T-cell receptor repertoire sequencing in health and disease

James Malcolm Heather

Division of Infection and Immunity

University College London

A thesis submitted for the degree of Doctor of Philosophy

2015

I, James Heather, 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.

Signature: _____

Date: _____

"In a varying environment, changes in the character of the population take place. In every instance we are concerned to understand (a) whence the change in character of the individual components arises and (b) the nature of the process by which the constitution of the population is modified to conform to the requirements of the changing environment."

Frank Macfarlane Burnet
From 'The Clonal Selection Theory of Acquired Immunity'.
1958

"... [A] knowledge of sequences could contribute much to our understanding of living matter."

Frederick Sanger
From the autobiography given to the Nobel Foundation upon receiving his second Prize.
1980

Abstract

The adaptive immune systems of jawed vertebrates are based upon lymphocytes bearing a huge variety of antigen receptors. Produced by somatic DNA recombination, these receptors are clonally expressed on T- and B-lymphocytes, where they are used to help detect and control infections and help maintain regular bodily function. Full understanding of various aspects of the immune system relies upon accurate measurement of the individual receptors that make up these repertoires. In order to obtain such data, protocols were developed to permit unbiased amplification, high-throughput deep-sequencing, and error-correcting bioinformatic analysis of T-cell receptor sequences.

These techniques have been applied to peripheral blood samples to further characterise aspects of the TCR repertoire of healthy individuals, such as V(D)J TCR gene usage and pairing distributions. A large number of sequences are also found to be shared across multiple individuals, including sequences matching receptors belonging to known and proposed T-cell subsets making use of invariant rearrangements. The resolution provided also permitted detection of low-frequency recombination events that use unexpected gene segments, or contained alternative splicing events.

Deep-sequencing was further used to study the effect of HIV infection, and subsequent antiretroviral therapy, upon the TCR repertoire. HIV-patient repertoires are typified by marked clonal inequality and perturbed population structures, relative to healthy controls. The data presented support a model in which HIV infection drives expansion of an subset of $CD8^+$ clones, which – in combination with the virally-mediated loss of $CD4^+$ cells – is responsible for driving repertoires towards an idiosyncratic population with low diversity. Moreover these altered repertoire features do not significantly recover after three months of therapy.

Deep-sequencing therefore presents opportunities to investigate the properties of TCR repertoires both in health and disease, which could be useful when analysing a wide variety of immune phenomena.

Acknowledgements

Many people have helped me throughout the course of my PhD in one form or another, for which I am truly thankful.

First and foremost among those is Professor Benny Chain, who in his supervision of me through this project has displayed all of the qualities that make up not just great scientists but great supervisors, qualities that I hope to emulate in my own career. A clerical error meant that I was not assigned a secondary supervisor. Benny meant that I did not need one.

Similar thanks also go to Dr Maddy Noursadeghi, my *de facto* secondary supervisor. The group environment fostered between the collaboration of Chain and Noursadeghi has proven immensely fruitful to my academic development, and I am grateful to have the opportunity to work in it for a little longer.

While I am indebted to many current and past lab members for help, there are two that merit particular mention. Dr Theres Oakes and Katharine Best have made immeasurable contributions throughout my PhD, as sounding boards and collaborators for the development of the wet lab and bioinformatic protocols respectively.

Very large thanks also go to Dr Dan Depledge and Tony Brooks, both of whom very patiently answered my silly questions and helped me find my feet in the sequencing work. Thanks also go to Professor Paul Kellam and his group for sequencing several of the early experiments in this thesis.

My innumerable discussions and frantic schematic-drawing-sessions with Samir Aoudjane must also be noted, for helping me actually realise that I might know what I was talking about. I would also like to thank Professor Av Mitchison for his encouraging words before and after graciously looking over this thesis.

Finally, I must thank my parents, my partner Christy, and all the rest of my friends and family for supporting me throughout. You'll all be very pleased to know that it's done now (well, almost), so I can finally come out and play. Just after the next paper is written.

Contents

List of Figures

List of Tables

Glossary

6N A random hexamer (i.e. six random bases) in an oligonucleotide sequence

AC Aho-Corasick

Amplicon The section of DNA amplified during an amplification reaction

Antigen A substance to which the adaptive immune system can respond

APMI As per manufacturer's instructions

ART Antiretroviral therapy

AS Alternative splicing

Barcode A random stretch of nucleotides incorporated before amplification to allow 'collapsing' of data back to original cDNA molecules

BCR B-cell receptor

BLR Barcoded ligation RACE (amplification protocol)

bp Base pair

BSA Bovine serum albumin

BV Barcoded V (amplification protocol)

CD Cluster of differentiation

cDNA Complementary deoxyribonucleic acid

CDR Complementarity determining region

DCR Shorthand for the five-part TCR classifier produced by `Decombinator`

`Decombinator` The software used to perform basic TCR analysis in this thesis, in which each TCR chain sequence is classified using a five-field identifier from which the entire nucleotide sequence can be recapitulated

DN Double negative

DNA Deoxyribonucleic acid

dNTP Deoxyribose nucleoside triphosphate

DOI Digital Object Identifier

DP Double positive

Epitope The part of an antigen to which a TCR , BCR or antigen binds

FACS Flow assisted cell sorting

Functional IMGT definition: A germline V(D)J gene predicted to be able to contribute to productive rearrangements

gDNA Genomic DNA

GEM T-cell Germline-Encoded Mycolyl-reactive T-cells

HF High fidelity buffer for Phusion polymerase

HLA Human leukocyte antigen

HSCT Haematopoietic stem cell transplantation

HTS High-throughput sequencing

Ig Immunoglobulin

IMGT ImMunoGeneTics; the organisation responsible for collating and annotating TCR and Ig information

Index One of a list of pre-determined nucleotide sequences that is incorporated to all amplicons of a given sample before sequencing, to allow multiplexing of many samples on one run

iNKT cell Invariant natural killer T-cell

JSD Jensen-Shannon divergence

kb Kilobase (1000 nucleotide base pairs)

L-PART1/2 The two sections of the leader sequence, which lie immediately upstream of each V gene and are spliced together in mature TCR mRNA

Library Set of DNA molecules prepared to be sequenced

m7G 7-methylguanosine

MACS Magnetic assisted cell sorting

MAIT cell Mucosal associated invariant T-cell

Mbp mega base pair (1,000,000 base pairs)

MHC Major histocompatibility antigen

mRNA Messenger ribonucleic acid

N N represents an ambiguous, unknown or random nucleotide

NEB New England Biolabs

NGS Next-generation sequencing

NKT cell Natural killer T-cell

NP (Non-productive) IMGT definition: A rearrangement predicted to not be capable of forming valid receptors

ORF IMGT definition: A germline V(D)J gene predicted to be unlikely to be able to contribute to productive rearrangements

P (Pseudogene) IMGT definition: A germline V(D)J gene predicted to not be able to contribute to productive rearrangements

P5 One of two sequences required for Illumina bridge amplification

P7 One of two sequences required for Illumina bridge amplification

PBMC Peripheral blood mononuclear cells

PCR Polymerase chain reaction

PE Paired end

PEG Polyethylene glycol

pMHC Peptide major histocompatability complex

Productive IMGT definition: A rearrangement predicted to be capable of forming valid receptors

RACE Rapid amplification of cDNA ends

Regex Regular expression; a sequence of characters used to find strings that match certain criteria

Rep-Seq Repertoire sequencing

RNA Ribonucleic acid

SBS Sequence-by-synthesis

SMS Single molecule sequencing

SP Single positive

SP1 Illumina's sequencing primer 1

SP2 Illumina's sequencing primer 2

SRA Sequence Read Archive

TCR T-cell receptor

TCR gene/segment A section of DNA that can be assembled through V(D)J recombination to produce a TCR

Tdt terminal deoxynucleotidyl transferase

TEMRA T effector memory CD45RA+

TRA T-cell receptor alpha

TRAJ T-cell receptor alpha chain J gene

TRAV T-cell receptor alpha chain V gene

TRB T-cell receptor beta

TRBJ T-cell receptor beta chain J gene

TRBV T-cell receptor beta chain V gene

TRD T-cell receptor delta

TRG T-cell receptor gamma

UCL University College London

vDCR `Verbose Decombinator`, the specific version of `Decombinator` developed during this thesis to permit error-correcting TCR analysis

w/v Weight per volume

1

Introduction

1.1 Adaptive immunity

Various mechanisms have evolved across the domains of life to protect organisms from the negative consequences of being infected or parasitised by other organisms. In vertebrates these defences are typically grouped into one of two functional distinctions. Germline-encoded mechanisms that provide generic barriers and rapid responses to conserved elements of infectious agents constitute innate immunity. Working in conjunction with these innate mechanisms is the adaptive immune system, that *de novo* generates novel protein-coding genes in certain lymphocyte cell types to provide defence from a potentially near-infinite array of different biological and chemical entities. While typically less rapid than innate mechanisms, adaptive responses develop immunological memory, leading to faster responses during subsequent exposures. Despite these differences, both innate and adaptive immune mechanisms involve the dynamic regulation of different tissues, cells and effector molecules.

One of the most important immunological questions is how the body distinguishes what is foreign – and potentially harmful – from its own component parts, many of which will be composed of the same basic chemical structures. This problem was famously addressed by Frank Macfarlane Burnet in 1949 with his theory of immunological tolerance (1). This theory states that during embryogenesis the developing immune system should essentially only be exposed to the 'self', i.e. the sum total of identifiable molecular entities expressed throughout the body. Developing animals become 'tolerant' to these self-features, and can therefore reserve immunological reactions to anything else that is 'non-self', thus targeting foreign material while limiting damage to themselves. While modern theories have moved on – not least due to the knowledge that the adaptive immune system does indeed mount responses towards self molecules – this concept of 'self' and 'non-self' remains central to our understanding of immune processes.

The mode of adaptive immunity that humans enjoy is shared by all gnathostomes (jawed-vertebrates), having evolved approximately 500 million years ago in our shared ancestors (2). In this system a series of DNA recombination events is used to generate

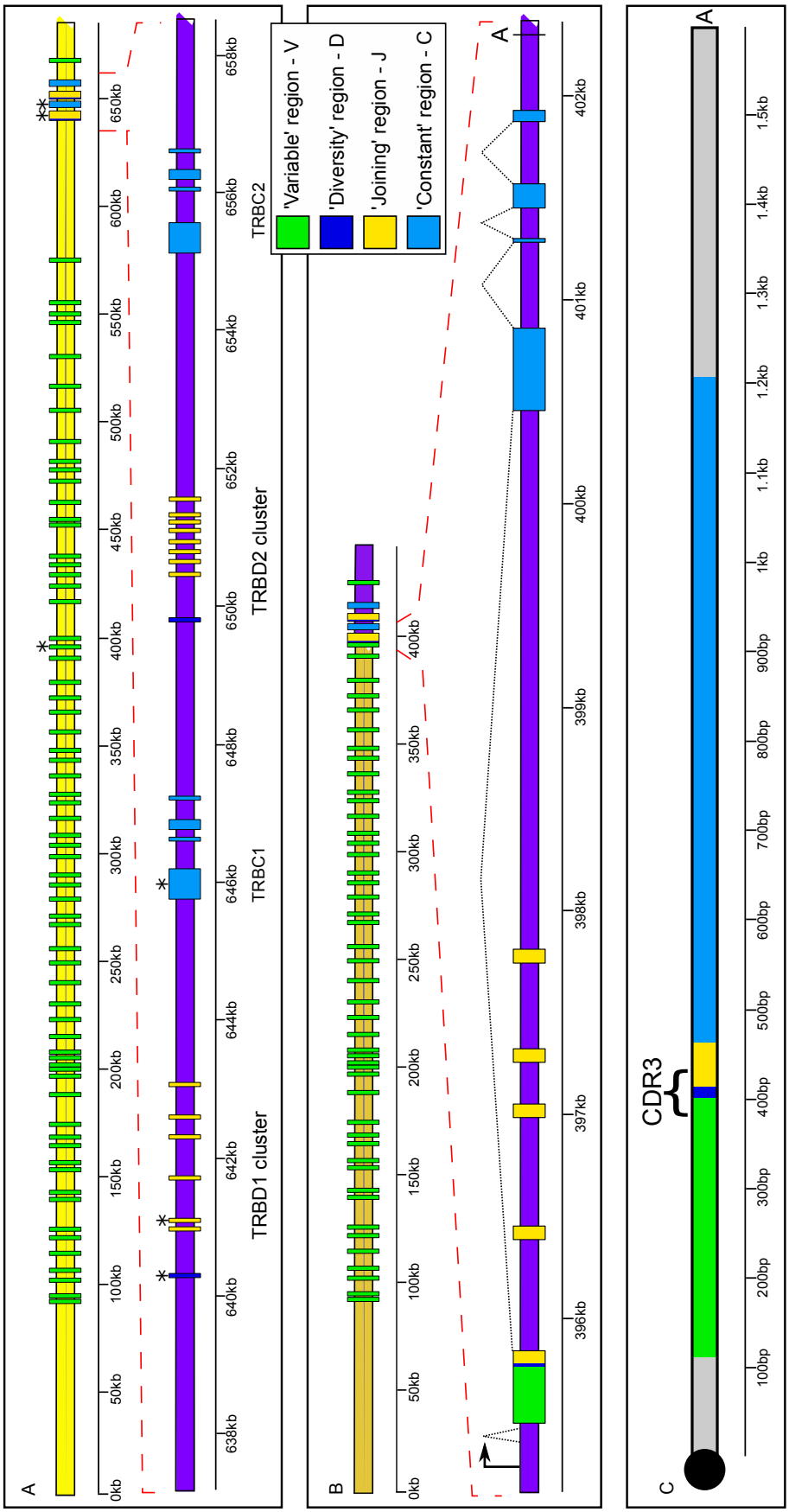
populations of cells bearing a wide-range of different receptors, which can detect a huge assortment of different molecules in order to trigger immune responses. These cells – B and T lymphocytes, or B-cells and T-cells – are derived from a common haematopoietic progenitor (3, 4), and express a variable recombined receptor based on their cell type. B-cells express B-cell receptors (BCRs) and secrete soluble variations as antibodies (or immunoglobulins), while T-cells express T-cell receptors (TCRs), which are the focus of this thesis.

1.2 V(D)J recombination

TCRs exist as surface-expressed glycoproteins composed of two polypeptide chains. In humans, most T-cells express a heterodimer consisting of an α and a β chain (although a second lineage accounting for 5% of total T-cells makes use of an alternative TCR, composed of a γ chain paired with a δ (5)). Unlike the majority of genes, TCRs (and BCRs) are not functionally encoded in the germline. Instead, the loci encoding each chain contain a number of non-contiguous sections of DNA that can be recombined together to form complete coding sequences. This somatic DNA recombination was first observed in B-cells (6), and was found to be due to a number of related gene segments (7, 8) bordered by specific recombination sequences (9, 10) being brought together.

These recombinable genes fall into one of three categories, or types: 'variable' (V), 'diversity' (D) or 'joining' (J). These can be physically joined together through the action of a recombinase complex which is recruited to conserved recognition signal sequences (RSS) at the gene boundaries (11). This complex is the product of two recombinase activating genes (RAG), RAG-1 and RAG-2, which work co-operatively to mediate recombination between RSS (12, 13). Each RSS consists of a conserved nonamer and heptamer nucleotide sequence, separated by either 12 or 23 nucleotides; the correct order of assembly is promoted by recombination only occurring efficiently between pairs of genes that are bordered by RSS with different length spacers (the '12/23 rule') (14). The complement of genes types are not equally distributed between chains, as only one per heterodimer contains a D gene in its rearrangement: in the TCR it is the β chain[1]. The distribution of RSS therefore directs recombination between Dβ and Jβ, and Vβ with Dβ (see Figure 1.1). There is only one type of recombination event in the α chain, that between a Vα and a Jα. However recombination at this locus is complicated as the panel of V genes is shared between both α and δ rearrangements;

[1]D regions are also components of TCR δ and BCR heavy chain rearrangements.


A
TRBD1 cluster
TRBC1
TRBD2 cluster
TRBC2
'Variable' region - V
'Diversity' region - D
'Joining' region - J
'Constant' region - C
B
A
C
CDR3

the δ locus resides within the α (see Figure 1.2) (15, 16). Following transcription, rearranged V(D)J sections are spliced onto a constant region, which is required for surface expression and intracellular transmission of TCR binding signals.

