of species in an ecological system (118). As the species within the TCR repertoire can be likened to the species within an ecological system, the index has been borrowed to describe TCR repertoires. It measures the species richness through the calculation of the probability that two random samples of individuals within a dataset are the same species, while assuming that the first individual drawn is not replaced to the dataset before the second draw. It is calculated according to the following equation;

$$D(X) = \frac{N(N-1)}{\sum n\,(n-1)}$$

Where "D(X)" is the Simpson index, "n" is the number of individuals in one species, "N" is the total number of individuals in all species. For T-cells, the clonotypes are the species. The minimum value occurs when there is only one species and the maximum value occurs when there is evenness of representation across all species. The inverse Simpson index (1/D(X)), is often used instead to demonstrate divergence in diversity. Therefore, the higher the value of the inverse Simpson index, the lower the diversity of the population (119).

4.1.2 Shannon Entropy

Shannon entropy is also used to measure species diversity, even though it was originally introduced to measure entropy, or uncertainty, in strings of text in the field of information theory (120). Shannon entropy calculates the uncertainty associated in predicting the identity of an individual when drawn at random from a dataset, based on the weighted mean of the proportional abundances of each individual within that dataset (121). For T-cells, this relates to the uncertainty in predicting which clonotype is selected, when drawn at random from the repertoire. Therefore, when this function is applied to a sequencing dataset, the diversity can be estimated. The equation used to calculate the Shannon entropy is as follows;

$$H(X) = -\sum_{i=1}^{s} p_i \, ln \, p_i$$

Where "H(X)" is the Shannon entropy, "p" is the proportion of individuals from one species, "s" is the number of species and "ln" is the natural log (121). It may also be referred to as the natural log of the true diversity (104). As the dataset becomes less diverse, the value of the Shannon entropy approaches zero. When there is no diversity in the repertoire, the Shannon entropy measures exactly zero, as there is no uncertainty in the random selection of the next individual, or clonotype, from the dataset.

4.1.3 Gini Index

The Gini index, or Gini coefficient, has been more commonly used in economics than ecology. It is calculated by plotting the Lorenz curve, a graphical representation of the inequality of the distribution of wealth and calculating the area between the line and the curve or calculating the integral. The greater the area and therefore the greater the inequality, which is equivalent to diversity for T-cell populations (122). A schematic representation of the Gini index can be seen in Figure 29.

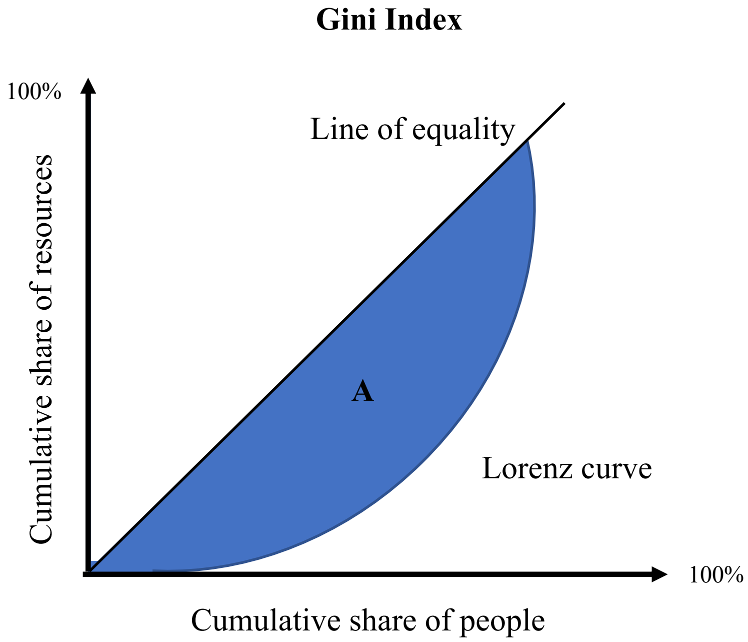


Figure 29 Schematic representation of how to calculate the Gini index. The area under the curve, A, is the Gini index. The closer this is to 0, the more diverse the population.

For TCRs, when the Gini index is calculated, the lower the value, the more diverse the population is. As this value represents a proportion, the maximum value, the value where diversity represents zero, or where there only one clonotype is identified, is 1.

4.2 Methods

Having established the diversity indices as above, the values for the inverse Simpson index, Shannon entropy and Gini index were calculated from the number of total and unique clonotypes identified through the Decombinator and MiTCR. This was done by inputting clonotype information into tcR, a package in R used for advanced TCR repertoire analysis (123).

4.3 Results

The samples discussed in Chapter 3 are the same that have undergone analysis in the current chapter. However, eight representative samples H8 to H15 will be focused on for simplicity. Each of these samples was derived from healthy controls, aged between 23 and 43.

4.3.1 Clonotype Identification by Different Pipelines

Both MiTCR and the Decombinator identified a different number of clonotypes for each healthy control, from the same datasets, which can be seen in Figure 30 and Figure 31 below. Both programs are able to analyse data for all four TCR chains α, β, γ and δ.

Statistical analysis on the results in Figure 30 using an unpaired t-test showed that differences between the numbers of total clonotypes identified by MiTCR and Decombinator were statistically significant for α, β, and δ chains ($p < 0.05$), but not for the γ chain ($p = 0.0547$). In contrast, the numbers of unique clonotypes identified by MiTCR and Decombinator were statistically significant for all TCR chains (Figure 2; $p <0.05$).

This was also true for the numbers of unique clonotypes identified. Statistical analysis on the results in Figure 31 using an unpaired t-test showed that differences between the numbers of total clonotypes identified by MiTCR and Decombinator were statistically significant for α, β, and δ chains ($p < 0.05$), but not for the γ chain ($p = 0.0530$). In contrast, the numbers of unique clonotypes identified by MiTCR and Decombinator were statistically significant for all TCR chains (Figure 2; $p <0.05$).

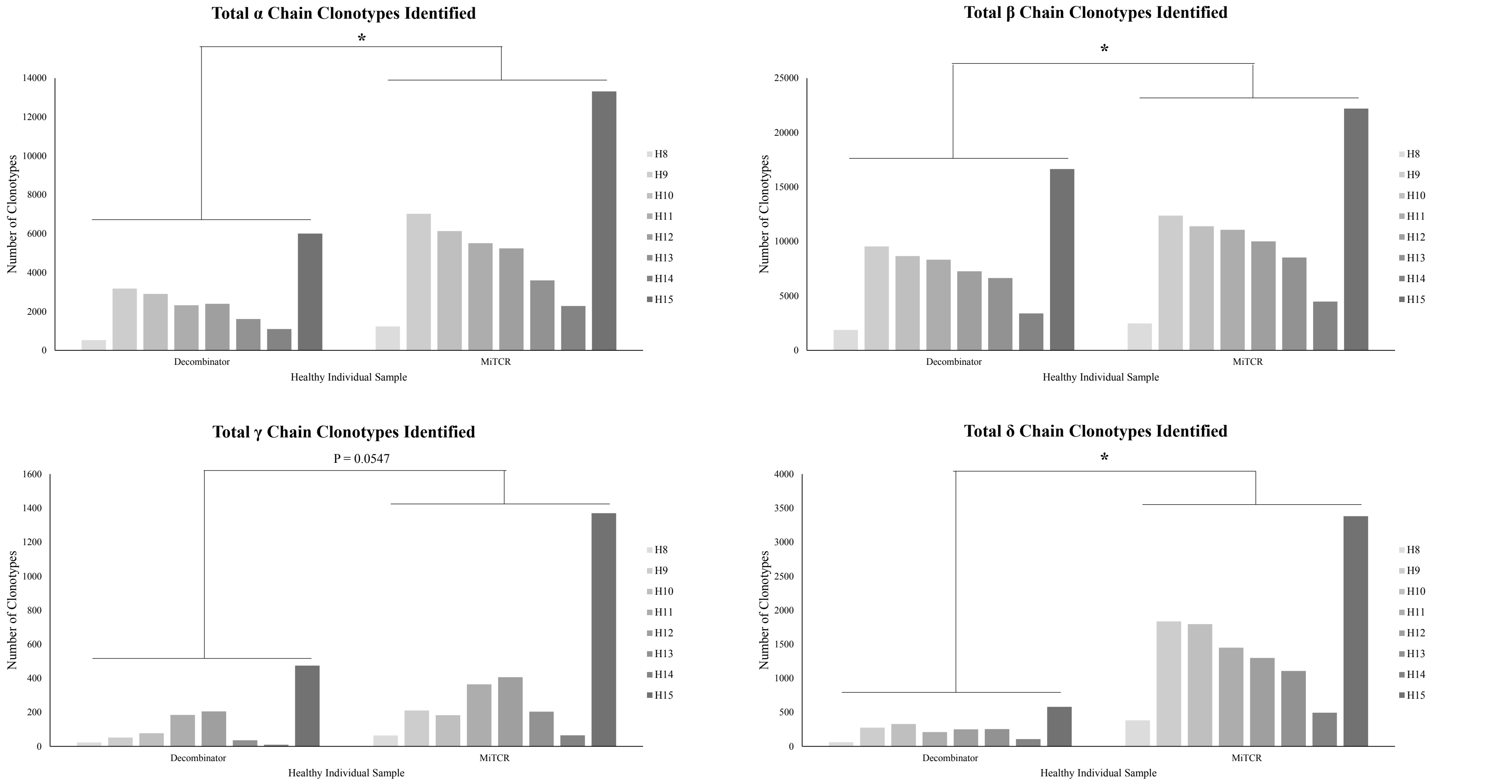


Figure 30. Number of total clonotypes identified by repertoire analysis pipeline. Decombinator and MiTCR were both used to analyse matched datasets for eight samples, H8 to H15. Both analysis pipelines consistently identified significant different numbers of clonotypes for α, β, and δ chains. *, $p < 0.05$.

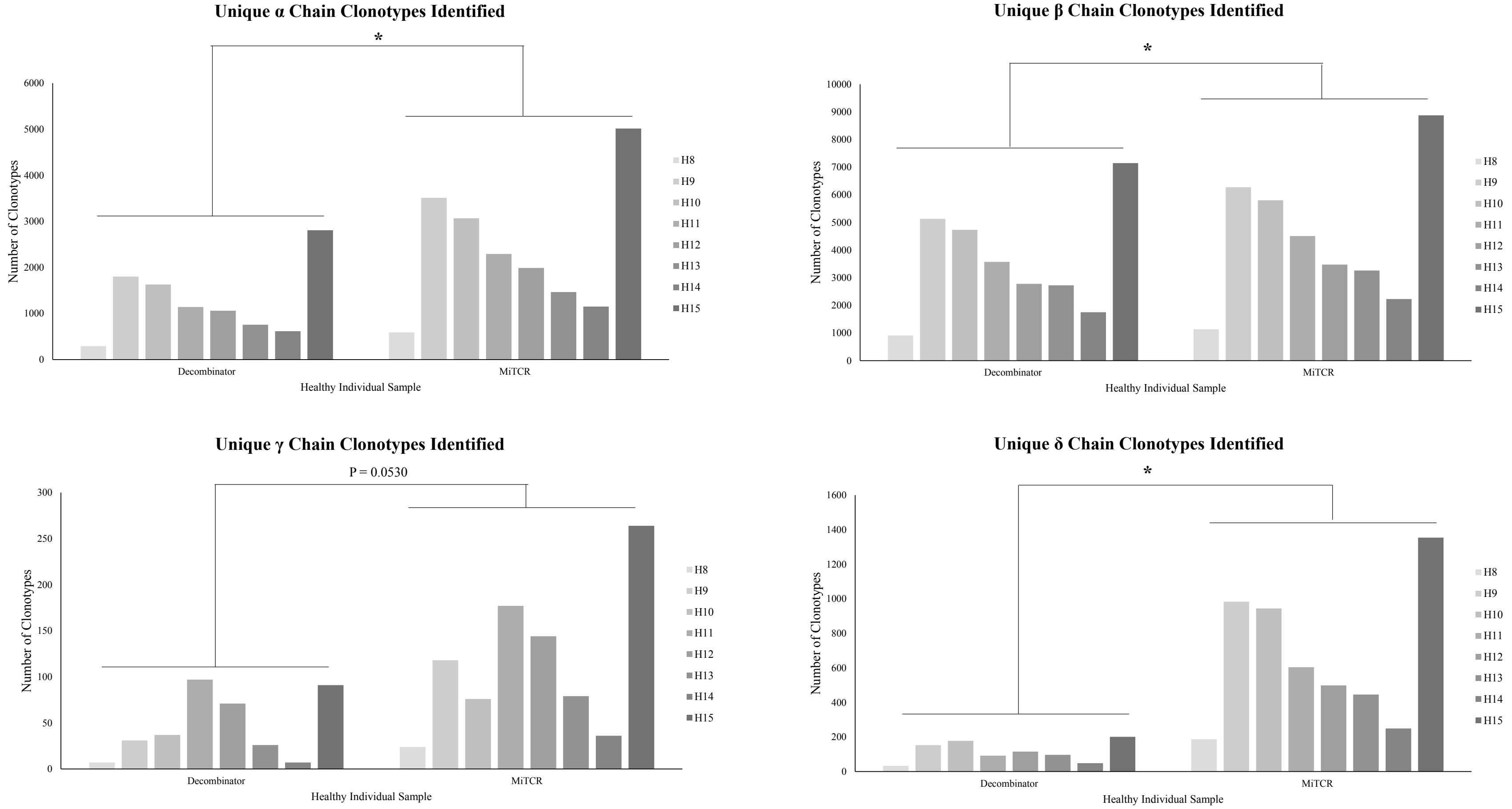


Figure 31. Number of unique clonotypes identified by repertoire analysis pipeline. Decombinator and MiTCR were both used to analyse the same datasets for eight samples, H8 to H15. Both analysis pipelines consistently identified significant different numbers of unique clonotypes for α, β, and δ chains. *, $p < 0.05$.

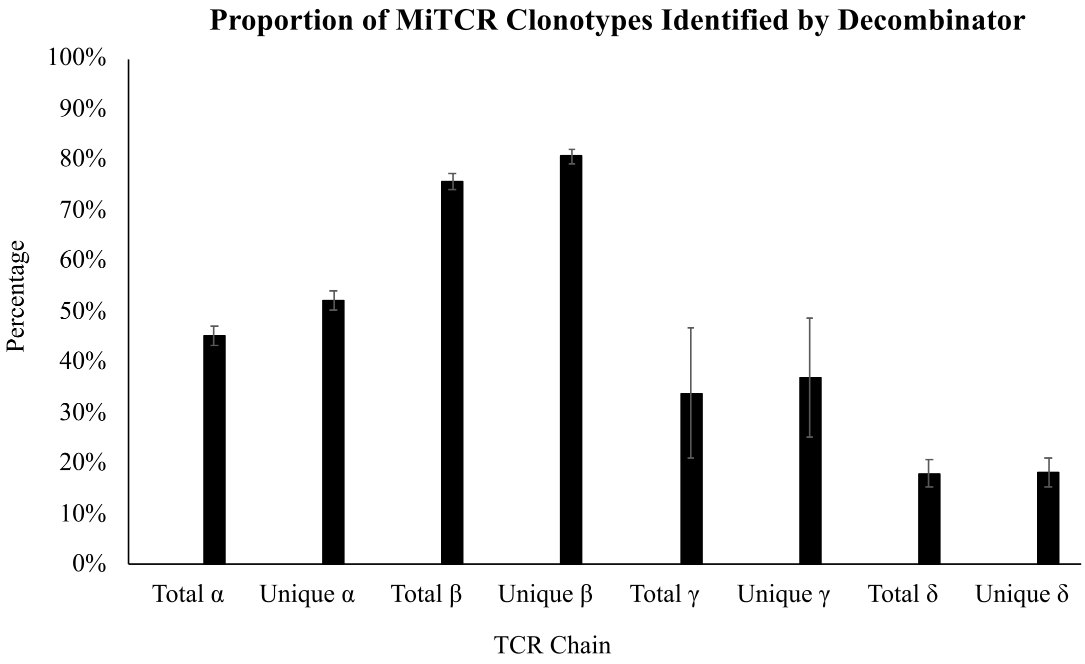


Figure 32 Percentage of clonotypes identified by both MiTCR and Decombinator. The mean values and respective standard deviations from samples H5a, H5b and H5c were calculated from the figures in Table 40.

Decombinator consistently identified fewer clonotypes than MiTCR, but this was not even across the chains, as shown in Figure 32. The Decombinator identified up to 75% of β chain, 45% of α chain, 35% of γ chain and 18% of δ chain total clonotypes that were identified by MiTCR.

No.	**Total α**	**Unique α**	**Total β**	**Unique β**	**Total γ**	**Unique γ**	**Total δ**	**Unique δ**
H8	43.1%	49.7%	75.8%	80.5%	35.9%	29.2%	16.0%	17.6%
H9	45.2%	51.3%	77.2%	81.8%	24.6%	26.3%	14.9%	15.6%
H10	47.3%	53.1%	76.0%	81.5%	42.1%	48.7%	18.2%	18.9%
H11	42.2%	49.8%	75.2%	79.3%	50.8%	54.8%	14.5%	15.2%
H12	45.7%	53.3%	72.6%	79.9%	50.5%	49.3%	19.2%	23.0%
H13	44.8%	51.6%	78.0%	83.5%	17.6%	32.9%	22.9%	21.7%
H14	48.3%	53.5%	75.8%	78.6%	15.4%	19.4%	21.4%	19.3%
H15	45.1%	55.9%	75.0%	80.5%	34.6%	34.5%	17.2%	14.8%
Mean	45.2%	52.3%	75.7%	80.7%	34.0%	36.9%	18.0%	18.3%

Table 40. Percentage of clonotypes that were identified by MiTCR that were also identified by the Decombinator for each chain.

Table 40 demonstrates that there was a consistency in the percentage of MiTCR clonotypes identified by the Decombinator for each chain. The range of percentages across the samples is narrow.

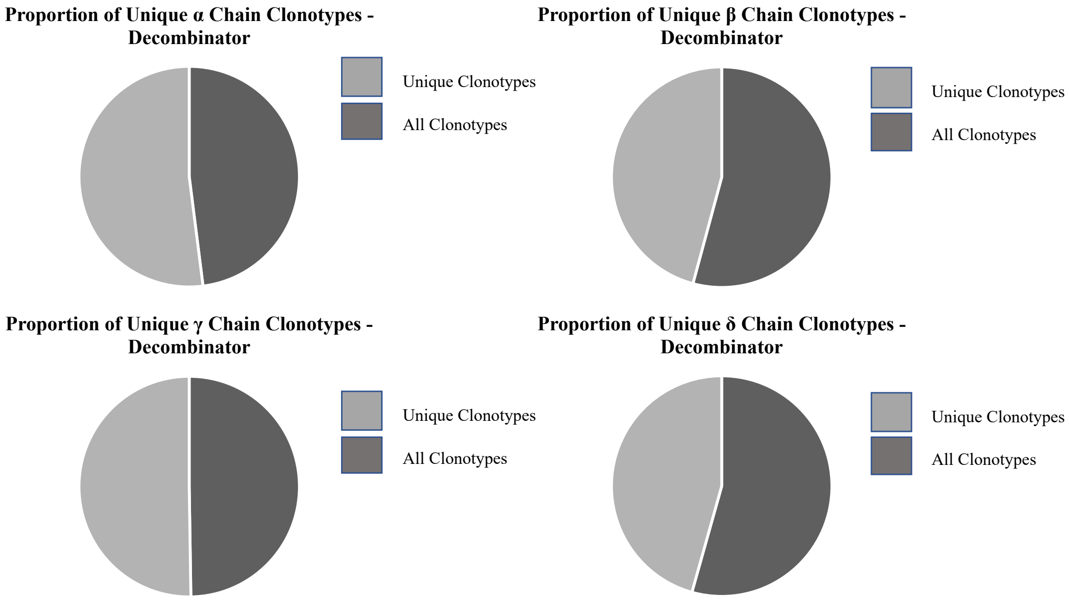


Figure 33. Unique Decombinator clonotypes identified as a proportion of the total for the α, β, γ and δ chains. Figures are calculated as an average of the proportions for each samples H8 to H15.

Figure 30, Figure 31 and Figure 32 show that MiTCR consistently identifies more clonotypes overall than the Decombinator. This effect is stronger for the α, γ, and δ chains than for the β. The average ratio of total clonotypes to unique clonotypes also varied depending on the analysis pipeline used, with the proportion of unique clonotypes identified by Decombinator higher than by MiTCR as shown in Figure 33 for Decombinator and Figure 34 for MiTCR. This suggests that MiTCR is able to extract more information from the data generated using the capture method than Decombinator with respect to clonotypes. Figure 33 shows slightly more variation in the Decombinator proportions, and the proportions higher than the MiTCR. This is consistent with the results obtained, as the Decombinator identifies fewer repeated clonotypes. To ensure that all potential clonotypes were identified, MiTCR was used for clonotype analysis for most of this study, but both analysis platforms are likely to be valid, and either could be applied depending on the requirements.

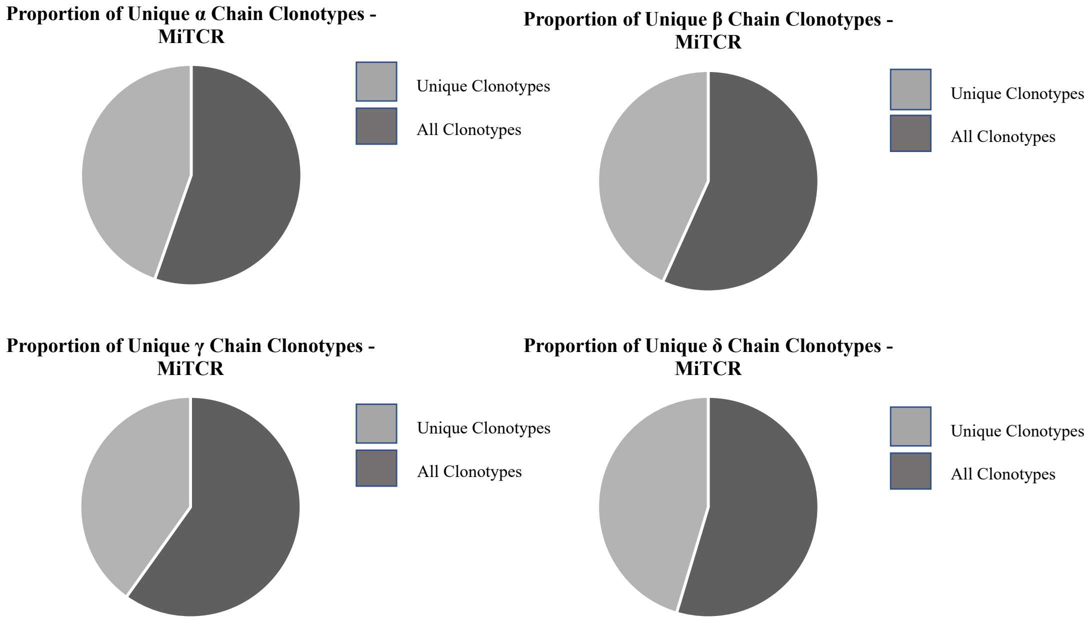


Figure 34. Unique MiTCR clonotypes identified as a proportion of the total for the α, β, γ and δ chains. Figures are calculated as an average of the proportions for each samples H8 to H15.

4.3.2 Functional and Non-Functional Clonotypes

Not all clonotypes that are identified are functional. Functional clonotypes are the RNA sequence from which a protein is translated. Some clonotypes in the data were found to be out of frame, contain stop codons, or both. The figures below (Figure 35 and Figure 36) show the breakdown of clonotypes discarded from further analysis due to non-functionality.

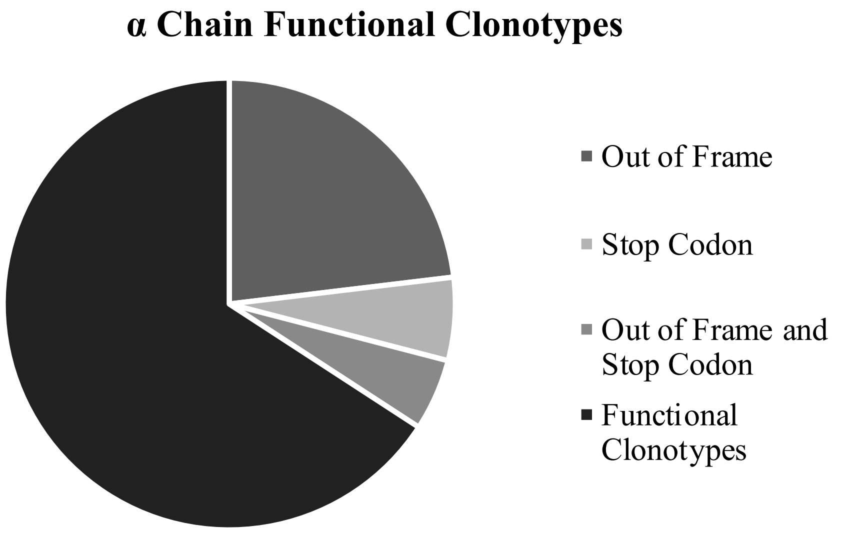


Figure 35. α chain functional clonotypes. Breakdown of α chain clonotypes discarded from the datasets due to non-functional sequences, based on the averages calculated from the values of samples H8 to H15.

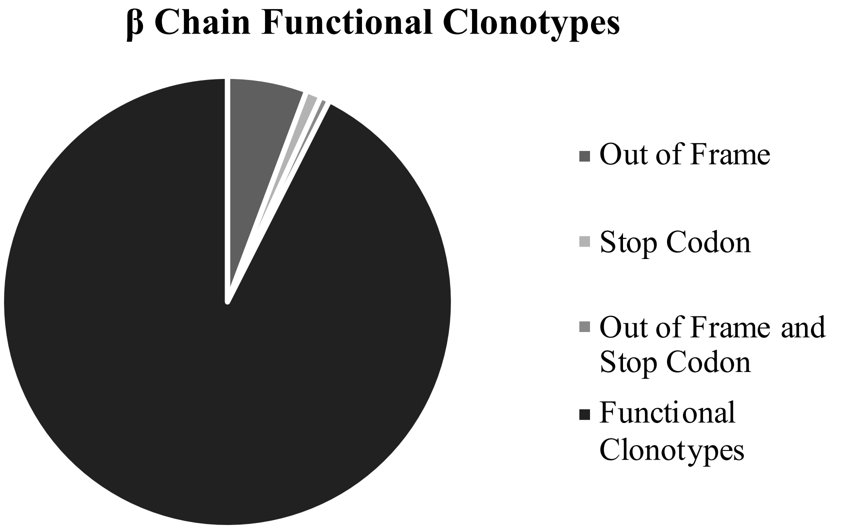


Figure 36. β chain functional clonotypes. Breakdown of β chain clonotypes discarded from the datasets due to non-functional sequences, based on the averages calculated from the values of samples H8 to H15.

Figure 35 and Figure 36 identified a large difference between the number of clonotypes discarded from the α and β chains. Figure 35 shows 23.10% of α sequences were discarded as they were out of frame, 5.95% because they contained stop codons, and 5.14% due to both. Figure 36 shows 5.71% of β sequences were discarded as they were out of frame, 1.09% because they con contained stop codons, and 0.66% due to both. This leaves 92.54% of functional clonotypes remaining.

4.3.3 Diversity Estimates of the TCR Repertoire

Three different measures of species richness were applied to the data to approximate diversity of the TCR repertoire; the inverse Simpson index, Shannon entropy and the Gini index. MiTCR was used to identify clonotypes for comparison using different diversity measures. All diversity indices described here were calculated in R.

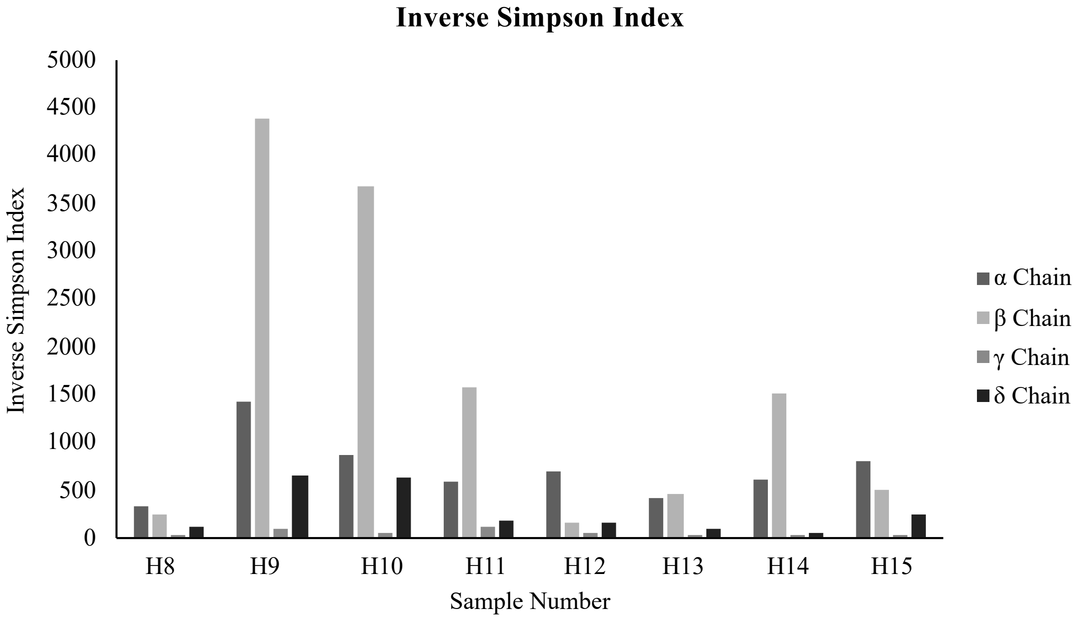


Figure 37. Inverse Simpson index for the α, β, γ and δ chains of eight healthy individuals.

The inverse Simpson index is shown for eight healthy individuals in Figure 37. The lower the value of the inverse Simpson index the greater the diversity. In two of the samples above, H8, H12 and H15, the β chain was more diverse than the α chain. The γ chain was consistently more diverse than the δ chain. The γ and δ chains were more diverse than the α and β chains. The mean inverse Simpson index values and their standard deviation for these 8 individuals is shown in Figure 38.

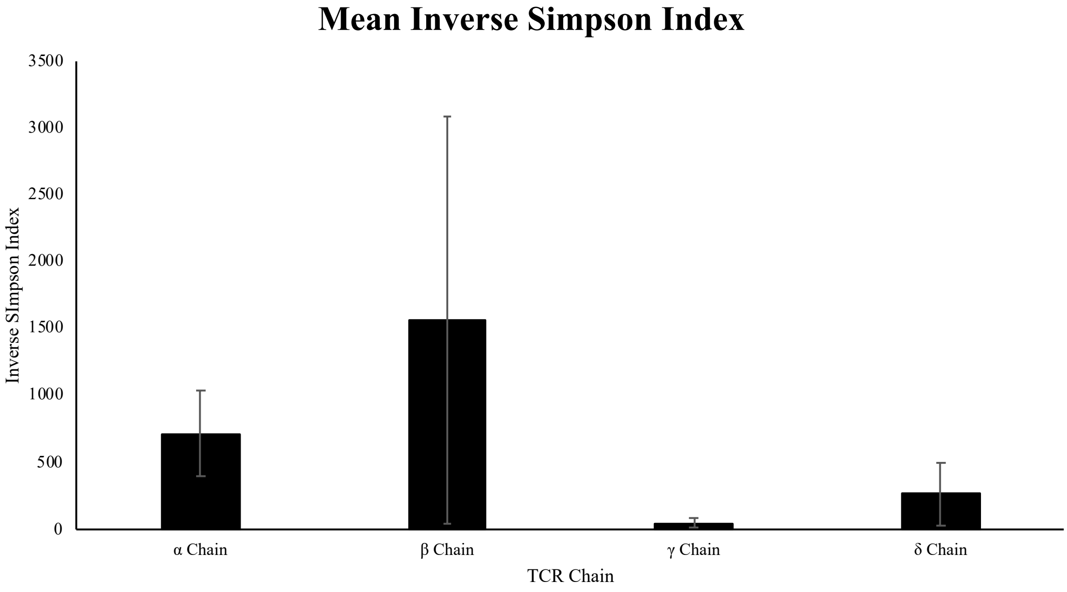


Figure 38 Mean inverse Simpson index of 8 healthy individuals. The mean values of diversity values for individual TCR chains and respective standard deviations were calculated.

The Shannon entropy of the eight individuals is shown in Figure 39. The higher the value of Shannon entropy the greater the diversity. For each individual, the β chain was the most diverse, followed by the α chain, then the δ chain and the γ chain. This pattern was consistent for all individuals in this experiment. The mean Shannon Entropy values and their standard deviation for these 8 individuals is shown in Figure 40.

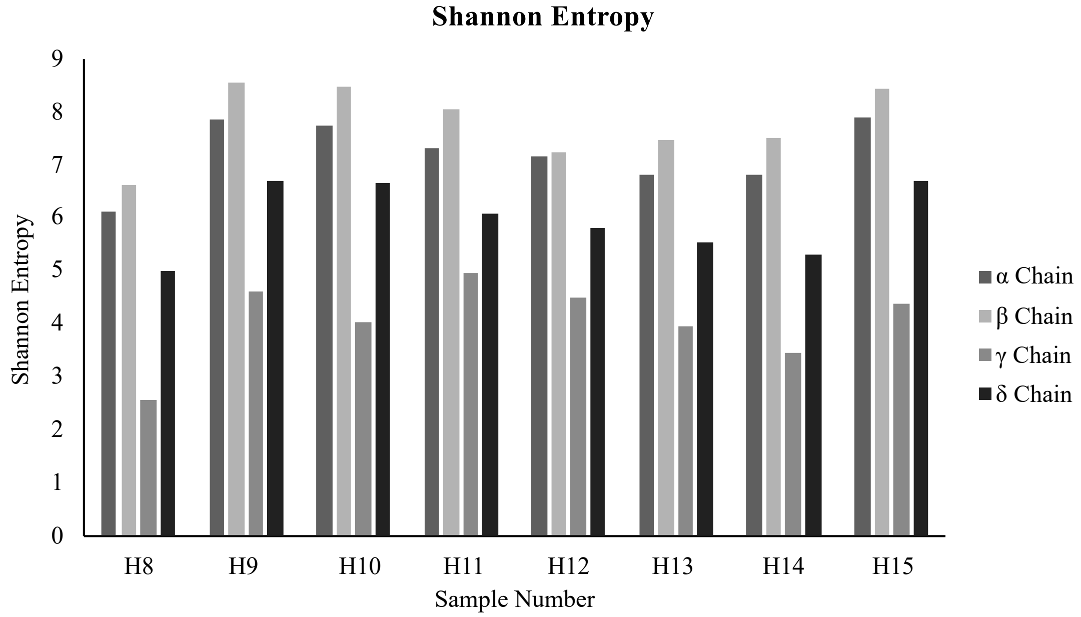


Figure 39. Shannon entropy of the α, β, γ and δ chains of eight healthy individuals. The α and β chains demonstrated greater diversity than the γ and δ chains. In all cases, the β chain was more diverse than the α chain.

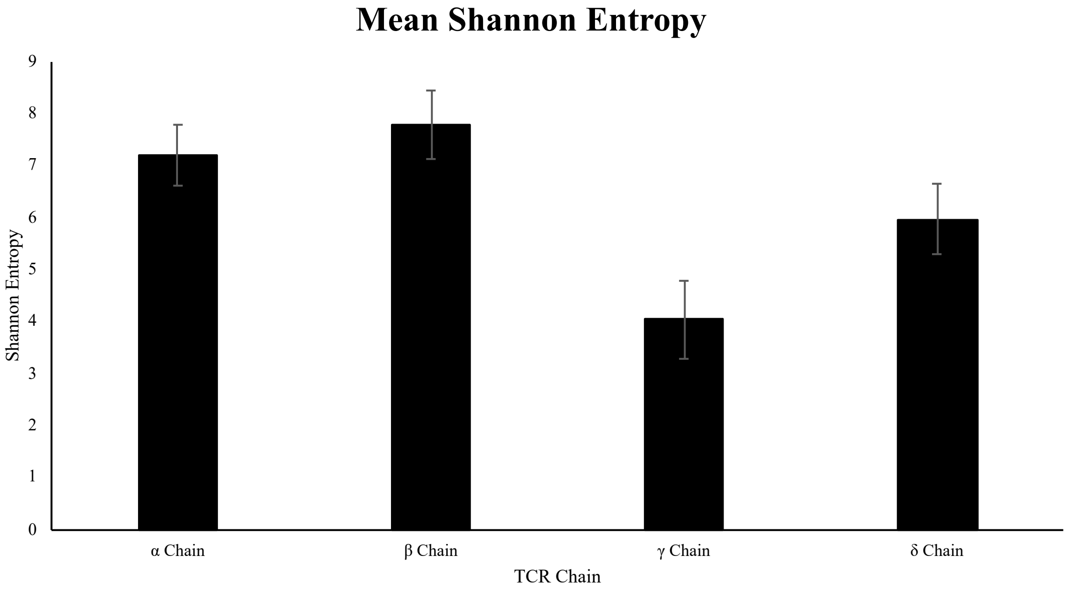


Figure 40 Mean Shannon entropy of 8 healthy individuals with standard deviation. The mean values of diversity values for individual TCR chains and respective standard deviations were calculated.

The Gini index of the eight individuals is shown in Figure 41. The lower the value of the Gini index the greater the diversity. In five of the samples above, H9, H10, H11, H14 and H15, the β chain was more diverse than the α chain. In half of the samples above, H8, H10, H12 and H15 the δ chain was more diverse than the γ chain. The mean Gini values and their standard deviation for these 8 individuals is shown in Figure 42.

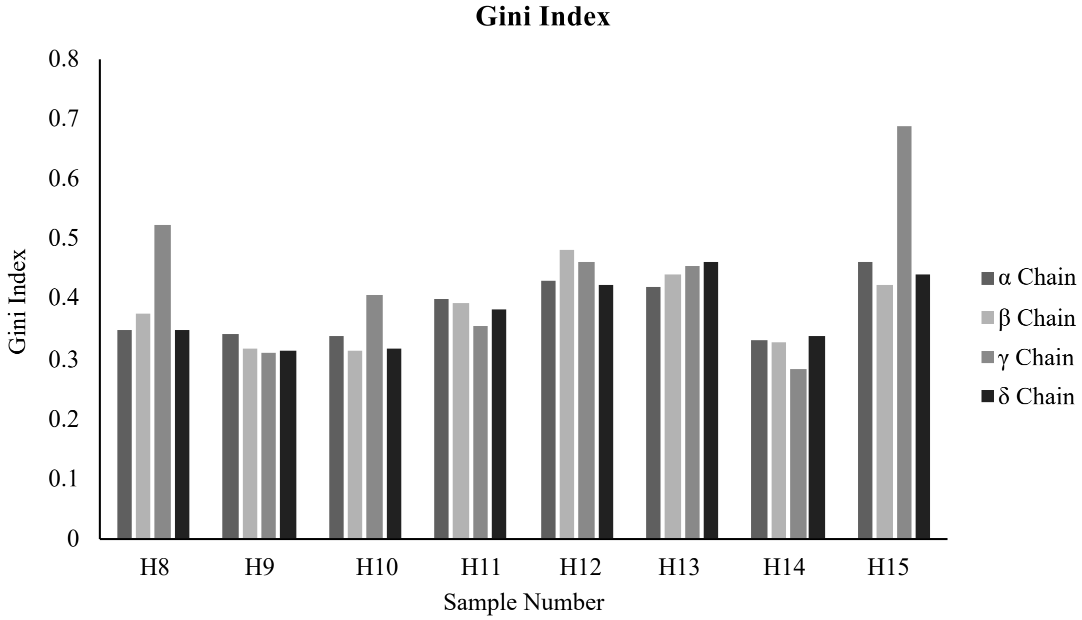


Figure 41. Gini Index of the α, β, γ and δ chains of eight healthy individuals. The α and β chains did not demonstrate greater diversity than the γ and δ chains with the Gini index.

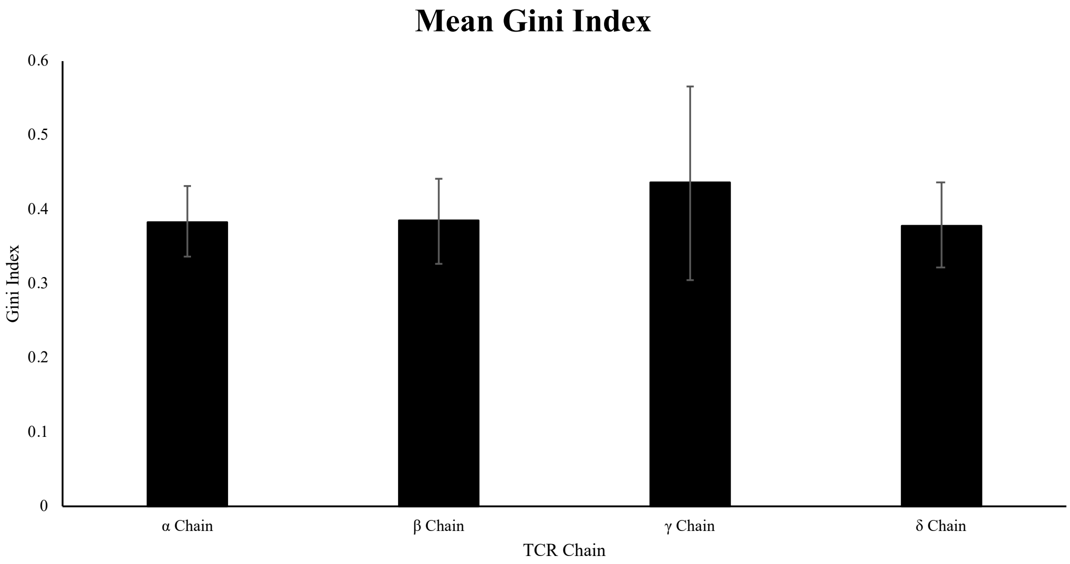


Figure 42 Mean Gini Index of 8 healthy individuals with standard deviation. The mean values of diversity values for individual TCR chains and respective standard deviations were calculated.

4.3.4 CDR3 Length Distribution

The sequencing data obtained was used to investigate diversity through comparison of the length of the CDR3 region, which has previously been done using spectratyping and mostly for the β chain. This was achieved through the use of software that analyses the T-cell sequence data and identifies the start and at the termination of the CDR3 region.

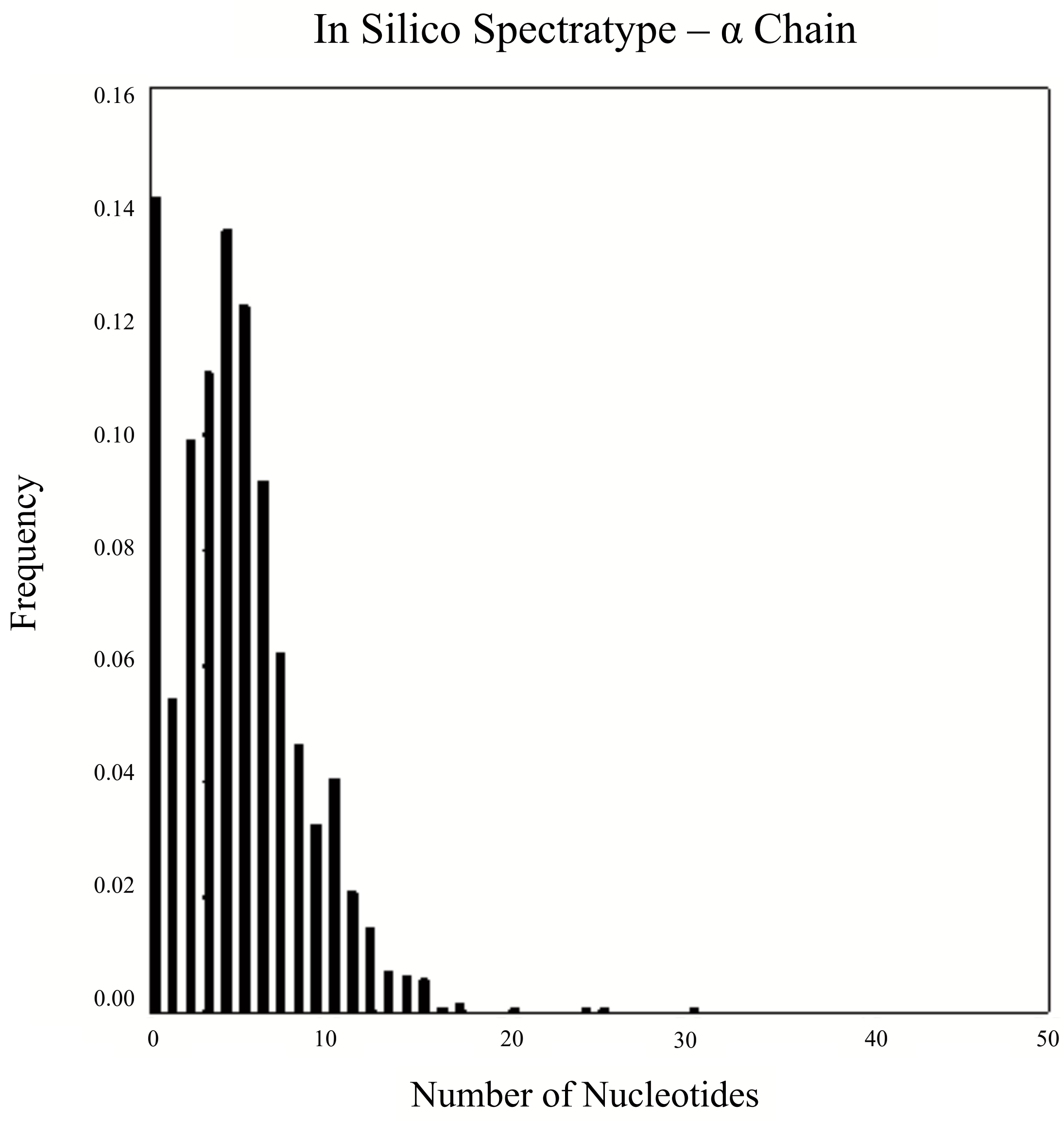


Figure 43. In silico generated spectratype data for the α, β, γ and δ chains of a representative healthy individual, H15.

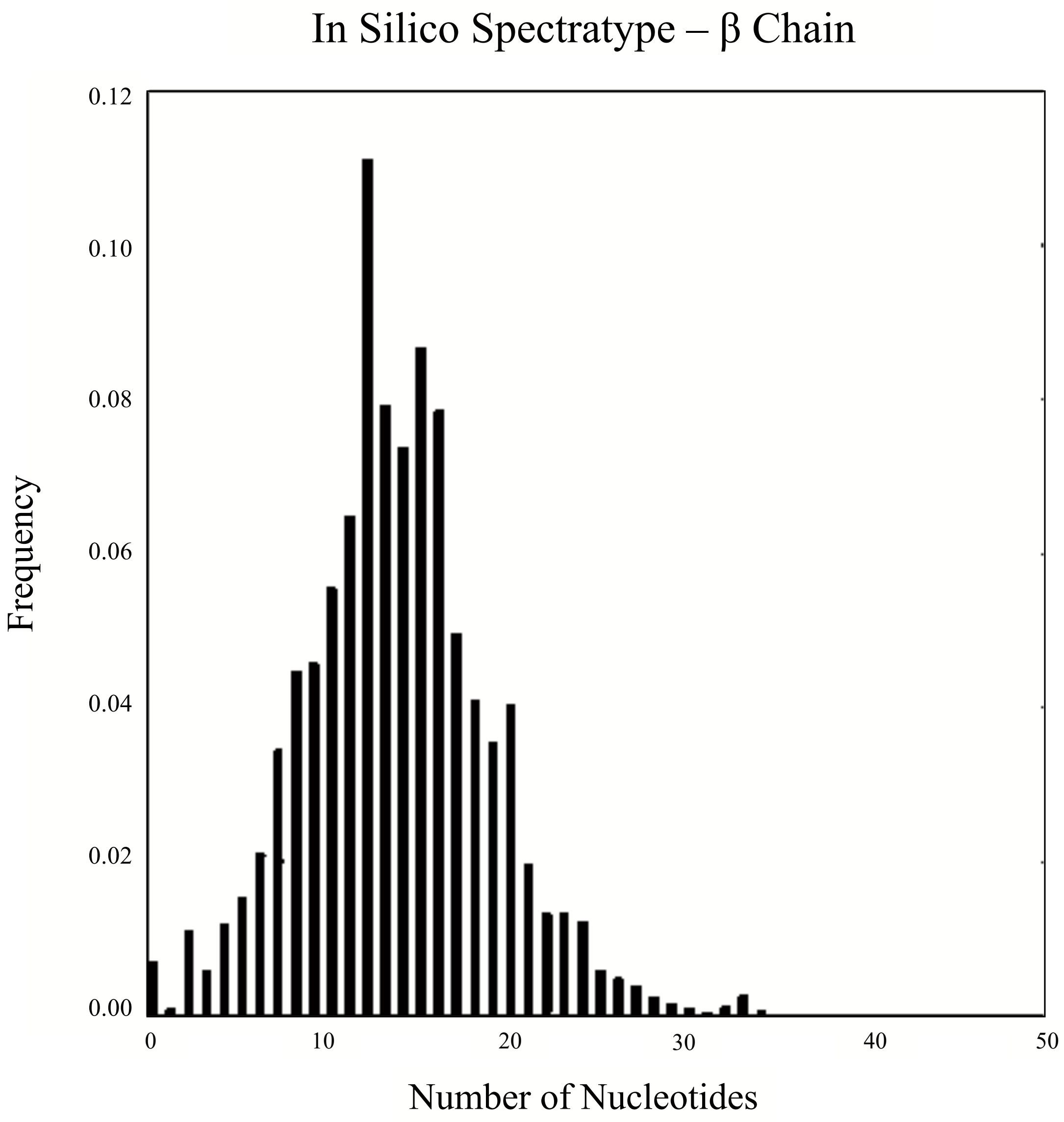


Figure 44. In silico generated spectratype data for the β chain of a representative healthy individual (H15).

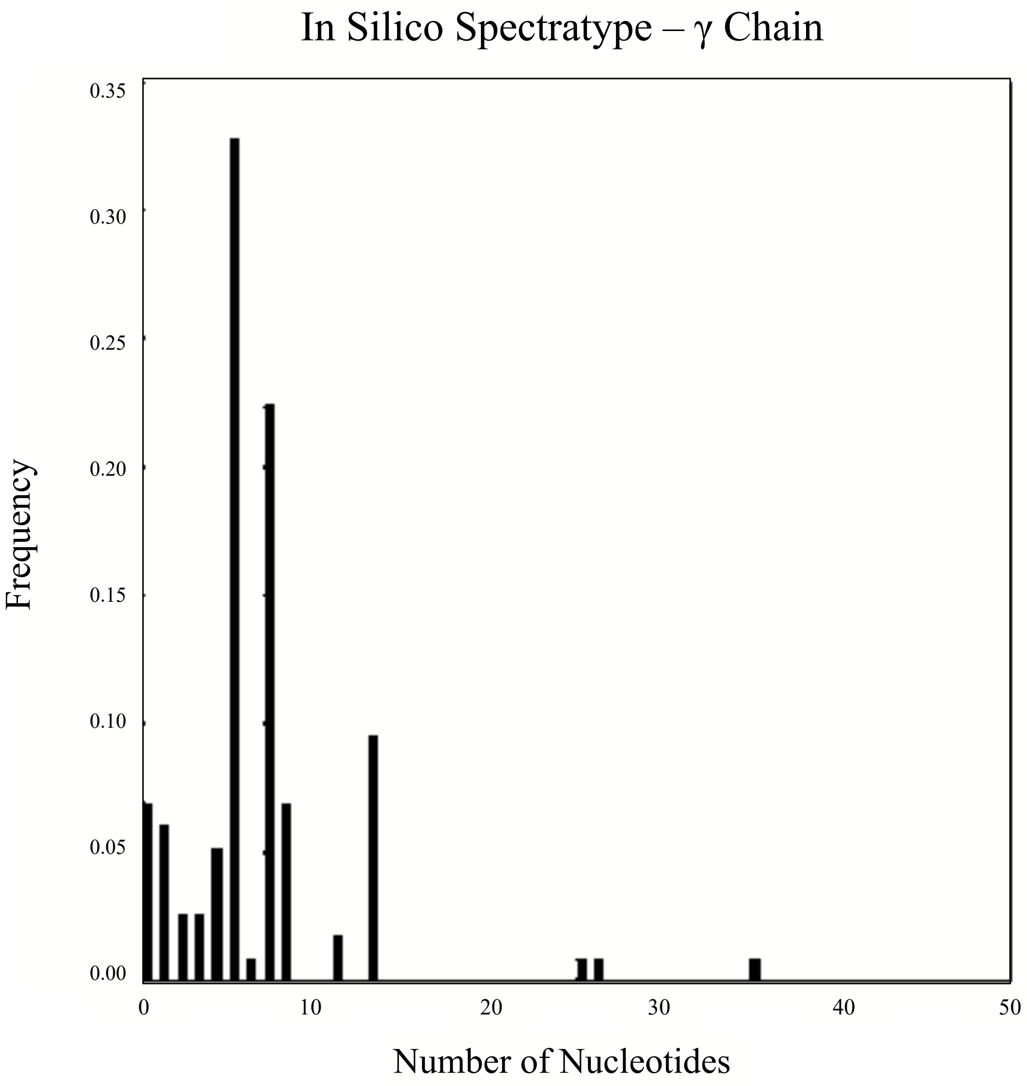


Figure 45. In silico generated spectratype data for the γ chain of a representative healthy individual (H15).

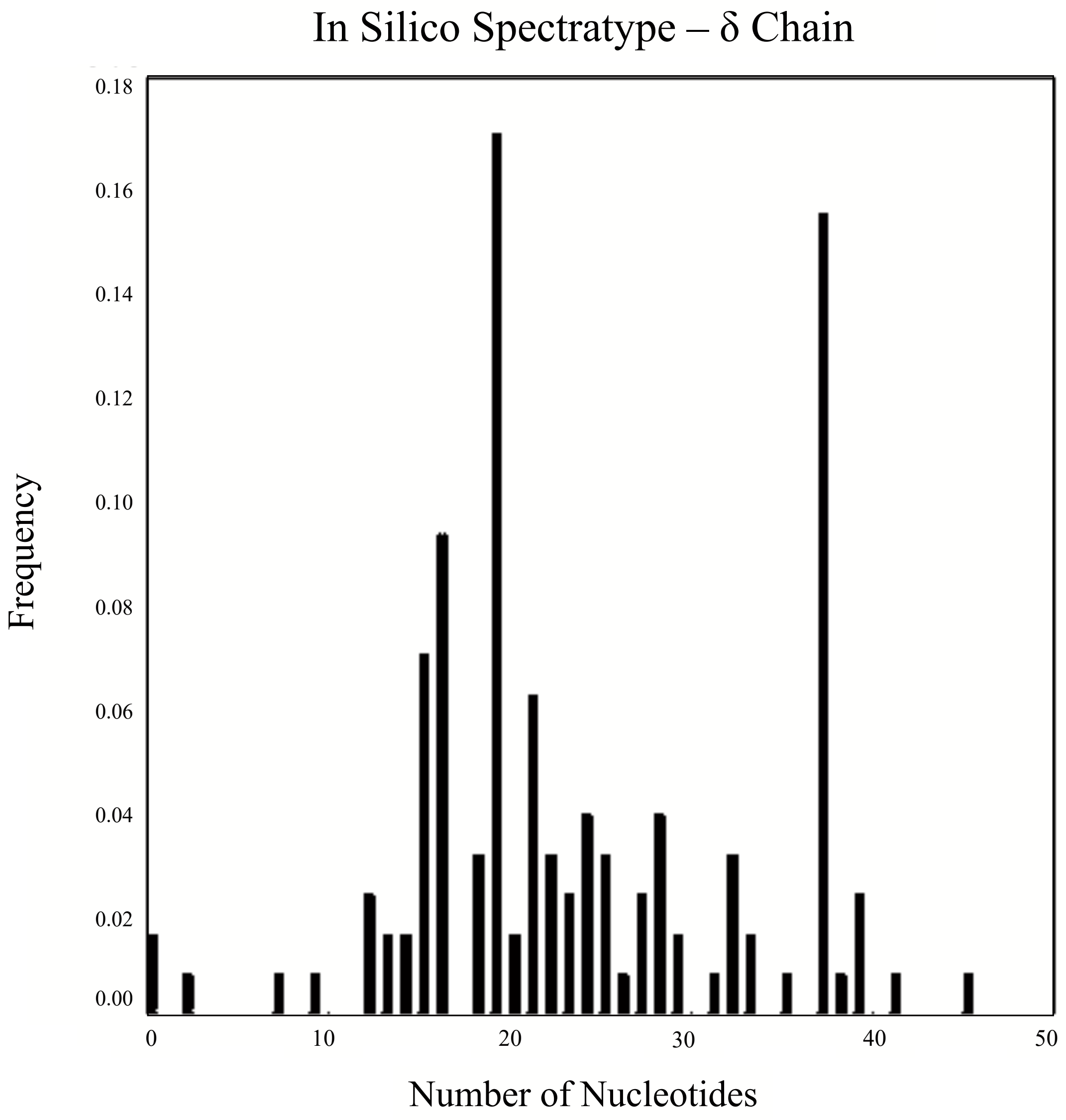


Figure 46. In silico generated spectratype data for the δ chain of a representative healthy individual (H15).

The β chain CDR3 lengths showed a distribution pattern, which was closest to the Gaussian distribution seen in the spectratype data of healthy individuals. The next most normally distributed was the α chain. The average length of the α chain CDR3 regions was shorter than the β-chain which is expected as the α chain (VJ) does not comprise the diversity gene segment like the β chain (VDJ) does. There were more individual peaks in both the α and the β lengths, which could reflect clonal expansion, especially as there was one prominent spike in each that may reflect a particular clonotype that had expanded. The γ and δ CDR3 length distributions showed less resemblance to a Gaussian distribution. This may reflect lower diversity in γδ+ T-cells than in αβ+ T-cells, or lower numbers of cells in the individual's blood sample.

The present study is the first to report the spectratype data for α, β, γ and δ chains from the same blood sample. The α and β distribution profiles observed are likely to be a result of clonal expansion in the individual at the time of sampling. The distribution may also reflect lower diversity highlighted by the high-throughput sequencing which shows higher resolution of the CDR3 lengths that are not observed with spectratyping.

4.3.5 Effect of Sequencing Platform on Clonotype Identification

To investigate the effect of sequencing platform on repertoire diversity, the number of clonotypes identified using MiTCR from Ion Torrent PGM and Illumina MiSeq data was compared. The total and unique clonotypes are shown in Table 41, along with the number of clonotypes that were only represented once and the percentages at which different groups of clonotypes were represented.