The consequence of using V(D)J recombination to produce variable antigen receptors is the breadth of molecules that it can produce. Recombinatorial diversity is generated by merit of there being multiple V, D and J genes, and from the final protein being a heterodimer of two recombined chains: without any other mechanism in place this system could produce over 2.8 million different TCRs[2]. However the majority of the potential diversity arises due to mechanisms that result in the non-templated addition and deletion of nucleotides at the recombination junction. The RAG enzymes catalyse formation of a DNA hairpin intermediate during recombination; action of the DNA repair enzymes recruited to this site can lead to exonucleolytic 'nibbling' away of nucleotides, and formation of 'P-nucleotides' (where non-symmetrical hairpin resolution causes insertion of a short palindromic repeat) at the recombining edges (22, 23).

However the greatest source of diversity is the non-templated addition of nucleotides ('N-nucleotides') (24) via the action of terminal deoxynucleotidyl transferase (Tdt) (25, 26, 27). Hypothetically, allowing up to ten non-templated additions and deletions at each recombination junction raises the potential diversity to more than 10^{16} possible different TCR heterodimers. While these diversities will not even be approached in an individual – not least because this theoretical estimate is orders of magnitude greater than the total number of eukaryotic cells in a body – they serve to illustrate how a few megabases of DNA are able to produce a vast array of antigen receptors that can be used to detect and protect from a far greater number of immune challenges than a static genome could.

[2] $(45\,V\alpha \times 50\,J\alpha) \times (48\,V\beta \times 2\,D\beta \times 13\,J\beta) = 2,808,000$. Values taken from IMGT/GENE-DB (19), considering only the prototypical allele (thus ignoring the effect of polymorphism), for those V(D)J genes predicted to be functional. Note that this value is two orders of magnitude greater than the number of genes in the rest of the human genome, the estimated value of which varies between 19,000 and 31,000 (depending on the definition of 'gene') (20, 21).

Figure 1.1 ***(preceding page)*****: V(D)J recombination in the human TRB locus. A**: Overview of the germline layout of the locus, with inset zoomed in on the DJC section. **B**: Whole locus view of the rearrangement found in the β chain of the Jurkat cell line (genes used marked with asterisks), with inset zoomed in on the rearrangement. **C**: Fully spliced mRNA of the Jurkat β rearrangement. 'A' represents the poly-A signal. Complementarity determining region 3 (CDR3) formed at the recombination site is marked.

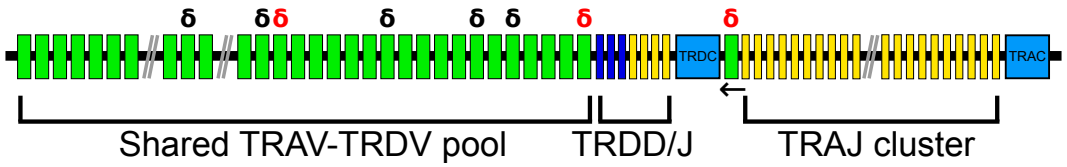


Figure 1.2: The human alpha-delta germline locus. Schematic representation of the human TRA/TRD locus. Successful recombination of an α chain prevents expression of a δ chain by excision. Genes are coloured as per the IMGT standard system (described later in Section 2.1): V genes are green, D genes dark blue, J genes yellow and constant regions lighter blue. Black 'δ' characters indicate a IMGT-designated TRA/DV gene, while red 'δ' characters indicate those which are supposed to only contribute towards delta chain rearrangements. Note the furthest 3' V gene is underscored by a leftwards arrow: this signifies TRDV3, which appears downstream of the TRD cluster, forcing it to recombine by inversion (as opposed to standard deletional joining) (17, 18). Double grey lines indicate omitted genes.

1.3 TCR structure, function and selection

T-cells broadly fall into one of two cell types that perform separate immune functions. Cytotoxic T-lymphocytes (CTLs) survey cells of the body for indications of viral infection or tumour development, and kill such potentially dangerous cells. In constrast, T helper (T_H) cells are non-cytolytic and release cytokines to modulate the local immune response of other cells. These contrasting abilities are both mediated by their TCRs binding to specific ligands. These cell types can generally be distinguished by the mutually exclusive expression of a glycoprotein co-receptor: CTLs express CD8 while T_H cells express CD4.

The exact nature of the TCR ligand was gradually revealed by a number of important experiments, starting with the discovery that CTLs from one animal could only lyse cells derived from an animal which shares a major histocompatibility complex (MHC) allele (29); TCRs are thus 'MHC-restricted'. Unlike antibodies, TCRs were not found to be able to bind 'native' antigens; instead their protein antigens must first be processed and presented by cells (30). In combination with the solved crystal structure of an MHC molecule[3] (31), the observation that addition of appropriate synthetic peptides to cells could cause T-cells to specifically activate against them (32, 33) lead to the realisation that TCR ligands consist of short peptides, bound in the membrane-distal groove of MHC molecules (34, 35) (Figure 1.3A). CD4 and CD8 contribute by binding specific families of MHC molecules: CD8 on CTLs binds to MHC class I, while CD4 binds to MHC class II, directing T-cell restriction (36). Further structural data revealed that the region of each TCR polypeptide chain encoded by the hypervariable DNA at

[3]Human leukocyte antigen (HLA)-A2.

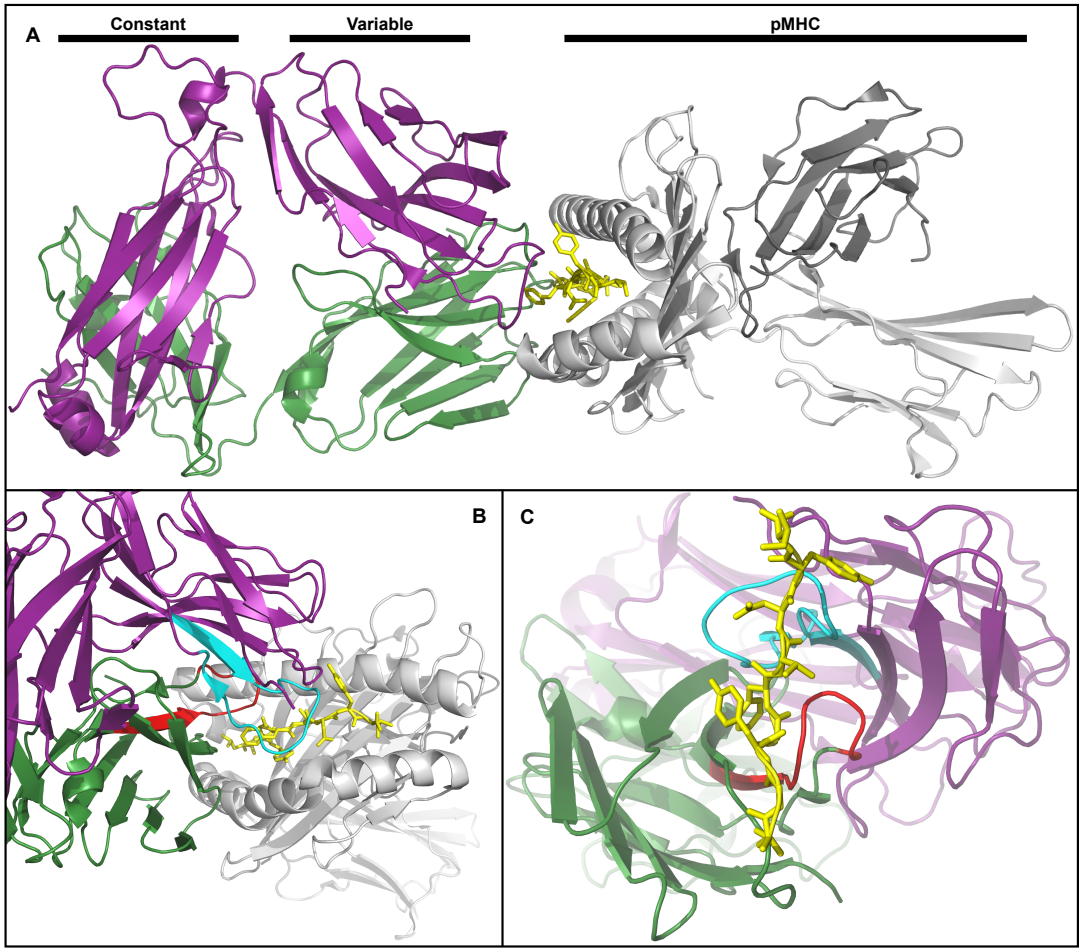


Figure 1.3: TCR binding to peptide-MHC (pMHC). Images of the A6 TCR recognising a self-peptide from neuronal protein HuD, bound in the context of the class I MHC HLA A2 (28). **A**: Whole structure profile revealing the peptide (yellow), bound in the groove of the MHC molecule (light grey, β2-microglobulin in dark grey). Constant and variable domains of the TCR are indicated above. α chain is coloured green, β chain is coloured purple. **B** and **C**: Demonstration of the proximity of the α and β chain CDR3 regions (coloured red and aqua respectively) with the MHC-bound peptide (note that C is taken from the point of view of the MHC, not shown). Crystal structure obtained via the PDB accession 3PWP, images made using `PyMol`.

the site of recombination – named complementarity determining region 3 (CDR3) – forms the primary contacts with the peptide bound in the MHC groove (Figure 1.3B and C), while other (germline encoded) variable sites (CDR1 and CDR2) usually make contacts elsewhere on the MHC (37).

T-cells are derived from lymphoid progenitor cells that originate in the bone marrow, which must travel to the thymus to develop[4]. Thymocytes begin their thymic development with unrearranged TCR gene loci, double negative (DN) with respect to the TCR co-receptors. As they navigate through various thymic compartments they undergo various genetic events and checkpoints including the rearrangement of their β chain, which must be confirmed by successfully pairing with an invariant pre-TCR to pass 'β-selection', and expression of both CD4 and CD8 during a double positive stage (DP), followed by recombination of their α chain (39). DP cells must then undergo two rounds of selection to ensure their TCR has the required properties, through interaction with thymic epithelial cells expressing a wide variety of self peptide-MHC (pMHC) antigens: cells must demonstrate the ability to bind to pMHC (positive selection), while not binding so strongly to self antigens so as to pose an autoimmune risk (negative selection) (40). DP cells that pass through thymic selection lose expression of one of the co-receptors, becoming CD4 or CD8 single positive (SP), and can leave the thymus to enter the periphery as a naïve T-cell.

This system of T-cell development therefore uses a somatic DNA recombination process to produce a diverse anticipatory array of different TCRs – the TCR repertoire – that stands to activate in response to antigens derived from a world of unknown and potentially pathogenic organisms.

1.4 TCR discovery

Study of the TCR began with its discovery in the early 1980s. By this time it was already known that different T lymphocyte clones (or 'clonotypes') could respond to specific antigens, producing the cellular immune response (41). However the quest to find the receptor responsible for this specificity proved difficult, perhaps hampered by initial expectations that such receptors would resemble the B-cell style immunoglobulins (42). In the space of one year, three groups had raised monoclonal antibodies against T-cell lines which recognised clonotype-specific surface-expressed disulphide-linked heterodimeric glycoproteins, in both mouse and man (43, 44, 45). These heterodimers

[4]Jacques Miller's discovery of this immune function for the thymus not only finally answered the long standing mystery of what role the organ fulfilled, but also began the chain of events that lead to the identification of T-cells and B-cells as separate lymphocyte lineages (38).

associated with CD3[5], which was already known to be a marker of T lymphocytes that promoted activation when targeted by antibodies (44). These clonotypic protein structures – that were of similar structure but comprised of different protein sequences – were found on T-cells, but not other cells such as B-cells, providing the first clue that these discoveries might correspond to the TCR (46, 47). These findings began the search for the transcripts and genes encoding these proteins. Two groups headed by Mark Davis and Tak Mak published back to back papers describing this achievement, using similar strategies working on the hypothesis that TCR genes would possess qualities of rearranged immunoglobulin genes (having both a variable, recombined region for antigen binding and a constant region for signalling) yet only be expressed in T-cells, not B-cells (48, 49). These assumptions informed experimental designs that allowed isolation of the first human and murine TCR sequences through a process of cDNA screening called subtraction[6]. Sequence analysis revealed homology to immunoglobulin genes (49, 50), as well as a comparable arrangement scheme consisting of a constant and a variable region, which was further divided into separate segments (51). The first TCRβ sequences had been found, beginning the modern age of TCR analyses. Two additional chains were cloned soon afterwards using similar techniques and proposed as potential binding partners of TCRβ (52, 53). Partial peptide sequences of the TCRβ-associated molecules revealed that one of these chains (the second to be discovered) was the partner of the β chain, and therefore the α chain (54)[7]. The remaining chain would become known as the γ chain, whose contribution to TCR receptors would remain unknown until pulsed-field gel electrophoresis experiments revealed an additional recombining locus nestled within that encoding the α chain (15), which was found to produce δ chains, the binding partner of γ.

[5]Then known as 'T3'.

[6]Subtraction consists of incubating labelled cDNA from the cell of interest (T-cell clones) with unlabelled RNA from a related cell (B-cells), which removes the sequences that are common to both by separating single and double stranded sequences.

[7]The chains had already been named arbitrarily upon initial discovery of the protein heterodimer, following electrophoresis under denaturing conditions: the larger chain was named α, while the smaller chain was named β (55).

1.5 Historical methods of investigating TCR repertoires

Having identified the individual TCR chains, the challenge then became how to study the distribution of different rearrangements within a sample. However, the complexity produced by V(D)J recombination presents a number of technical challenges to study. Various technologies have been employed to monitor the TCR repertoire with varying throughput, resolution, and capabilities.

Since their role in the discovery of the TCR proteins, monoclonal antibodies directed against specific V families have been widely used to measure T-cells bearing TCRs using particular V genes via flow cytometry (56). Flow cytometry is able to screen millions of cells for surface expression of specific Vs, both quickly and at relatively low cost (57). Modern flow cytometers can detect up to 18 colours, theoretically giving simultaneous recording of large swathes of the repertoire in one experiment; the advent of mass cytometry – in which mass spectrometry is used to detect antibody-bound rare-earth elements, instead of fluorescence detection via CCDs – could even allow detection of all Vs of a given chain (58). Cytometric techniques can be used to further characterise V-specific T-cells by staining for phenotypic and lineage cell markers. However, in practice even modern studies using flow cytometry tend to only monitor the β chain, using a panel of antibodies designed to cover the majority of TRBV genes (57, 59), and there are currently no published accounts of mass cytometry being employed for TCR repertoire studies. Furthermore cytometry can only provide information as to the frequency of cells using particular V genes in a population, providing no further information about the nature of the rearrangement. Therefore despite the large throughput, flow cytometry produces very low resolution data. The availability of antibodies covering the majority of V genes can also be limiting (47, 60), particularly in non-model organisms, and V gene polymorphisms may prevent binding.

The discovery of TCR genes was contemporaneous with the development of polymerase chain reaction (PCR) (61, 62); this technology was rapidly adopted by immunologists to quantitatively monitor the expression of V gene families (63, 64). PCR has the advantage of being able to selectively amplify nucleic acids bearing specific sequences, allowing one to target TCR chains at various levels. One can target and amplify individual clonotypes (to track leukaemic clones (65), for instance), individual V or J gene families (64) or even all TCRs within a chain. PCR has also facilitated the development of almost every other technique to measure TCR repertoires which uses genetic information.

One of the more widely used PCR-based techniques to study repertoires of variable immune receptors involves amplifying from RNA using specific V gene family and

constant region primers, before visualising the profile of CDR3 lengths produced by electrophoresis (66, 67)[8]. These 'spectratypes' give a great deal of information about the spread of clonalities within a V family; expansions of clones can be visualised as departures from the typical Gaussian-like distribution of CDR3 lengths. Spectratyping therefore provides a relative increase in resolution with respect to flow cytometry, providing information regarding clonal expansions within a V family. Variations upon spectratyping such as heteroduplex analysis (69) or single-strand conformation polymorphism analysis (70) exist that further separate amplicons based on their nucleotide conformational properties, which can offer resolution beyond CDR3 length distributions.

Having played a key role in the identification of TCR chains, sequencing of cloned TCR cDNA molecules remains an important means of assessing TCR identity and diversity (71). Sequencing produces the highest resolution data of all methods discussed thus far: nucleotide level information, often for the complete variable region. This information typically comes at the cost of throughput, as cloning and sequencing are relatively time, labour and expense intensive. However, cloning and sequencing from single cells allows extraction of paired α and β chain sequences, giving the actual complete clonotype of each cell (72, 73).