β Chain	**Ion Torrent PGM**		**Illumina MiSeq**	
	H1	**H2**	**H12**	**H15**
Total Clonotypes	55080	61764	10006	22202
Unique Clonotypes	30377	21158	3476	8877
Clonotypes Represented Once	45938	50937	8112	17410
% Unique Clonotypes	55.15%	34.26%	34.74%	39.98%
% Highest Frequency Clonotype	0.35%	0.49%	4.92%	3.70%
% Representation of Clonotypes Represented Once	83.40%	82.47%	81.07%	78.42%

Table 41. Total, unique and singularly represented β chain clonotypes identified by the Decombinator. Samples HC1 and HC2, which were sequenced using the Ion Torrent PGM, and HC7 and HC8, which were sequenced using the Illumina MiSeq, are included. Statistical analysis was not performed due to sample size (n = 2).

Table 41 shows total and unique clonotypes identified. Also shown is the percentage of the total clonotypes that were unique, which does not appear to vary greatly between the sequencing platforms although the percentage was higher for H1 than the remaining three, the percentage representation of the highest frequency clonotype, which was lower for the Ion Torrent PGM samples than for the Illumina MiSeq samples, and the proportion of all clonotypes that were only represented once, which was slightly higher for Ion Torrent PGM than the Illumina MiSeq.

The majority of Ion Torrent PGM reads were only represented once. This was 83.40% for H1 and 82.47% For H2. For the Illumina MiSeq, it was still a majority but lower than the Ion Torrent PGM, 81.07% for H12 and 78.42% for H15. Sequencing error is a potential problem for all datasets, as there may be an artificial overrepresentation of unique clonotypes creates the illusion of a highly diverse repertoire, with a high diversity index. Therefore, it may be problematic that the Ion Torrent PGM datasets demonstrate a higher diversity than the Ilumina MiSeq datasets.

4.3.6 Effect of Sequencing Platform on Repertoire Diversity

Following the comparison of numbers of clonotypes identified, the diversity values were investigated. The Inverse Simpson Index, Shannon Entropy and Gini Index were calculated as previously for H1, H2, H12 and H15.

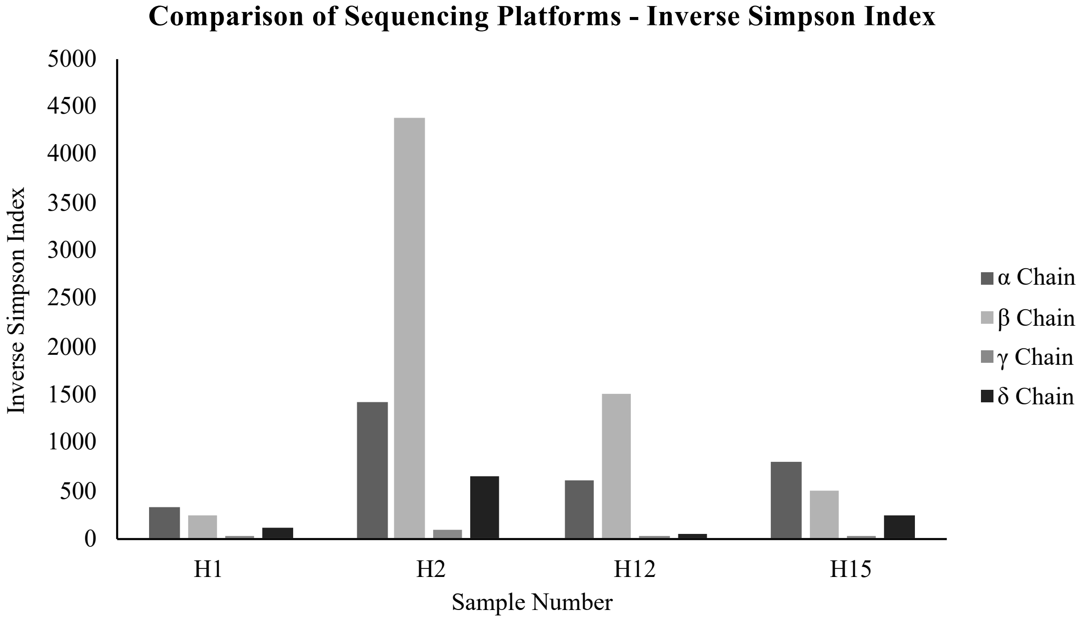


Figure 47. Inverse Simpson Index for Ion Torrent PGM and Illumina MiSeq samples. Samples HC1 and HC2 were sequenced using the Ion Torrent PGM platform, whereas HC12 and HC15 were sequenced using the MiSeq. Diversity indices were calculated using tcR from data analysed using MiTCR. Statistical analysis was not performed due to sample size (n = 2).

The inverse Simpson index in Figure 47 shows that the β chain is more diverse than the α chain for H1 and H15 and that the γ chain is more diverse than the δ chain for all four samples. It does not show a consistent difference between the Ion Torrent PGM and the Illumina MiSeq platforms.

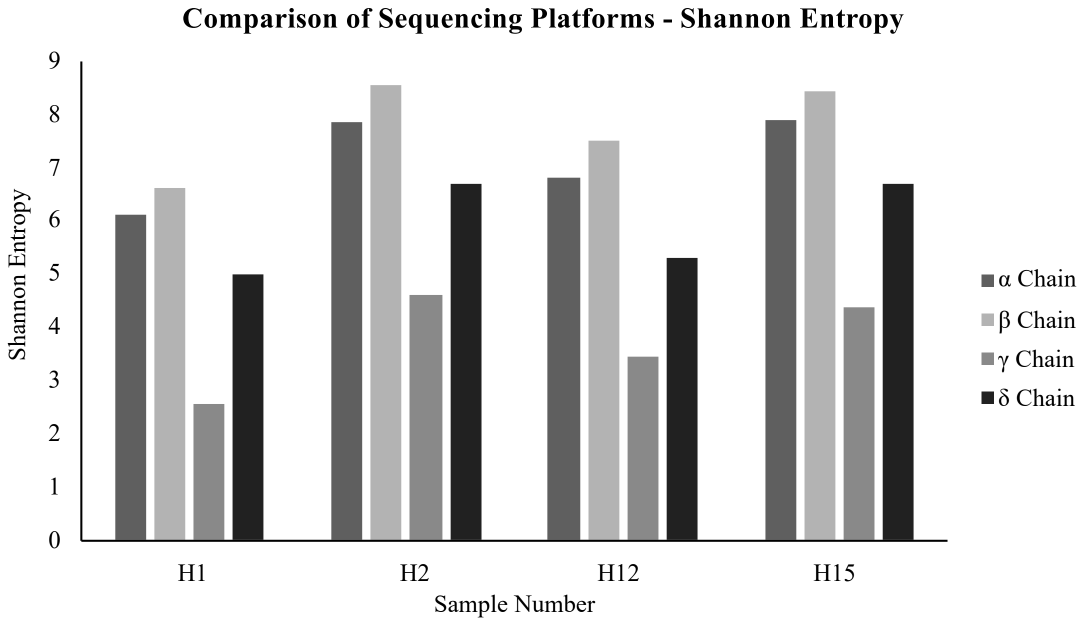


Figure 48. Shannon Entropy for Ion Torrent PGM and Illumina MiSeq samples. Samples H1 and H2 were sequenced using the Ion Torrent PGM platform, whereas H12 and H15 were sequenced using the MiSeq. Diversity indices were calculated using tcR from data analysed using MiTCR. Statistical analysis was not performed due to sample size (n = 2).

Figure 48 shows that the Shannon Entropy is consistent in its representation of diversity of each TCR chain. Across all four samples, the β chain shows higher diversity than the α chain, which is higher than the δ chain, which is higher than the γ chain. It does show a consistent difference between the Ion Torrent PGM and the Illumina MiSeq platforms.

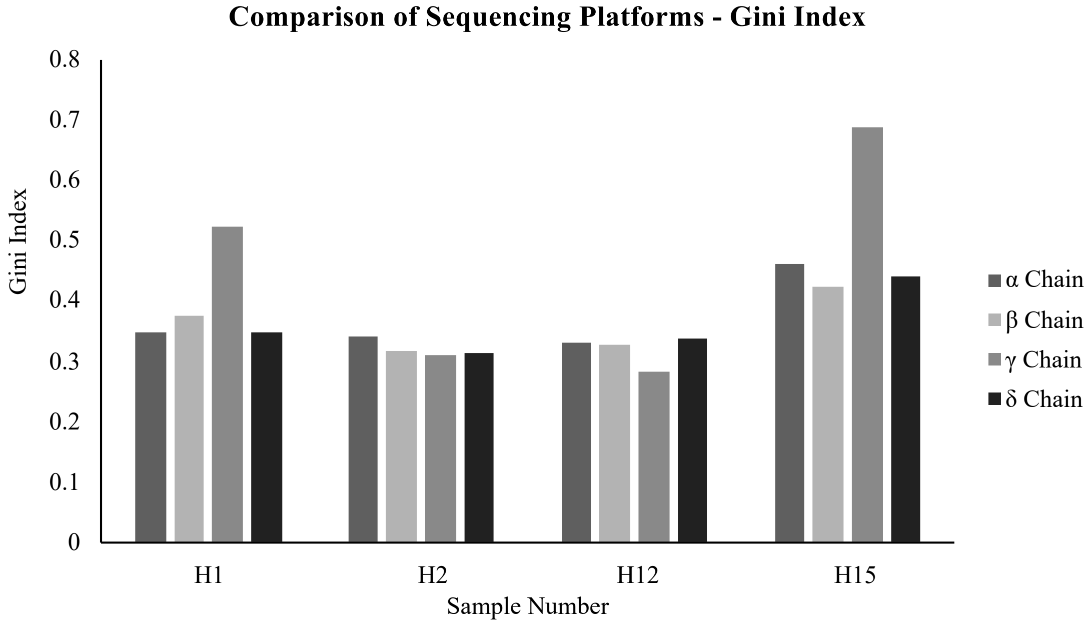


Figure 49. Gini Index for Ion Torrent PGM and Illumina MiSeq samples. Samples H1 and H2 were sequenced using the Ion Torrent PGM platform, whereas H12 and H15 were sequenced using the Illumina MiSeq. Diversity indices were calculated from data analysed using MiTCR. Statistical analysis was not performed due to sample size (n = 2).

Figure 49 shows that the Gini Index is less consistent in its representation of diversity of each TCR chain than the Shannon Entropy. Across three samples, H2, H12 and H15, the β chain shows higher diversity than the α chain. The δ chain shows greater diversity in the γ chain for half of the samples, H1 and H15. It does not show a consistent difference between the Ion Torrent PGM and the Illumina MiSeq platforms.

4.3.7 Effect of Analysis Platform on Repertoire Diversity

The distribution of the frequencies at which clonotypes are identified by the different pipelines was very similar. However, the choice of analysis pipeline may have an effect on the identification of clonotypes and therefore on the repertoire diversity, as each pipeline had a different computational approach. Diversity indices for 6 samples (H3 to H8), each sequenced on the Illumina MiSeq are shown in Figure 50 and Figure 51. For each sample there are 2 sets of indices; 1 for data analysed using the Decombinator and 1 for data analysed using MiTCR.

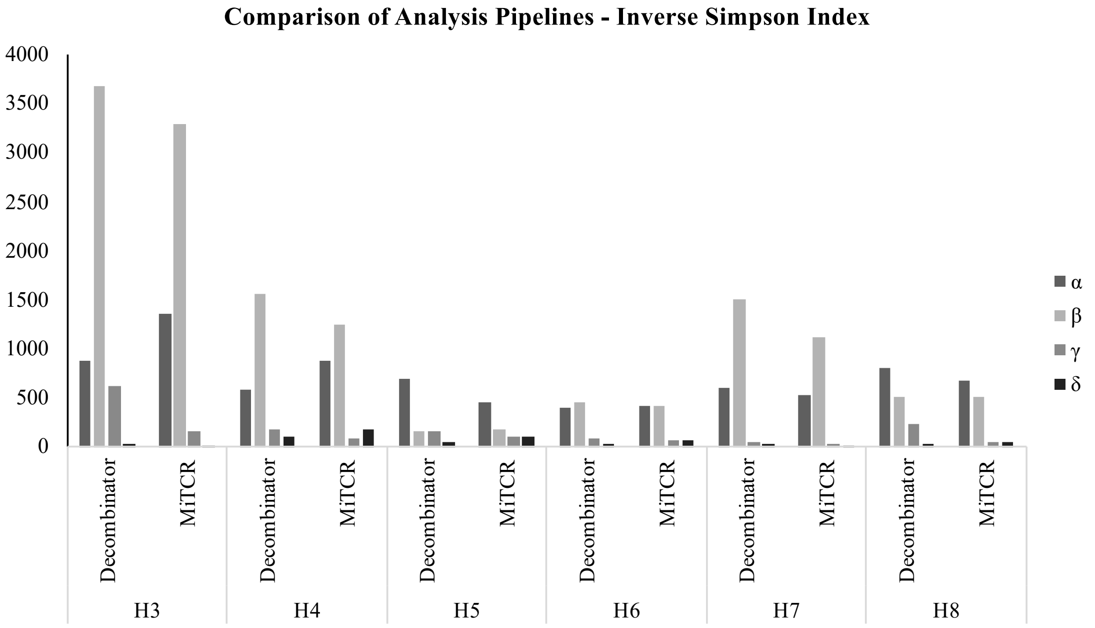


Figure 50 Inverse Simpson Index Values for α, β, γ and δ chains calculated from Decombinator and MiTCR data.

The values of the inverse Simpson index varied across all four chains and the values are different for each sample H3 to H8. Both the Decombinator and MiTCR generated similar values of this diversity index for all α, β, γ and δ chains of each sample.

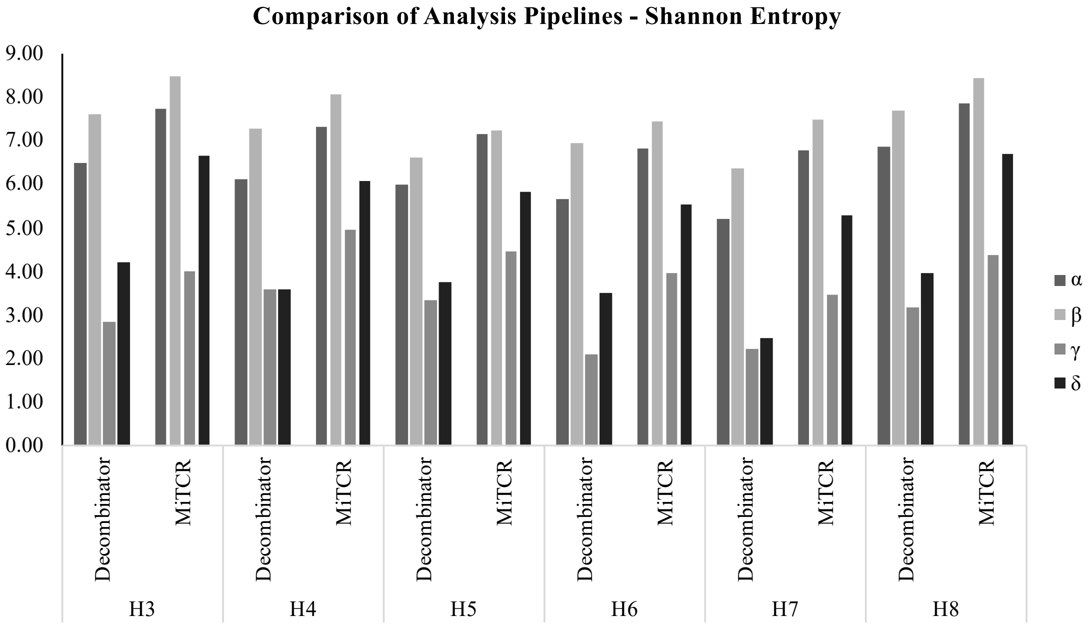


Figure 51 Shannon Entropy Values for α, β, γ and δ chains calculated from Decombinator and MiTCR data.

The values of the Shannon entropy varied across all four chains and the values are different for each sample H3 to H8. Both the Decombinator and MiTCR generated similar values of this diversity index for all α, β, γ and δ chains of each sample.

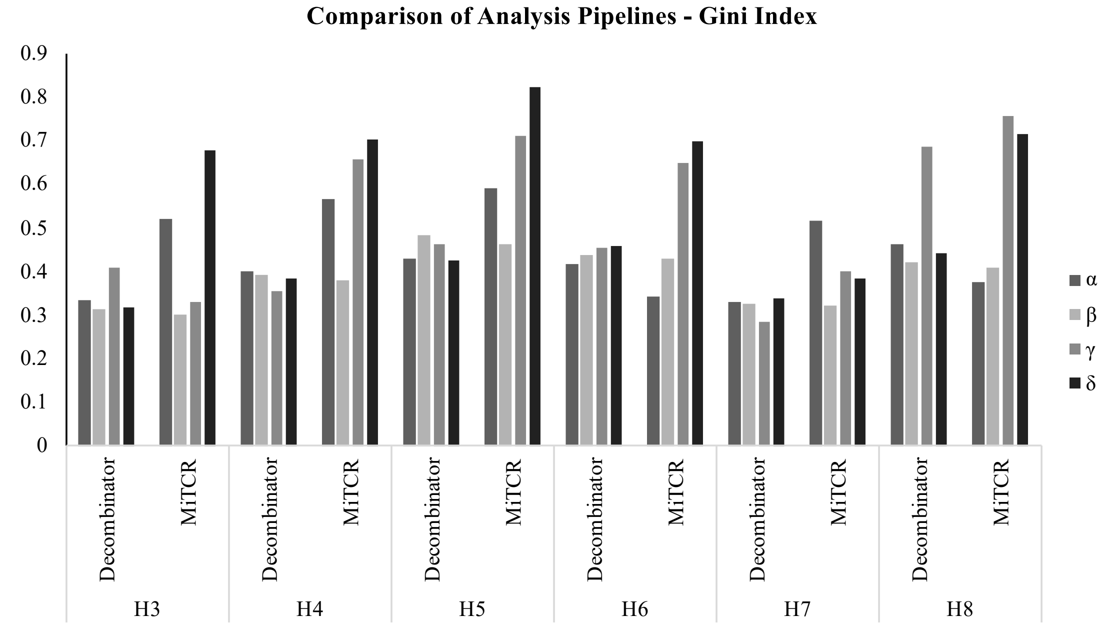


Figure 52 Gini Index Values for α, β, γ and δ chains calculated from Decombinator and MiTCR data.

The values of the Gini index varied across all four chains and the values are different for each sample H3 to H8. However, Figure 52 shows a difference between the values calculated from the Decombinator data compared to the values calculated from the MiTCR data of the α, β, γ and δ chains for samples H3, H5 and H8.

Figure 50, Figure 51 and Figure 52 demonstrate that the diversity of each sample is similar depending on which TCR identification software was employed.

4.3.8 Investigation of Sequencing Depth Using Subsampling

The previous chapter suggested that read depth of 1,000,000 reads per sample was sufficient to represent V and J usage. This usage was investigated further by looking at diversity measures of subsampled data. This was done using the same FASTQ script (as in Chapter 3). Each of these subsampled output files were analysed using the Decombinator and the results are shown below.

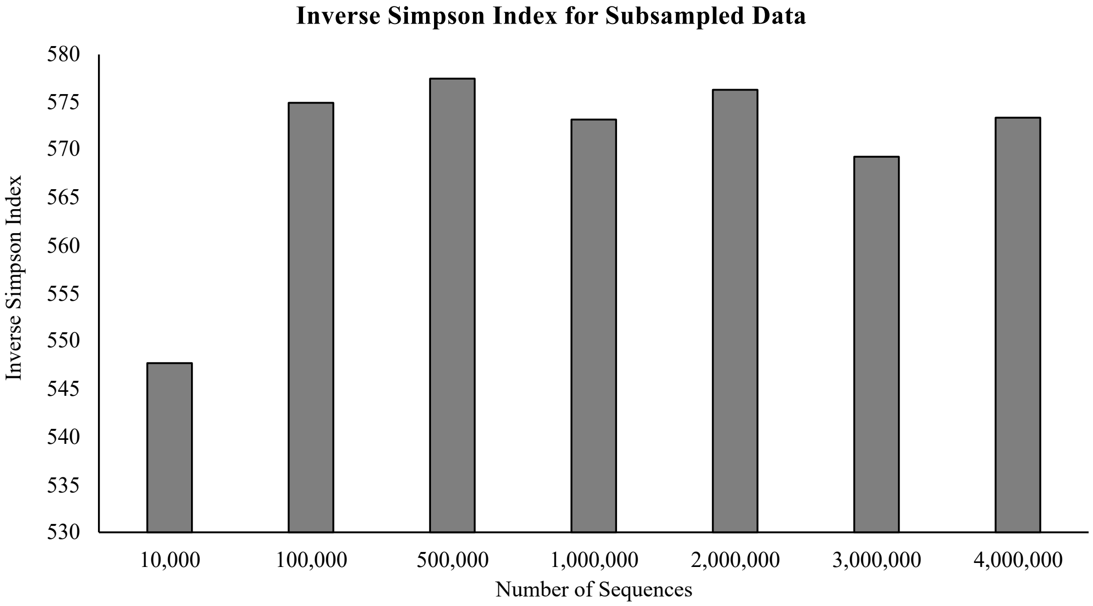


Figure 53. Inverse Simpson index for β chains of seven sequentially subsampled datasets. The diversity remains almost consistent across the different sized data files from H3. Diversity is lower at 10,000 sequences. Statistical analysis was not performed due to sample size (n = 1).

The data obtained supports the hypothesis that the diversity measures are mostly independent of sample size. With high-throughput sequencing, it was difficult to predict exactly how many sequences will be obtained from one run, or for each particular indexed sample within that run. Therefore, it was important to be able to standardise analysis for different sized datasets. Within sample datasets taken from a larger dataset, values for the inverse Simpson index did not differ greatly, as shown in Figure 53. This suggests that the estimate of diversity obtained is stable, provided the number of sequencing reads was >100,000 reads, again suggesting that depth of sequencing was a major factor when interpreting results. Samples with <100,000 reads were likely to provide a distorted perspective of the repertoire.

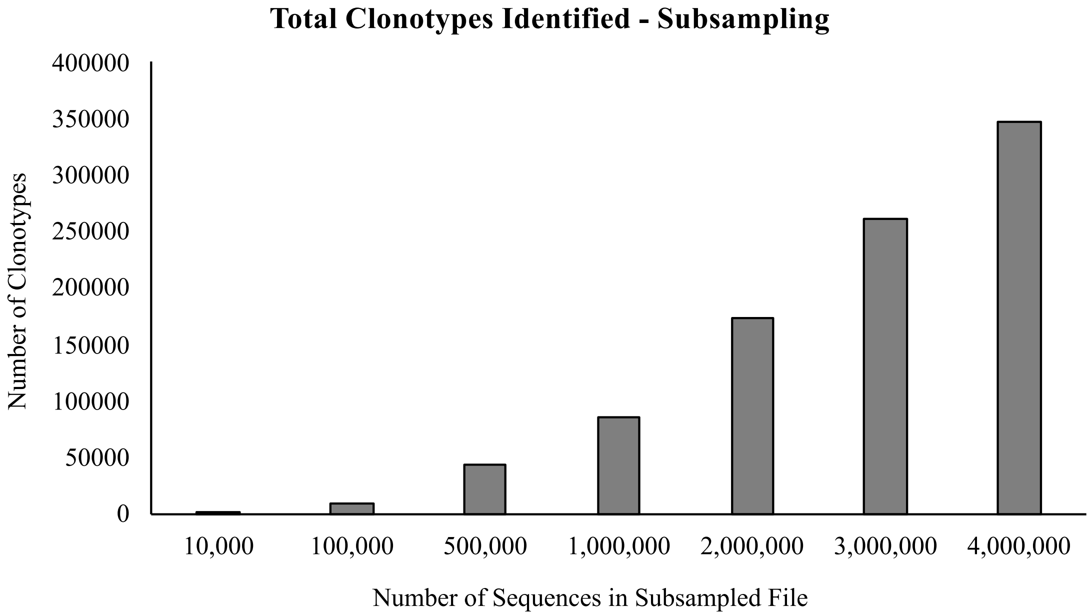


Figure 54. Total clonotypes identified in subsampled data. The total number of clonotypes increases proportionally with the number of reads. Statistical analysis was not performed due to sample size (n = 1).

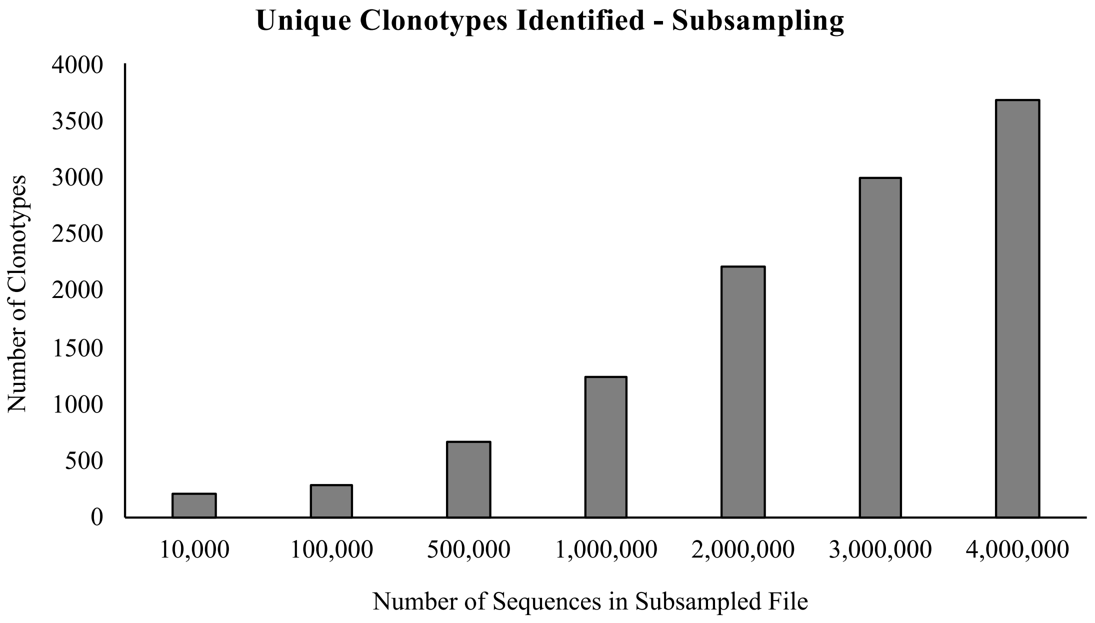


Figure 55. Unique clonotypes identified in subsampled data. The number of unique clonotypes increases proportionally with the number of reads. However, the number of unique clonotypes identified increases at a greater rate to start, but then the rate of identification of new unique sequences begins to level off. Statistical analysis was not performed due to sample size (n = 1).

When the sample size was plotted against the number of unique clonotypes identified within that subsample, more unique clonotypes were found in larger files, as shown in Figure 54. However, it was not clear from this whether all clonotypes present in the original blood draw were represented by this sample. It was hypothesised that the curve of this graph would approach an asymptote, representing the total number of clonotypes. This suggests that there were more species in the sample than have been detected, which demonstrates the sampling issue with regards to repertoire data.

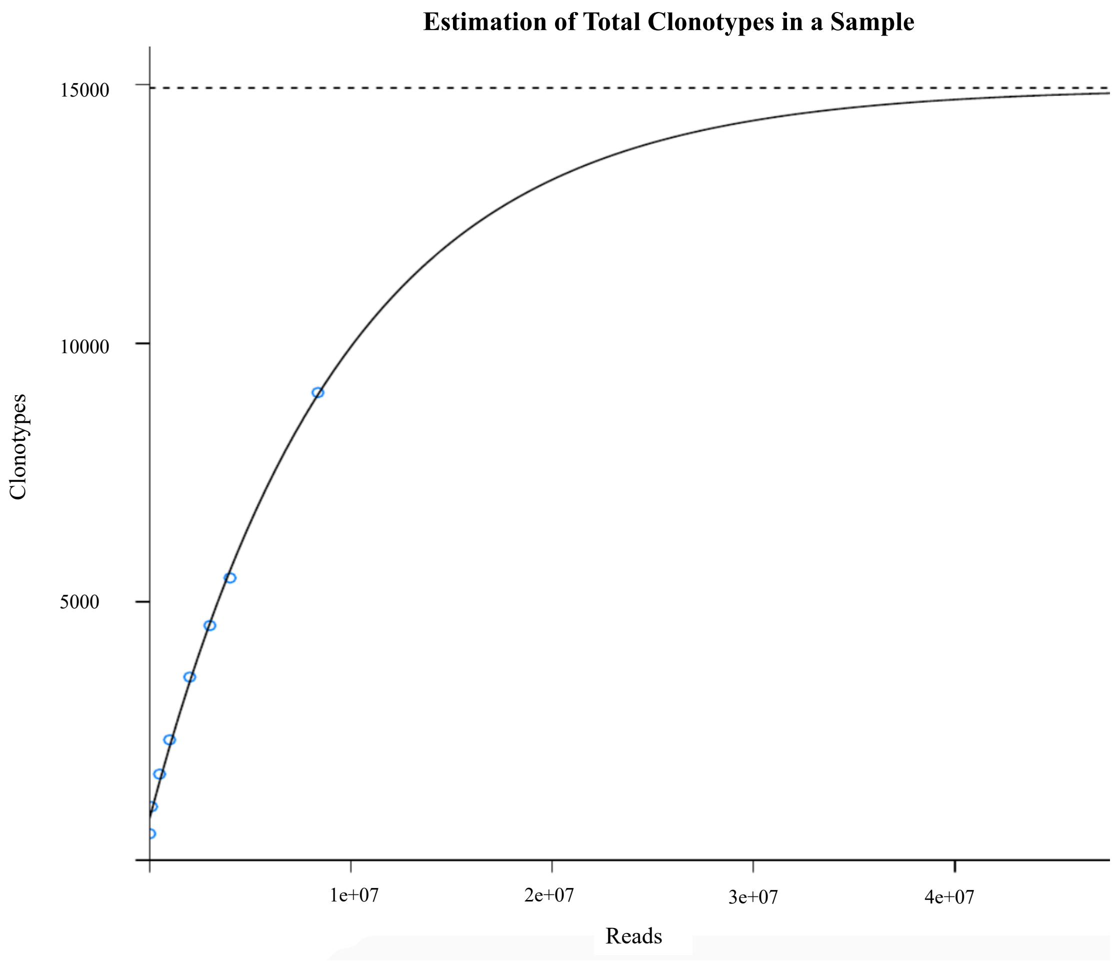


Figure 56. Extrapolation of subsampled datasets to estimate the total clonotypes in a given sample (H3). Taking the subsampled dataset and, the 1-exponential regression was plotted. This estimates that the total number of clonotypes in the sample was 14,128.