1.6 TCR repertoire analysis by deep-sequencing

As with the emergent technologies described above, the field of immunology has begun to adopt high-throughput sequencing (HTS) (described in detail in Chapter 5) in order to study TCR repertoires. These technologies offer both high-throughput and high-resolution data, as they can produce nucleotide sequence information for thousands to millions of clones in one experiment. Whilst still a technique in its infancy, repertoire sequencing (or 'RepSeq' (74)) has revealed a number of interesting features of T-cell biology.

The earliest TCR deep-sequencing papers set out to parameterise the TCRβ repertoires of healthy individuals, measuring clonotype frequencies, CDR3 length and V(D)J usage (75, 76). Early reports used such data to estimate the lower bound of number of distinct peripheral TCRβ rearrangements. Values ranged from one to four million sequences (76, 77), in agreement with the oft-cited estimate from Arstila *et al* (78), although at least one group claims it could be two order of magnitude greater (79). A

[8]It is also possible to perform this procedure using gDNA as a template, but this requires using primers directed against the J genes instead of the constant region, due to the presence of the large J-C intron(68).

profound decrease in repertoire diversity has also been reported to occur with increased age, consistent with findings from previous technologies (79, 80). The repertoires of T-cell responses to Epstein-Barr Virus (EBV) and cytomegalovirus (CMV) have been found to be very stable, with the hierarchy of expanded sequences being maintained over several years (81).

TCR RepSeq has also shown its use in answering various other immunological research questions. For instance, $\alpha\beta$ and $\gamma\delta$ T-cells were sorted from healthy donors and both TCRβ and γ chains were sequenced; while only a small number of β sequences were present in the $\gamma\delta$ cell libraries, recombined γ chain sequences were found at high frequency in $\alpha\beta$ populations, supporting the hypothesis that TCRγ rearrangement occurs prior to T-cell lineage commitment, while TCRβ rearranges afterwards (82). Similarly, the presence of some sequences in multiple different T-helper subsets of the same individual argues against the possibility that TCRs are deterministic with regards to T-helper fate (83). Extensive profiling of repertoires of multiple different tissues from organ donors reveals considerable compartmentalisation of different clonotypes to different tissues (84). Patients with various immunodeficiencies, including aplastic anaemia (85), Wiskott-Aldrich syndrome (86, 87), and Omenn syndrome (88) have been sequenced, invariably revealing restriction or skewing of the TCR repertoire relative to healthy controls.

One of the first major clinical applications of TCR RepSeq is in identifying and tracking the clonal sequences of T-cell derived leukaemias and lymphomas (89, 90, 91, 92), as the clonal rearrangement of the original cell acts a specific marker of neoplastic cells. Different groups have employed sequencing of β and/or γ chains, depending on the nature of progenitor cell's lineage [9]. Deep-sequencing has shown itself to be a particularly sensitive tool for detecting minimal residual disease (MRD), i.e. the continued presence of neoplastic cells during therapy, out-performing the existing techniques of flow-cytometry (89) allele-specific oligonucleotide PCR(90), and (in a model system of MRD) capillary sequencing (96). In a recent study by Logan *et al*, MRD detected by HTS was completely predictive of relapse to disease following haematopoietic stem cell transplantation (HSCT) (95). In addition to detecting MRD, TCR sequencing produces other data that could be of diagnostic or prognostic value in cancer patients, such as monitoring the dynamics of the repertoire fluctuation on treatment.

Measuring immune system reconstitution following HSCT is another important clinical application of repertoire sequencing, which has been explored beyond treatment of

[9] Immunoglobulin rearrangements have also been sequenced to track B-cell neoplasms (90, 91, 93, 94, 95).

cancer. One patient with ankylosing spondylitis who received HSCT has been extensively studied in order to evaluate immune recovery (97, 98). Not only did sequencing reveal the emergence of a skewed but stable repertoire, but a number of clones that were present at low frequency pre-transplantation were found to be both retained and expanded at later timepoints. A later study aimed to test whether whether TCR sequencing might stratify responses in patients receiving transplants from different stem cell sources (99). Six months after transfusion, cord blood stem cell recipients developed repertoires with diversities approaching that of healthy donors, while those who received stem cells derived from adult peripheral blood were far less diverse, potentially leaving them at greater risk of opportunistic infections and cancer relapse. Graft-versus-host disease (GVHD), a major risk factor for HSCT patients, has also been investigated with HTS. Steroid-refractory patients showed greater repertoire conservation between samples from multiple gastrointestinal (GI) sites (100). Moreover, clones identified in these GI samples were found to be expanded in longitudinal peripheral blood samples. Therefore TCR RepSeq potentially offers both a means of quantifying immune reconstitution in HSCT patients and stratifying those who might fail to respond to therapy.

1.7 Aims of this thesis

TCR repertoire sequencing represents a powerful tool that can be applied in the study of diverse immunological and clinical phenomena. However it is still a developing technology, and its full potential is yet to be fully realised. Therefore the first aim of this thesis was to develop protocols for the amplification, high-throughput sequencing and bioinformatic analysis of TCR repertoires from human blood samples. These protocols should be robust, repeatable, and incorporate as little bias as possible. Having established repertoire-sequencing techniques, the next aim was to apply them to blood samples taken from healthy volunteers. This should allow both testing of the capabilities and limitations of the method, as well as produce an understanding of the nature of the repertoire features in healthy adult individuals using such high-resolution data. The final aim of this thesis was to sequence the repertoires of patients suffering from a clinically relevant infection, to assess the merit of using such technologies to understand the effect of infection upon the immune system. As a pathogen of global clinical importance, that is known to have a large impact on the T-cell compartment, human immunodeficiency virus (HIV) was chosen as a suitable model. Patient samples were to be sequenced and compared to those of healthy, uninfected donors, to allow measurement of the differences between the two groups.

2

Materials and methods

2.1 The international ImMunoGeneTics (IMGT) information system and nomenclature

The IMGT system was established in the Montpelier Laboratoire d'Immuno Génétique Moléculaire under the direction of Professor Marie-Paule Lefranc, to provide standardised references for immunogenetic and immunogenomic research (101). I have made frequent use of IMGT system throughout the generation of my data and the writing of this thesis. All annotated TCR gene sequences were obtained from the IMGT database IMGT/GENE-DB (19), and IMGT/LocusView was very useful in the production of a number of schematics in this thesis.

The IMGT-defined and Human Genome Organisation (HUGO)-approved TCR nomenclature has been used throughout this document (102, 103). In this system, TCR (and immunoglobulin) loci are first divided into gene types, i.e. V, D and J[1] Any genes within a given locus of a species that share 75% nucleotide sequence identity with each other are designated as subgroups. Specific genes can furthermore exist as one of several alleles. Genes are referred to with identifiers that incorporate all of this information. For example, TRAV8-4*01: TR indicates a TCR gene, AV indicates that it's an α chain V gene; 8-4 signifies that it's the fourth gene in the eighth subgroup in that locus, while *01 indicates that it's the prototypical allele of that gene[2].

Where possible, I have also adopted the IMGT colour conventions for use in schematic illustrations, as detailed in the IMGT Scientific Chart (104, 105). Accordingly, V genes have been coloured green (hexadecimal colour code #00EE00), D genes dark blue (#0000E4), J genes yellow (#FFE400), constant regions light blue (#0099FA), N nucleotides red (#FF0000) and where featured, leader sequences in light yellow (#FFFF00 and bordered in black to distinguish from J sequences).

[1]The recombinable DNA segments are referred to as 'genes' under IMGT terminology, for both practical and biological reasons. Practically, as labelling them as genes was a requirement for them to be indexed in various databases, and biologically on the grounds that each can be transcribed independently as they have their own promoters.

[2]Note that in this system variable antigen receptor gene names are not italicised, in contrast to the standard convention for human genes.

2.2 General considerations and RNA extraction

In order to minimise the risk of contamination or biohazards, all molecular or cell biology was undertaken with appropriate control measures. Handling of nucleic acids prior to amplification was undertaken in UV- and chemically-treated PCR hoods, and highly-concentrated PCR or cloning products were purified in separate rooms, to minimise the potential for DNA contamination. Peripheral blood and cell lines were manipulated in laminar flow hoods. Bacterial cloning was carried out at the bench under asceptic conditions.

RNA samples were obtained from consenting adults in one of two ways. RNA for protocols that did not use molecular barcoding (i.e. simple V and poly-T amplification), was derived from peripheral blood mononuclear cells (PBMC) or sorted cell fractions, obtained by density centrifugation of up to 120 ml of peripheral blood using Lymphoprep (Axis-Shield). PBMC were then lysed in RLT buffer (Qiagen) and RNA extracted using the Qiagen RNeasy kits and eluted in 60 μl of water. RNA for the protocols using random barcodes (BV and BLR amplification) was extracted by other researchers (see relevant sections) by drawing 2.5 ml of peripheral blood into Tempus tubes (Life Technologies) and eluting RNA in approximately 90 μl of water. All cell lysates, RNA and Tempus tubes were stored at -80°C, while cDNA and amplified DNA products were stored at -20°C. RNeasy and Tempus extractions, and Lymphoprep separation were performed as per the manufacturer's instructions (APMI).

2.3 Amplification and deep-sequencing protocols

The development and application of the following protocols made up the major component of all practical molecular biology undertaken during this body of work; the rationale and details of this development is covered in detail in Chapter 5. Primer sequences for all protocols are detailed below in Section 2.3.7.1.

2.3.1 Reverse transcription

After extraction, residual genomic DNA (gDNA) was removed from RNA samples using commerically available DNase I (or derivative engineered forms). For all protocols except the BLR protocol, RNA was incubated with 0.1 U/μl of RQ1 DNase (Promega) in 1X reaction buffer for 30 min at 37°C. Samples processed using the BLR protocol had gDNA removed using the TURBO DNase kit, and globin mRNA was depleted using GLOBINclear[3] (both Life Technologies, APMI).

[3]Globin message was removed so that this RNA could be used for additional microarray studies.

Reverse transcription was carried out using oligonucleotides named αRC2 and βRC2, which are complementary with sequences towards the 5' end of the α and β constant regions respectively. βRC2 targets a region that is conversed between the two human β constant regions, TRBC1 and TRBC2. RNA (typically 500 ng) was mixed with 1 μl of each RC2 primer (10 μM) and 1 μl of a 10mM dNTP mix and heated at 65°C, 5 min and cooled rapidly on ice. 1 μl of DTT (0.1 μM), 1 μl of RNasin (RNase inhibitor, 20-40U/μl, Promega), 1 μl of SuperScript III reverse transcriptase (RT) (200 U/μl, Life Technologies) and 4 μl of 5X FS buffer was added, before incubation at 55°C, 20 min, then 70°C, 15 min for heat inactivation. Complementary DNA (cDNA) from RT reactions was purified and eluted into either 10 μl or 30 μl of buffered Tris-Cl using MinElute or QIAquick columns respectively (Qiagen).

2.3.2 Simple V amplification

Primers were designed against TRAV8-4 and TRBV12-3 (named αF1 and βF3) and the very 5'-most end of TRAC and TRBC (αRC1 and βRC1). The two β chain constant region genes differ at two positions in this region, therefore βRC1 contains two degenerate bases. Also note that while αRC1 ends at the most 5' nucleotide of the constant region, the beta primer actually overlaps with the last nucleotide of the J gene, as all TRBJ genes end with the same nucleotide.

1 μl reverse transcribed cDNA from lysed PBMC or sorted cells underwent a multiplex PCR using all four primers (αF1, βF3, αRC1 and βRC1). Reagent final concentrations in 50 μl reaction volumes were: 1X GoTaq Green buffer (Promega); 0.1 mM each dNTP; 0.2 μM each primer; 0.5 mM additional $MgCl_2$ and 2.5 U GoTaq DNA Polymerase (Promega). PCR mixtures were subjected to 30 cycles of: 94°C, 30 sec; 60°C, 2 min; 72°C, 1 min, with a 1 min final extension. PCR products were visualised on 1% agarose gels, and gel purified before submission for 100 bp paired-end (PE) Illumina HiSeq sequencing. Sequencing was performed at the Wellcome Trust Sanger Institute, in collaboration with Professor Paul Kellam's group.

2.3.2.1 Magnetic assisted cell sorting and flow cytometry

PBMC from two donors – D3 and D4 – were sorted into $CD4^+CD45RA^+$ and $CD4^+CD45RO^+$ populations (broadly representative of $CD4^+$ naïve and memory cells) prior to lysis, using negative-selection magnetic assisted cell sorting (MACS). Negative-selection refers to the targets of the selection: antibodies are used to target beads to cells other than those of interest by binding surface antigens that they do not express, and collecting unbound enriched desired cells from the eluate.

After isolation, PBMC were divided in two and subjected to either the Naïve or Memory $CD4^+$ T cell Isolation Kit (Miltenyi Biotech). Cells were counted and suspended in 40 μl of MACS buffer, before adding 10 μl of relevant antibody cocktail and incubating at 4°C for 10 min, after which the protocols diverge.

- Naïve: cells were washed in 20 ml of MACS buffer, spun down and supernatants removed. Cells were then resuspended in 80 μl of buffer before adding 20 μl of microbeads, before mixing and incubation at 4°C for 15 min. Another wash with 20 ml of MACS buffer followed, before cells were spun down and supernatants removed. Sorted cells were resuspended in 500 μl of buffer.
- Memory: 30 μl MACS buffer was added, followed by 20 μl of microbeads, before mixing and incubation at 4°C for 15 min. Cells were then washed in 20 ml buffer, spun down and supernatants were discarded. Sorted cells were finally resuspended in 500 μl buffer.

Cells underwent two rounds of negative selection to achieve greater purification, which was assessed by flow cytometry. Staining was performed in PBS with 10% goat serum. The following antibodies were used: CD3-FITC (Miltenyi), CD4-RPE (Dako), CD45RO-PeCy7 (BD) and CD45RA-APC (BD). Cells were incubated with antibodies at 4°C for a minimum of 30 min, before being washed and resuspended in a suitable volume. Flow cytometry was performed on a FACS Calibur (BD Biosciences) and analysed in `FlowJo`.

2.3.3 Poly-T RACE amplification

Rapid amplification of cDNA ends (RACE) was performed by addition of an adenosine tail to cDNA followed by ampification using a poly-T containing primer. First, 5 μl of cDNA was polyadenylated using terminal deoxytidyl transferase (Tdt, Promega). Final concentrations were: 1X Tdt reaction buffer; 0.4 mM ATP and 30 U Tdt (Promega). Reagents were mixed and incubated at 37°C for one hour, heat inactivated at 75°C for 15 min and purified using QIAquick PCR purification columns (Qiagen).

5 μl of poly-adenylated cDNA was then amplified using a combination of two different poly-T containing oligonucleotides (αL2 and βL2), and the constant region primers αRC1 and βRC1. Unequal representation of TRBJ genes from the two TRBD clusters in the 454 data was observed, leading to the degenerate βRC1 oligonucleotide being replaced with an equimolar pool of two primers that specifically target each TRBC gene (βRC1.1 and βRC1.2). This allows for greater control of the concentration of each

desired oligonucleotide sequence in the mixture, and removes the additional unwanted DNA species that would be produced in the original degenerate βRC1 formulation.

Final concentrations for the poly-T PCR were: 0.2 mM each dNTP; 0.5 μM each primer; 1 U Phusion polymerase (New England Biolabs, or NEB) and 1X HF buffer. Due to the reported strong 3'→5' endonuclease activity of Phusion, all reagents save enzyme and buffer were mixed separately first, followed by rapid addition of a mixture of enzyme in buffer on ice immediately before commencing thermal cycling, APMI. Thermal cycling conditions were: 98°C, 3 min; 54°C, 1 min; 72°C, 1 min, then 35 to 40 cycles of: 98°C, 15 sec; 70.7°C, 30 sec; 72°C, 20 sec, before final extension at 72°C, 5 min. Amplified TCRs were then selected and purified by gel extraction from 1-2% agarose gels. Amplicons for 454 sequencing were sent to the Wellcome Trust Sanger Institute for sequencing on the GS FLX instrument.