A 1-exponential regression was applied to the subsampled data, to estimate the total number of clonotypes in the sample, as shown in Figure 56. The calculation was done in R, as shown in Appendix 3. This takes the numbers of unique clonotypes identified for each subsampled file and extrapolates. However, this method may only approximate the clonotypes in the sequencing library, without taking into account material lost at different stages of library preparation. RNA was extracted from one million PBMCs, and at each stage only a proportion of the material was used to progress. Estimating the total number of clonotypes in the entire PBMC sample, or in a human remains challenging.

4.4 Discussion

This method has the capacity to simultaneously collect data from the α, β, γ and δ TCR chains, which has not been done previously. However, results obtained here using our capture technique appear to result in low yields of identified clonotypes. A significant number of clonotypes identified were non-functional, and the proportion was not consistent across the different TCR chains. A higher proportion of α chain clonotypes were non-functional than β chain clonotypes.

The diversity values differ across the different TCR chains, but this was not always consistent. For example, β chain diversity was not always higher than α chain diversity. Diversity values also demonstrate variation between healthy control individuals. This may be as a result of sub-clinical illness in individuals and/or a clonotype expansion in response to a challenge from pathogens. It would have been advantageous to acquire longitudinal samples from the same individuals. Factors such as age are also likely to play a role in diversity. However, the values calculated here may reflect a normal range of diversity that is required for a sufficient response to pathogens to maintain T cell immunity in health.

The diversity indices present here also varied in their robustness. The inverse Simpson index was selected first due to the simplicity of calculating it but the values in this chapter are not supported by other sources of data on TCR diversity. The values of each of the diversity values was verified by manual calculation of a few representative samples. Based on the results shown in this chapter, it is not likely that the inverse Simpson index would be recommended for use in this situation, as the results do not fit with known diversity of TCR repertoires, whereas the Shannon entropy and Gini index are in support of current understandings. Moving forward, inverse Simpson index will not be presented in this work.

As previously shown (Chapter 3), a depth of 1,000,000 FASTQ sequences using this method is considered sufficient to represent TCR diversity within an individual. However, factors such as the cost of multiplexing, the inability to predict how many sequences will be obtained from a run on the MiSeq, will play a role in determining the sequencing depth.

The consistency in the percentage of MiTCR clonotypes identified by the Decombinator for each chain suggest that this may be due to differences between the ways in which each of the programs identifies TCR sequences.

The difference in retention of β over α clonotypes (Figure 35 and Figure 36) may reflect an issue with the sequencing or analysis pipeline, or suggest that there is a higher rate of error when the α chain is transcribed compared to the β chain. In addition, for reasons unclear, there be a lower yield of α chain sequences, a feature that requires further validation.

Many clonotypes were detected only once in each dataset. This suggests that many further rare clonotypes may be present. This is partly a sampling issue, as it is not possible to sequence the entire TCR repertoire of a human, only a subsample. The total repertoire may contain up to 10^{12} T-cells, and therefore many of the rarer clonotypes may not be identified at all (68). There is also a difficulty in modelling clonotype extrapolation from subsampled datasets, highlighting future work that can be done in this area, as supported by conclusions drawn by other groups working on modelling TCR repertoires (116).

Sequencing errors, such as base substitution, addition and deletion, were incurred by both the Ion Torrent and Illumina platforms. The Ion Torrent platform has also been shown to have a homopolymer issue, which may contribute to the error rate. More sequencing errors may lead to a greater diversity, but further work is required in this area to understand the size of this effect and whether it is uneven across samples.

Overall, the work described in this chapter highlights the need for further development of analysis techniques, in particular analytic pipelines tailored to this approach. Considering this and rates of PCR and sequencing error, it is clear that conclusions from these datasets must be made with caution. However, it is possible to gain a large amount of information about the TCR repertoire using this technique.

It would be expected that the β chain would be more diverse than the α chain, and the δ chain to be more diverse than γ, as both β and δ comprise an additional diversity segment (D) between the V and the J. This provides two further junctional regions for nucleotide addition and deletion, which is not found in either the α or the γ chains. The β chain may appear more diverse as there are greater numbers of β chain sequences but the diversity measures used are independent of sample size. The differences observed between samples may also be due to the suitability of the diversity indices. Overall, the current study demonstrated a range of values that can be expected from healthy individuals, and average values for diversity, which can contribute to the understanding of TCR diversity in human health.

Chapter 5

Clinical Applications of T-Cell Receptor Sequencing

5 Clinical Applications of T-Cell Receptor Sequencing

5.1 Introduction

The TCR repertoire can become distorted in many disease states. Therefore, the ability to investigate immune repertoires at high resolution is applicable to many clinical situations and the results may influence treatment regimens or help to discover new therapies. The majority of this chapter will focus on the use of TCR repertoire sequencing in paediatric patients undergoing hematopoietic stem cell transplantation (HSCT). There will be an additional section related to sequencing expanded γδ+ T-cell populations as part of the development of novel immunotherapies targeting neuroblastoma.

5.1.1 Hematopoietic Stem Cell Transplantation

Slow recoveries of T-cell numbers and repertoire diversity have been associated with higher post-transplant infection and/or viral reactivation and incidence of relapse following HSCT for leukaemia (124-126).

Allogeneic HSCT (allo-HSCT) is a potentially curative treatment for a range of malignant and non-malignant diseases, including haematological malignancies and primary immunodeficiency diseases (PIDs). Here, we addressed the use of allo-HSCT to treat acute myeloid leukaemia (AML) in paediatric patients. Allo-HSCT is, in most cases, an end of the line treatment, following the failure of chemotherapy. In 1968 the world's first successful paediatric HSCT was carried out using a matched sibling donor (127). Historically, bone marrow has been the source of stem cells for transplant, but the discovery and utilisation of granulocyte colony-stimulating factor (G-CSF) and granulocyte-macrophage colony-stimulating factor (GM-CSF) to increase circulating stem cell numbers has meant that the stem cells can be collected less invasively from the peripheral blood of donors (128).

Only 30% of patients eligible for transplant have a matched sibling donor, so the use of other sources of stem cells have become increasingly commonplace, such as matched unrelated donors (MUD), mismatched cord blood and haploidentical related donors, usually a parent. Stem cells for transplantation are isolated through positive selection of CD34+ cells or through T-cell depletion (TCD) (129). TCD may involve the depletion of CD3+ cells, CD3+/CD19+ cells, or TCRαβ+/CD19+ cells (130). In the case of the latter, it is hypothesised that the remaining γδ+ T-cells may be able to provide some residual anti-infection and anti-tumor protection for the host, whilst being non-alloreactive and not increasing the risk of graft versus host disease (GVHD) (131-133).

Transplant-related complications, including GVHD, opportunistic viral infections, relapse, secondary malignancies and immune disorders, happen early post-transplant and have a severe effect on patient recovery (134-136). Recent consensus is that these events can attribute to delayed immune reconstitution (137-139). Current treatment regimens for GVHD and viral infections, such as CMV, EBV and adenovirus, include repletion of T-cells collected from the patient prior to total body irradiation (TBI). In the case of viral infections, these T-cells are exposed to antigen and the subsequent expanded clones are considered to be specific to the virus. These T-cells can then be returned to the patient in case of uncontrolled viraemia. Alternatively, adoptive transfer, where donor-derived virus-specific T-cells are transferred directly to the patient, may be carried out (140, 141).

AML patients undergo a pre-transplant conditioning regime primarily for anticancer purposes and immunosuppression to reduce immune responses from the host to the transplant. This comprises drug administration in addition to the myeloablative TBI. These drugs may include alkylating agents such as busulphan, cyclophosphamide, thiotepa and treosulphan, or analogues of purine such as fludarabine. Most patients also receive chemotherapy treatments after transplant to reduce the activity of their immune system, including cyclosporin, mycophenolate (also known as mycophenylate mofetil) and corticosteroids. The effect of these drugs reduces the number of circulating T-cells and their ability to proliferate, which will have a downstream effect on TCR diversity.

One of the primary goals of HSCT is the restoration of TCR diversity and subsequent immune protection. Limited TCR diversity following HSCT has been associated with increased rates of infection and relapse. Hsieh et al. found that lower thymic output and restricted patterns of TCR diversity were linked to an increased likelihood of acquiring opportunistic infections following HSCT in children with PID, although these were not indicators of disease severity (142).

T-cell reconstitution following HSCT is dependent on two factors; *(a)* the output of naïve T-cells from the thymus and *(b)* proliferation of T-cells in the periphery. Thymic output requires five to six months to recover in paediatric patients following HSCT, as defined by the increase in proportion of CD45RO- naïve T-cells with high counts of

TCR thymic excision circles (TRECs), and this timeframe is likely to increase with the age of the recipient (143-145). Current evidence supports the idea that most early T-cell reconstitution, including clonal response to infection, is due to expansion of mature donor cells (146). This early expansion is crucial in providing immediate immune protection following transplant and the expansion of regulatory cells may prevent adverse host reactions to the graft. However, this effect is short-lived, and Okamoto et al. have subsequently suggested that long-term TCR diversity is actually dependent on *de novo* thymic output (147). Patients shown to have higher rates of thymic output before transplantation have been noted to have better outcomes, which highlights the importance of thymic-dependent T-cell reconstitution (148).

Differences in T-cell reconstitution between different sources of stem cells have also been reported. Enhanced T cell reconstitution in umbilical cord transplants differs from typical HSCT (149, 150). The bone marrow transplant team at GOSH demonstrated that the naïve CD4+ cells reconstituted after cord blood transplant were most similar to foetal CD4+ T cells, which was done by the comparison of the transcription profiles of the different CD4 T cells. This has important consequences for the T-cell repertoire as the transcription profile of reconstituting naive CD4$^+$ T cells from cord blood transplant (CBT) recipients was upregulated in the TCR signaling pathway and its transcription factor activator protein-1 (AP-1) (149). Early data indicate that diversity does increase over time but that those in receipt of CBT had greater diversity at 6 months than TCD. Clinical reports suggest there is a protective benefit to this, as the rates of infection and relapse are lower in the population of CBT recipients (151-153). Possible explanations for this maybe the fact that ~7000 times as many T-cells are given in a CBT transplant and in addition, the procedure involves the absence of T-cell depleting steps. Furthermore, immune reconstitution in patients receiving TCD transplants may be more rapid in the case of CD3+/CD19+ depletion compared to CD3+ depletion alone (154).

The data presented in this chapter does show an increase in CD4+ T-cell count at the second time point and onwards for two of the patients that received CBT, Patient 1 and Patient 2, but not Patient 3. However, this is preliminary data and a larger cohort would be needed to draw conclusions. In future, there is scope for many studies to investigate further the effects of CBT on T-cell reconstitution and TCR diversity.

Several methods measuring repertoire diversity have been used post-HSCT, and clinical labs routinely perform spectratype, FACS and TRECs analysis (142, 147). High-throughput sequencing has more recently been used to investigate the reconstitution of the adult immune system following HSCT, through the use of 5' RACE PCR to assess the TCR-β chain repertoire of CD4+ cells (9). Applications of high-throughput sequencing technologies following HSCT have more commonly been for the detection of minimal residual disease (MRD) to predict disease relapse in haematological malignancies (155, 156).

Studies to date have involved small samples sizes and few have utilised high-throughput sequencing technology. Therefore, our knowledge on the recovery of the TCR repertoire following HSCT can still be considered limited. There are many opportunities for the application of immune repertoire sequencing in comparing reconstitution and patient benefit achieved through different clinical paths, *e.g.* through the use of cord blood or haploidentical donors as stem cell sources. Herein, data was obtained through the high-throughput sequencing and targeted capture of peripheral T-cell samples from pediatric patients' post-transplantation. The main objective of the current study was to test the capture technique in a clinical setting and, if appropriate, the technique may form the basis of a larger study in the future. TCR sequencing has the potential to be adopted in the clinical laboratory for routine assessment of patient repertoire diversity in a research capacity.

5.1.2 IPP Expansion of γδ+ T-Cells

The most frequently identified and best characterised subset of γδ+ T-cells are the Vγ9Vδ2 subset. Vγ9Vδ2 T-cells expand in the presence of phosphoantigens as a component of the mevalonate-independent metabolic pathway, which is how isopentyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) are synthesised. These molecules are important both because they are isoprenoid precursors and they act as ligands to Vγ9Vδ2, thus stimulating expansion of this cell population. Isoprenoids are metabolically important and in mammals include molecules such as cholesterol, bile acids and some hormones, although many more have been described (157).

Recent evidence has also suggested that γδ+ T-cells may have antitumor properties, through antibody-dependent cell-mediated cytotoxicity (ADCC) and these effects increase in the presence of APCs presenting antigen derived from malignant cells (158, 159). However, Vγ9Vδ2 cells represent only a proportion of all γδ+ T-cells present in the periphery and therefore, it is also helpful to investigate the other populations. Therefore, this capture technique was applied to the sequencing of expanded γδ+ T-cells to both look at the expanded populations and to demonstrate the utility of the technique.

5.2 Methods

Samples used in the HSCT experiments presented here were obtained from GOSH, with the support and collaboration of Dr Stuart Adams, Professors Paul Veyes and Graham Davies. Samples used in the γδ+ T-cell experiments were obtained from Dr Jonathan Fisher and Professor John Anderson. Data presented here was published by Fisher et al. in 2014 (37). All samples were received as RNA aliquots, which were extracted using Qiagen RNeasy Mini Kits (haematopoetic stem cell transplant patients) and TRIzol (γδ+ T-cell experiments).

5.2.1 HSCT Patient Samples

Samples from four patients who had recently received HSCTs for the treatment of AML were sequenced using this high-throughput sequencing and targeted capture method (T1, T2, T3 and T4). The first three patients (T1, T2 and T3) were recipients of stem cells derived from cord blood. The fourth patient (T4) was the recipient of a haploidentical TCR αβ+/CD19+ depleted transplant, one of the first to be carried out in the United Kingdom (160). Samples include the cord blood used in the transplant, where applicable, plus clinical follow-ups from patients at approximately day 30, 60, 90 and 180, as shown in Table 42. Twelve samples were sequenced in total in this group, multiplexed on one MiSeq run, numbered TCR14. Statistical analysis was not performed on results in this section as each patient presented with different conditions, treatment regimens and time points and were therefore not directly comparable.

Patient Number	Sample Reference	Patient Condition	Patient Transplant	Date of Transplant	Starting Material	Date of Sample	Time Point	Sequencing Index
Patient 1 (T1)	T1.1	AML	CBT	22/08/2013	Cord blood	22.08.2013	Day 0	1
	T1.2				PBMC	21.09.2013	Day 30	2
	T1.3				PBMC	21.10.2013	Day 60	3
	T1.4				PBMC	04.02.2014	Day 170	4
Patient 2 (T2)	T2.1	AML	CBT	11/10/2013	Cord blood	11.10.2013	Day 0	5
	T2.2				PBMC	08.12.2013	Day 60	6
	T2.3				PBMC	09.04.2014	Day 160	7
Patient 3 (T3)	T3.1	AML	CBT	07/11/2013	Cord blood	07.11.2013	Day 0	8
	T3.2				PBMC	10.12.2013	Day 37	9
	T3.3				PBMC	12.02.2014	Day 99	10
Patient 4 (T4)	T4.1	AML	TCD	06/11/2013	PBMC	09.12.2013	Day 33	11

Table 42 Patient samples and time points sequenced.

5.3 Results

5.3.1 Sample Representation in Sequencing Run

All of the HSCT patient samples were sequenced on the same run, TCR14. Samples were indexed using Illumina barcodes 1 to 12, as shown below. Although the samples from HSCT patients were normalised for concentration before being sequenced in the same run, Figure 57 shows that there was uneven representation of indexed samples in the run. Illumina Base Space was used to demultiplex samples and this gave the number of sequences acquired according to index.

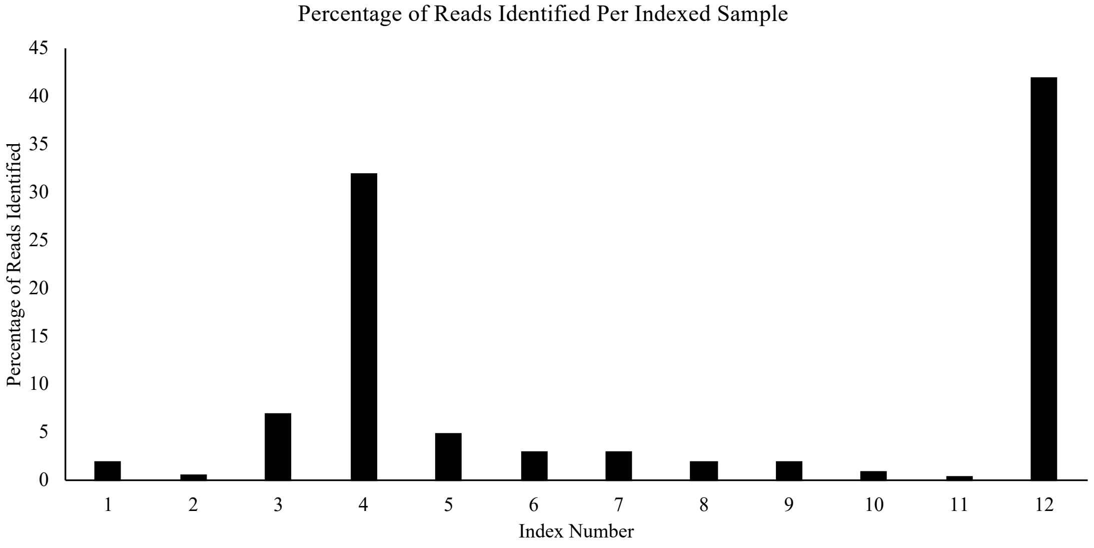


Figure 57. Representation of samples within sequencing run. Statistical analysis was not performed due to sample size (n = 1).

The two samples with the highest representation in Figure 57 were T1.4 (index 4) and T4.2 (index 12), which represent samples taken at time points + 170 and + 61 days' post-transplant respectively. Therefore, this may also be an indication of repertoire recovery over time and a consequent increase in quality and quantity of TCR transcripts, providing more template for the capture reaction. The uneven representation of samples shown in Figure 57 may indicate bias in the pre capture PCR amplification, the post capture PCR amplification, the capture reaction or the sequencing. It may also reflect the need for more accurate quantification of sample concentration. These issues may arise from sample quality, or a different proportion of target sequences in the original sample material. The different representations of samples and different numbers of reads obtained could affect data analysis downstream.

5.3.2 Recovery of White Blood Cells Post-Transplant

T-cell counts in patients' tested were low, which is frequently seen following HSCT. Myeloablative treatment and immune conditioning pre-transplant eliminated the majority of T-cells in the periphery, which left the patients with negligible to low cell counts at day +0. Samples at day +30 recorded very low cell counts, which gradually increased over time, but some white cell counts also decreased over time as shown in Table 43. In this first group, cell counts did not appear to stabilise within the time frame investigated, contributing to the low sequence numbers of some samples. In future sections, the focus will be on the lymphocyte counts.

Patient	Time Point	White Cells /10^9/L	Neutrophils /10^9/L	Lymphocytes /10^9/L	Monocytes /10^9/L	Eosinophils /10^9/L	Basophils /10^9/L
T1	+0	1.13	1.06	0.04	0.02	0.01	0
	+30	8.64	4.17	2.31	1.92	0.22	0.03
	+60	5.65	4.19	0.98	0.29	0.18	0.01
	+170	6.08	0.15	2.34	0.94	0.15	0.01
T2	+0	3.62	3.4	0.08	0.03	0.1	0.01
	+60	1.45	0.65	0.49	0.3	0	0.01
	+160	5.37	3.79	1.01	0	0	0.01
T3	+0	1.62	1.45	0.02	0.01	0.13	0.01
	+37	5.96	2.02	0.72	1.37	1.83	0.02
	+99	9.19	5.44	2.27	1.01	0.45	0.02
T4	+0	0.07	ND*	ND	ND	ND	ND
	+33	5.59	3.64	0.51	1.3	0.11	0.03
	+61	4.08	2.41	1.22	0.37	0.07	0.01

Table 43. Patient white blood cell counts at all time points made available. Cells were counted by the Clinical Haematology Laboratory at Great Ormond Street Hospital. Data released with the permission of the attending physician. *ND refers to “not done” where tests were not done.

5.3.3 Bias in V and J Gene Usage post-transplantation

High-throughput sequencing data from patient samples was aligned to V and J regions, using both MiTCR and the Decombinator. Heat maps demonstrating V and J combinations were generated using the Decombinator. The profile of V regions identified may be skewed following HSCT compared to a healthy individual. The haploidentical TCR αβ+/CD19+ depleted transplant recipient, T4, was used to demonstrate V and J usage at + 33 and + 61 days post-transplant in Figure 58 and Figure 59 respectively.

Previous data shows that there is also a dominance of certain VJ combinations in healthy individuals and that this can vary between individuals. It holds true that in these HSCT patients there is also a bias in VJ combinations. Patient T4 was used to illustrate how the repertoire may change over time following transplant. Figure 58 and Figure 59 show the VJ pairings at + 33 and + 61 days' post-transplant.

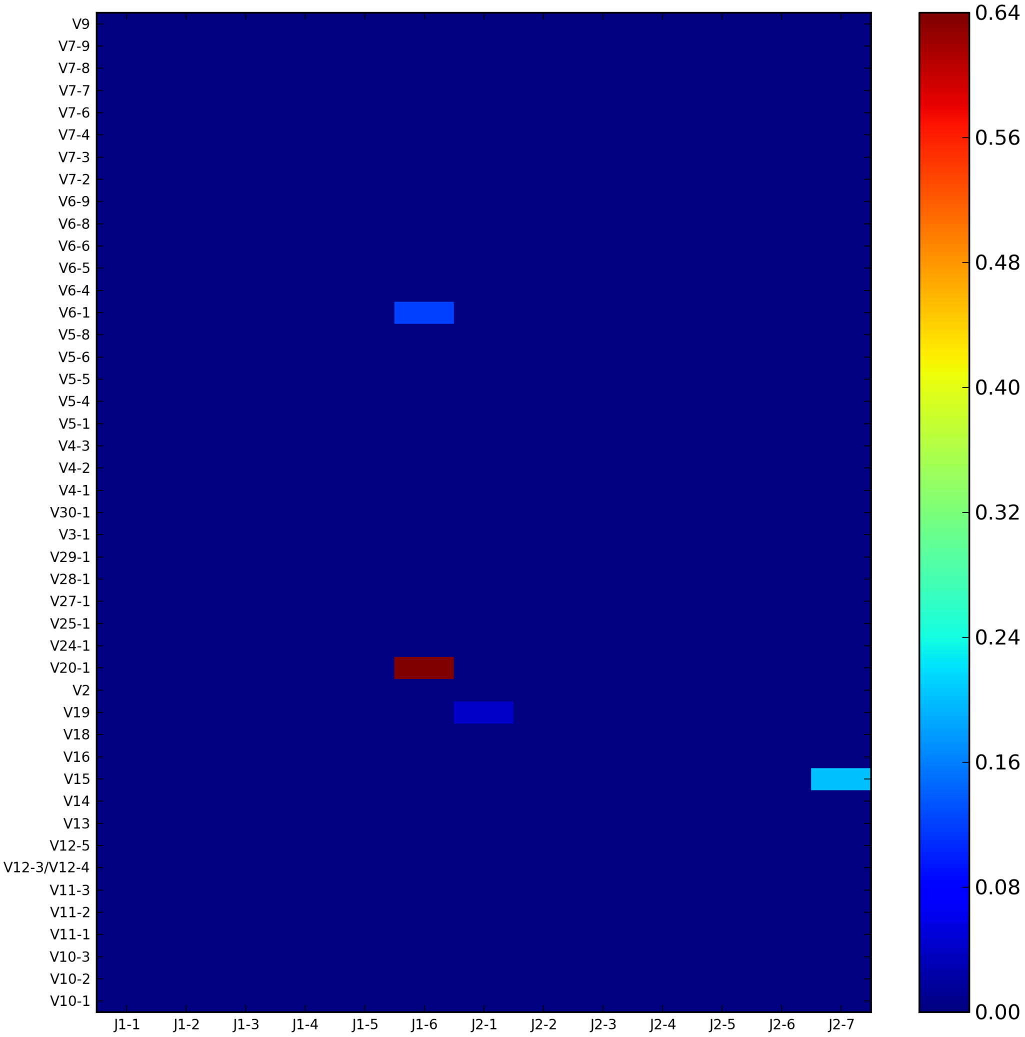


Figure 58. Pairings of VJ β chain segments for patient T4 at day + 33 post-transplant.

Figure 58 shows that the combination of V20-1 and J1-6 is dominant at +33 days post-transplant. The representation of all other gene segments is low in comparison. There were very few cells in this sample (0.4 x 10^9/L), and very few reads were obtained from this sample (Figure 57), which is visualised here by the high proportion of dark blue squares, representing zero of these combinations detected.

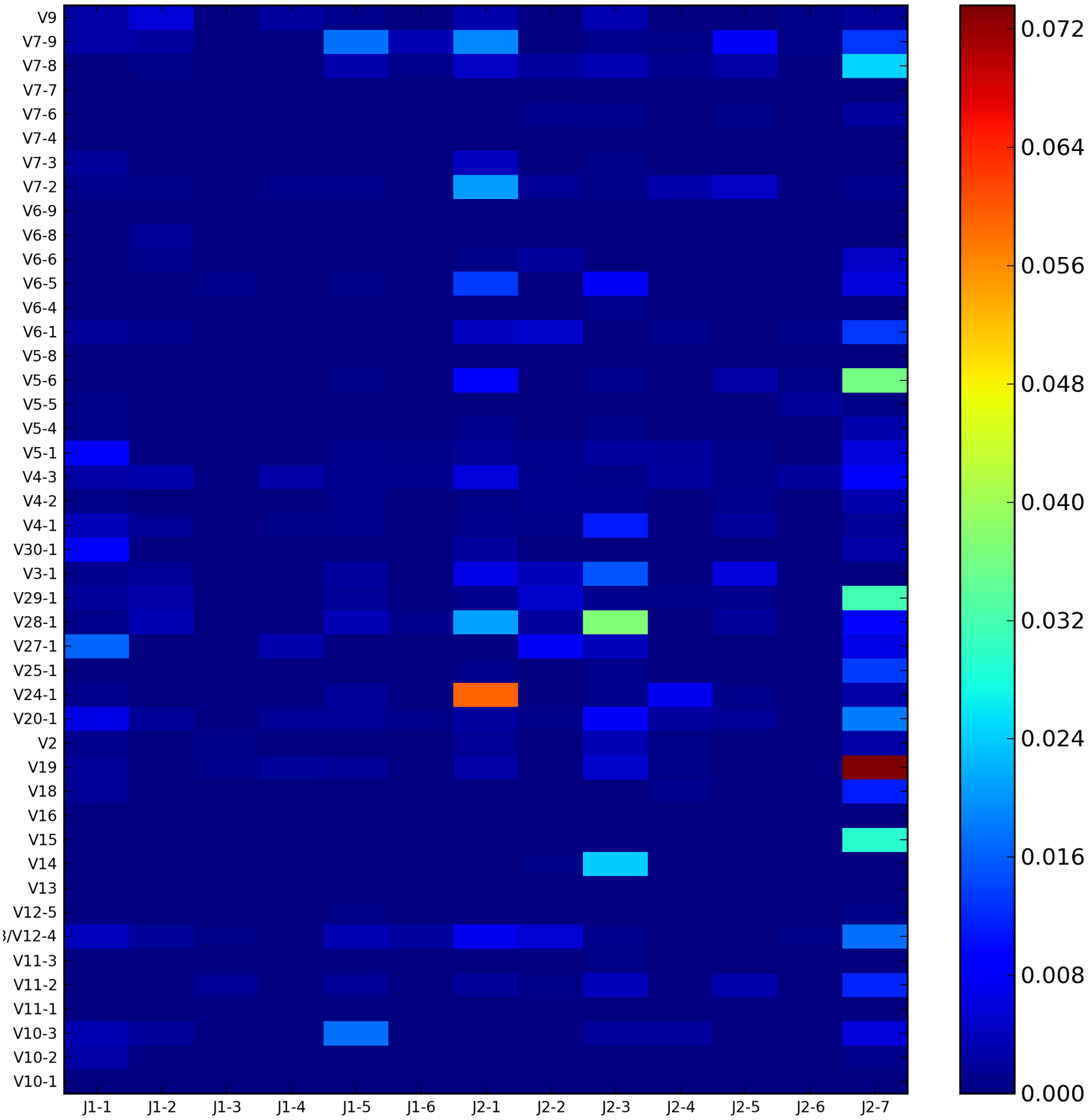


Figure 59. Pairings of VJ β chain segments for patient T4 at day + 61 post-transplant.