2.3.3.1 Poly-T RACE MiSeq library preparation

A second PCR was developed to add the required Illumina sequences specifically to the ends of the poly-T amplicons to allow processing on the MiSeq. Sequencing Primer 1 and 2 (SP1 and SP2) – the binding sites which define the start of the two sequencing reads – must reside immediately adjacent to amplicons sequences. P5 and P7 sequences are required immediately adjacent to the sequencing primers, as these bind to the flow-cell bound oligonucleotides to facilitate bridge amplification. Therefore I designed a set of primers to add each of these elements in one reaction by adding the innermost primers at limiting dilutions: as the reaction progresses these should be outcompeted by the higher concentration outermost primers, promoting the assembly of full length amplicons. SP1 and SP2 were added to the constant regions and poly-T sequences respectively with long primers SP1-N6-αRC1 or SP1-N6-βRC1.1+2 (where N6 represents a random hexamer to augment nucleotide diversity), and SP2-pT. P5 and P7 elements were then added to these by extension with long primers P5-SP1 and P7-X-SP2, where X is a six-base index to allow sequencing of multiple samples. This index is read in a separate read from a primer targeting the reverse complement of the SP2 sequence. Finally, two primers at the highest concentration target the outermost sequences: a 3' shortened version of P5 (P5s, or short) and P7. α and β reactions were carried out separately.

Final concentrations were: SP1-N6-α/βRC1 and SP2-pT, 0.05 μM; P5-SP1 and P7-X-SP2 0.1 μM; P5s and P7, 0.5 μM; 0.2 mM each dNTP, 1 U Phusion and 1X HF buffer (enzyme in buffer added last). Thermal cycler steps were: 98°C, 3 min; 3 cycles of 98°C for 15 sec, 56°C for 45 sec, 72°C for 45 sec; then 17 cycles of: 98°C, 15 sec; 69°C,

30 sec; 72°C, 30 sec; final extension at 72°C, 5 min. The first three cycles with lower annealing temperatures allow annealing of homopolymer RACE sequences, which have low melting temperatures. Amplified TCRs were purified by gel extraction from 1-2% agarose gels and processed for MiSeq sequencing.

2.3.3.2 Cloning

Initial clones produced to test the efficacy of the Tdt-RACE PCR were produced through blunt-ended cloning of PCR products into pUC19. Plasmid DNA was linearised using the restriction enzyme SmaI and dephosphorylated using calf intestinal alkaline phosphatase (both NEB, APMI) to prevent recircularisation, before gel excision and purification. Phusion DNA polymerase produces blunt-ended products and therefore required no further treatment post-purification. Ligation reactions using 2 U T4 DNA ligase (Promega), 1X reaction buffer and varying ratios of purified insert and linearised vector DNA were incubated overnight at 4°C. 10 μl of ligation mixtures were used to transform 50 μl of α-Select Silver competent *E. coli* (Bioline). Ligation mixes were added to bacteria (thawed on ice), before being briefly flick-mixed and incubated on ice for 30 min. Cells were then heat-treated (42°, 30 sec), immediately cooled on ice (2 min) and diluted in 1 ml antibiotic-free lysogeny broth (LB, Sigma-Aldrich) before shaking at 37°C, 220 rpm for 1 hour. Cells were then spread on AIX LB agar plates (0.1 g/ml ampicillin, 0.04 g/ml IPTG and 0.03 g/ml X-gal) and incubated overnight at 37°C. Potential recombinants were selected by blue/white screening and digested to check for the presence of inserts.

Clones generated to test for the correct incorporation of MiSeq elements were ligated into pGEM-T-Easy (Promega) to allow TA-cloning, facilitated by amplification using GoTaq polymerase (which produces amplicons with a single adenosine overhang). Purified amplicons produced from the Phusion reaction were re-amplified using GoTaq. PCR mixtures consisted of: 1X GoTaq Green buffer (Promega); 0.1 mM each dNTP; 0.2 μM of P5s and P7 primers; 0.5 mM additional $MgCl_2$ and 2.5 U GoTaq DNA Polymerase (Promega). PCR mixtures were subjected to 35 cycles of: 95°C, 30 sec; 64°C, 30 sec; 72°C, 1 min, with a 5 min final extension. PCR products were gel-purified and ligation into pGEM-T-Easy in an equivalent protocol as used above.

All plasmids were Sanger sequenced by the Wolfson Institute for Biomedical Research DNA Sequencing Service at UCL[4], using AB capillary sequencing technology. Sequencing traces were processed using `FinchTV` (Geospiza); poor quality bases were trimmed before analysis.

[4]Now defunct.

2.3.4 Randomly barcoded V (BV) amplification

The BV srategy targets two more V genes than the simple V amplification protocol (described in Section 2.3.2), while improving upon the MiSeq library preparation protocol developed for the poly-T RACE (Section 2.3.3.1). Long primers were designed to add SP2 elements and random hexamers to four specific V genes: those targeted in the simple V amplification, TRAV8-4 and TRBV12-3 (those used by Jurkat), and TRAV21 and TRBV20-1 (which had both consistently appeared at high frequency in previous experiments). These primers were used to peform a second strand extension on our cDNA, making double stranded DNA. A further extension reaction step – which I refer to as a 'third strand' reaction – then uses long oligonucleotides to add SP1 and random hexamer elements immediately downstream of the constant region sequence, allowing bidirectional sequencing through two random sequences, which can be used as molecular barcodes. Two further PCRs are then performed, the first adding P5 and P7 elements, the second amplifying to required concentrations. AMPure XP bead purifications (Agencourt Bioscience) are performed in between each round of DNA polymerisation using 1:1 bead:solution ratios to remove primers (APMI), eluting into 30 μl of water each time. These purifications make use of the propensity of high concentrations of polyethylene glycol (PEG) and salt to flocculate DNA (106) onto carbonyl-coated paramagnetic beads, allowing contaminants or low-molecular weight DNA to be removed by washing.

5 μl of cDNA was carried into 25 μl separate α or β second strand reactions, using: 0.2 mM each dNTP; 0.5 μM of each chain-specific primer pair (SP2-N6-TRAV21 and SP2-N6-TRAV8-4 or SP2-N6-TRBV12-3 and SP2-N6-TRBV20-1); 0.4 U Phusion, and 1X HF buffer. Reactions underwent the following program: initial denaturation at 95°C, 3 min, followed by a slow ramp to 80°C, 10 sec and another slow ramp to 58°C, 45 sec, and final extension for 5 min at 72°C. Slow ramping involves descending at 0.5°C/sec; this slow descent aims to encourage optimal annealing of the V region specific component of the long oligonucleotides to anneal at the highest possible temperature, minimising the chance of off-target priming. After purification, all 30 μl of double-stranded DNA is carried into 50 μl third strand synthesis reactions. These use the same reagent concentrations and cycling conditions as the second strand reaction, save that the SP2 primers are replaced by SP1-adding primers SP1-N6-I-αRC1 or SP1-N6-I-βRC1.1+2, where 'I' represents a second indexing sequence, allowing greater flexibility when planning sample multiplexing.

All of the purified third strand DNA undergoes 'PCR1' to add the P5 and P7 elements: 0.2 mM each dNTP; 0.5 μM of each primer (P5-SP1 and P7-X-SP2) and

1 U Phusion in 1X HF buffer. Mixtures were then thermally cycled: initial denaturation occurred at 95°C for 3 min, before ramping slowly to 69°C, 15 sec, and extended at 72°C, 1 min, before 3 cycles of 98°C, 10 sec to 72°C, 1 min. Final extension was at 72°C for 5 min. DNA was purified and all carried into the final reaction, 'PCR2', to amplify sufficient product for sequencing. Reaction conditions were the same as in PCR1, save that the primers used were just the outermost sequences, P5s and P7. PCR program began with initial denaturation at 95°C, 3 min, before 22 cycles of: 98°C, 10 sec; 69°C, 15 sec; 72°C, 40 sec, before final extension at 72°C for 5 min. Samples thus underwent a total of 26 cycles of PCR. Samples were then bead purified before dilution to equilibrate concentrations.

2.3.5 Barcoded ligation RACE (BLR) amplification

For each donor sequenced using the BLR protocol, 500 ng of RNA was reverse transcribed and purified using MinElute RNeasy columns (Qiagen). Half of the purified solution (5 μl of the 10 μl produced) was used in each experiment. The SP2 sequence, followed by a hexamer of random bases, was ligated to the V region end of the cDNA. T4 RNA ligase was used to ligate single stranded oligonucleotides to cDNA in a 30 μl reaction mix containing: 1X T4 RNA ligase buffer; 20 U T4 RNA Ligase 1 (NEB), 0.33 mM dATP; 25% (w/v) PEG 8000; 1 mM hexamine chloride; 0.1 mg/ml acetylated-BSA, and 0.33 μM SP2-6N ligation oligo. Master mixes were vortexed for several seconds to ensure mixing with the PEG, added to cDNA and mixed thoroughly. Ligations were incubated for 23 hours at 16°C, heat inactivated for 10 min at 65°C, diluted in 70 μl water and bead purified.

All purified ligated DNA was subjected to second strand synthesis, using just the SP2 sequence as the priming target. Final concentrations in the 50 μl reactions were: 0.2 μM of each dNTP; 0.5 μM SP2 primer and 1 U Phusion, in 1X HF buffer. The single strand extension program was the same as that described for the BV protocol in Section 2.3.4. Purified double-stranded samples were split in two and each half underwent either α or β chain specific third strand synthesis. Reaction conditions were the same as the second strand reaction, except the SP2 primer was replaced with either SP1-6N-I-αRC1 or SP1-6N-I-βRC1. PCR1 and PCR2 were then performed to add flowcell-binding elements and amplify as in Section 2.3.4, with the exception that PCR2 instead contained 23 instead of 22 cycles. BLR samples thus underwent a total of 27 PCR cycles. Samples were purified before normalisation.

Blood samples from HIV patients for BLR sequencing were obtained by the research nurse team based in the Mortimer Market Centre Clinic, Bloomsbury, and RNA

extraction was performed by Dr. Eleanor R. Gray. Samples from healthy donors were obtained and had RNA extracted by Drs. Nandi Simpson and Jennifer Roe. RNA in both cases was extracted from approximately 2.5 ml of whole peripheral blood lysed in Tempus tubes (Life Technologies), APMI.

2.3.5.1 Flow assisted cell sorting

PBMC from two healthy donors were sorted into $CD4^+$ and $CD8^+$ populations by FACS. Cells were isolated through density-gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare Lifescience) and incubated in 10% goat serum at 4°C for 30 min to block Fc-receptors. Cells were subsequently stained with CD8-FITC, CD3-PE, CD4-PerCp-Cy5.5 (BD Biosciences) and LIVE/DEAD fixable far-red dead cell stain (Life Technologies) APMI. $CD3^+CD4^+CD8^-$ and $CD3^+CD8^+CD4^-$ T-cells were sorted on a BD FACS Aria machine (BD Biosciences). 7.2×10^6 $CD4^+CD8^-$ cells and 6×10^6 $CD8^+CD4^-$ T-cells were lysed in RLT buffer (Qiagen) prior to RNA extraction. Cell sorting was performed by Dr. Theres Oakes, at the UCL Flow Cytometry Core Facility.

2.3.5.2 HIV patient criteria

Adult HIV+ HAART-naive patients due to start treatment imminently were recruited to the study. Exclusion criteria included: febrile or AIDS-related illness; tumours; coinfection with Hepatitis B or C; immunomodulatory therapy, or recent vaccination. Interruption or adverse reactions to ART were further grounds for exclusion.

2.3.6 MiSeq operation

Apart from initial libraries amplified using the poly-T protocol (which were sequenced commercially via UCL Genomics), all libraries were sequenced by myself. Purified amplified samples were first normalised to a suitable concentration. Concentrations were assessed using a Qubit fluorometer, using the dsDNA HS Kit (Life Technologies). Approximately 2 ng of DNA per sample was then inspected on a Bioanalyzer using the High Sensitivity DNA Analysis Kit (Agilent Technologies), allowing determination of amplification success and peak band size. Amplicon length and concentration were then used to normalise samples to 4 nM before pooling. Typically 10 to 12 amplified samples (α and β samples from five to six donors) were pooled per run. MiSeq flowcells were prepared as per Illumina's instructions (107). Denatured libraries were loaded at 11-12 pM, using a 1% PhiX spike in. 2x250PE kits with version 2 chemistry were used.

Read quality was assessed using in-built MiSeq metric software in combination with `FastQC` (108).

2.3.7 Primers and indexes

2.3.7.1 Primer sequences

Primer	Sequence
αRC2	GAGTCTCTCAGCTGGTACACG
βRC2	ACACAGCGACCTCGGGTGGGAA
αF1	AGCTTGGCAGCCACGTCTCTG
βF3	CCAATTTCAGGCCACAACTCC
αRC1	ACGGCAGGGTCAGGGTTCTGGATAT
βRC1	GGTGGGAACACSTTKTTCAGGTCCTC
αL2	TCACACAGGAAACAGCTATGACTTTTTTTTTTTTTTTTTTTTTVN
βL2	GTCATAGCTGTTTCCTGTGTGATTTTTTTTTTTTTTTTTTTTTVN
βRC1.1	GGTGGGAACACCTTGTTCAGGTCCTC
βRC1.2	GGTGGGAACACGTTTTTCAGGTCCTC
SP1-N6-αRC1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNACGGCAGGGTCAGGGTTCTGGATAT
SP1-N6-βRC1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNGGTGGGAACACCTTGTTCAGGTCCTC
SP1-N6-βRC1.2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNGGTGGGAACACGTTTTTCAGGTCCTC
SP2-pT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTVN
P5-SP1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7-X-SP2	CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
P5s	AATGATACGGCGACCACCGAGATC
P7	CAAGCAGAAGACGGCATACGAGAT
SP2-N6-TRAV8-4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNTCATCGTCTGTTCCACCATATCTCTTCTG
SP2-N6-TRAV21	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNAGACTTAATGCCTCGCTGGA
SP2-N6-TRBV12-3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNATGCCCGAGGATCGATTCTCAG
SP2-N6-TRBV20-1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNTTTGGTATCGTCAGTTCCCG
SP1-N6-I-αRC1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNxxxxxxACGGCAGGGTCAGGGTTCTGGATAT
SP1-N6-I-βRC1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNxxxxxxGGTGGGAACACCTTGTTCAGGTCCTC
SP1-N6-I-βRC1.2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNxxxxxxGGTGGGAACACGTTTTTCAGGTCCTC
SP2-6N (ligation)	[Phos]NNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[SpcC3]

Table 2.1: Primer sequences used in the protocols described above. IUPAC nucleotide codes used: 'N' = any base; 'V' = any base apart from T (A, C or G); 'S' = C or G; 'K' = G or T. Variable bases for indexes are represented by 'x' (see Section 2.2). [Phos] represents a 5' phosphate group, to permit ligation with the 3' hydroxyl group of the cDNA. [SpcC3] = Spacer C3, a blocking moiety to prevent polymerisation of ligation oligonucleotides.

2.3.7.2 Indexing sequences

Number	P7-X-SP2 indexes	SP1-N6-I-α/βRC1 indexes
1	ATCACG	ATCACG
2	CGATGT	CGATGT
3	TTAGGC	TTAGGC
4	TGACCA	TGACCA
5	ACAGTG	ACAGTG
6	GCCAAT	GCCAAT
7	CAGATC	
8	ACTTGA	
9	GATCAG	
10	TAGCTT	
11	GGCTAC	GGCTAC
12	CTTGTA	CTTGTA
13	AGTCAA	TAGACT
14	AGTTCC	ACACGG

Table 2.2: Index sequences used to multiplex samples. Note that the hexamer index sequences ('X') in the P7-X-SP2 sequences are actually the reverse complement of the equivalent standard Illumina indexes of the same number. This makes pooling libraries with others easier (as it minimises the chance of sharing indexes) while maintaining the edit distances between the sequences.

2.4 Bioinformatic analysis

2.4.1 General considerations

The bioinformatic work described in this thesis was almost exclusively run on a standard work-range laptop computer, with a quad-core Intel i5 CPU (4×2.67GHz), with 7.6 GiB of RAM, running 64-bit Ubuntu Linux. The majority of analyses were performed using `Python` (v2.7.6) or `R` (v3.0.2) and custom scripts, while general text file exploration and manipulation were performed on the command line using `bash`, `awk` or `sed` commands. All manipulation of FASTQ data was performed using `Biopython` v1.63 (109). `Python` was typically employed for any rearrangement- or sequence-specific analysis, while `R` was used for plotting numerical features of whole files. The `Python` library `collections` was regularly employed for storing TCR features, the `difflib` module was used to compare sequences, and statistical analysis was achieved through a combination of `NumPy` and `SciPy` libraries, with results plotted using `Matplotlib`. The

`R` package `lattice` was additionally used in the production of heatmaps. Additional sequence file manipulation was performed using the FASTX-toolkit (110). Schematics were made and figures collated using Inkscape.

Unless otherwise stated, analyses are described in this section as they were applied to the data produced using the BLR protocol, as this constitutes the majority of the results of this thesis.

2.4.2 Demultiplexing BLR MiSeq data

High-throughput sequencing (HTS) machines produce data in FASTQ format. FASTQ files are text files, where every four lines corresponds to a different sequencing read, providing an identifier, the sequence, and quality data. Multiple samples are typically processed on one run of a machine – or 'multiplexed' – by introducing sample-specific index sequences prior to pooling and sequencing.

Data produced using the BLR protocol requires additional processing so that both random hexamer sequences are present at known positions in the first read, to allow identified TCRs to be linked to the barcode sequence when analysing a single read file. This pre-processing occurs at the same time as demultiplexing, in one of two ways. The earliest samples were first subjected to the standard Illumina demultiplexing procedure that occurs by default on the MiSeq during FASTQ generation, which deconvolutes samples based on their P7-X-SP2 index sequence. This produces two FASTQ files (one starting from each end of the amplicon) for every P7-X-SP2 index used. A custom script (named **`AddAllR2Hex`**) then pastes the random hexamer sequence (and associatd quality scores) from the second read to the start of the first. This produces a read where the first twelve bases are made up of the two random hexamers, followed by the rest of the read, which could contain V(D)J information. These reads can then be further demultiplexed based on the presence of the relevant SP1-N6-I-α/βRC1 index immediately downstream of the barcode, which was done using the `grep` command in `bash` to pull out direct matches.

The second technique implemented (used on the majority of runs) involves simultaneously demultiplexing on both indexes, allowing complete control over index string matching criteria. This method involves first obtaining FASTQ data for all three reads, i.e. including the index read (which is not produced by default). This was achieved in one of two ways, depending on which of the two MiSeq machines I had access to was available. One of the machines had an altered configuration file which overrode various default settings, making the MiSeq output the index read simultaneously with the two sample reads. To obtain the index read for data from the other machine, the entire

output run folder (containing the raw base-call data) was copied and reprocessed using the Illumina software `CASAVA/bcl2fastq` to produce all three FASTQ files. All three reads could then be simultaneously processed to enable the production of reads containing: both random hexamer sequences (to become the barcode), both hexamer indexing sequences (to allow demultiplexing) and the rest of the read one sequence (containing the V(D)J information). A single script (**`DualIndexDemultiplexing`**) scrolls through all three FASTQs simultaneously, moving the relevant sequences to known locations in one output read, which is then written to a specific file based on its sample-specific index sequence.

Demultiplexed reads can then be analysed for TCR rearrangements, while still retaining any barcode information without needing to refer to multiple files.

2.4.3 TCR assignment and error-correction

Our group has previously released software called `Decombinator`, which rapidly scans HTS data for recombined TCR sequences (111). 'Decombined' data stores TCR rearrangements as a five-part classifier, where the fields detail which V and J genes were used, how many deletions there were from each of these genes, and the nucleotide sequence (or 'insert') between the remaining ends of the two genes. As described in Chapter 5, a modified version of `Decombinator` called 'verbose `Decombinator`' (**`vDCR`**) was developed and employed in the work of this thesis. `vDCR` was built on `Decombinator` version 1.4, and incorporates additional error-filtering capabilities and outputs additional read information for downstream processes. Note that the most recent release of `Decombinator` (v2.0) incorporates a number of the general improvements made in `vDCR`, and thus can be considered broadly functionally equivalent (112). An extended set of V and J gene tags (relative to the standard `Decombinator` tag set) was used in this work.

Rearrangements detected by `vDCR` are output in 'n12' file format. If the libraries were produced using either the BV or BLR protocols – which introduce hexamers of random nucleotides to both ends of each cDNA molecule prior to amplification – then the additional information contained in the n12 file can be used for further data correction. By using the 12 random nucleotides as a barcode of unique cDNA molecules, comparison of TCR sequences sharing barcodes allows removal of those containing errors. Furthermore, counting the number of barcodes associated with an error-corrected TCR (after clustering to remove probable errors in the barcode) provides an estimate of cDNA frequency. Barcodes were clustered if they possessed an edit

distance of three or fewer to all other sequences in that cluster. Error- and frequency-corrected data is said to be 'collapsed', therefore the script that performs this function is named **`CollapseTCRs`**. The development of this process is described in greater detail later, in Chapter 5.

By default the 'OmitN' option in the `vDCR` code is set to 'False' (unlike in standard `Decombinator`), thus allowing reads through even if they contain ambiguous base calls. Reads that contain Ns in their inter-tag region (or barcode sequence, where applicable) are instead filtered after Decombining by using `bash/awk` commands. Reads with erroneously long inter-tag distance greater than 130 nucleotides were removed at the same time. If the data was not generated using a barcoded protocol, then the following command was used, which only filters out reads with ambiguous base call in the inter tag region (the seventh field of the n12 files):

```
awk '{FS=`, `} {if ($7~/N/){next} else{ if (length($7) < 130){print}}}' \
  FILE.n12 > temp && mv temp FILE.n12
```

If the data had been generated using a barcoded protocol, then the command was altered to also filter out n12 entries containing an ambiguous base in their barcode field as well (the ninth field), as this would impact upon the clustering process:

```
awk '{FS=`, `} {if ($7~/N/ || $9~/N/){next} else{ if (length($7) < 130) \
  {print}}}' FILE.n12 > temp && mv temp FILE.n12
```

Note that in the BV analysis of the cord blood data, low-frequency rearrangements found to not be using the primed V regions (as a result of off-target priming or misassignment) were similarly filtered out prior to further analysis.

2.4.4 TCR productivity and CDR3 amino acid determination

TCR productivity – i.e. the likelihood of producing polypeptide chains that could functionally serve in a surface-expressed receptor complex – was determined for raw or collapsed `Decombinator` identifier files. This is achieved by inferring the complete nucleotide sequence, determining the reading frame, translating the sequence and searching for the presence of particular TCR motifs in the appropriate locations. These tasks are all performed by a single `Python` script (**`CDR3ulator`**), which makes use of functions adapted from equivalent scripts written for similar purposes by other group members Dr. Niclas Thomas and Katharine Best. In order to be considered productive a TCR classifier must produce a sequence that meets the required criteria:

- Sequences must be in frame.

  - This is determined by the number of nucleotides from the first base of the V region to the last base of the J region minus one being a multiple of three (as the last nucleotide of the J region contributes to the first codon of the constant region).

- Sequence must contain no stop codons.
  - Using `Biopython`, this equates to discounting any translated sequences that contain the '*' character.

- Sequences must contain the conserved CDR3 cysteine residue.
  - This is the second conserved cysteine residue that marks the start of the CDR3 sequence (113).

- Sequences must contain the conserved FGXG motif.
  - This is the motif that ends the CDR3 sequence (113). Note that in human $\alpha\beta$ TCR chains a number of germline J genes predicted to be functional diverge from this motif by one residue; in order to be correctly assigned each gene must contain the appropriate motif.

Productive and non-productive reads are then written into separate files. In this thesis, I have used the standard CDR3 definition that runs inclusively from the second conserved cysteine to the phenylalanine in the conserved FGXG motif. Note that these statuses are predicted, and may not reflect the true productiveness of a given cDNA. It is possible that transcripts assigned as productive contain non-sequenced post-transcriptional modifications that prevent their function, or that some non-productive sequences are able to contribute to some biological function.

2.4.5 Sorting of poly-T homopolymer containing reads

Data produced using the poly-T RACE protocol (described above in Section 2.3.3) were sorted into reads that did or did not contain poly-T sequences in their 5' (i.e. those which were sequenced in the same orientation as transcription or not) through use of a custom `Python` script (**`SortToPolyTREGEX`**). This script makes use of the `re` package to search for various regular expressions (regex) that look for the presence of poly-T tracts. Regex analysis is required as polymerase slippage and homopolymer-induced error results in exact string matching of long homopolymers being unfeasible. Three regexes were employed: `'TTTTTTTT'`, `'TT+[ACGN]T+'` and `'TT+[ACGN]T*[ACGN]T+'`. The search was configured to detect any poly-T tract of eight or more bases in the first

70 basecalls of a read, making allowances for up to two non-T bases to be included. Reads were sorted into one of two output files depending on the presence of poly-T homopolymers being detected in their 5'.

2.4.6 Splicing analysis

2.4.6.1 Mapping of reads to TCR germline loci

Alternative splice events were detected by mapping reads to the germline genomic TRA/D and TRB loci. The TRA/D germline sequence consisted of nucleotides 3080000-4030000 from the GRCh37.p5 primary assembly of the *Homo sapiens* chromosome 14. The TRB sequence used was the NG_001333.2 version of the curated locus sequence from chromosome 7, which is 684973 bp long (accession number 114841177).

Reads were mapped against these references using the RNA sequence aligner `STAR` (114). Bam files produced by `STAR` (v2.3.0) were sorted and indexed using `SAMtools` (v0.1.19) (115), and visualised using the Broad Institute's Integrative Genomics Viewer (IGV, v2.3.32) (116). Mapped reads were visually compared against mapped germline sequences for L-PART1, L-PART2 as well as V and J genes, downloaded from IMGT/GENE-DB (19).

2.4.6.2 Quantification of intra-V gene splicing events

Having identified intra-V splice sites through manual inspection of one donor's reads, these sequences were searched for in a number of donors. Paired-end reads were first merged using `FLASH` (117), which finds those which overlap in their 3' sequences, reverse complements one and combines them, producing reads with the longest possible coverage over the V region. This process typically merged around one third of all reads per file of data produced using the BLR protocol.

Merged reads were then analysed with a single `Python` script (**`SpliceAnalysis`**) that performed a number of tasks, including: collapsing, to reduce errors; Decombining and translation, to assess the presence and productive status of rearrangements, and checking for the presence of previously identified intra-V splice events. Collapsing of merged reads occurs slightly differently to collapsing barcoded `Decombinator` output, as it is the whole read that is being collapsed, not just the section that has been determined to cover the rearrangement. As before, the most common sequence within a barcode is taken as the likely-actual sequence, and any others but error-containing PCR-descendants. Only barcodes with a minimum of three separate reads are considered. If there is no sequence within a barcode that is more prevalent than the others, then a consensus sequence is created by taking the most common nucleotide at each

position of each equivalent read: if two or more nucleotides are equally present at a position then that barcode is ignored. All barcode sequences shared by a particular corrected sequence are then clustered as in `CollapseTCRs` to estimate the frequency of sequences. Corrected sequences were then inspected for rearrangements using the core functions of `vDCR`. The original, published V and J gene tag set was used, as these tend to lie closer to the recombining edge. Corrected sequences with rearranged TCRs are then translated by inference from the `Decombinator` classifier and predicted productivity is recorded. Amino acid sequences are then also translated from the complete actual sequence (not that inferred from the classifier). Correct reading frames were deduced by finding the frame that contains the amino acids of the constant region covered by the RC1 primer used during amplification: 'IQNPDPA' for TRAC[5], 'DLKNVF' for TRBC1 and 'DLNKVF' for TRBC2. The productiveness of the transcript can therefore be inferred from the whole sequence itself, i.e. whether there is an in-frame start codon relative to the required constant region residues, and the absence of an interrupting stop codon. Sequences were also searched for the exact string matches of splice events as determined from the mapping described in Section 2.4.6.1. All findings are input to a dictionary, recording for a given corrected transcript sequence: its frequency; the presence of a rearranged TCR; the predicted CDR3 (or criteria for being deemed non-productive); the actual open-reading frame of the complete sequence, and whether or not it contains any identified intra-V splice events.

2.4.7 Calculation of diversity and sharing values

2.4.7.1 Gini index

The Gini index is a measure of inequality or dispersion of a distribution, often employed in economics to represent the distribution of wealth in a system. Gini indexes are calculated in reference to a Lorenz curve, where the proportion of resource allocation is plotted as a function of the proportion of the population (Figure 2.1):

$$\mathrm{G} = \frac{A}{A+B}$$

Gini index values can range from zero, representing total equality or maximum dispersion (e.g. all money is shared equally) to one, representing total inequality (e.g.

[5]This sequence actually starts from the second amino acid of the constant region: the last nucleotide of the J gene starts the first codon of the constant region, and TRAJ genes do not all end in the same nucleotide, thus after splicing the first codon of the constant region can encode one of four different amino acids.

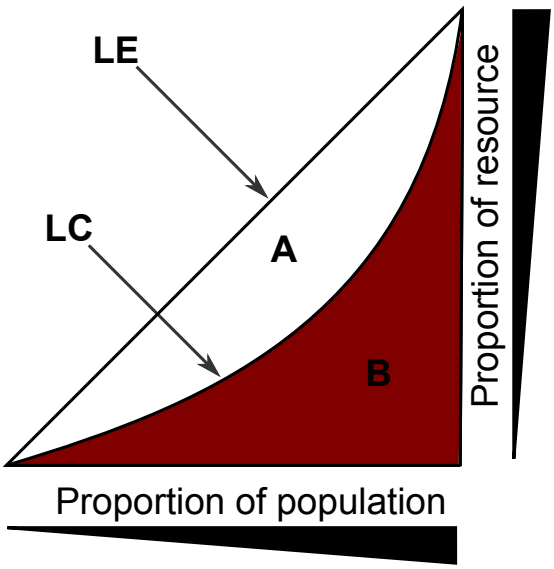


Figure 2.1: Gini index calculation from the Lorenz curve. The Gini index is defined as the ratio between the area under the Lorenz curve (LC) and the total area, under the line of equality (LE, or perfect distribution line).

one entity has all of the money). Gini indexes were calculated using the `ineq` library in `R`.

2.4.7.2 Shannon entropy

Shannon entropy is a measure borrowed from information theory, where it is used to quantify the 'entropy' of a system. It can also be considered to be measuring either the uncertainty or information content, where higher entropies correspond to more uncertainty or information. Consider a comparison of a fair dice roll and a fair coin toss: there is more uncertainty as to the output of the die than the coin, as there are more options (more entropy). The greater array of possible values produced from an equal number of rolls compared to tosses similarly could contain more information (X^6 as opposed to X^2). When Pi is the probability of instance i occurring in a distribution (and b is the base of the logarithm), the Shannon entropy is defined as:

$$H = -\sum_{i} P_i \, log_b \, P_i$$

Shannon entropies were calculated using the `entropy` library in `R`, using the base 2 logarithm to produce output in bits.

2.4.7.3 Jaccard index

The Jaccard index is a normalised measure of sharing between two sets of items. It is defined as the size of the intersection of those two sets (i.e. how many unique items are found in both) divided by the size of the union (how many different unique items are found in total between the two sets):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In this thesis, Jaccard indexes were calculated using a custom function in `R`, using the default functions `intersect` and `unique`.

2.4.7.4 Jensen-Shannon divergence

The Jensen-Shannon divergence (JSD) value is a modification of the Kullback-Leibler divergence (KLD) – to make it symmetric and produce finite values – that relates to the amount by which two distributions (P and Q) differ:

$$I(P||Q) = \tfrac{1}{2}\left[D\left(P||\ \frac{P+Q}{2}\right) + D\left(Q||\ \frac{P+Q}{2}\right)\right]$$

Here D represents the KLD using the base 2 logarithm, defined as:

$$D(P||Q) = \sum_i P(i) log\ \frac{P(i)}{Q(i)}$$

The KLD is therefore the logged difference between distributions P and Q, using P as the expected distribution. By taking the average of KLDs calculated in either orientation, the measure becomes symmetric. In this work, JSD was calculated in `Python` using a function written by Jonathan Friedman (118).

2.5 Commitment to open science

I adhere to the principles of open science, under the expectation that verification, reproduction and proper dissemination of research can only be properly achieved through adequate sharing of the processes and results involved. Accordingly, the following components of the content of this thesis have been made publicly available.