Figure 59 shows that the dominance of the V20-1 and J1-6 pairing, as seen at +33-day post-transplant, is no longer apparent. There is a more diverse repertoire at this time point. The pairing of V19 and J2-7 is dominant here, with V24-1 and J2-1 the second most dominant. However, this increase in diversity may be due to the higher number of cells in the sample and the higher number of reads obtained (Figure 57). Therefore Figure 58 and Figure 59 may not be directly comparable.

There is a dominant pairing at + 33 days that disappears by + 61 days, V20-1 and J1-6, which shows a greater amount of diversity in the V and J usage. However, a new dominant pairing is seen, V19 and J2-7. It is unclear as to whether these pairings are due to random expansion or are a true representation of diversity of the repertoire in response to immune challenges.

5.3.4 Clonotypes Identified Following HSCT

The number and frequency of clonotypes identified from patients who have received HSCT may help to demonstrate the recovery of the TCR repertoire following transplant. After sequencing, the data was analysed using the MiTCR software and the numbers of clonotypes identified for each of the α, β, γ and δ chains were plotted (Figure 60-Figure 66).

These figures show that the number of clonotypes increase after transplant for each chain, but the number did not stabilise across the time points shown. The three cord blood samples (T1.1, T2.1 and T3.1) demonstrated a difference in the number of clonotypes identified from each of the samples given.

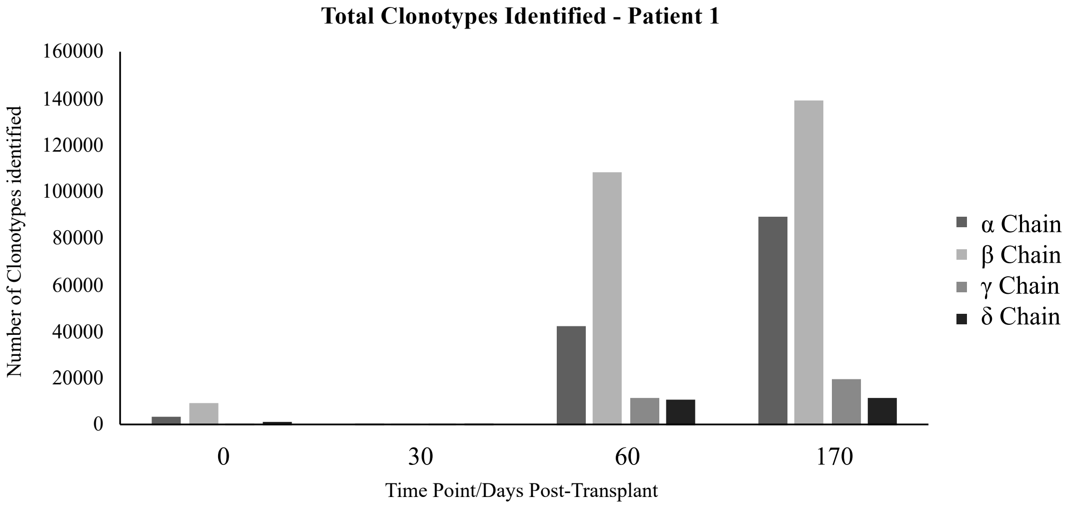


Figure 60. Total α, β, γ and δ clonotypes identified from Patient 1.

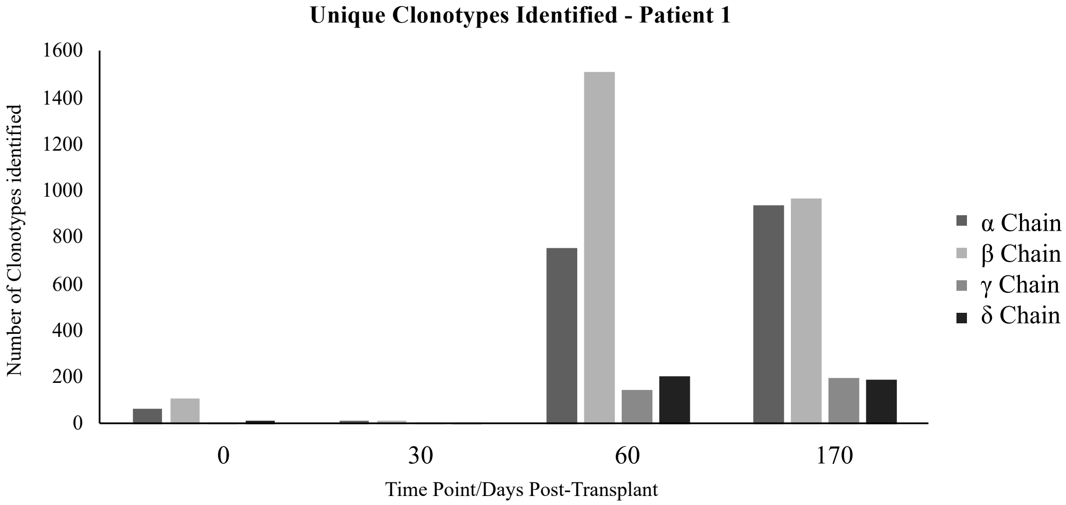


Figure 61. Number of unique α, β, γ and δ clonotypes identified from Patient 1.

Patient samples were sequenced on the MiSeq and MiTCR was used to identify clonotypes. Day + 0 samples represent the cord blood used to provide the transplant.

For Patient 1 (T1), shown in Figure 60 and Figure 61, the + 0 day cord blood sample does not contain many clonotypes from any of the four chains, +30 days after transplant shows very few. The repertoire appears to recover from + 60 days and decrease slightly at + 170 days. The peak in β chain total clonotypes may represent PCR bias or clonal expansion. However, there is also a peak in β chain unique clonotypes, suggesting a more even proliferation of cells.

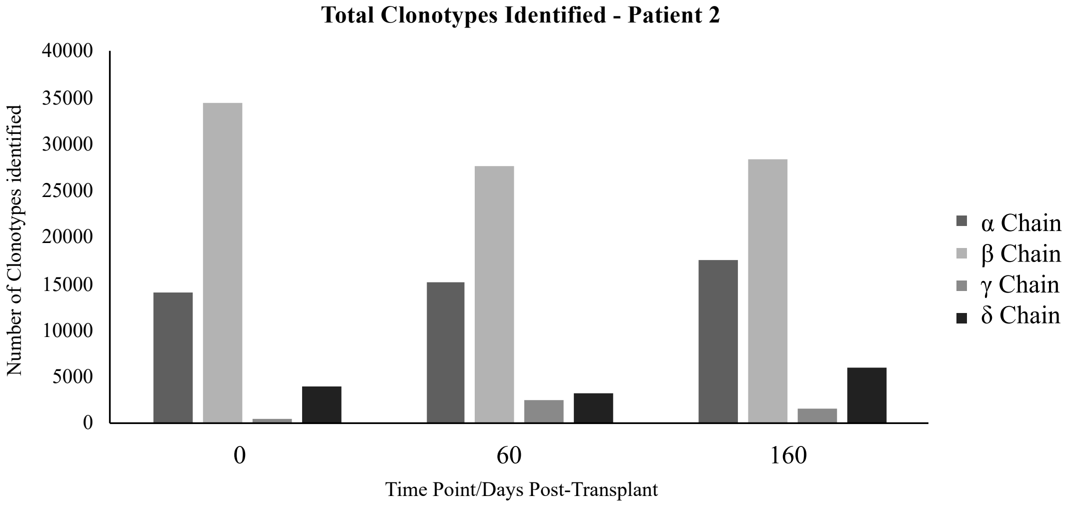


Figure 62. Total α, β, γ and δ clonotypes identified from Patient 2.

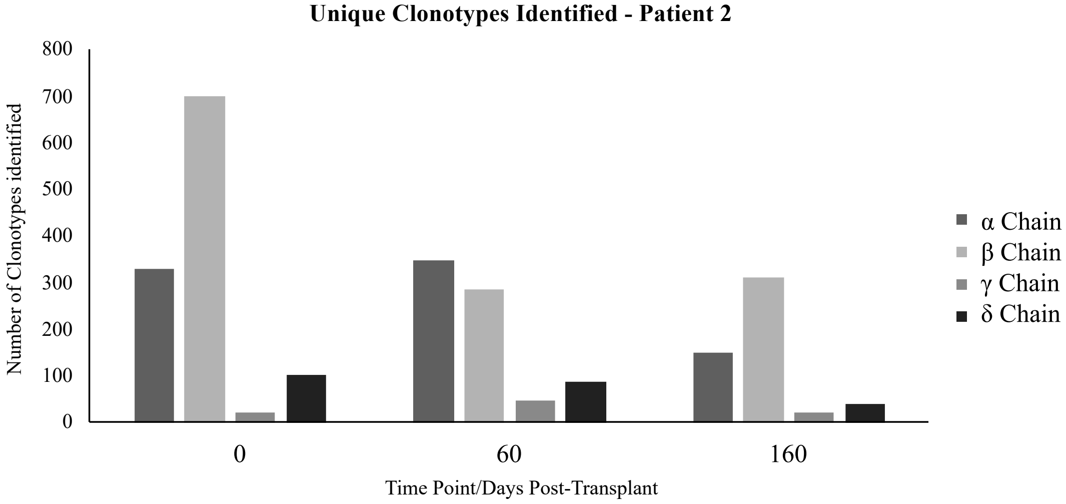


Figure 63. Number of unique α, β, γ and δ clonotypes identified from Patient 2.

For Patient 2, shown in Figure 62 and Figure 63, the + 0 day cord blood sample contained more identified clonotypes than the equivalent sample for patient T1. By +

60 days, there are lots of clonotypes identified, more α and β clonotypes than γ and δ. However, the count of clonotypes from each chain drops at + 160 days.

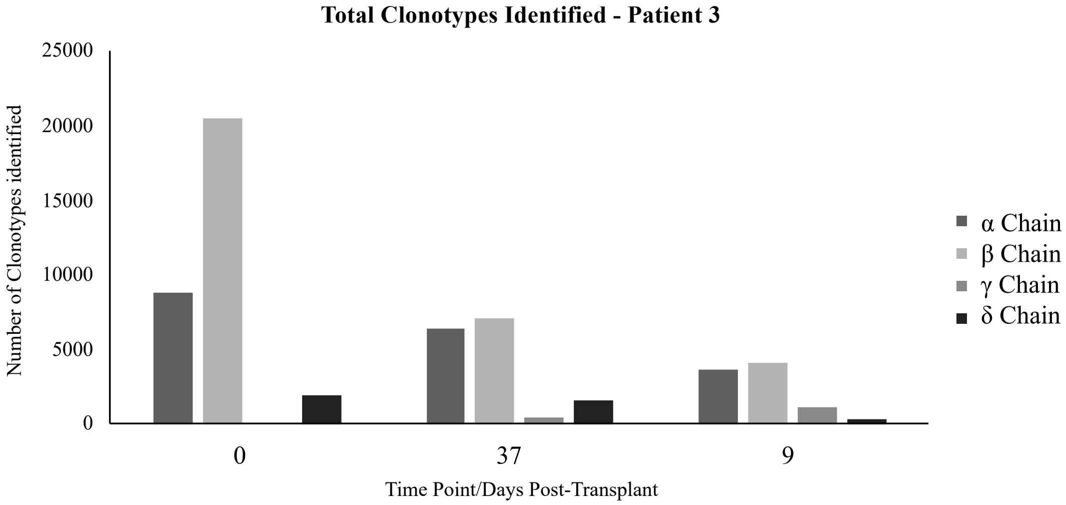


Figure 64. Total α, β, γ and δ clonotypes identified from Patient 3.

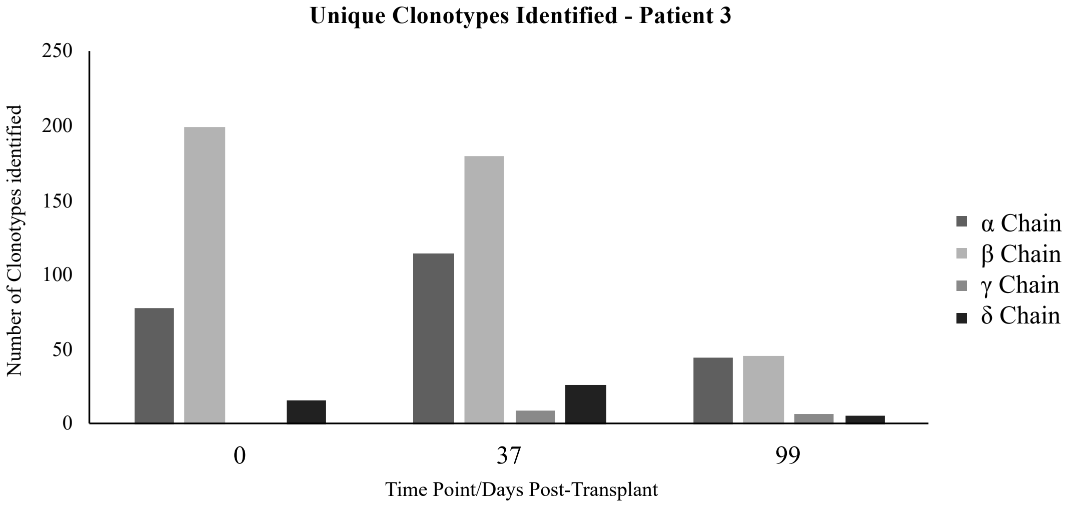


Figure 65. Number of unique α, β, γ and δ clonotypes identified from Patient 3.

For Patient 3 (T3), the + 0 day cord blood sample contains more clonotypes than the transplant received by patient T1, but fewer than the transplant received by patient T2.

At + 30 days post-transplant, there were significant numbers of clonotypes identified, but this number appears to drop by + 90 days post-transplant.

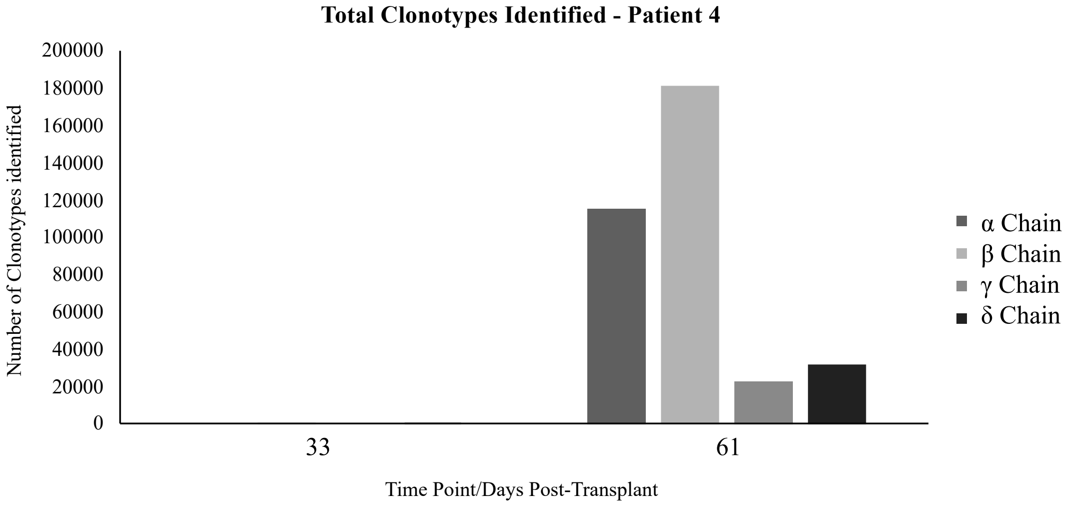


Figure 66. Total α, β, γ and δ clonotypes identified from Patient 4.

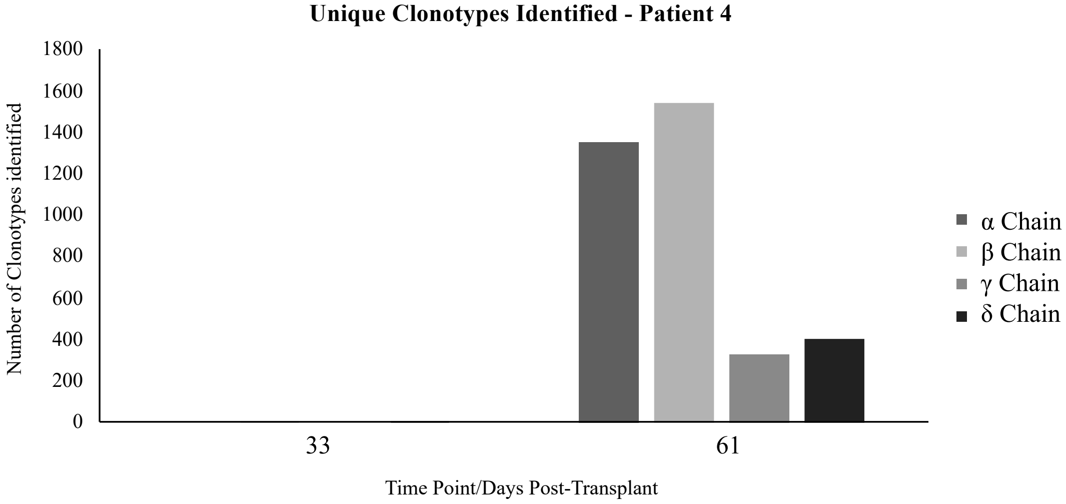


Figure 67. Numbers of unique α, β, γ and δ clonotypes identified from Patient 4.

Patient 4 (T4), shown in Figure 66 and Figure 67, received a haploidentical αβ+/CD19+ depleted transplant, rather than a cord blood transplant. Therefore, the original transplant likely contained a high proportion of γδ+ T-cells from the donor. Figure 66 shows a higher proportion of both γ and δ clonotypes identified at day + 61 post-transplant than in the other patients. This may reflect the proliferation of γδ+ T-cells from the transplant over the recovery period. There may be an additional protective benefit to this. However, it may also be an artifact due to the uneven acquisition of reads for this sample (Figure 57).

5.3.5 Diversity of Patient Repertoires Following HSCT

Diversity measures can also be used to monitor patient outcomes. The Gini index was used to investigate diversity of patient samples following the recovery period after HSCT. The figure below (Figure 68) shows how diversity changes over time for the α, β, γ and δ chain repertoires.

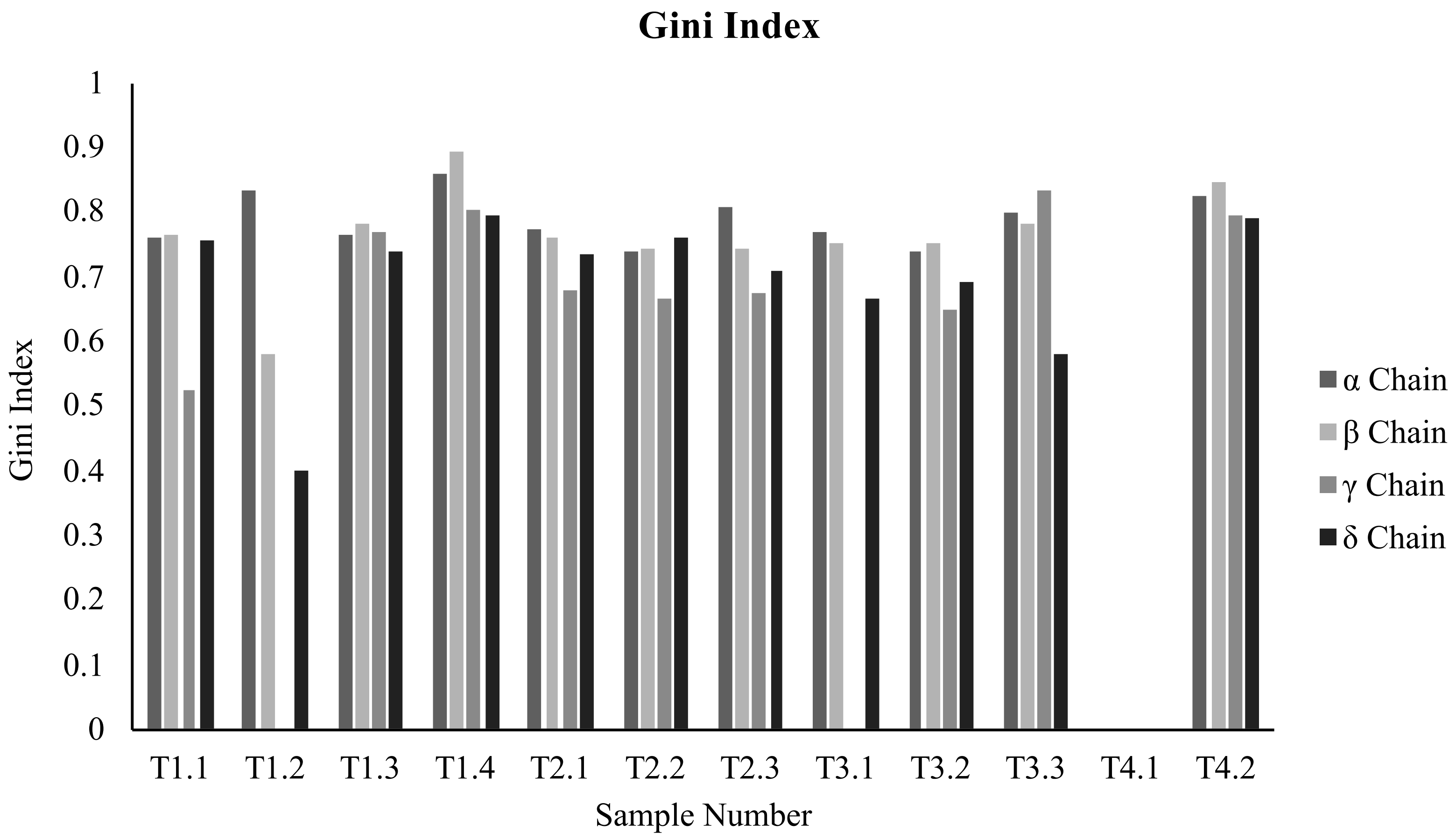


Figure 68. Gini index for patients T1 to T4, for each time point sequenced. Missing values indicate there were not enough clonotypes to calculate the Gini index. Statistical analysis was not performed as patients are not directly comparable.

The Gini index values are shown in Figure 68. They demonstrate that the diversity of the TCR repertoire at each time point is not strictly mirrored across the α, β, γ and δ chains. The diversity of both T2 and T3 increases slightly post-transplant, as the value of the Gini index decreases. At 6 months post-transplant, the diversity of the repertoire still seems to be dynamic, possibly decreasing, despite an increase in overall cell count.

Gini Index	T1.1	T1.2
α chain	N/A	0.822
β chain	N/A	0.846
γ chain	N/A	0.793
δ chain	N/A	0.792

Table 44. Diversity of patient T4. The Gini Index was calculated from sequencing data analysed using MiTCR for the two time points available, +33 days and + 61 days, from the patient who received the haploidentical αβ+/CD19+ depleted transplant.

The fourth patient, T4, received a haploidentical αβ+/CD19+ depleted transplant and therefore was not directly compared to the other three. In Table 44 the Gini Index is shown for the + 33 and + 61-day time post-transplant time points. For the first time point, the diversity is not available across all chains, as no clonotypes were identified from that sample. However, by day + 61, the values of diversity for all four TCR chains increased. The β chain showed the lowest diversity, followed by the α, with the γδ slightly higher.

5.3.6 Cord Blood Diversity

Cord blood diversity was also demonstrated across this data set, as each time point at +0 days post-transplant represents the cord blood sample itself, as patient cell counts were too low to sequence.

TCR Chain	**Shannon Entropy**		**Gini Index**	
	Healthy Donors	**Cord Blood**	**Healthy Donors**	**Cord Blood**
α	7.203	3.511	0.383	0.768
β	7.786	4.351	0.383	0.756
γ	4.042	1.706	0.435	0.400
δ	5.967	2.293	0.378	0.719

Table 45. Shannon Entropy and Gini Index for healthy donors and cord blood samples. The mean value of diversity was calculated from eight healthy donor data sets and three cord blood sample data sets for each α, β, γ and δ chain.

The average values of Shannon entropy and Gini Index are shown in Table 45, Figure 69 and Figure 70. Values for Shannon entropy are higher in all chains for the healthy donors and values for Gini index are consistently lower in the cord blood samples than in the healthy donors in all four TCR chains. This suggests that the TCR repertoire is more diverse in the periphery of a healthy adult than in cord blood. However, this does not agree with currently available literature and is not expected due to the high level of naïve T-cells in cord blood.

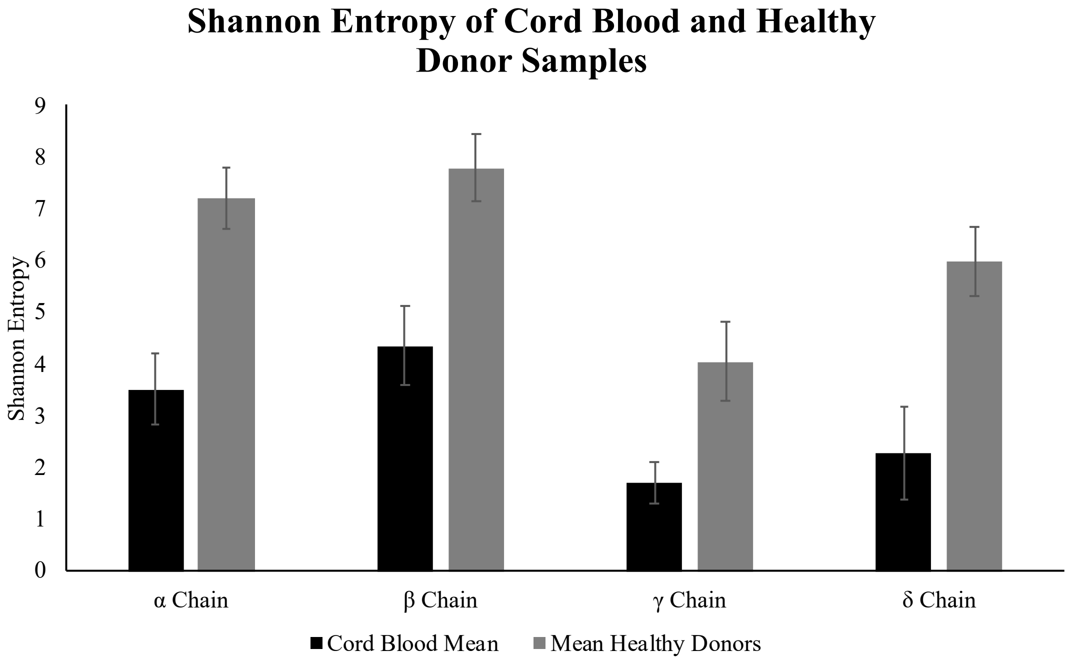


Figure 69 Shannon entropy of cord blood and healthy donor samples, calculated from the mean values with standard deviation. The cord blood diversity was significantly lower than the healthy donors in the α, β, γ and δ chains ($p < 0.05$; unpaired t-test).

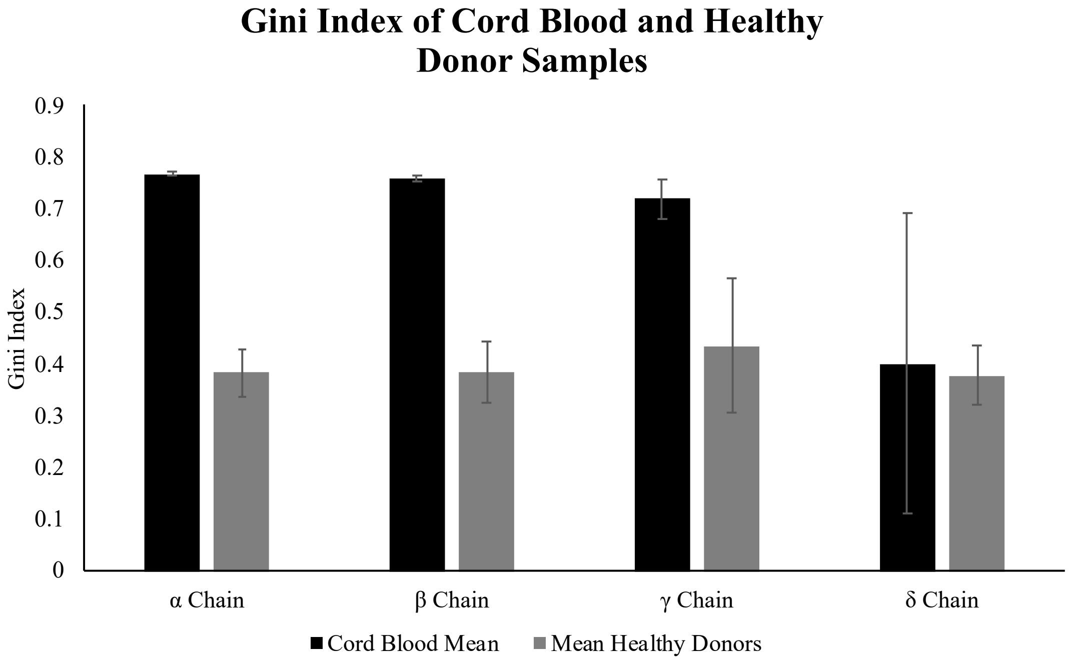


Figure 70 Gini index of cord blood and healthy donor samples, calculated from the mean values with standard deviation. The cord blood diversity was significantly lower than the healthy donors in the α, β and γ chains ($p < 0.05$; unpaired t-test), but not the δ chain.