2.5.1 Data deposition

The vast majority of the raw sequence data produced and analysed in this data has been uploaded to the NCBI Sequence Read Archive (SRA). The SRA accession numbers of these datasets are:

- Simple V amplification HiSeq data: **SRP055031**
- Poly-T RACE 454 data: **SRP055040**
- Barcoded V MiSeq data: **SRP055080**
- Barcoded ligation RACE MiSeq data: **SRP045430**

2.5.2 Source code

The scripts required for the standard analysis of TCRs from high-throughput sequencing data obtained using these protocols can be downloaded from a freely accessible GitHub repository at https://github.com/JamieHeather/tcr-analysis (119). This repository contains `vDCR`, `CDR3ulator`, the extended tag sets as well as the associated demultiplexing and basic plotting scripts.

Another repository, available from https://github.com/JamieHeather/ThesisScripts was generated to contain a sample of the more specific analyses that were used during this thesis. In addition to the scripts listed in bold in preceding sections, this repository also contains scripts that:

- Randomly draw a given number of sequences (either DCR classifiers or CDR3s) from a larger file, for the purpose of size-matching data for fair comparison (**`SizeMatch`** and **`CDRSizeMatch`**)
- Construct FASTQ files from the inter-tag region data contained in n12 files, and assess the average quality score and converted error probabilities (**`FastqInterTag`**, **`PrintTotalAvgQuals`** and **`PrintErrorProbabilities`**)
- Take two DCR index files and output a comma-delimited table, giving the proportional frequency of each TCR sequence in both of the two files (**`ProportionPrePlot`**), allowing for convenient correlation measurements
- Quantify and compare the presence of published TCRs sequence in multiple samples (**`HIVPublished_CDR3s`** and **`InvariantTCR`**)

- Measure TCR repertoire population level differences across multiple samples, such as publicness, translational convergence and similarity of TCR gene usage (**`Publicness`**, **`HIVTranslational_Convergence`** and **`LigRaceJSDs`**)
- Quantify the presence of certain genes or subsequences (**`GeneUsage`** and **`FindDs`**)
- Mix TCR classifiers from multiple subsets to generate *in silico* simulated repertoires of mixes populations (**`HIVCD4_CD8_Mixing_CDR3s`**)

Note that the sequences of invariant and antigen specific TCRs harvested from the literature are contained within the `HIVPublished_CDR3s` code.

3

Exploring healthy T-cell receptor repertoires

3.1 The global TCR repertoire at steady state

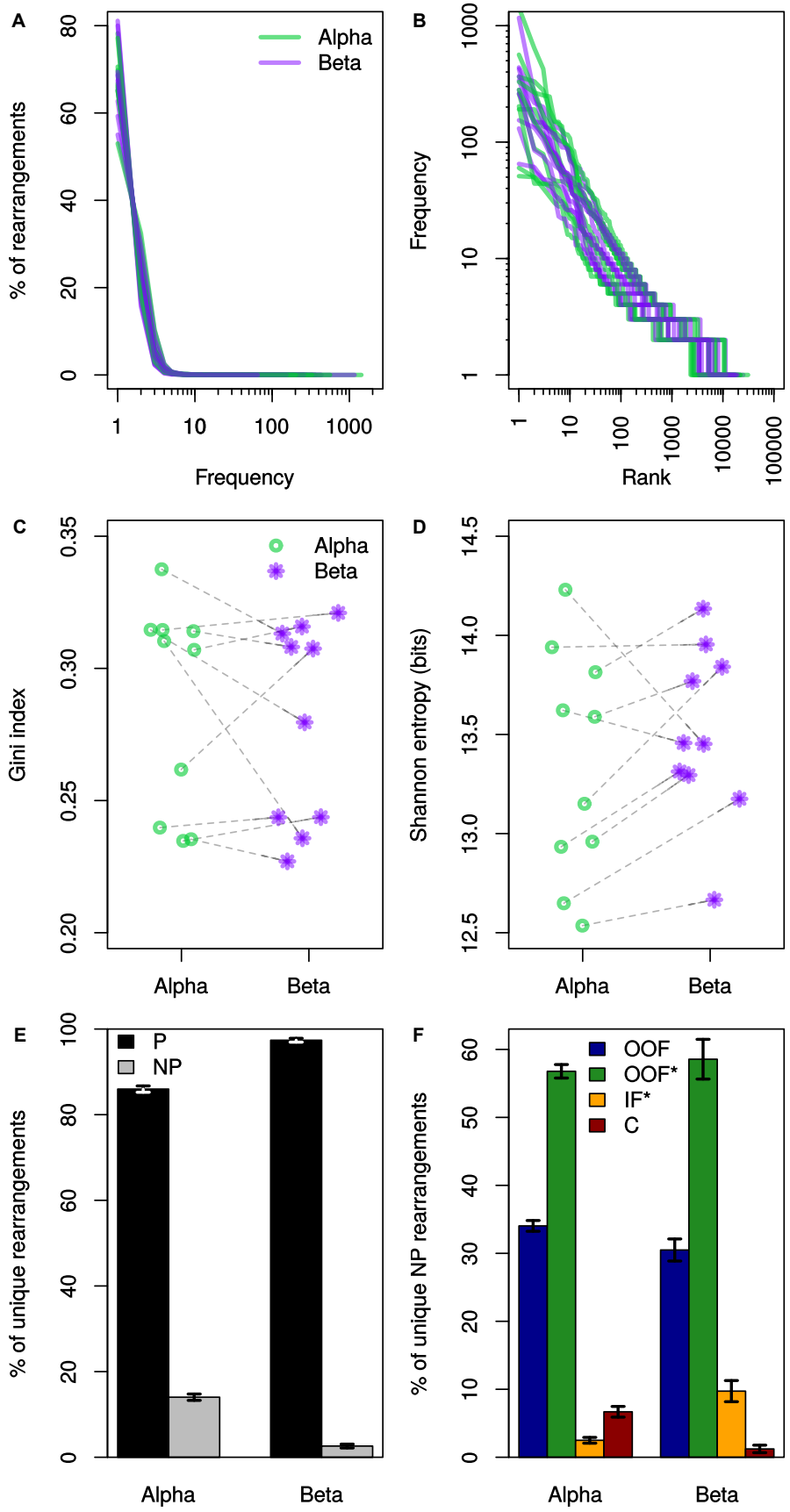
In this section I explore the data generated from sequencing ten healthy donors' whole peripheral blood repertoires using the barcoded ligation RACE strategy (or 'BLR')[1]. In these experiments a fraction of RNA extracted from a 2.5 ml sample of whole blood served as the template for reverse transcription, TCR amplification and sequencing. Sequences produced were analysed using `Decombinator`, the TCR analysis software developed in our group, and further error-corrected using `CollapseTCR`. The error-corrected TCR classifiers produced by `Decombinator` ('DCRs') – which are equivalent to unique nucleotide sequences – are the basis for the following analyses.

3.1.1 Clonal distribution of the healthy TCR repertoire

The first aspect of the repertoire considered is its basic clonal composition. The distribution of TCR rearrangement frequencies is an intrinsically difficult property to visualise, as the vast majority of unique DNA recombinations – 50 to 80% – only occur once in a given sample, while a small fraction of sequences are found repeated tens, hundreds, or even thousands of times (Figure 3.1A). This produces a characteristic 'L' shaped distribution, which makes the data somewhat difficult to interpret. By plotting the logged rank (with the most common rearrangement as rank one) against the logged frequency (Figure 3.1B) we observe an approximately linear relationship between the two (at least for sequences with more than five copies), indicating that a given sequence will be more frequent than the next most common sequence by a fixed proportion. Thus there are a small number of highly frequent rearrangements, and a large number of sequences that only occur very rarely in these samples.

The frequency distribution of different TCR chains relates to another property of a repertoire, that of diversity. This is thought to be a very important repertoire characteristic, as the number of different types of antigen receptor in a system should

[1]The development of this protocol, and all others, is covered in detail later in Chapter 5.


A
% of rearrangements
Frequency
Alpha
Beta
B
Frequency
Rank
C
Gini index
Alpha
Beta
D
Shannon entropy (bits)
Alpha
Beta
E
% of unique rearrangements
P
NP
Alpha
Beta
F
% of unique NP rearrangements
OOF
OOF*
IF*
C
Alpha
Beta

relate to how many antigens that system can respond to. TCR receptor diversity can be measured in many different ways (120); in this thesis I make use of two indexes of diversity borrowed from other fields. The Gini index is a measure of inequality (or unevenness) frequently used in economics, as it permits an overview of the distribution of a resource such as wealth. For example, a Gini index value of one represents total inequality, in which all of the wealth belongs to one entity, while a Gini index of zero represents total equality, in which all entities possess an equal share of wealth (121). With respect to TCRs, Gini index values of zero would correspond to all rearrangements being equally present in a sample, and as oligoclonality increases – i.e. as subsets of sequences began to become more common and dominate the repertoire – values would tend towards one. Figure 3.1C shows that the α and β repertoires of healthy donors have Gini indexes towards the low end of the scale, indicative of a generally even distribution structure. The other measure employed is Shannon entropy, borrowed from information theory. Shannon entropy is a more complicated measure that incorporates both aspects of diversity – the number and frequency of different species – to produce a value that can be interpreted as the information content or complexity of a system (122). More species (or TCRs) being present increases the total information content, as does increasing the evenness. Thus the higher the Shannon entropy the more information content a system has, and the harder it is to predict to what species additional observations are likely to belong. Conversely, a distribution where all species are the same has no information

Figure 3.1 ***(preceding page)*****: Clonal composition of the healthy human $\alpha\beta$ TCR repertoire.** **A**: Percentages of each of the ten donor's repertoires that are accounted for by rearrangements of different frequencies (i.e. how many times that rearrangements appears). **B**: Logged rearrangement rank versus logged frequency reveals a linear distribution for all but the lowest frequency sequences. **C**: Gini indexes of the distribution of unique rearrangements of healthy donor repertoires; dotted lines connect the α and β values of the same donor. There is no significant difference between the paired α and β Gini index values (p-value = 0.38, Wilcoxon signed rank test). **D**: Shannon entropies of the healthy donor samples. There is no significant change between the $\alpha\beta$ paired Shannon entropy values (p-value = 0.08), although the majority of donors tend to have greater values in their β chains. **E**: The proportion of unique rearrangements (i.e. ignoring the frequency of each) that are predicted to be productive (P) or non-productive (NP). **F**: The proportions of the criteria by which the NP rearrangements shown in E were determined to be so. OOF = out-of-frame; OOF* = out-of-frame, also containing a stop codon; IF* = in-frame but containing a stop codon; C = lacking the conserved cysteine residue at the start of the CDR3. Note that the translation software only generates sequence up until the end of the J region, thus OOF transcripts without stop codons in this analysis could contain stop codons further downstream in the sequence. All error bars indicate standard deviation around the mean value calculated using all ten donors.

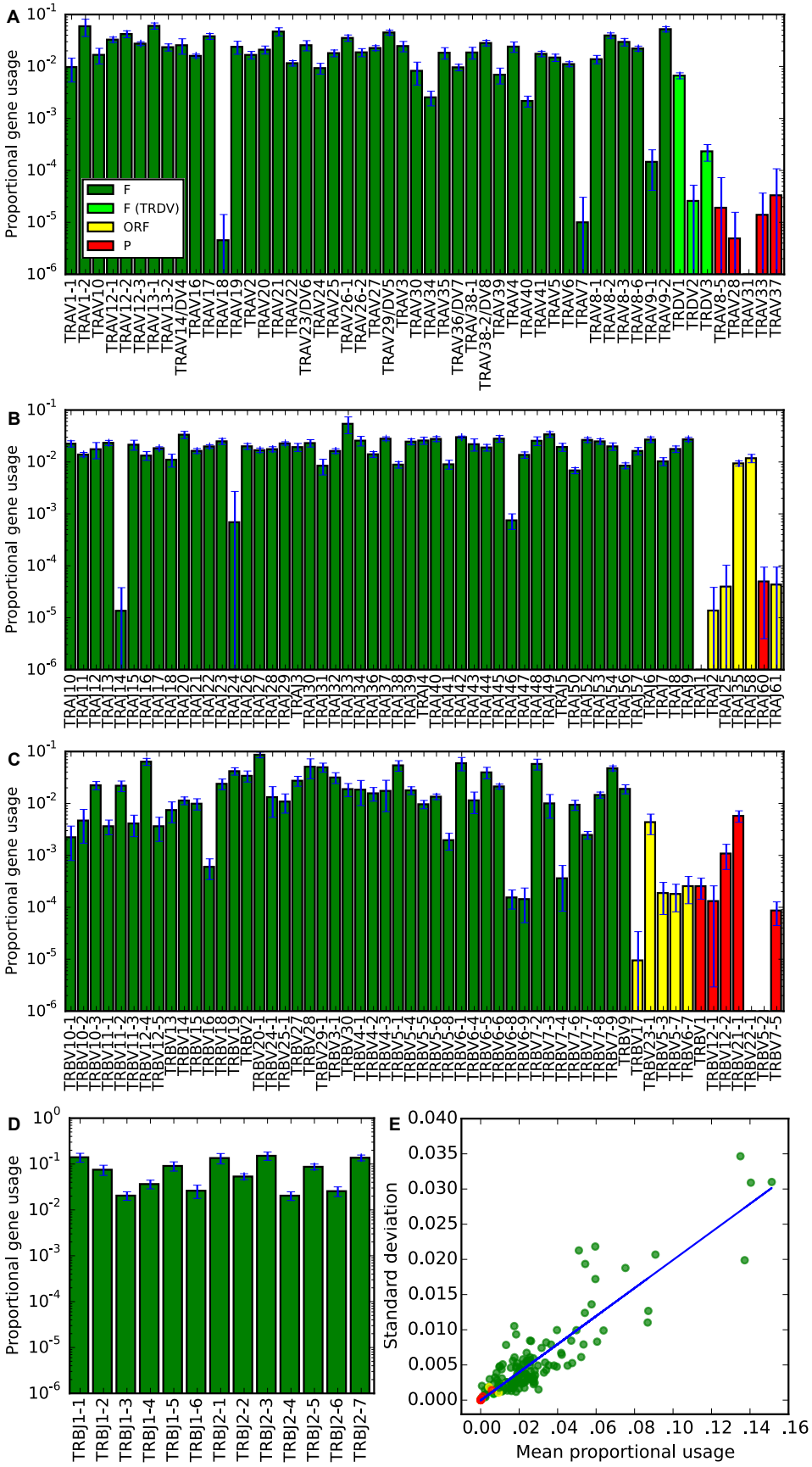
per se, and thus would have a Shannon entropy of zero. Figure 3.1D shows the Shannon entropies of the healthy donor repertoires, showing that typical healthy repertoires (of these sample sizes) possess in excess of twelve bits of information.

V(D)J recombination will not always produce a recombined gene that is capable of being translated into a functional polypeptide. Due to the random nature of the recombination process, rearrangements may fail to be in the correct frame downstream of the junction, or contain premature stop codons, preventing translation of valid TCR chains. When developing the software to translate TCR nucleotide sequences and extract CDR3 regions, I counted TCR sequences that are expected to fail to produce a valid CDR3, and are thus considered non-productive. When considering unique rearrangements (i.e. ignoring how prevalent each rearrangement is) I observed that on average 16% of all α chain rearrangements and 2.6% of β were non-productive (Figure 3.1E). Examination of these non-productive sequences reveals that the majority of them are out-of-frame and contain stop codons (Figure 3.1F). A small number are in-frame yet contain a stop codon, and a similar number lack the conserved cysteine residue in the V gene which marks the N-terminal end of the CDR3 region. Note that there was a similar requirement for detection of the FGXG-based motif in the J gene that marks the C-terminal CDR3 end, however no sequences were deemed non-productive due to lacking this sequence. This difference is likely due to the fact that the conserved cysteine in the V is much closer to the recombining edge than the FGXG is in the J, and thus is more likely to be lost through base deletion during recombination.

3.1.2 V and J gene usage

Another important feature that will define a given repertoire is the frequency with which different V and J genes are used. Figure 3.2A-D shows the mean proportional usage of all the different TRAV, TRAJ, TRBV and TRBJ genes in our ten healthy donor repertoires respectively. The distributions are non-uniform; some genes are used in a far higher proportion of rearrangements than others. Figure 3.2E shows that the more a given V or J gene is used, the more deviation there is between how much different individuals use it.

During this analysis I observed a large number of unexpected recombination events. The first class of unexpected rearrangements involved detection of supposedly TCR δ chain specific V genes rearranged onto α chains, using the α constant region in combination with α chain J genes. Despite α and δ V genes sharing a common pool of V genes (Figure 1.2) it is currently reported that not all genes can be used by either chain: of the eight V regions known to contribute to δ chain receptors, five are also known


A
B
C
D
E
Proportional gene usage
F
F (TRDV)
ORF
P
Standard deviation
Mean proportional usage

to participate in α chain production (termed 'TRAV/DV' genes) (123). According to IMGT, the three remaining genes – TRDV1, TRDV2 and TRDV3 – are supposedly exclusively limited to δ chain rearrangements. However as Figure 3.2 demonstrates, all three can be found in α chain recombinations, even present at greater amounts than some of the officially-recognised TRAV genes. In order to confirm these findings, I also looked for such sequences in the only other publicly available high-throughput α chain data set: Figure A.1 in Appendix A shows that rearrangements using all three genes can also be found among these independently generated sequences.

The second class of unexpected TCRs are those that do not use genes designated as functional. For various reasons, several of the genes in the germline receptor gene loci are predicted to either not participate in rearrangements, or not be able to encode working polypeptides if they do (124). These non-functional germline genes come in two varieties: open-reading frame genes (ORFs), or pseudogenes. ORFs are predicted to not participate in rearrangements due to alterations in regulatory sequences, such as splice sites or recombination signal sequences, or in conserved amino acid sequences (such as the cysteine and phenylalanine residues that bound the CDR3 region), while pseudogenes are those that contain stop codons or frameshift mutations. However examples of both V and J ORFs and pseudogenes can be found for both TCR chains in the healthy donor data, with the exception TRBJ (which only has one such gene, a pseudogene). Some of these genes account for a considerable proportion of the repertoire, such as TRAJ35 and TRAJ58 (Figure 3.2C) or TRBV21-1 and TRBV23-1 (Figure 3.2D), approximately occurring more frequently than one in every 500 sequences.

In addition to monitoring how frequently each V and J gene appears, I also measured how frequently each combination of V and J gets used across our healthy donors. The mean frequencies and standard deviations are shown for α and β in Figures 3.3 and 3.4 respectively. In terms of average α chain rearrangements (Figure 3.3A) most V-J combinations account for small, roughly equivalent proportions of the repertoire, with one particular V-J pair – TRAV1-2 - TRAJ33 – being far more prevalent than any other. However comparison with the standard deviation of those averages (Figure 3.3B) reveals an increased number of mid-level hot-spots, indicating that at least in some donors certain V-J combinations are used preferentially. β chain VJ pairing shows

Figure 3.2 *(preceding page)*: V and J gene usage in healthy human $\alpha\beta$ TCR rearrangements. Mean proportional usage of the different genes for TRAV (**A**), TRAJ (**B**), TRBV (**C**) and TRBJ (**D**) in the ten healthy donors. F, ORF and P stand for functional, open-reading frame and pseudogene genes respectively, as described above. Error bars indicate standard deviation.

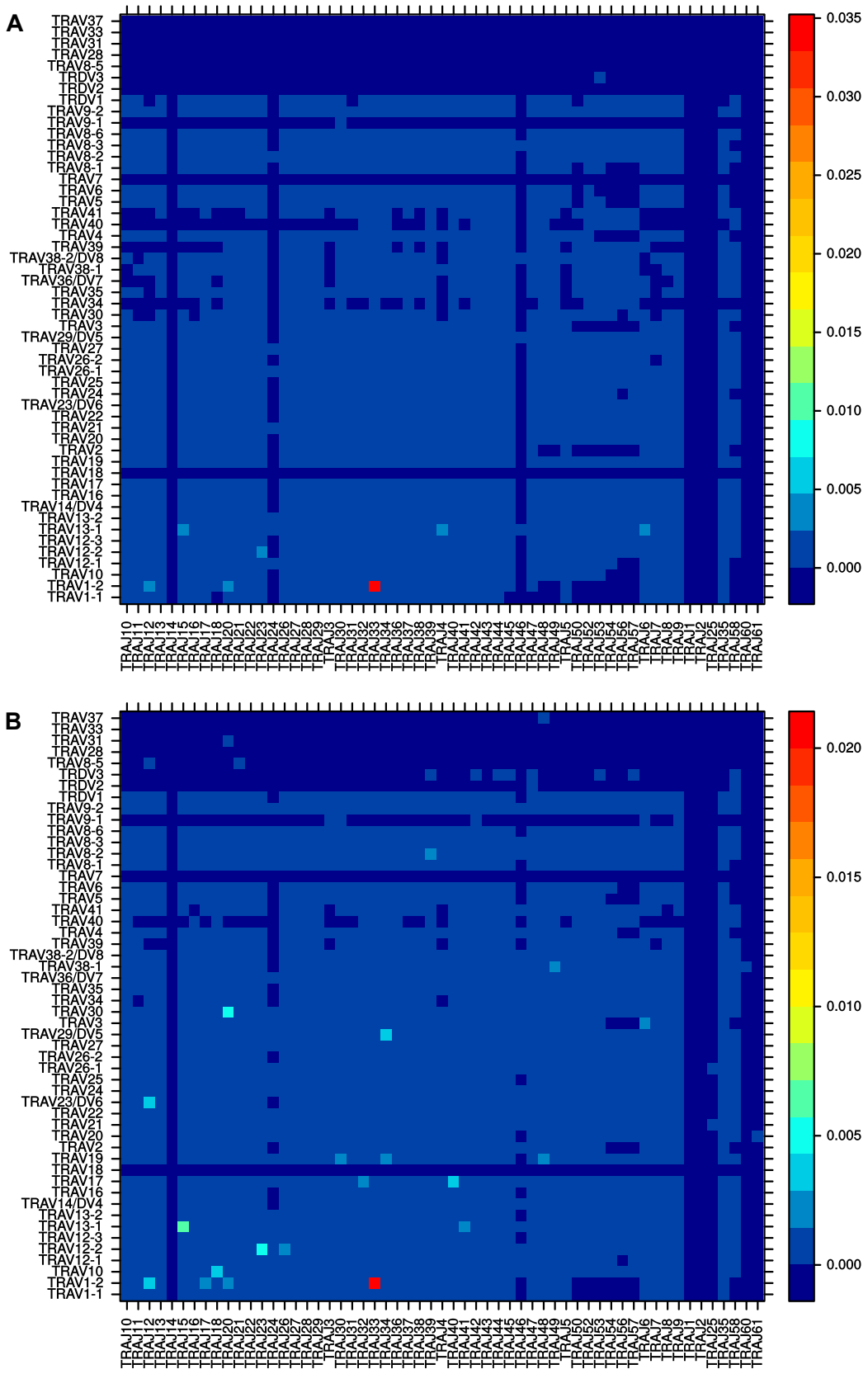


Figure 3.3: V and J gene pairing in healthy human α TCR rearrangements. **A**: Heat-map showing the mean frequency with which the different α chain V and J genes appear together in rearrangements, averaged across our ten healthy donor samples (taking the frequency of each rearrangement in each donor into account). **B**: Heat-map showing the standard deviation of the averages in **A**.

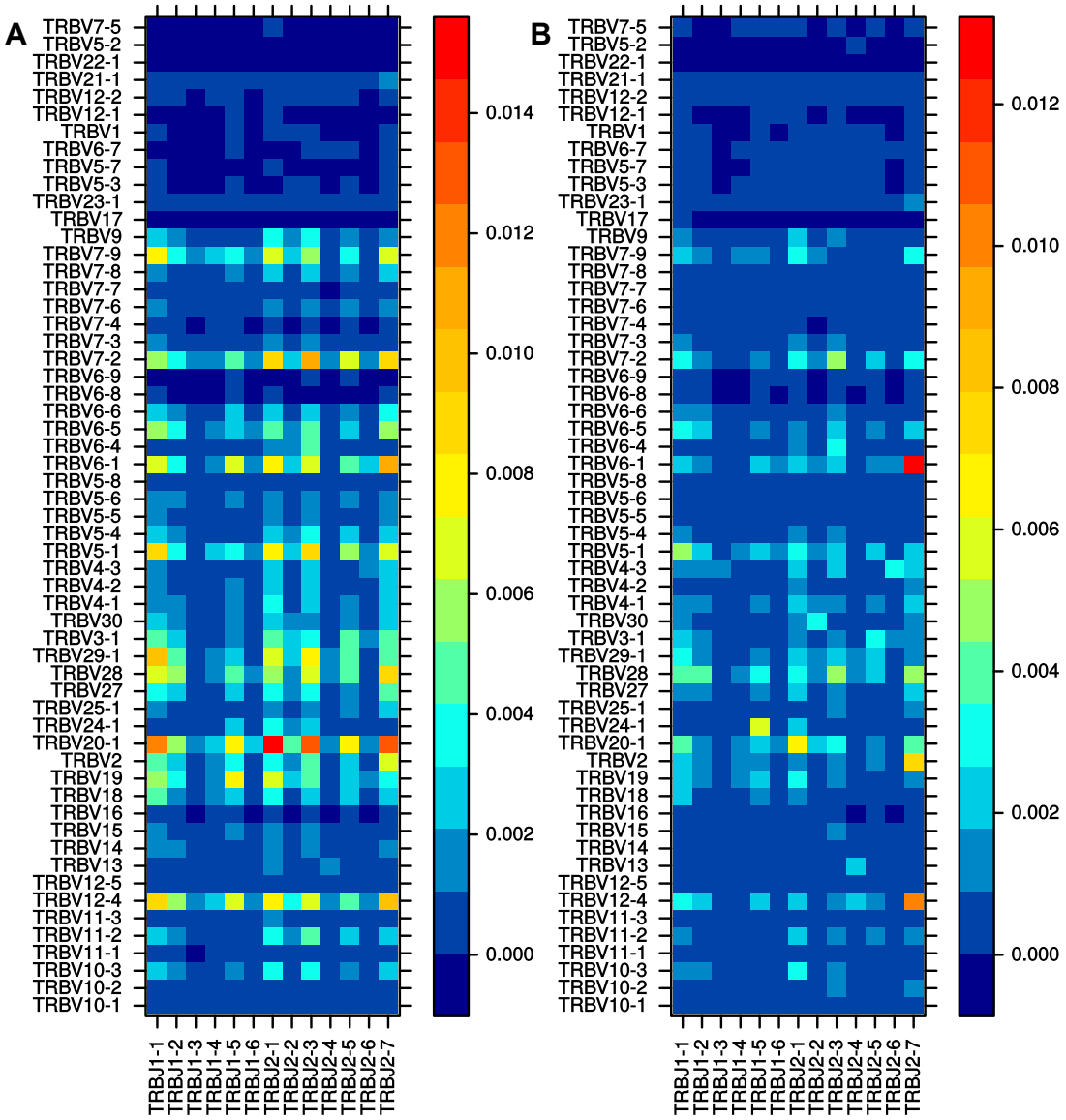


Figure 3.4: V and J gene pairing in healthy human β TCR rearrangements. **A**: Heat-map showing the mean frequency with which the different β chain V and J genes appear together in rearrangements, averaged across our ten healthy donor samples (taking the frequency of each rearrangement in each donor into account). **B**: Heat-map showing the standard deviation of the averages in **A**.

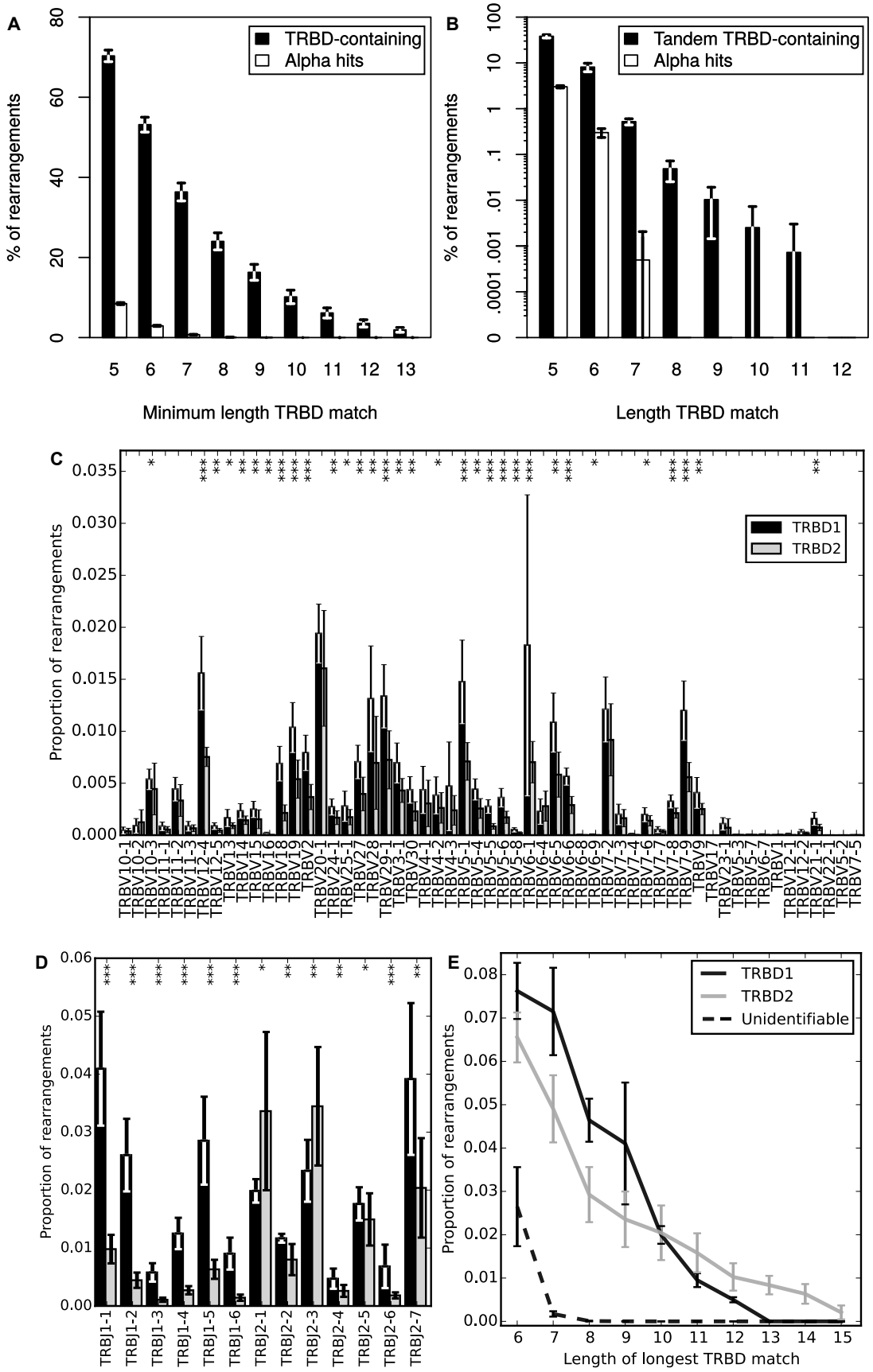
a number of hotspots in usage, particularly at the convergence of specific V and J genes which appear to be highly expressed relative to the others (e.g. TRBV20-1 - TRBJ2-1), while standard deviations appear relatively lesser in comparison. The upper limit of the scale for the α chain average proportions is more than twice as much as that in β, indicating that the TRAV1-2 - TRAJ33 rearrangements account for a greater proportion of the α repertoire than the most frequent β V-J pairs do in theirs.

3.1.3 TRBD detection and usage

The two human TRBD genes are both very short and highly similar (Table A.1, Appendix A) and subject to nucleolytic degradation from both ends, making their assignation by the tag-based gene identification system of `Decombinator`[2] not feasible. I developed additional methods to ascertain what the measurable contribution of the different D genes to β chain recombinations was, by searching the `Decombinator` classifier insert string – the nucleotides sequence found between the end of the V gene and the start of the J gene – for residual TRBD sequences. Figure 3.5A shows what percentage of total rearrangements contain detectable sequences from either TRBD gene. This search was performed multiple times, each time with an increased minimum length of matching sequence required. By applying the same search to the inserts of α chain rearrangements (which should not contain TRBD sequences except by chance) we can get an estimate of the false positive rate. For example, if we look in β chain data for matches of only five nucleotides or longer we observe that seemingly 70% of reads contain TRBD sequences; however by the same criteria, almost 8% of α reads appear to as well. As the threshold increases the rate of detection falls – as you would expect, given that there will be fewer sequences of the required length due to nucleolytic 'nibbling' – along with the false positive α assignation rate. This method represents an improvement in measuring TRBD usage over existing techniques, both in the inclusion of false positive detection as well as the separation of different length matching; existing published analyses often just use thresholds of five (125) or six (126, 127) nucleotides.