5.3.7 Clinical Features of Patient 1

Sample Number	Timepoint	Cell Counts/10^9/L							
		CD3	CD4	Naïve CD4	CD4 Memory	CD8	Naïve CD8	CD8 Memory & Effector	Lymphocyte count
T1.1	Day 0	ND	ND	ND	ND	ND	ND	ND	0.04
T1.2	Day +30	0.67	0.56	0.19	0.20	0.15	0.03	0.13	2.93
T1.3	Day +60	0.60	0.39	0.04	0.20	0.24	0.03	0.13	1.02
N/A	Day 104	2.46	1.05	0.01	0.09	1.42	0.21	1.04	3.08
T1.4	Day +170	1.82	0.81	ND	ND	1.01	ND	ND	2.46

Table 46. Cell counts for Patient 1 for time points following CBT. Patient 1 received their HSCT on the 22nd August 2013. They received pre-transplant conditioning including busulphan, cyclophosphamide and melphalan. Note that if the total lymphocyte count is <0.1 it is not possible to run lymphocyte subsets as the small sample volume (100ul) means that the results will not be representative of the proportions of cells seen.

The total lymphocyte count shown in Table 46 show that cell numbers increase rapidly in the first 30 days, from 0.04 x 10^9/L; they then decrease at day +60 post-transplant before increasing again. The same pattern is seen in total T-cells (CD3+ cells) and CD4+ T-cells, but not CD8+ T-cells. However, there are no samples shown here to represent the lymphocyte counts of a healthy age-matched control. In addition, there are not enough time points to show the long-term recovery of circulating T-cells.

Patient 1 received cyclosporin and mycophenylate mofetil, and corticosteroids from day +31 to around day +100. Therefore, the decrease in cell counts at day +60 in Table 46 may be due to the immune suppressing effects of the corticosteroids.

Patient 1 tested positive for CMV at both 30 days and 104 days after the transplant. It was possible that diversity seen in this patient's samples was driven by clonal proliferation.

Read Count	Percentage Representation	CDR3 nucleotide sequence
18572	20.81%	TGCGGCACAGTTACCACTTCTGGTTCTGCA
4551	5.10%	TGTGCCGTGAATTTCTATAACCAGGGAGG
2470	2.77%	TGTGCCTGGAGTTCCGGGCCCTCGAACAC
2119	2.37%	TGCGGCACAGTAGCCCAGGCAGGAACTGC
1549	1.74%	TGTGCTCTTGGGGAACTCTGGGCTGGTGGT

Table 47. Most common α chain clonotypes identified using MiTCR for patient T1 at +170 days post-transplant.

Read Count	Percentage Representation	CDR3 nucleotide sequence
31970	22.94%	TGCGCCAGCAGCCGCGTGGGCTCGGAAGC
8424	6.04%	TGTGCCAGCAGCTCAACACCGGGTACGCA
6299	4.52%	TGCGCCAGCAGCTCACAGGGGACCGGAAC
5152	3.70%	TGTGCCAGCAGCGTAGGAACGGGGGACTA
5043	3.62%	TGCGCCAGCAGCCGGGACAGCTCCTACAA

Table 48. Most common β chain clonotypes identified using MiTCR for patient T1 at +170 days post-transplant.

Table 47 and Table 48 show the 5 most common clonotypes identified for the α and β chains of patient T1 at +170 days' post-transplant. The most frequently identified clonotype for the α and β chains represented 20.81% and 22.94% of all clonotypes identified respectively. It is highly likely that these two sequences pair together to create the full clonotype.

5.3.8 Clinical Features of Patient 2

Sample Number	Timepoint	Cell Counts/10^9/L							
		CD3	CD4	Naïve CD4	CD4 Memory	CD8	Naïve CD8	CD8 Memory & Effector	Lymphocyte count
T2.1	Day 0	ND	ND	ND	ND	ND	ND	ND	0.08
T2.2	Day +30	0.20	0.20	0.04	ND	0.08	0.04	ND	0.40
T2.3	Day +60	0.65	0.35	0.08	0.27	0.28	0.21	0.07	1.01

Table 49 Cell counts for Patient 2 for time points following CBT. Patient 2 received their HSCT on the 11th October 2013. They received pre-transplant conditioning including busulphan, cyclophosphamide and melphalan. Note that if the total lymphocyte count is <0.1 it is not possible to run lymphocyte subsets as the small sample volume (100ul) means that the results will not be representative of the proportions of cells seen.

Patient 2's lymphocyte count was the lowest of all the patients considered here, as seen in Table 49. This patient received cyclosporin and mycophenylate mofetil, and corticosteroids from day +5 to around day +120. The low T-cell counts in Table 49 may be due immune suppressing effects of these medications. The changes in diversity of Patient 2 may also be due to viraemia. This patient tested positive for adenovirus at day +0, day +60 post-transplant, and positive for BK virus at day +0, day +60 and day +160 post-transplant.

5.3.9 Clinical Features of Patient 3

Sample Number	Timepoint	Cell Counts/10^9/L							
		CD3	CD4	Naïve CD4	CD4 Memory	CD8	Naïve CD8	CD8 Memory & Effector	Lymphocyte count
T3.1	Day 0	ND	ND	ND	ND	ND	ND	ND	0.02
T3.2	Day +37	1.30	1.22	ND	ND	0.08	ND	ND	1.67
T3.3	Day +99	0.64	0.59	0.20	0.36	0.05	ND	ND	2.27

Table 50 Cell counts for Patient 3 for time points following CBT. Patient 3 received their HSCT on the 7th November 2013. They received pre-transplant conditioning including treosulphan, fludarabine and thiotepa. Note that if the total lymphocyte count is <0.1 it is not possible to run lymphocyte subsets as the small sample volume (100ul) means that the results will not be representative of the proportions of cells seen.

The total lymphocyte count of Patient 3, shown in Table 50, increases over time from day +0 to day +99 post-transplant. However, the T-cell count (CD3+ cells), both CD4+ and CD8+, decreases. The patient was given ciclosporin and mycophenylate mofetil, with no corticosteroids. These medications may have contributed to selective suppression of T-cell replication. No positive tests were recorded for this patient for any viral infections throughout the time period observed.

5.3.10 Clinical Features of Patient 4

Sample Number	Timepoint	Cell Counts/10^9/L							
		CD3	CD4	Naïve CD4	CD4 Memory	CD8	Naïve CD8	CD8 Memory & Effector	Lymphocyte count
N/A	Day 0	ND	ND	ND	ND	ND	ND	ND	ND
T4.1	Day +33	0.20	0.20	0.04	ND	0.08	0.04	ND	0.40
T4.2	Day +61	0.41	0.18	0.00	0.05	0.06	0.00	0.08	1.22
N/A	Day +77	0.37	0.40	0.00	0.07	0.16	0.01	0.11	1.17

Table 51 Cell counts for Patient 4 for time points following αβ+ TCD transplant. Patient 4 received their haploidentical αβ+ T-cell depleted transplant on the 22nd August 2013. They received pre-transplant conditioning including fludarabine, treosulphan, thiotepa and anti-thymocyte globulin (ATG).

Patient 4 received a haploidentical αβ+ TCD transplant, so the transplant they received contained γδ+ T-cells. However, Table 51 shows that the CD3+ T-cell counts were low at all time points, suggesting that the γδ+ T-cells did not proliferate following transplant.

The patient received ciclosporin but no corticosteroids. In addition, they received an infusion of adenoviral cytotoxic lymphocytes on day +63 post-transplant due to persistent adenovirus viraemia. Although no samples were sequenced beyond this time point, this may drive a decrease in TCR diversity due to monoclonal anti-adenovirus expansion. This infusion of lymphocytes may have contributed to the increase in the cell counts observed in Table 51.

The number of T-cells and diversity increased between the two time points for patient 4, but the patient was positive for adenovirus infection for all time points observed in Table 51. This may suggest that the increase in diversity was not sufficient to provide protection.

5.3.11 Isopentenyl Pyrophosphate and Artificial Antigen Presenting Cell Expansion of the γδ+ T-cell Receptor Repertoire

As well as investigating the reconstitution of the TCR repertoire in conditions such as HSCT, high-throughput sequencing of TCR repertoires can be used for multiple other applications. To demonstrate the utility and future potential applications of this capture technology, γδ+ T-cells from healthy individuals were expanded using IPP stimulation or artificial antigen presenting cells (aAPCs), and then sequenced. Table 52 lists the samples that are discussed in this section. Heat maps were generated to show the V and J segment usage both before and after expansion of healthy donor PBMCs using IPP or aAPC (Figure 71, Figure 72, Figure 73 and Figure 74).

Sample Reference	Donor Number	Starting Material
G1	25	PBMC
G2	25	IPP expansion
G3	26	PBMC
G4	26	IPP expansion
G5	26	Sorted Vδ1
G6	26	Sorted DN
G7	28	PBMC
G8	28	Sorted Vδ1
G9	28	Sorted Vδ2
G10	29	PBMC
G11	29	Sorted DN

Table 52. Samples sequenced to demonstrate another use of the capture kit and high-throughput sequencing of TCR repertoires.

PBMCs from four healthy donors (donor numbers 25, 26, 28 and 29) were stimulated with IPP or an aAPC, as shown in Table 52. The expanded T-cell populations were then sequenced. The aAPC expanded T-cell populations were first sorted using fluorescence activated cells sorting (FACS) and then sequenced.

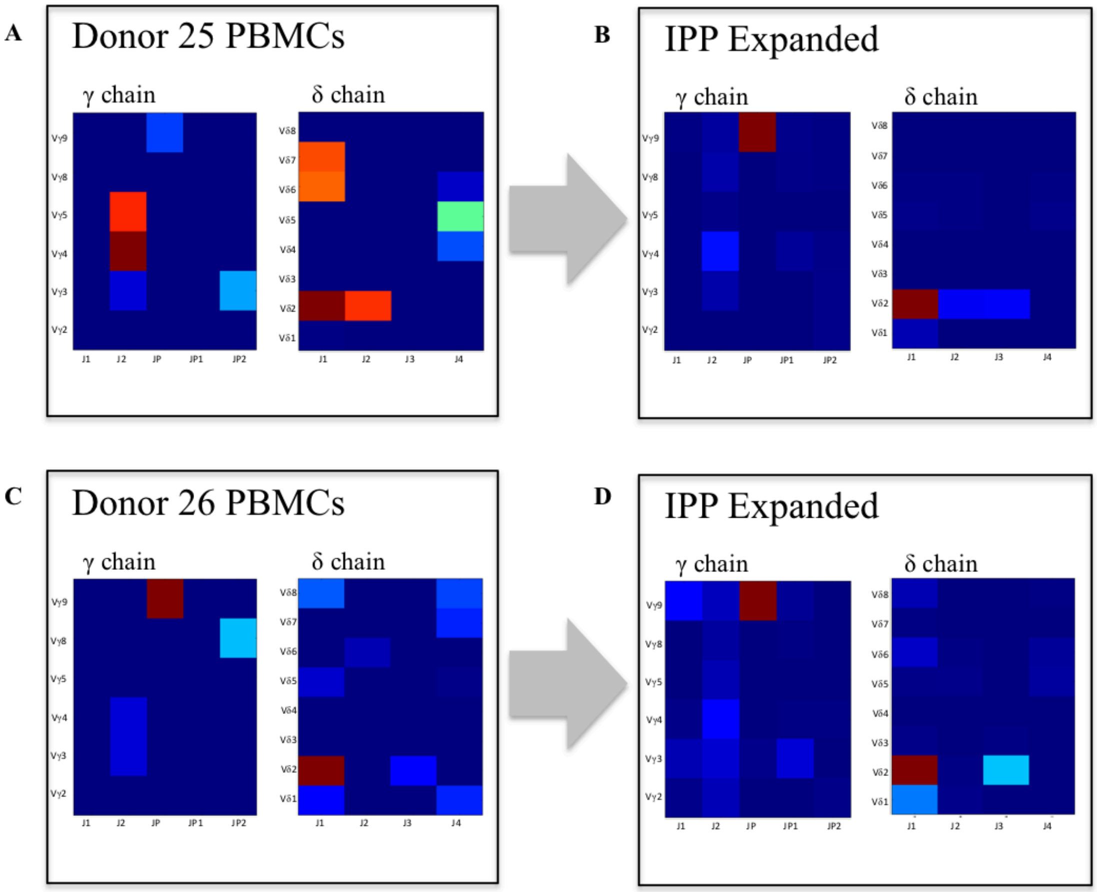


Figure 71. Decrease in diversity of V and J segment usage following IPP expansion of γδ+ T-cells. PBMCs from two healthy donors, numbers 25 and 26, were treated with IPP and samples pre and post expansion were sequenced.

The heat maps in Figure 71 show V and J usage of the γ and δ chain TCRs in PBMCs from two healthy donors, numbers 25 and 26, and the V and J usage of the IPP expanded populations. Relative abundance of parings is shown from red (high) to blue (low).

The effect of IPP is more apparent for donor 25, which demonstrates more diversity pre-expansion that donor 26, which is already dominated by Vγ9 and Vδ2. In both post-expansion samples, the dominant segments are Vγ9 and Vδ2, as expected from IPP.

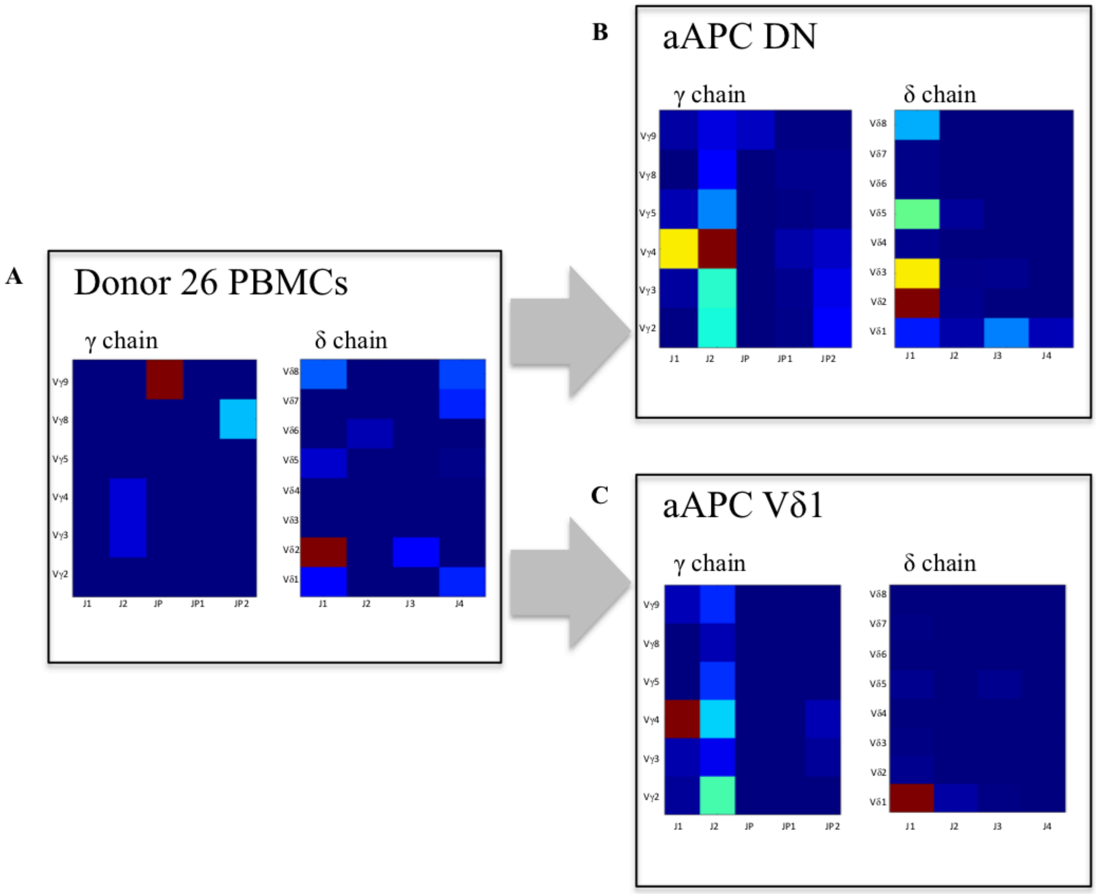


Figure 72. Change in diversity of V and J segment usage following stimulation of γδ+ T-cells with an aAPC.

PBMCs from healthy donor number 26, were also treated with aAPC. Post-expansion samples were then sorted using FACS into two populations that were positive for Vδ1 and negative for both Vδ1 and Vδ2 (double negative, DN). Pre-expansion PBMCs and post expansion sorted populations were sequenced. These heat maps show V and J usage of the γ and δ chain TCRs in the expanded populations. Relative abundance of parings is shown from red (high) to blue (low). However, the sorted cells are not truly DN, as there is still representation of Vδ2 which may reflect lack of specificity of the antibody used in the sorting process, or a specificity to only Vδ2 when it is paired with Vγ9.

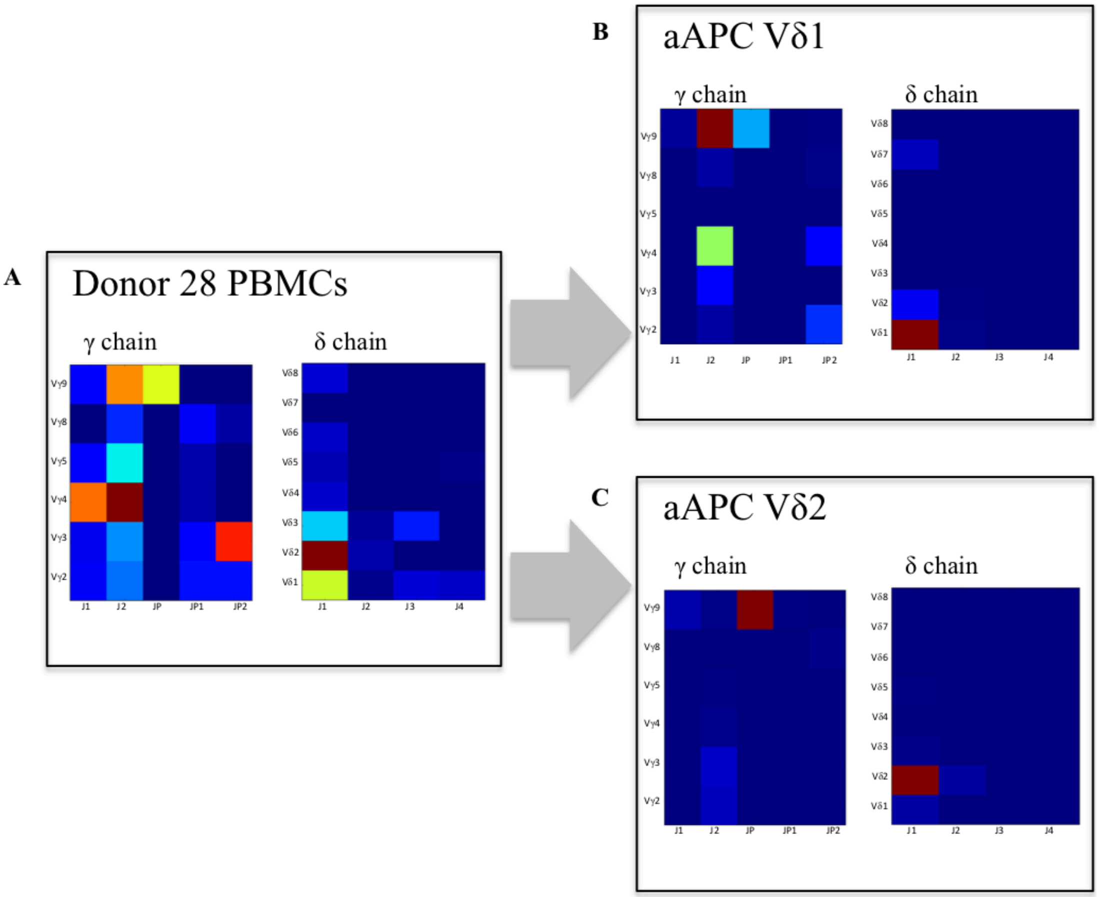


Figure 73. Decrease in diversity of V and J segment usage following stimulation of γδ+ T-cells with an aAPC.

PBMCs from healthy donor number 28 were treated with aAPC. Post expansion samples were then sorted using FACS into two populations, the Vδ1+ populations and the Vδ2+. Pre-expansion PBMCs and post expansion sorted populations were sequenced. Data was analysed using the Decombinator. These heat maps show V and J usage of the γ and δ chain TCRs in the expanded populations. Relative abundance of parings is shown from red (high) to blue (low). The Vδ1+ population demonstrates more diversity post expansion than the Vδ2+ population.

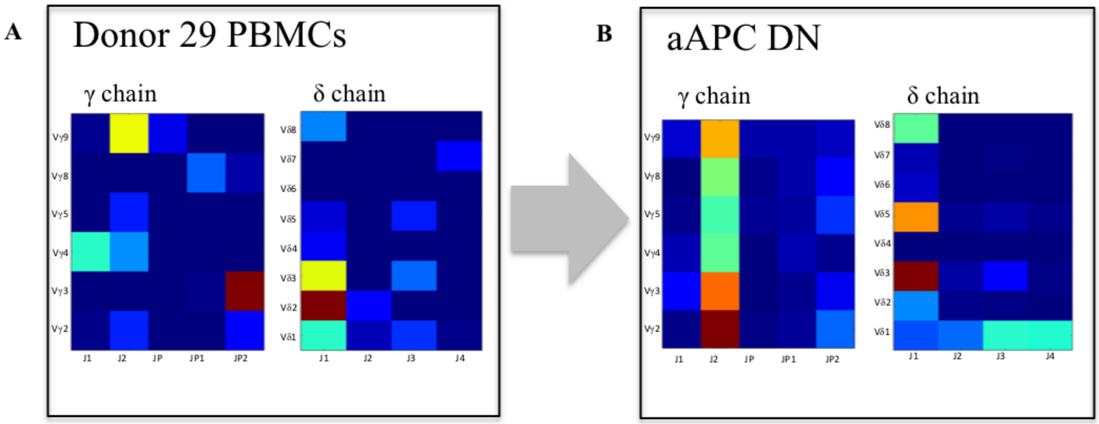


Figure 74. Decrease in diversity of V and J segment usage following stimulation of γδ+ T-cells with an aAPC.

PBMCs from healthy donor number 29 were treated with aAPC. Post expansion samples were then sorted using FACS for the Vδ1- and Vδ2- populations (DN). Pre-expansion PBMCs and post expansion sorted populations were sequenced. Data was analysed using the Decombinator. These heat maps show V and J usage of the γ and δ chain TCRs in the expanded populations. Relative abundance of parings is shown from red (high) to blue (low). The pre-expansion population shows notable diversity, but high frequencies of Vγ3 and Vδ2. The post expansion DN population shows higher diversity, but still with high frequencies of particular segments, in this case Vγ2 and Vδ3.

The heat maps of the PBMCs from healthy donors demonstrate that there is diversity in the γδ+ T cell population, which has not been fully characterised previously, and which varies from donor to donor. For example, donor 29 has a highly heterogeneous γδ+ population (Figure 74), but donor 26 does not (Figure 71). This may mean that the post IPP or aAPC stimulation populations are more or less diverse depending on the diversity of the initial PBMC samples. Following IPP stimulation, the dominant gene segments in the post-expansion samples are Vγ9 and Vδ2 (Figure 71). This is as expected, as IPP has been previously demonstrated to expand this population (161).

5.4 Discussion

Despite the limitations of the capture technique and analysis pipelines highlighted in previous chapters, the data presented above suggests that this method could potentially be used to investigate immune reconstitution following HSCT in future. The reconstitution of the TCR repertoire following transplantation has been described, and the ability to show changes in repertoire has been demonstrated through the sequencing of different αβ+ and γδ+ T-cell populations. In the small dataset presented, we cannot draw significant conclusions for this cohort of patients however, it is possible to make inferences about the utility and potential benefits of immune repertoire sequencing.

The different number of reads obtained for different patient samples could be due to a number of reasons. Inaccurate measurements of concentration given by the Qubit may have meant that samples were unevenly pooled into the hybridisation. Therefore, there would be more T-cell and B-cell material from some samples than other, introducing a bias in sample representation post-capture. It is difficult to determine retrospectively whether this was the case, but in future it may be necessary to introduce a qPCR step before samples are pooled for capture, following the pre-capture amplification step.

As described previously, the quality of starting material and sequencing depth is important (116). All 12 samples comprising TCR14 were analyzed on the Bioanalyzer prior to library preparation. The quality of the initial starting RNA obtained from the clinical laboratory, according to the RIN score of the sample (not shown), does not correlate with the proportion of reads assigned to that sample during the run. The unevenness of representation of samples within run TCR14 may have introduced bias into the analysis of diversity as different sized FASTQ files result in different yields of TCR sequences identified.

Following TBI and HSCT, lymphocyte counts of all four patients were low with regards to paediatric reference values, but dates of birth of patients were not made available so it is not possible to directly compart (162). The lower the lymphocyte count, the fewer target transcripts available for capture. Sample amounts were standardised to 1μg of total input RNA, and 10ng of prepared library before being pooled prior to hybridisation. Therefore, if the time points sequenced here were low in T-cell and B-cell transcripts, the capture would introduce a bias in sample representation, as the few existing transcripts would have undergone multiple PCR replication cycles and this may introduce an artificial appearance of evenness across the repertoire, provided they were amplified at an equal rate (163).

At early time points post-transplant, the new T cells have yet to engraft and thymic output does not contribute significantly to the circulating repertoire until around six months after the transplant (164). Therefore, increases seen in diversity values across post-transplant time points are most likely driven by the expansion and more comprehensive representation of mature donor cells. A high level of diversity may suggest proliferation of all T cell-types, rather than the clonal expansion of a limited few.

Clonal expansions were observed across all patients. These expansions correlated with clinical events. For example, Patient T1 had high CMV viraemia between day 60 and day 170. Clonotype analysis from day 170 showed one clonotype which represented

~20% of all clonotypes detected, for both α and β chains. α chain clonal expansions correlated with β chain expansions within the same sample, so these are unlikely to be due to preferential PCR amplification. It is highly likely that the two CDR3 sequences that account for 20% of patient T1d α and β clonotypes represent the same TCR. Furthermore, one may speculate that expansion of this clonotype is likely due to the CMV viraemia at this point. This clonotype may therefore be a response to CMV. A database search of the β chain nucleotide/amino-acid sequence of this clonotype identified the same sequence in a Galaxy immunosequencing dataset, increasing the probability that this has arisen in response to a common CMV antigen (165). In recent advancements, single cell sequencing has been used to identify pairs of αβ and γδ chains (166).

Diversity may be an indicator of post-transplant recovery. All patient samples fell below the threshold of normal TCR repertoire diversity, based on the healthy control data reported in this study. However, this technique was unable to definitively describe the healthy repertoire, so more work should be done to characterise this. The heatmaps in Figure 58 and Figure 59 appear to show an increase in diversity between pre and post-transplant samples as the representation of V and J pairings changes. They represent the number of pairings identified within a single sample, rather than comparisons between samples, therefore statistical analysis is not applicable in this context. However, these pairings may be due to random expansion rather than a true representation of an increase in repertoire diversity. In addition, the hotspots that show high proportional representation of specific VJ pairings may be due to an immune response, for example to a pathogen, and subsequent clonal expansion, or they could be random non-significant events. In order to investigate this further, sequencing of more patient samples could be undertaken to determine if diversity or bias of VJ pairings is consistent across recovery. This may also be dependent on disease, treatment regimen and exposure to pathogens in the environment during convalescence.

Each of these transplants took place in 2013, and in the five years since transplant, each patient has survived and none of the patients has relapsed. Data from more timepoints and more patients is required to determine whether the diversity of the repertoire is a driver of recovery following HSCT. Additionally, longer term follow-up may be required for each patient, particularly in the first six to twelve months after thymic output begins to contribute more significantly to the TCR repertoire.

The sample size of patients and the number of time points here is too small to determine whether there is a significant difference between the cord blood transplant and haploidentical αβ+/CD19+ transplant recipients. However, by day 60, the diversity of the haploidentical transplant recipient (patient T4), had increased to a higher value than the values reached by other patients at any time. This may suggest no significant clinical disadvantage to the use of a haploidentical donor as a source of stem cells. This could potentially be beneficial to many patients without a matched sibling, as another close family donor would be more convenient than finding a MUD, a factor that is especially important for patients of particular ethnic groups that are underrepresented on current donor registers.

The value of the Gini index is not dependent on the size of the sample. Therefore, decrease in diversity is not due to the decrease in numbers of sequencing reads obtained.

This may support the idea that the low numbers of reads are actually due to a compromised TCR repertoire. The changes in diversity across samples reflect the incidences of clonal expansions, as clonal expansions represent a lower diversity sample. A more prominent factor in determining the diversity is the clinical course of the patient. All 3 cord blood recipients appear to have a contraction in repertoire diversity in all TCR chains, during which time clonal expansions arise.

It has been shown previously that thymic output is initially very low post transplant and can take up to 6 months to recover and for new CD31+ RTEs to begin to repopulate the periphery (164). Unfortunately, the thymic output data for these patients was not made available, and therefore it cannot be commented on whether clonal expansion, both homeostatic and in response to infection, is the driving influence in repertoire diversity in the early months. In future experiments, the thymic output should also be compared to cell counts and diversity, as well as monitoring the prevalence of TRECs to determine proliferation. This would provide more information and help to further characterise the nature of the repertoire.

The low diversity of the cord blood samples compared to diversity in healthy adults was not expected and is concerning. Cord blood is generally thought to be highly diverse, due to the high proportion of naïve T-cells and low levels of clonal expansion. Two measures of diversity were used to compare, the Gini index and Shannon entropy, and both measures were in agreement. However, there are not enough samples in this study to conclude as only three cord blood samples were sequenced. These may have been samples of unusually low diversity, or there may be an issue with consistency of the sequencing. In addition, this occurrence may have been introduced due to low cell counts in the cord blood samples that the RNA was extracted from. Further investigation into this will be of benefit to the field as cord blood diversity is well established and could be used in future as a metric to validate methods.

This data set has provided interesting preliminary results. However, the sample size in this study was small and therefore no definitive conclusions can be drawn at this stage without more data and statistical analysis. The individuals cannot be statistically analysed due to the number of replicates, but we can observe and compare patterns across these results. As the patients included here were all paediatric, the analysis may also benefit from sequencing age-matched controls. Furthermore, all sequence data has been obtained from RNA isolated from total PBMCs. More detailed information could be obtained about T-cell reconstitution across different cell populations if naïve, memory, CD4+ and CD8+ T-cells were isolated before sequencing. This might not be financially feasible to perform routinely in the clinical department, as this would quadruple the number of samples per patient. However, an initial study of isolated cell compartments might elucidate whether certain populations yield more useful information than others. For example, it may be more important clinically and of more direct benefit to the patient to monitor the diversity of the naïve T-cell pool, and at different time points, rather than including repertoire analysis with all routine blood draws.

In addition to investigating individual T-cell compartments, the conditioning and treatment regimens of patients should be considered. Each of these patients received drugs that may have had immunosuppressive effects and therefore may have impacted on the diversity of the repertoire. Since this work was carried out, advancements have been made in this field and this context should also be considered. However, the major limitation of this data is that not enough patients or time points were studied. Therefore, future studies should be carried out in a larger cohort, with the specific aim of investigating TCR diversity, with appropriate age-matched controls. This can be expanded to look at HSCT recipients with other conditions other than AML, and to look more comprehensively at the different in immune reconstitution between cord blood recipient and TCD recipients. Further studies should also be carried out to investigate the effects of different treatments and pre-transplant conditioning on TCR reconstitution and diversity.

The additional information presented here regarding the repertoire and expansion of γδ+ T-cells in response to IPP and aAPC demonstrates that this technique has multiple research applications, but again that more data is needed. It also provides an example of the use of investigating all TCR chains simultaneously, as it supported data demonstrating the capability of γδ+ T-cells of antibody dependent and antibody independent cytotoxicity. Clinical applications may include development of γδ+ T-cell based immunotherapies for cancers such as neuroblastoma[1].

The DN populations still demonstrate a high abundance of Vδ2 sequences, despite being sorted through FACS to exclude Vγ9Vδ2+ cells (Figure 72 and Figure 73). This may suggest that there was a high concentration of Vδ2+ cells in the samples or that the sorting process did not yield a completely pure sample. It could also suggest that the antibody used to label Vδ2+ cells is only specific for the Vγ9 Vδ2 combination and therefore that where cells expressed Vδ2 with a different Vγ segment they were sorted as the negative population. This indicates the need for further research and development into antibodies for γδ+ T-cells.

Chapter 6

Discussion & Conclusion

6 Discussion and Conclusion

6.1 Discussion

The preliminary aims of this project have been met; information about gene expression, clonal frequency and clone size for each of the α, β, γ and δ chains of the TCR has been obtained simultaneously. Significant doubts have been raised regarding the practicality of this technique, as it is time consuming and does not produce consistent results.

However, there is much more information that can be obtained from these datasets in future, including all of the equivalent Ig data from these samples. The use of the capture kit and high-throughput sequencing has contributed to the knowledge base for immune repertoire. Targeted capture has since been investigated (167). To date there have been no studies that have included all α, β, γ and δ chains of the TCR and κ, heavy and light chains of the Ig and therefore this method remains novel.

6.1.1 Capture Kit Deficiencies

This project demonstrated the success of a novel target enrichment method for high-throughput sequencing of TCR and Ig transcripts. The proportion of on-target sequences is acceptable but could be improved upon in future iterations of the capture technique. Most of the off-target sequences were from transcripts that are very highly expressed, such as mitochondrial sequences and zinc finger domains. In future, ways of depleting these off-target sequences before the capture stage could be investigated, such as adding blocking sequences to the hybridisation reaction, but this should not necessarily be a deterrent from using the kit in the meanwhile.

However, it was surprising to find that even when pan T-cells were isolated from PBMCs, alignment of all reads obtained to TCR sequences did not increase by much, while the number of off-target sequences dramatically increased. This suggested that the process of isolating T-cells had served only to deplete the data of B-cell sequences and that a limit for the proportion of TCR sequences acquired had been reached. This is potentially an issue with saturation of TCR sequences within the kit, leading to non-specific binding of off-target sequences during hybridisation. Saturation may pose other issues such as limiting the representation of some sequences due to a lack of binding sites during hybridisation. At this stage it is not recommended that this technique be used in a clinical setting.

6.1.2 Advances in High-Throughput Sequencing

The number of unique clonotypes detected differs according to a number of factors. These factors include the sequencing platform and analysis technique. Samples that were sequenced using the Ion Torrent PGM appear to contain far more CDR3 sequences that are only represented once, accounting for up to half of all clonotypes detected, whereas this figure was lower for the Illumina MiSeq. This could imply that the Ion Torrent PGM has a much higher error rate and that many sequences that appear to be only represented once within a sample, and the subsequent higher diversity value are actually the result of sequencing errors. It is recommended that all samples be sequenced on the Illumina MiSeq both for consistency as it seems to introduce fewer artificial CDR3 sequences. Similar issues are seen with the computational analysis, where MiTCR appears to be more sensitive.

Although consistency was mentioned above, it is important to acknowledge that sequencing technologies and analysis platforms are relatively new and are being constantly refined. Therefore, a set protocol may only be relevant for a few years or months at a time and so new technologies should be introduced when appropriate. These advances will continue to improve the accuracy of data and the understanding of immune repertoires.

6.1.3 Analysis of Data

Once aligned to specific TCR V and J genes using Decombinator or MiTCR, similar patterns of biased TRBV gene usage were seen as have been established through decades of spectratyping T-cell samples and are being reinforced by other reports on immune repertoire sequencing (8, 87, 88, 90). This can confirm that the selection of at least some V and J segments during recombination is not random and neither are the combinations of these genes spliced together in TCRs. In future it will be interesting to investigate factors that may influence VJ pairing, and whether or not it is connected to the immune response of different individuals to common antigens. The capture technique allows for simultaneous investigation of all VJ pairings across α, β, γ and δ chains, in contrast to many experiments that focus only on the β chain, taking into account perturbations across the entire repertoire.

Although mathematical analysis of the diversity of TCR repertoires in still in the early stages, diversity measures have enabled the comparison of data and describe relative variation between individuals. There are many concerns that must be taken into account such as the need to standardise the size of samples to perform direct comparisons between data, and the need for larger sample sizes. However, so far it has strengthened the investigation and is therefore a strong starting point for further work in this area.

The work described here on γδ+ T-cells has provided a valuable insight into this lesser understood population. As seen with αβ+ populations, gene usage and VJ pairing are not random. The results have also highlighted the diversity present in γδ+ T-cell populations. This may be an artifact of the method or analysis, and γδ+ T-cell samples of more individuals should be sequenced in order to confirm whether this is a genuine phenomenon. γδ TCR diversity may have implications on cancer research and immunotherapy.

Confidence in this method has been achieved through the experiment involving IPP expansion of γδ+ T-cells. The expectation was that certain populations of γδ+ T-cells would increase in frequency, and this method was able to demonstrate this. Populations of Vδ1+, Vδ2+ and double negative cells have individually shown tumor killing properties, following an induced expansion and this work has elucidated the clonal repertoire of these populations, which could be used in future to further examine the anticancer potential.

However, there are still concerns with the capture kit regarding data analysis. The RNA capture baits are 120bp long, whereas the average read length of the data described in this work is around 200bp. The MiTCR and Decombinator pipelines assign a V and a J tag to a read in order to identify a CDR3 sequence. Therefore, this system may be biased towards picking up the shorter reads, as they are more likely to be represented as whole sequences in our data. Fragmenting samples into larger sizes and increasing the read length to 400bp could mitigate this issue. However, this would only reduce this bias, not eradicate it. Therefore, assembly of these fragments into entire TCR sequences and aligning the original FASTQ file to these constructed sequences might be a more successful way of analysing this data and obtaining genuine frequencies of CDR3 clones. Most short-read assemblers in existence are designed for *de novo* assembly of whole genomes, therefore assume all reads in a sample fit together in one long genomic sequence. As this is not the case here, assembly of TCR sequences could be carried out computationally, for example using a program called iSSAKE that has been specially designed for immune receptor sequences (113, 168).

6.2 Conclusions & Future Work

Meaningful data on all TCR chains has been obtained. The use of targeted capture produces much more data than previous methods, as T-cell and B-cell information can be acquired in a single experiment across all chains. High-throughput immune repertoire sequencing has huge research and clinical potential. There are multiple conditions that can be investigated, and the data output tailored according to the population of cells. This capture technique is largely magnetic bead based and therefore has potential to be automatable and scaled up to allow for fast, high-throughput analysis of patient samples. With pre-capture pooling and multiplexing of samples the cost can be driven down to potentially make it a viable, cost-effective clinical investigation.

However, there are still many inconsistencies in the data. Therefore, additional steps should be taken to optimise this technique before it is used in a clinical setting. At present, the use of targeted capture for high-throughput sequencing of TCR repertoires needs further investigation and the experimental method should be refined. Until this is achieved, the use of other more established methods of sequencing TCR repertoires should be considered first for practical applications.

6.2.1 Future Work

The results here have highlighted the need to further test and refine the capture technique and protocol, if it is to be used again, which can be done in complement to the improvements in sequencing technologies and analysis pipelines. There are multiple ways in which the capture could be further investigated. For example, bait saturation and spiking of samples with a known TCR sequence to test sensitivity could be tried.

So far, this work has been carried out using the total T-cell population, as isolated from the periphery. This does not give an accurate reflection of the TCR repertoires of the individual T-cell subsets. For example, it would be expected that the repertoire of the naïve T-cell pool would be more diverse than that of the memory pool. A comparison of CD4+ and CD8+ T-cell repertoires would also generation information that may be particularly useful in the context of certain disease states. Following future iterations of the capture technique, it may be used for further investigation of immune conditions. Applications of TCR sequencing work could include more detailed investigations into repertoire reconstitution following HSCT, antiviral treatment introduction for HIV, or following thymic transplants in patients with DiGeorge syndrome.

6.2.2 Recent Advances in TCR Repertoire Analysis

It is important to note that the experimental work described in this thesis was undertaken between September 2011 and December 2014. As such, since this work began, many advances have been made in parallel across the field of immune repertoire sequencing. For example, MiTCR has been iterated multiple times and is now available as MiXCR, which allows for more modifications of parameters, enabling customization of the analysis workflow (169). There are more groups than ever utilising high-throughput sequencing technology and some companies have now commercialised the process, providing a full repertoire sequencing and analysis service, or specialised kits (170, 171). Therefore many more research groups now have access to this technology, without the need to develop their own techniques, to investigate a wide range of conditions including cancers, aging, HIV, atherosclerosis and rheumatoid arthritis (172-175). Therefore, as with advances in sequencing technologies, it is important to consider and reevaluate which method of immune repertoire investigation is appropriate for individual applications.

It is also true that the statistical analysis of TCR repertoire data has advanced greatly since this work was undertaken. Many new software packages are now available for statistical analysis and modelling of TCR repertoire data (123, 176, 177). These packages allow for the comparison between samples and individuals and many models are now able to distinguish between covariables including timepoints, which would be helpful to apply to the paediatric stem cell transplant patient samples.

Much of the analysis done to date, including in this work, focuses on the comparison of gene usage frequencies, CDR3 length distributions (178, 179). It is possible to define characteristics of the repertoire through distribution of identified amino acid motifs or kmers (nucleotide sequences of length, k). The frequency and distribution of kmers within a repertoire can be used as a marker to identify disease states, immunization status, or other characteristic (180-182). However, these approaches packages still require larger sample sizes than what is available within this work. Therefore, a strong recommendation of this project is to identify follow up patient groups of interest and obtain ethical approval and consent for the sequencing and analysis of their immune repertoires, in line with the technology now available.

The bioinformatic and statistical analysis described in this these was appropriate for the period in which this work was carried (2011 to 2014). During this time there was not the abundance of software tools available and very few options available for immune repertoire analysis; both the Decombinator and MiTCR were considered cutting edge and highly advanced options. Therefore, although it is agreed by bioinformaticians in the field that this work still contributes to the global understanding of the TCR repertoire, particularly through the use of the capture technique and simultaneous acquisition of the α, β, γ and δ TCR chains and the Ig chains, the experimental aspects and data analysis could be approached differently. However, since this work suggests that the standard multiplex and RACE PCR techniques in use may be more efficient in producing data, there would be little value in the re-analysis of this dataset, as supported by Dr. James Heather in Appendix 4.

Due to the uncertainty within the data, and the dearth of computational analysis designed with capture data as input, this technique was eventually discarded in favour of TCR sequencing techniques that used primer sets designed to amplify sequences (183, 184). The work contained within this thesis was used in support of this, although this does not mean that targeted capture will not be helpful in future. In particular, the project was sustained using the method developed by Professor Benny Chain's team at UCL, and the TCR repertoire of 16 CBT patients were investigated in 2018, influenced by the work included in this thesis (94, 104, 185).

7 Publications

Fisher J, Yan M, Heuijerjans J, **Carter L**, Abolhassani A, Frosch J, Wallace R, Flutter B, Capsomidis A, Hubank M, Klein N, Callard R, Gustafsson K, Anderson J. *et al.* Neuroblastoma killing properties of V-delta2 and V-delta2 negative gamma delta T cells following expansion by artificial antigen presenting cells. ***Clin Cancer Res***. (2014), Nov 15;20(22):5720-32. doi: 10.1158/1078-0432.CCR-13-3464.

Bashford-Rogers RJ, Palser AL, Idris SF, **Carter L**, Epstein M, Callard RE, Douek DC, Vassilliou GS, Follows GA, Hubank M, Kellam P. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. ***BMC Immunol***. (2014), Aug 5;15:29. doi: 10.1186/s12865-014-0029-0

8 References

1. Charles A Janeway, Jr., Travers P, Walport M, Shlomchik MJ. Immunobiology. 2001.

2. Iwasaki A, Medzhitov R. Control of adaptive immunity by the innate immune system. Nat Immunol. 2015;16(4):343-53.

3. J. ABJAL, et al. Molecular Biology of the Cell. 4th Edition. New York: Garland Science; 2002 2002.

4. Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Gartland GL, Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature. 2004;430(6996):174-80.

5. Pancer Z, Saha NR, Kasamatsu J, Suzuki T, Amemiya CT, Kasahara M, et al. Variable lymphocyte receptors in hagfish. Proc Natl Acad Sci USA. 2005;102(26):9224-9.

6. Mackay LK, Rahimpour A, Ma JZ, Collins N, Stock AT, Hafon M-L, et al. The developmental pathway for CD103(+)CD8+ tissue-resident memory T cells of skin. Nat Immunol. 2013;14(12):1294-301.

7. Casey KA, Fraser KA, Schenkel JM, Moran A, Abt MC, Beura LK, et al. Antigen-independent differentiation and maintenance of effector-like resident memory T cells in tissues. J Immunol. 2012;188(10):4866-75.

8. Goronzy JJ, Qi Q, Olshen RA, Weyand CM. High-throughput sequencing insights into T-cell receptor repertoire diversity in aging. Genome Med. 2015;7.

9. van Heijst JWJ, Ceberio I, Lipuma LB, Samilo DW, Wasilewski GD, Gonzales AMR, et al. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. Nat Med. 2013;19(3):372-7.

10. Colten HR. Expression of the MHC class III genes. Philos Trans R Soc Lond, B, Biol Sci. 1984;306(1129):355-66.

11. Clevers H, Alarcon B, Wileman T, Terhorst C. The T cell receptor/CD3 complex: a dynamic protein ensemble. Annu Rev Immunol. 1988;6:629-62.

12. Smith CA, Williams GT, Kingston R, Jenkinson EJ, Owen JJ. Antibodies to CD3/T-cell receptor complex induce death by apoptosis in immature T cells in thymic cultures. Nature. 1989;337(6203):181-4.

13. Brenner MB, McLean J, Dialynas DP, Strominger JL, Smith JA, Owen FL, et al. Identification of a putative second T-cell receptor. Nature. 1986;322(6075):145-9.

14. Fu H, Ward EJ, Marelli-Berg FM. Mechanisms of T cell organotropism. Cell Mol Life Sci. 2016;73:3009-33.

15. Harrington LE, Hatton RD, Mangan PR, Turner H, Murphy TL, Murphy KM, et al. Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. Nat Immunol. 2005;6(11):1123-32.

16. Park H, Li Z, Yang XO, Chang SH, Nurieva R, Wang Y-H, et al. A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. Nat Immunol. 2005;6(11):1133-41.

17. Deo SS, Mistry KJ, Kakade AM, Niphadkar PV. Role played by Th2 type cytokines in IgE mediated allergy and asthma. Lung India. 2010;27(2):66-71.

18. Milstein O, Hagin D, Lask A, Reich-Zeliger S, Shezan E, Ophir E, et al. CTLs respond with activation and granule secretion when serving target for T cell recognition. Blood. 2010:blood-2010-05-283770.

19. Krzyzowska M, Cymerys J, Winnicka A, Niemiałtowski M. Involvement of Fas and FasL in Ectromelia virus-induced apoptosis in mouse brain. Virus Res. 2006;115(2):141-9.

20. Strasser A, Jost PJ, Nagata S. The many roles of FAS receptor signaling in the immune system. Immunity. 2009;30(2):180-92.

21. Gattinoni L, Klebanoff CA, Restifo NP. Paths to stemness: building the ultimate antitumour T cell. Nat Rev Cancer. 2012;12(10):671-84.

22. Thomas ML. The leukocyte common antigen family. Annu Rev Immunol. 1989;7:339-69.

23. Seddiki N, Santner-Nanan B, Tangye SG, Alexander SI, Solomon M, Lee S, et al. Persistence of naive CD45RA+ regulatory T cells in adult life. Blood. 2006;107(7):2830-8.

24. Shen Y, Zhou D, Qiu L, Lai X, Simon M, Shen L, et al. Adaptive immune response of Vgamma2Vdelta2+ T cells during mycobacterial infections. Science (New York, NY). 2002;295(5563):2255-8.

25. Hoft DF, Brown RM, Roodman ST. Bacille Calmette-Guérin vaccination enhances human gamma delta T cell responsiveness to mycobacteria suggestive of a memory-like phenotype. J Immunol. 1998;161(2):1045-54.

26. Worku S, Gorse GJ, Belshe RB, Hoft DF. Canarypox vaccines induce antigen-specific human gammadelta T cells capable of interferon-gamma production. J Infect Dis. 2001;184(5):525-32.

27. Chien YH, Jores R, Crowley MP. Recognition by gamma/delta T cells. Annu Rev Immunol. 1996;14:511-32.

28. Sciammas R, Johnson RM, Sperling AI, Brady W, Linsley PS, Spear PG, et al. Unique antigen recognition by a herpesvirus-specific TCR-gamma delta cell. J Immunol. 1994;152(11):5392-7.

29. Allison TJ, Garboczi DN. Structure of gammadelta T cell receptors and their recognition of non-peptide antigens. Mol Immunol. 2002;38(14):1051-61.

30. Eberl M, Hintz M, Reichenberg A, Kollas A-K, Wiesner J, Jomaa H. Microbial isoprenoid biosynthesis and human gammadelta T cell activation. FEBS Lett. 2003;544(1-3):4-10.

31. Brown AC, Eberl M, Crick DC, Jomaa H, Parish T. The nonmevalonate pathway of isoprenoid biosynthesis in Mycobacterium tuberculosis is essential and transcriptionally regulated by Dxs. J Bacteriol. 2010;192(9):2424-33.

32. al WXe. MYC-Regulated Mevalonate Metabolism Maintains Brain Tumor-Initiating Cells. - PubMed - NCBI.

33. Qiu L, Yang H, Lv G, Li K, Liu G, Wang W, et al. Insights into the mevalonate pathway in the anticancer effect of a platinum complex on human gastric cancer cells. Eur J Pharmacol. 2017;810:120-7.

34. Tanaka Y, Morita CT, Tanaka Y, Nieves E, Brenner MB, Bloom BR. Natural and synthetic non-peptide antigens recognized by human γδ T cells. Nature. 1995;375:155-8.

35. Fisher J, Kramer A-M, Gustafsson K, Anderson J. Non-V delta 2 gamma delta T lymphocytes as effectors of cancer immunotherapy. Oncoimmunology. 2015;4(3):e973808.

36. Zocchi MR, Ferrarini M, Migone N, Casorati G. T-cell receptor V delta gene usage by tumour reactive gamma delta T lymphocytes infiltrating human lung cancer. Immunology. 1994;81(2):234-9.

37. Fisher J, Yan M, Heuijerjans J, Carter L, Abolhassani A, Frosch J, et al. Neuroblastoma killing properties of V-delta 2 and V-delta2 negative gamma delta T cells following expansion by artificial antigen presenting cells. Clin Cancer Res. 2014.

38. Nielsen JS, McNagny KM. Novel functions of the CD34 family. J Cell Sci. 2008;121(Pt 22):3683-92.

39. Allman D, Sambandam A, Kim S, Miller JP, Pagan A, Well D, et al. Thymopoiesis independent of common lymphoid progenitors. Nat Immunol. 2003;4(2):168-74.

40. Kondo M, Weissman IL, Akashi K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell. 1997;91(5):661-72.

41. Schwarz BA, Bhandoola A. Trafficking from the bone marrow to the thymus: a prerequisite for thymopoiesis. Immunol Rev. 2006;209:47-57.

42. Petrie HT. Role of thymic organ structure and stromal composition in steady-state postnatal T-cell production. Immunol Rev. 2002;189:8-19.

43. Anderson G, Harman BC, Hare KJ, Jenkinson EJ. Microenvironmental regulation of T cell development in the thymus. Semin Immunol. 2000;12(5):457-64.

44. Anderson G, Moore NC, Owen JJ, Jenkinson EJ. Cellular interactions in thymocyte development. Annu Rev Immunol. 1996;14:73-99.

45. Germain RN. T-cell development and the CD4–CD8 lineage decision. Nat Rev Immunol. 2002;2(5):309-22.

46. Godfrey DI, Kennedy J, Suda T, Zlotnik A. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. J Immunol. 1993;150(10):4244-52.

47. Yui MA, Feng N, Rothenberg EV. Fine-scale staging of T cell lineage commitment in adult mouse thymus. J Immunol. 2010;185(1):284-93.

48. Xu X, Zhang S, Li P, Lu J, Xuan Q, Ge Q. Maturation and emigration of single-positive thymocytes. Clin Dev Immunol. 2013;2013:282870.

49. Teng F, Zhou Y, Jin R, Chen Y, Pei X, Liu Y, et al. The molecular signature underlying the thymic migration and maturation of TCRαβ+ CD4+ CD8 thymocytes. PLoS One. 2011;6(10):e25567.

50. Stritesky GL, Xing Y, Erickson JR, Kalekar LA, Wang X, Mueller DL, et al. Murine thymic selection quantified using a unique method to capture deleted T cells. Proc Natl Acad Sci USA. 2013;110(12):4679-84.

51. Haynes BF, Markert ML, Sempowski GD, Patel DD, Hale LP. The role of the thymus in immune reconstitution in aging, bone marrow transplantation, and HIV-1 infection. Annu Rev Immunol. 2000;18:529-60.

52. Spits H. Development of alphabeta T cells in the human thymus. Nat Rev Immunol. 2002;2(10):760-72.

53. Sallusto F, Lenig D, Förster R, Lipp M, Lanzavecchia A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature. 1999;401(6754):708-12.

54. Radtke F, Wilson A, Stark G, Bauer M, van Meerwijk J, MacDonald HR, et al. Deficient T cell fate specification in mice with an induced inactivation of Notch1. Immunity. 1999;10(5):547-58.

55. Washburn T, Schweighoffer E, Gridley T, Chang D, Fowlkes BJ, Cado D, et al. Notch activity influences the alphabeta versus gammadelta T cell lineage decision. Cell. 1997;88(6):833-43.

56. Ciofani M, Knowles GC, Wiest DL, von Boehmer H, Zúñiga-Pflücker JC. Stage-specific and differential notch dependency at the alphabeta and gammadelta T lineage bifurcation. Immunity. 2006;25(1):105-16.

57. Van de Walle I, Waegemans E, De Medts J, De Smet G, De Smedt M, Snauwaert S, et al. Specific Notch receptor–ligand interactions control human TCR-αβ/γδ development by inducing differential Notch signal strength. J Exp Med. 2013;210(4):683-97.

58. Pomés A, Chruszcz M, Gustchina A, Minor W, Mueller GA, Pedersen LC, et al. 100 Years later: Celebrating the contributions of x-ray crystallography to allergy and clinical immunology. J Allergy Clin Immunol. 2015;136(1):29-37.e10.

59. Allison TJ, Winter CC, Fournié JJ, Bonneville M, Garboczi DN. Structure of a human gammadelta T-cell antigen receptor. Nature. 2001;411(6839):820-4.

60. Gao GF, Jakobsen BK. Molecular interactions of coreceptor CD8 and MHC class I: the molecular basis for functional coordination with the T-cell receptor. Immunol Today. 2000;21(12):630-6.

61. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT(R), the international ImMunoGeneTics information system(R). Nucleic Acids Research. 2009;37(Database):D1006-D12.

62. Oettinger MA, Schatz DG, Gorka C, Baltimore D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science (New York, NY). 1990;248(4962):1517-23.

63. Schatz DG, Oettinger MA, Baltimore D. The V(D)J recombination activating gene, RAG-1. Cell. 1989;59(6):1035-48.

64. Krangel MS. Mechanics of T cell receptor gene rearrangement. Curr Opin Immunol. 2009;21(2):133-9.

65. Arnaud J, Huchenq A, Vernhes MC, Caspar-Bauguil S, Lenfant F, Sancho J, et al. The interchain disulfide bond between TCR alpha beta heterodimers on human T cells is not required for TCR-CD3 membrane expression and signal transduction. Int Immunol. 1997;9(4):615-26.

66. Hockett RD, de Villartay JP, Pollock K, Poplack DG, Cohen DI, Korsmeyer SJ. Human T-cell antigen receptor (TCR) delta-chain locus and elements responsible for its deletion are within the TCR alpha-chain locus. Proc Natl Acad Sci USA. 1988;85(24):9694-8.

67. Michie AM, Zúñiga-Pflücker JC. Regulation of thymocyte differentiation: pre-TCR signals and beta-selection. Semin Immunol. 2002;14(5):311-23.

68. Janeway Ca Jr TPWM, et al. Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001.

69. Ma Y, Pannicke U, Schwarz K, Lieber MR. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell. 2002;108(6):781-94.

70. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature. 1988;334(6181):395-402.

71. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. J Immunol. 2000;164(11):5782-7.

72. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A Direct Estimate of the Human αβ T Cell Receptor Diversity. Science. 1999;286(5441):958-61.

73. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, et al. The influence of age on T cell generation and TCR diversity. J Immunol. 2005;174(11):7446-52.

74. Palmer DB. The effect of age on thymic function. Front Immunol. 2013;4:316.

75. Ouyang Q, Wagner WM, Walter S, Müller CA, Wikby A, Aubert G, et al. An age-related increase in the number of CD8+ T cells carrying receptors for an immunodominant Epstein-Barr virus (EBV) epitope is counteracted by a decreased frequency of their antigen-specific responsiveness. Mech Ageing Dev. 2003;124(4):477-85.

76. Ouyang Q, Wagner WM, Wikby A, Walter S, Aubert G, Dodi AI, et al. Large numbers of dysfunctional CD8+ T lymphocytes bearing receptors for a single dominant CMV epitope in the very old. J Clin Immunol. 2003;23(4):247-57.

77. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. 2014;24(10):1603-12.

78. Krell PFI, Reuther S, Fischer U, Keller T, Weber S, Gombert M, et al. Next-generation-sequencing-spectratyping reveals public T-cell receptor repertoires in pediatric very severe aplastic anemia and identifies a β chain CDR3 sequence associated with hepatitis-induced pathogenesis. Haematologica. 2013;98(9):1388-96.

79. Pannetier C, Cochet M, Darche S, Casrouge A, Zöller M, Kourilsky P. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. Proc Natl Acad Sci USA. 1993;90(9):4319-23.

80. Gorski J, Yassai M, Zhu X, Kissela B, Kissella B, Keever C, et al. Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. J Immunol. 1994;152(10):5109-19.

81. Baum PD, Young JJ, Schmidt D, Zhang Q, Hoh R, Busch M, et al. Blood T Cell Receptor Diversity Decreases During the Course of HIV Infection but the Potential for a Diverse Repertoire Persists. Blood. 2012.

82. Holt RA, Jones SJM. The new paradigm of flow cell sequencing. Genome Res. 2008;18(6):839-46.

83. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45.

84. Wang C, Sanders CM, Yang Q, Schroeder HW, Jr., Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci USA. 2010;107(4):1518-23.

85. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010;2(47):47ra64.

86. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009;114(19):4099-107.

87. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009;19(10):1817-24.

88. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. 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 Res. 2011.

89. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics. 2011;12(1):106.

90. Bolotin D. Next Generation Sequencing for TCR Repertoire Profiling: platform-specific features and correction algoriithms.

91. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research. 2011;39(13):e90.

92. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011;12(11):R112.

93. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8(3):186-94.

94. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics (Oxford, England). 2013;29(5):542-50.

95. Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, et al. MiTCR: software for T-cell receptor sequencing data analysis. Nature Methods. 2013;10(9):813-4.

96. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

97. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.

98. Gharizadeh B, Akhras M, Nourizad N, Ghaderi M, Yasuda K, Nyrén P, et al. Methodological improvements of pyrosequencing technology. J Biotechnol. 2006;124(3):504-11.

99. Purushothaman S, Toumazou C, Ou C-P. Protons and single nucleotide polymorphism detection: A simple use for the Ion Sensitive Field Effect Transistor. Sensors and Actuators B: Chemical. 2006;114(2):964-8.

100. Song L, Huang W, Kang J, Huang Y, Ren H, Ding K. Comparison of error correction algorithms for Ion Torrent PGM data: application to hepatitis B virus. Sci Rep. 2017;7.

101. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009;1(12):12ra23.

102. Klarenbeek PL, Tak PP, van Schaik BDC, Zwinderman AH, Jakobs ME, Zhang Z, et al. Human T-cell memory consists mainly of unexpanded clones. Immunol Lett. 2010;133(1):42-8.

103. Conrad JA, Ramalingam RK, Duncan CB, Smith RM, Wei J, Barnett L, et al. Antiretroviral therapy reduces the magnitude and T cell receptor repertoire diversity of HIV-specific T cell responses without changing T cell clonotype dominance. J Virol. 2012;86(8):4213-21.

104. Heather JM, Best K, Oakes T, Gray ER, Roe JK, Thomas N, et al. Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and following Antiretroviral Therapy. Front Immunol. 2016;6.

105. Hebels DGAJ, van Herwijnen MHM, Brauers KJJ, de Kok TMCM, Chalkiadaki G, Kyrtopoulos SA, et al. Elimination of heparin interference during microarray

processing of fresh and biobank-archived blood samples. Environ Mol Mutagen. 2014;55(6):482-91.

106. Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. Brief Bioinformatics. 2018;19(4):554-65.

107. Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, et al. Comparison of Commercially Available Target Enrichment Methods for Next-Generation Sequencing. J Biomol Tech. 2013;24(2):73-86.

108. Bashford-Rogers RJ, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, et al. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunol. 2014;15(1):29.

109. Babraham Bioinformatics BI. FastQC.

110. Yang M. Average RNA Yields2017 2017/06/24/.

111. Eikmans M, Rekers NV, Anholts JDH, Heidt S, Claas FHJ. Blood cell mRNAs and microRNAs: optimized protocols for extraction and preservation. Blood. 2013;121(11):e81-9.

112. He H-J, Stein EV, DeRose P, Cole KD. Limitations of methods for measuring the concentration of human genomic DNA and oligonucleotide samples. Biotechniques. 2018;64(2):59-68.

113. Warren RL, Nelson BH, Holt RA. Profiling model T-cell metagenomes with short reads. Bioinformatics. 2009;25(4):458-64.

114. Haile S, Corbett RD, MacLeod T, Bilobram S, Smailus D, Tsao P, et al. Increasing quality, throughput and speed of sample preparation for strand-specific messenger RNA sequencing. BMC Genomics. 2017;18(1):515.

115. Cline J, Braman JC, Hogrefe HH. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research. 1996;24(18):3546-51.

116. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnology. 2017;17(1):61.

117. Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Commun ACM. 1975;18(6):333-40.

118. Simpson EH. Measurement of Diversity. Nature. 1949;163:688.

119. Kitaura K, Shini T, Matsutani T, Suzuki R. A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains. BMC Immunol. 2016;17(1):38.

120. Chaara W, Gonzalez-Tort A, Florez L-M, Klatzmann D, Mariotti-Ferrandiz E, Six A. RepSeq Data Representativeness and Robustness Assessment by Shannon Entropy. Front Immunol. 2018;9:1038.

121. Gart JJ. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Volume IV: Contributions to Biology and Problems of Medicine. Held at the Statistical Laboratory. Jerzy Neyman. The Quarterly Review of Biology. 1963;38(4):447-8.

122. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee J-Y, et al. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci USA. 2014;111(36):13139-44.

123. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: an R package for T cell receptor repertoire advanced data analysis. BMC Bioinformatics. 2015;16:175.

124. Fishman JA. Infection in solid-organ transplant recipients. N Engl J Med. 2007;357(25):2601-14.

125. Blyth E, Withers B, Clancy L, Gottlieb D. CMV-specific immune reconstitution following allogeneic stem cell transplantation. Virulence. 2016;7(8):967-80.

126. Lion T, Baumgartinger R, Watzinger F, Matthes-Martin S, Suda M, Preuner S, et al. Molecular monitoring of adenovirus in peripheral blood after allogeneic bone marrow transplantation permits early diagnosis of disseminated disease. Blood. 2003;102(3):1114-20.

127. Gatti RA, Meuwissen HJ, Allen HD, Hong R, Good RA. Immunological reconstitution of sex-linked lymphopenic immunological deficiency. Lancet. 1968;2(7583):1366-9.

128. Körbling M, Freireich EJ. Twenty-five years of peripheral blood stem cell transplantation. Blood. 2011;117(24):6411-6.

129. Schumm M, Lang P, Taylor G, Kuçi S, Klingebiel T, Bühring HJ, et al. Isolation of highly purified autologous and allogeneic peripheral CD34+ cells using the CliniMACS device. J Hematother. 1999;8(2):209-18.

130. Chaleff S, Otto M, Barfield RC, Leimig T, Iyengar R, Martin J, et al. A large-scale method for the selective depletion of alphabeta T lymphocytes from PBSC for allogeneic transplantation. Cytotherapy. 2007;9(8):746-54.

131. Bonneville M, O'Brien RL, Born WK. Gammadelta T cell effector functions: a blend of innate programming and acquired plasticity. Nat Rev Immunol. 2010;10(7):467-78.

132. Chiplunkar S, Dhar S, Wesch D, Kabelitz D. gammadelta T cells in cancer immunotherapy: current status and future prospects. Immunotherapy. 2009;1(4):663-78.

133. Godder KT, Henslee-Downey PJ, Mehta J, Park BS, Chiang KY, Abhyankar S, et al. Long term disease-free survival in acute leukemia patients recovering with increased gammadelta T cells after partially mismatched related donor bone marrow transplantation. Bone Marrow Transplant. 2007;39(12):751-7.

134. Patel SR, Ridwan RU, Ortín M. Cytomegalovirus reactivation in pediatric hemopoietic progenitors transplant: a retrospective study on the risk factors and the efficacy of treatment. J Pediatr Hematol Oncol. 2005;27(8):411-5.

135. Shimoni A, Yeshurun M, Hardan I, Avigdor A, Ben-Bassat I, Nagler A. Thrombotic microangiopathy after allogeneic stem cell transplantation in the era of reduced-intensity conditioning: The incidence is not reduced. Biol Blood Marrow Transplant. 2004;10(7):484-93.

136. Deeg HJ, Antin JH. The clinical spectrum of acute graft-versus-host disease. Semin Hematol. 2006;43(1):24-31.

137. Liu C, He M, Rooney B, Kepler TB, Chao NJ. Longitudinal analysis of T-cell receptor variable beta chain repertoire in patients with acute graft-versus-host disease after allogeneic stem cell transplantation. Biol Blood Marrow Transplant. 2006;12(3):335-45.

138. Zorn E. CD4+CD25+ regulatory T cells in human hematopoietic cell transplantation. Semin Cancer Biol. 2006;16(2):150-9.

139. Barrett J. Improving outcome of allogeneic stem cell transplantation by immunomodulation of the early post-transplant environment. Curr Opin Immunol. 2006;18(5):592-8.

140. Feuchtinger T, Matthes-Martin S, Richard C, Lion T, Fuhrer M, Hamprecht K, et al. Safe adoptive transfer of virus-specific T-cell immunity for the treatment of

systemic adenovirus infection after allogeneic stem cell transplantation. Br J Haematol. 2006;134(1):64-76.

141. Feuchtinger T, Opherk K, Bethge WA, Topp MS, Schuster FR, Weissinger EM, et al. Adoptive transfer of pp65-specific T cells for the treatment of chemorefractory cytomegalovirus disease or reactivation after haploidentical and matched unrelated stem cell transplantation. Blood. 2010;116(20):4360-7.

142. Hsieh M-Y, Hong W-H, Lin J-J, Lee W-I, Lin K-L, Wang H-S, et al. T-cell receptor excision circles and repertoire diversity in children with profound T-cell immunodeficiency. J Microbiol Immunol Infect. 2013;46(5):374-81.

143. Müller SM, Kohn T, Schulz AS, Debatin KM, Friedrich W. Similar pattern of thymic-dependent T-cell reconstitution in infants with severe combined immunodeficiency after human leukocyte antigen (HLA)-identical and HLA-nonidentical stem cell transplantation. Blood. 2000;96(13):4344-9.

144. Patel DD, Gooding ME, Parrott RE, Curtis KM, Haynes BF, Buckley RH. Thymic function after hematopoietic stem-cell transplantation for the treatment of severe combined immunodeficiency. N Engl J Med. 2000;342(18):1325-32.

145. Myers LA, Patel DD, Puck JM, Buckley RH. Hematopoietic stem cell transplantation for severe combined immunodeficiency in the neonatal period leads to superior thymic output and improved survival. Blood. 2002;99(3):872-8.

146. Sarzotti M, Patel DD, Li X, Ozaki DA, Cao S, Langdon S, et al. T cell repertoire development in humans with SCID after nonablative allogeneic marrow transplantation. J Immunol. 2003;170(5):2711-8.

147. Okamoto H, Arii C, Shibata F, Toma T, Wada T, Inoue M, et al. Clonotypic analysis of T cell reconstitution after haematopoietic stem cell transplantation (HSCT)

in patients with severe combined immunodeficiency. Clin Exp Immunol. 2007;148(3):450-60.

148. Chen X, Barfield R, Benaim E, Leung W, Knowles J, Lawrence D, et al. Prediction of T-cell reconstitution by assessment of T-cell receptor excision circle before allogeneic hematopoietic stem cell transplantation in pediatric patients. Blood. 2005;105(2):886-93.

149. Hiwarkar P, Hubank M, Qasim W, Chiesa R, Gilmour KC, Saudemont A, et al. Cord blood transplantation recapitulates fetal ontogeny with a distinct molecular signature that supports CD4(+) T-cell reconstitution. Blood Adv. 2017;1(24):2206-16.

150. Chiesa R, Gilmour K, Qasim W, Adams S, Worth AJ, Zhan H, et al. Omission of in vivo T-cell depletion promotes rapid expansion of naive CD4+ cord blood lymphocytes and restores adaptive immunity within 2 months after unrelated cord blood transplant. Br J Haematol. 2012;156(5):656-66.

151. Ponce DM, Zheng J, Gonzales AM, Lubin M, Heller G, Castro-Malaspina H, et al. Reduced late mortality risk contributes to similar survival after double-unit cord blood transplantation compared with related and unrelated donor hematopoietic stem cell transplantation. Biol Blood Marrow Transplant. 2011;17(9):1316-26.

152. Sauter C, Abboud M, Jia X, Heller G, Gonzales A-M, Lubin M, et al. Serious infection risk and immune recovery after double-unit cord blood transplantation without antithymocyte globulin. Biol Blood Marrow Transplant. 2011;17(10):1460-71.

153. Jakubowski AA, Small TN, Young JW, Kernan NA, Castro-Malaspina H, Hsu KC, et al. T cell depleted stem-cell transplantation for adults with hematologic malignancies: sustained engraftment of HLA-matched related donor grafts without the use of antithymocyte globulin. Blood. 2007;110(13):4552-9.

154. Chen X, Hale GA, Barfield R, Benaim E, Leung WH, Knowles J, et al. Rapid immune reconstitution after a reduced-intensity conditioning regimen and a CD3-depleted haploidentical stem cell graft for paediatric refractory haematological malignancies. Br J Haematol. 2006;135(4):524-32.

155. Warren EH, Matsen FAt, Chou J. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. Blood. 2013;122(1):19-22.

156. Logan AC, Zhang B, Narasimhan B, Carlton V, Zheng J, Moorhead M, et al. Minimal residual disease quantification using consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic lymphocytic leukemia. Leukemia. 2013;27(8):1659-65.

157. Sacchettini JC, Poulter CD. Creating isoprenoid diversity. Science (New York, NY). 1997;277(5333):1788-9.

158. Himoudi N, Morgenstern DA, Yan M, Vernay B, Saraiva L, Wu Y, et al. Human γδ T lymphocytes are licensed for professional antigen presentation by interaction with opsonized target cells. J Immunol. 2012;188(4):1708-16.

159. Gertner-Dardenne J, Bonnafous C, Bezombes C, Capietto A-H, Scaglione V, Ingoure S, et al. Bromohydrin pyrophosphate enhances antibody-dependent cell-mediated cytotoxicity induced by therapeutic antibodies. Blood. 2009;113(20):4875-84.

160. Shah RM, Elfeky R, Nademi Z, Qasim W, Amrolia P, Chiesa R, et al. T-cell receptor αβ+ and CD19+ cell–depleted haploidentical and mismatched hematopoietic stem cell transplantation in primary immune deficiency. Journal of Allergy and Clinical Immunology. 2018;141(4):1417-26.e1.

161. Wesch D, Marx S, Kabelitz D. Comparative analysis of alpha beta and gamma delta T cell activation by Mycobacterium tuberculosis and isopentenyl pyrophosphate. Eur J Immunol. 1997;27(4):952-6.

162. Tosato F, Bucciol G, Pantano G, Putti MC, Sanzari MC, Basso G, et al. Lymphocytes subsets reference values in childhood. Cytometry A. 2015;87(1):81-5.

163. Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, et al. Systematic Comparative Evaluation of Methods for Investigating the TCRbeta Repertoire. PLoS One. 2016;11(3):e0152464.

164. Vianna PH, Canto FB, Nogueira JS, Nunes CF, Bonomo AC, Fucs R. Critical influence of the thymus on peripheral T cell homeostasis. Immun Inflamm Dis. 2016;4(4):474-86.

165. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 2016;44(W1):W3-W10.

166. Grigaityte K, Carter JA, Goldfless SJ, Jeffery EW, Hause RJ, Jiang Y, et al. Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. bioRxiv. 2017:213462.

167. Linnemann C, Heemskerk B, Kvistborg P, Kluin RJC, Bolotin DA, Chen X, et al. High-throughput identification of antigen-specific TCRs by TCR gene capture. Nat Med. 2013;19(11):1534-41.

168. Warren RL, Holt RA. Targeted assembly of short sequence reads. PLoS One. 2011;6(5):e19816.

169. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12(5):380-1.

170. Röhr C, Kerick M, Fischer A, Kühn A, Kashofer K, Timmermann B, et al. High-throughput miRNA and mRNA sequencing of paired colorectal normal, tumor and metastasis tissues and bioinformatic modeling of miRNA-1 therapeutic applications. PLoS One. 2013;8(7):e67461.

171. Morin A, Kwan T, Ge B, Letourneau L, Ban M, Tandre K, et al. Immunoseq: the identification of functionally relevant variants through targeted capture and sequencing of active regulatory regions in human immune cells. BMC Medical Genomics. 2016;9:59.

172. Hu F, Zhang W, Shi L, Liu X, Jia Y, Xu L, et al. Impaired CD27+IgD+B Cells With Altered Gene Signature in Rheumatoid Arthritis. Front Immunol. 2018;9:626.

173. Kallemeijn MJ, Kavelaars FG, van der Klift MY, Wolvers-Tettero ILM, Valk PJM, van Dongen JJM, et al. Next-Generation Sequencing Analysis of the Human TCRγδ+ T-Cell Repertoire Reveals Shifts in Vγ- and Vδ-Usage in Memory Populations upon Aging. Front Immunol. 2018;9:448.

174. Winkels H, Ehinger E, Vassallo M, Buscher K, Dinh H, Kobiyama K, et al. Atlas of the Immune Cell Repertoire in Mouse Atherosclerosis Defined by Single-Cell RNA-Sequencing and Mass Cytometry. Circ Res. 2018.

175. Mohme M, Schliffke S, Maire CL, Rünger A, Glau L, Mende KC, et al. Immunophenotyping of Newly Diagnosed and Recurrent Glioblastoma Defines Distinct Immune Exhaustion Profiles in Peripheral and Tumor-infiltrating Lymphocytes. Clin Cancer Res. 2018.

176. Olson BJ, Moghimi P, Schramm CA, Obraztsova A, Ralph D, Vander Heiden JA, et al. sumrep: A Summary Statistic Framework for Immune Receptor Repertoire Comparison and Model Validation. Front Immunol. 2019;10:2533-.

177. Ni Q, Zhang J, Zheng Z, Chen G, Christian L, Grönholm J, et al. VisTCR: An Interactive Software for T Cell Repertoire Sequencing Data Analysis. Front Genet. 2020;11:771-.

178. Miqueu P, Guillet M, Degauque N, Doré JC, Soulillou JP, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Mol Immunol. 2007;44(6):1057-64.

179. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol. 2012;189(6):3221-30.

180. Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics. 2014;30(22):3181-8.

181. Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. BMC Bioinformatics. 2017;18(1):401.

182. Cinelli M, Sun Y, Best K, Heather JM, Reich-Zeliger S, Shifrut E, et al. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. Bioinformatics. 2017;33(7):951-5.

183. Uddin I, Woolston A, Peacock T, Joshi K, Ismail M, Ronel T, et al. Quantitative analysis of the T cell receptor repertoire. Methods Enzymol. 2019;629:465-92.

184. Chain B, Greiff V, Textor J, Yaari G. Editorial: Methods and Applications of Computational Immunology. Front Immunol. 2019;10:2818.

185. Gkazi AS, Margetts BK, Attenborough T, Mhaldien L, Standing JF, Oakes T, et al. Clinical T Cell Receptor Repertoire Deep Sequencing and Analysis: An Application to Monitor Immune Reconstitution Following Cord Blood Transplantation. Front Immunol. 2018;9(2547).

Appendix 1

```
# Dickey3.py 19/10/2013
#to use me: Dickey3.run(sam file, file of transcript names, output file name)
# eg. Dickey3.run("file.sam","ensembl_transcript_list.txt","test3.txt")
print('to use me: Dickey3.run(sam file, file of transcript names, output file name)')
print('eg. Dickey3.run("file.sam","ensembl_transcript_list.txt","test3.txt")')
import fileinput
import datetime
printinterval = 5000 # set counter print interval
fileinput.close() # fixes any unclosed fileinput.input() from previous runs
def count_reads(filename):
        trcount = (169)
        counter = 0
        for line in fileinput.input(filename):
                L=line.strip()
                V=L.split("\t")
                transcript=V[2]
                marker = V[0]
                if marker[0] != "@" and marker[0] != "(" and marker[0] != "" and transcript[:4] == "ENST":
                        if transcript in trcount:
                                trcount[transcript]+=1
                        else:
                                trcount[transcript]=1
                counter+= 1
                if counter%(printinterval*500) == 0:
                    print("Transcr input so far: "+str(counter))
    print("Total Transcr input: "+str(counter))
    print("Transcr dictionary length: "+str(len(trcount)))
    fileinput.close()
    return trcount


# Dictionary listing 1st 3 letters of gene name to check against and info printed in 'Is
TCR?' column.
# gene names not listed here will print 'IS TCR' as FALSE
ids = {'TRA':'TCR-a', 'TRB':'TCR-b', 'TRG':'TCR-g', 'TRD':'TCR-d', 'IGH':'IG-h',
'IGL':'IG-l', 'IGK':'IG-k'}
def name_dict(namedicfile,countdic):
    names_dic = {}
    counter = 0
    for line in fileinput.input(namedicfile):
        L=line.strip()
        V=L.split("\t")
        enst = V[1]
        gname = V[2]
        if enst in countdic.keys():
            if gname[0:3] in ids.keys():
                istcr = ids[gname[0:3]]
            else:
                istcr = "False"
            if len(V)>=4:
                gdes = V[3]
            else:
                gdes = ""
            ginfo = gname+"\t"+istcr+"\t"+gdes
            names_dic[enst] = ginfo
            counter+= 1
            if counter%printinterval == 0:
                print("Gene names added so far: "+str(counter))
    print("Total Gene names added: "+str(counter))
    print("Gene name dictionary length: "+str(len(names_dic)))
    fileinput.close()
    return names_dic


def output_dic(countdic,namedict,name_of_file):
    counter = 0
    f=open(name_of_file,"w")
    f.write("ENST code\tFrequency\tGene Name\tIs TCR?\tGene Info\n")
    lis=sorted(countdic, key=countdic.get, reverse=True)
    for key in lis:
        f.write(key+"\t"+str(countdic[key])+"\t"+namedict[key]+"\n")
        counter+= 1
        if counter%printinterval == 0:
            print("File output: "+str(counter))
```

```
	print("Total File output: "+str(counter))
	f.close()
	print("BAAAM - "+str(counter)+" Done")


def run(dic_file,namedic,output_file):
	start_time = datetime.datetime.now()
	print("I'm thinking...")
	countdic = count_reads(dic_file)
	name_dic = name_dict(namedic,countdic)
	output_dic(countdic,name_dic,output_file)
	end_time = datetime.datetime.now()
	time_taken = end_time-start_time
	print "Time taken:",time_taken
```

Appendix 2

```
print('to use me: randsample("input.fastq", "output.fastq", samplesize)')

import linecache

import random

import datetime


def CountLines(filename):

    f = open(filename)

    try:

        lines = 1

        buf_size = 1024 * 1024

        read_f = f.read # loop optimization

        buf = read_f(buf_size)


        # Empty file

        if not buf:

            return 0


        while buf:

            lines += buf.count('\n')

            buf = read_f(buf_size)


        return lines

    finally:
```

```
    f.close()

def randsample(infile, outfile, samplesize = 1000):
    sttime = datetime.datetime.now()
    print("I'm thinking...")
    #get number of lines
    fastqlength = CountLines(infile)
    numseqs = fastqlength / 4
    if samplesize > numseqs:
        print("error: Sample size is greater than number of sequences in FASTQ")
        return
    timetaken = datetime.datetime.now() - sttime
    samplepercentage = (1.00 * samplesize / numseqs) * 100
    print("Counted "+str(numseqs)+" in: "+str(timetaken))
    print("Generating random sample. Sample size - "+str(samplesize)+" reads ("+str(samplepercentage)+"% of fastq)...")
    counter = 0
    if samplesize >= 10:
        countprint = samplesize / 10
    else:
        countprint = samplesize
    f=open(outfile,"w")
    seqrecord = [] #list of seqnums already used
    while counter < samplesize:
        #pick random 1st (of group of 4) data line

        seqnum = random.randint(1,numseqs)

        #check if seqnum already used

        if seqnum in seqrecord:

            continue #returns to start of loop

        #record seqnum

        seqrecord.append(seqnum)

        linenum = ((seqnum)*4)-3

        #get 4-line group starting with linenum

        line1 = linecache.getline(infile,linenum)

        line2 = linecache.getline(infile,linenum+1)

        line3 = linecache.getline(infile,linenum+2)

        line4 = linecache.getline(infile,linenum+3)

        #print line to file

        f.write(line1+line2+line3+line4)

        counter+=1

        if counter % countprint == 0:

            timetaken = datetime.datetime.now() - sttime

            print(str(counter)+" reads, "+str(100.*counter / samplesize)+"%.

t="+str(timetaken))

    linecache.clearcache()

    #close files

    f.close()

    seqoutput = CountLines(outfile)/4

    timetaken = datetime.datetime.now() - sttime
```

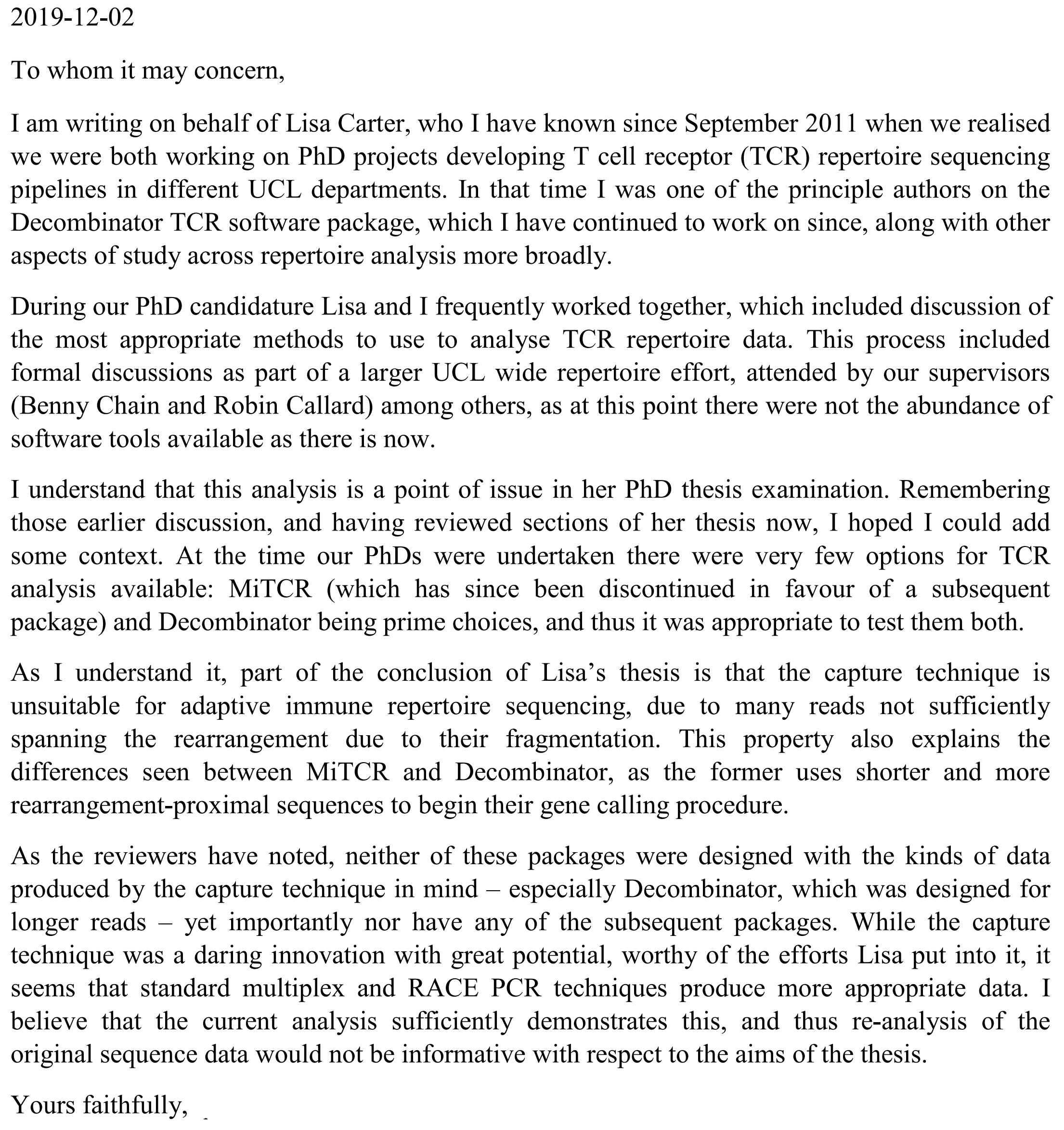
```
print("Done. "+str(counter)+" reads sampled, "+str(seqoutput)+" reads written in: "+str(timetaken))
```

Appendix 3

```
## R Script to calculate 1-exponential regression and plot it ###
library(lattice)


clonotypes=c(514,1037,1666,2329,3543,4540,5460,9045);
reads = c(10000,100000,500000,1000000,2000000,3000000,4000000,8366245)
df = data.frame(reads,clonotypes)



fit <- nls(clonotypes~A*(1-exp(-reads/C))+B,data=df,start=list(A=10000000,B=0,C=1000000),trace=TRUE)


xyplot(clonotypes~reads,df,ylim=c(0,1.1*(coef(fit)[1]+coef(fit)[2])),xlim=c(0,5*coef(fit)[3]),panel=function(...){
        panel.xyplot(...)
        panel.curve(coef(fit)[2] + coef(fit)[1]*(1-exp(-x/coef(fit)[3])))
        panel.abline(coef(fit)[1]+ coef(fit)[2],0,lty=2)})
```

Appendix 4

Letter from James M. Heather on behalf of Lisa Carter, regarding the bioinformatic analyses undertaken throughout this work.



2019-12-02

To whom it may concern,

I am writing on behalf of Lisa Carter, who I have known since September 2011 when we realised we were both working on PhD projects developing T cell receptor (TCR) repertoire sequencing pipelines in different UCL departments. In that time I was one of the principle authors on the Decombinator TCR software package, which I have continued to work on since, along with other aspects of study across repertoire analysis more broadly.

During our PhD candidature Lisa and I frequently worked together, which included discussion of the most appropriate methods to use to analyse TCR repertoire data. This process included formal discussions as part of a larger UCL wide repertoire effort, attended by our supervisors (Benny Chain and Robin Callard) among others, as at this point there were not the abundance of software tools available as there is now.

I understand that this analysis is a point of issue in her PhD thesis examination. Remembering those earlier discussion, and having reviewed sections of her thesis now, I hoped I could add some context. At the time our PhDs were undertaken there were very few options for TCR analysis available: MiTCR (which has since been discontinued in favour of a subsequent package) and Decombinator being prime choices, and thus it was appropriate to test them both.

As I understand it, part of the conclusion of Lisa's thesis is that the capture technique is unsuitable for adaptive immune repertoire sequencing, due to many reads not sufficiently spanning the rearrangement due to their fragmentation. This property also explains the differences seen between MiTCR and Decombinator, as the former uses shorter and more rearrangement-proximal sequences to begin their gene calling procedure.

As the reviewers have noted, neither of these packages were designed with the kinds of data produced by the capture technique in mind – especially Decombinator, which was designed for longer reads – yet importantly nor have any of the subsequent packages. While the capture technique was a daring innovation with great potential, worthy of the efforts Lisa put into it, it seems that standard multiplex and RACE PCR techniques produce more appropriate data. I believe that the current analysis sufficiently demonstrates this, and thus re-analysis of the original sequence data would not be informative with respect to the aims of the thesis.

Yours faithfully,

James M Heather