One previous study additionally reported the presence of low-frequency tandem TRBD gene usage, producing VDDJ transcripts, which were detected both in their own data (127) and a previously published dataset (77). Figure 3.5B shows the extent of such rearrangements in our data. We see at least an order of magnitude less fusion D-D usage than in the Liu *et al* report (an average 0.38% of rearrangements using a threshold of six nucleotides, as opposed to the lower bound of 1.73% stated in their

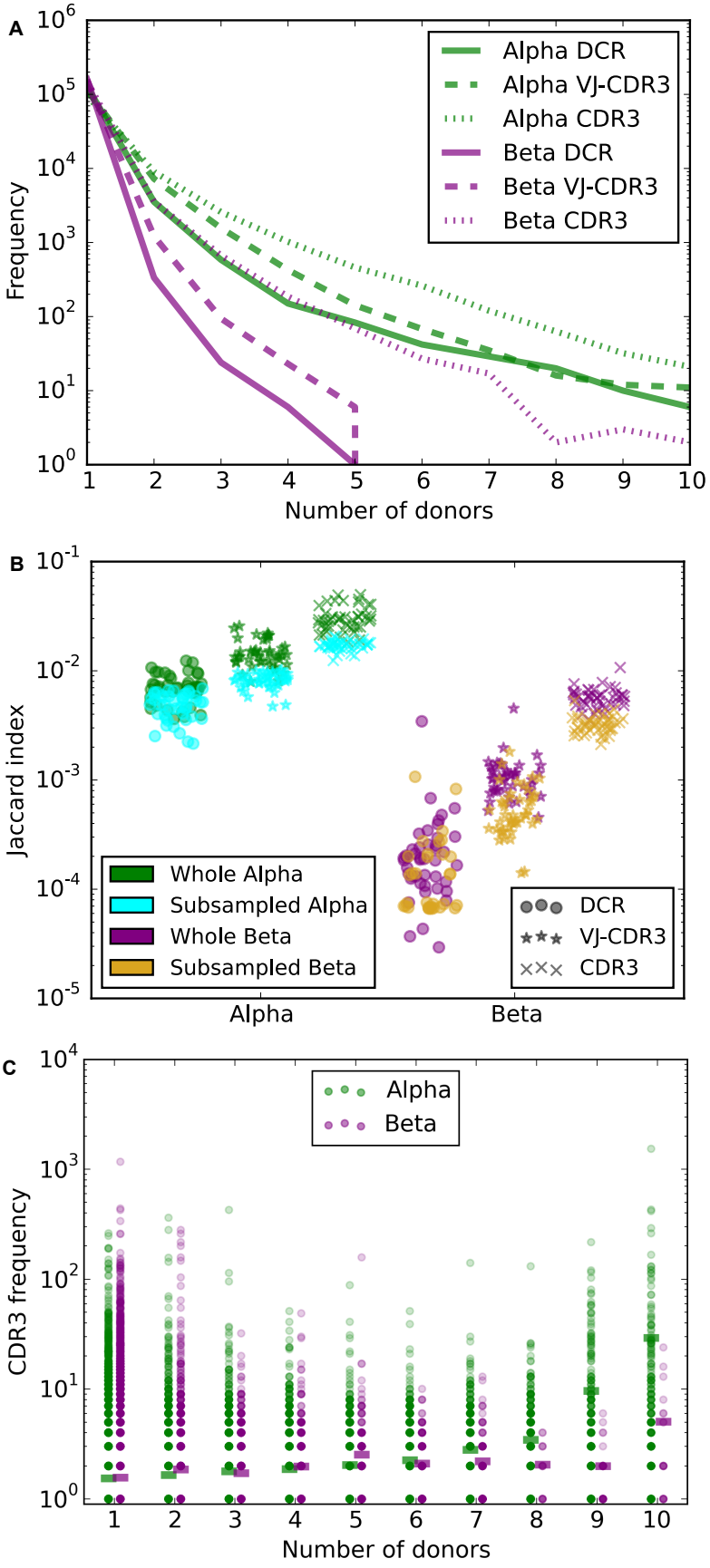
[2]This system is discussed in detail later in Section 5.3.1.


A
% of rearrangements
TRBD-containing
Alpha hits
Minimum length TRBD match
B
% of rearrangements
Tandem TRBD-containing
Alpha hits
Length TRBD match
C
Proportion of rearrangements
TRBD1
TRBD2
D
Proportion of rearrangements
E
Proportion of rearrangements
TRBD1
TRBD2
Unidentifiable
Length of longest TRBD match

publication). This is likely due to the current analysis requiring exact sequence matches, while that previously published allowed mismatches.

I also assessed the extent to which each TRBD gene associated with different TRBV and TRBJ genes. Figure 3.5C and D show that despite large deviations, there are both V and J genes which significantly pair more frequently with a particular β D gene. Of particular note is the separation of TRBD-TRBJ gene pairing between the two different TRBDJ gene clusters (Figure 3.5D). Despite pairing far less frequently with the TRBD1 cluster (TRBJ1-1 to TRBJ1-6, see Figure 1.1), TRBD2 does indeed pair with each of the TRBJ genes at some level. Extending this analysis, Figure 3.5E quantifies the contribution that each of the individual TRBD genes makes to the rearrangements that have detectable D regions, and how long the sequence match is. The majority of rearrangements with shorter D regions used TRBD1, while TRBD2 accounts for a greater proportion of the longer matches. This is not unexpected: the germline TRBD2 sequence is longer than TRBD1 (16 versus 12 nucleotides). Across all ten donors we note that TRBD1 is cumulatively used more frequently, accounting for 64% of all detectable TRBD usage. This is in contrast with a previous report of a more equitable distribution (75). This value is likely to be extremely cohort-specific, given that there is some evidence to suggest such patterns are influenced by both age and sex (128).

Figure 3.5 *(preceding page)*: TRBD detection and usage in healthy human TCR β rearrangements. **A**: Mean percentages of total healthy volunteer β chain rearrangements that contain exact contiguous matches to any TRBD gene sequences (see Table A.1, Appendix A). TRBD search was performing using an increasing threshold for the shortest sequence match to be counted. Each search was repeated using the α chain `Decombinator` rearrangements of the same donors as a specificity control, as these should only contain TRBD sequences as a result of non-templated nucleotide insertions and deletions. **B**: Mean percentages of total β chain rearrangements that contain tandem TRBD sequences, i.e. both a TRBD1 and TRBD2 sequence. **C** and **D**: Proportions of TRBV and TRBJ genes, respectively, that are found in combination with particular TRBD genes, for those rearrangements where one D gene was unambiguously assigned. **E**: The length distribution profiles for all of the TRBD matches given a particular match length threshold. **C**, **D** and **E** were all performed using a TRBD search with a minimum match length of six nucleotides. TRBD2 results include sequences that map to either allele (TRBD2*01 and TRBD2*02). All error bars show standard deviation. Asterisks indicate significance by Wilcoxon rank sum tests: * for p-values ≤ 0.05, ** for ≤ 0.01 and *** for ≤ 0.001.


A
Frequency
Number of donors
Alpha DCR
Alpha VJ-CDR3
Alpha CDR3
Beta DCR
Beta VJ-CDR3
Beta CDR3
B
Jaccard index
Whole Alpha
Subsampled Alpha
Whole Beta
Subsampled Beta
DCR
VJ-CDR3
CDR3
Alpha
Beta
C
CDR3 frequency
Alpha
Beta
Number of donors

3.1.4 TCR sharing and publicness

An emerging aspect of TCR repertoire research is the extent to which different TCR receptor sequences are shared between individuals. This phenomena is manifest at multiple levels, from any two individuals sharing a certain proportion of their sequences, up to so-called 'public' sequences, that are found across multiple individuals (129, 130, 131). I assessed the extent of such sequences in my healthy donor repertoire data using several techniques.

I first adopted a TCR-centric view, and measured how many productive sequences appear in what number of donors. This analysis was repeated at three levels of TCR sequence, considering: the nucleotide sequence level, represented by the `Decombinator` classifier (or 'DCR'); the CDR3 alone, and finally the CDR3 in combination with the V and J genes used in that rearrangement ('VJ-CDR3'). Figure 3.6A shows what we might expect; as we increase the number of donors we find fewer sequences that are shared across that many (for example, there are approximately 1×10^5 β chain DCR sequences that appear exclusively in one donor, but only one sequence that is shared by five donors). Therefore it may be more useful to consider 'publicness' as a scalar rather than a binary property. There is a general trend among the different types of sequence, where VJ-CDR3s are more public than DCR classifiers, and CDR3s are yet still more public. This may be explained by the possible convergence between these levels: multiple nucleotide sequences (DCRs) can encode the same VJ-CDR3, and the same CDR3 can be made in multiple V-J gene contexts.

I next took a donor-centric view, by performing pairwise comparisons of every combination of two donors and measuring by how much their repertoires overlap, using the Jaccard index (Figure 3.6B). This was performed both on complete repertoire

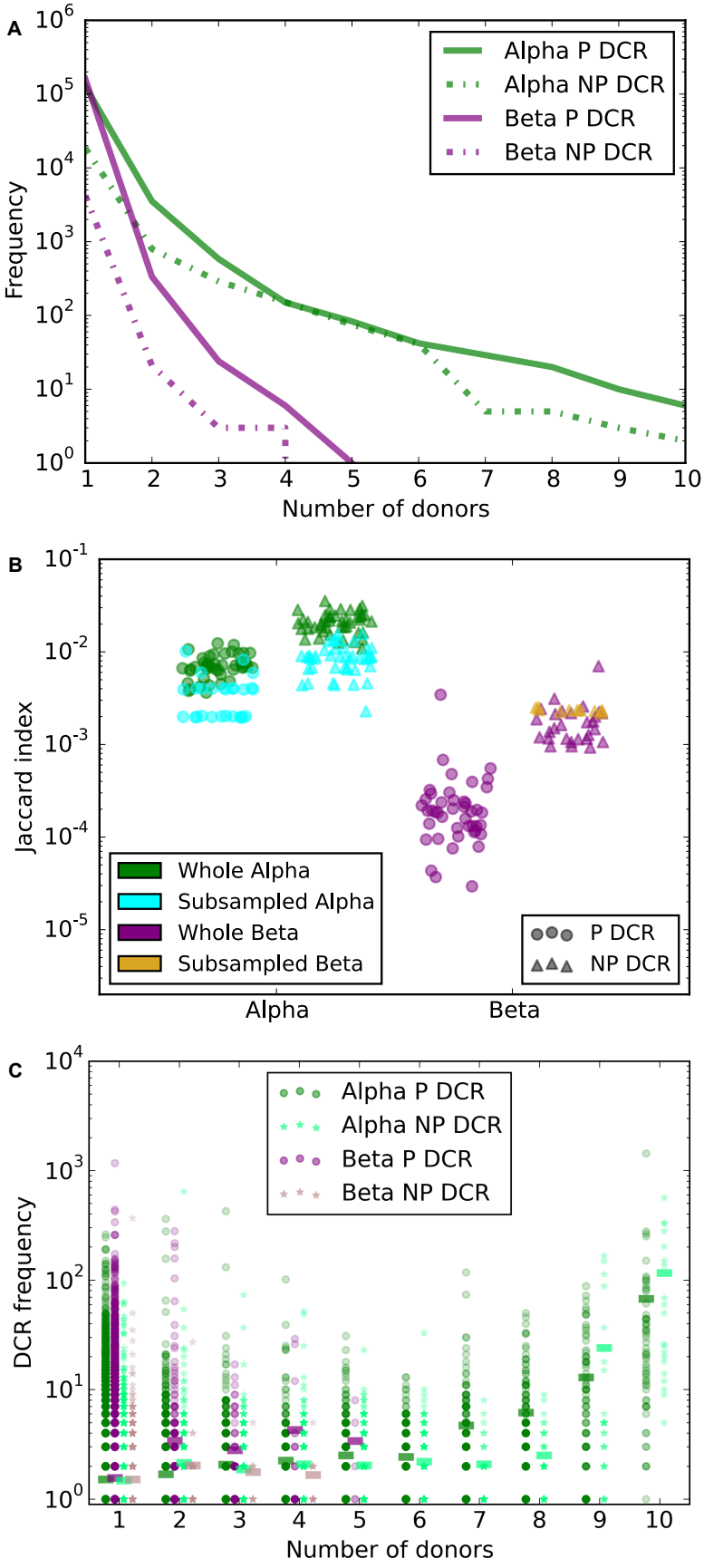
Figure 3.6 ***(preceding page)*****: TCR sharing and publicness of productive rearrangements.** **A**: Publicness of productive TCR sequences in healthy human donor repertoires, i.e. how many unique sequences were found only in one donor, or two donors, or three and so on. DCR = `Decombinator` classifier, which is representative of unique nucleotide sequences. VJ-CDR3 = a CDR3 amino acid sequence in combination with the V and J genes with which that CDR3 was found. CDR3 = the CDR3 region alone, from the conserved cysteine in the V region to the conserved phenylalanine in the J region. **B**: Sharing of unique sequences between pairwise comparisons of all donors, measured using the Jaccard index,for both complete and subsampled repertoires. Subsampling involves randomly selecting a fixed number of samples from each donor's repertoire before comparison. α chain repertoires were subsampled to 6000 unique sequences, β chain repertoires to 9000 (reflecting the lower bounds of number of sequences found in these groups). **C**: Frequencies of the CDR3s summarised in **A**. Horizontal dashes indicate group means.

samples and on subsamples, where a fixed and equal number of rearrangements were randomly selected from each donor before comparison, in order to control for differences in sequencing depth. We again see the sharing increase as we consider the different sequence types (CDR3 > VJ-CDR3 > DCR). As with Figure 3.6A, we also see that α chains are also far more shared than β chains. This probably reflects both the relative decrease in α chain diversity (Figure 3.1D) and the fact that there exist T-cell subpopulations that make use of so-called 'invariant' TCR α chains (discussed in more detail later in Section 3.1.5).

Shared and public clones are also reported to be present in higher frequencies in the repertoire (129). I investigated this in my data by plotting the proportional frequency of each CDR3 that was found shared among each number of donors (Figure 3.6C). Average values remain low in both chains up until the highest possible number of donors, indicating that the more public clonotypes are indeed more likely to be higher frequency than those that are less shared. However by looking at the underlying distributions we do see that there are equally high private clones (found in one or two donors), but these are just outweighed by the relative surplus of private clones that only occur at very low frequencies.

This analysis also revealed that there are non-productive nucleotide rearrangements (those expected to not translate correctly) that are also shared among multiple individuals (Figure 3.7A). Despite only a fraction of unique rearrangements being non-productive, there were still α chain sequences that were shared among all ten individuals, and β sequences that were found in four different donors. Pairwise comparisons revealed that between any two donors, non-productive sequences are actually significantly more shared than productive sequences, in either chain (Figure 3.7B). As with productive α chains, non-productive α sequences that were shared among more people showed larger mean frequencies than those shared among fewer (Figure 3.7C). Remarkably, for those rearrangements that were common to either nine or ten individuals, the non-productive recombinations averaged larger clone sizes than the productive.

Inspection of the overlap in repertoires between individuals (intra-donor) provides us information on TCR sharing and publicness. By taking multiple bleeds from one individual and comparing these (inter-donor), we are able to obtain information about the dynamics of the repertoire within a person. Four of our healthy donors were bled twice, three months apart, and processed. As before, I then calculated the Jaccard indexes of intra- and inter-donor pairwise repertoire comparisons (Figure 3.8A). As might be expected, two repertoires from the same donor share more TCR sequences


A
Frequency
Number of donors
Alpha P DCR
Alpha NP DCR
Beta P DCR
Beta NP DCR
B
Jaccard index
Whole Alpha
Subsampled Alpha
Whole Beta
Subsampled Beta
P DCR
NP DCR
Alpha
Beta
C
DCR frequency
Number of donors
Alpha P DCR
Alpha NP DCR
Beta P DCR
Beta NP DCR

than repertoires from two different donors. I did not observe non-productive rearrangements being more shared than productive for intra-donor as we do for inter-individual comparisons.

By plotting the proportional frequencies of rearrangements in the first bleed against those in the second we can see how stable the clonotypic hierarchy is within individuals. The examples shown in Figure 3.8B demonstrate the donors that had the best and worst correlation by R^2 value. Even for the samples with a poorer fit by linear regression show a clear pattern; low-frequency sequences (below 1×10^{-3}, or one in every 1000 recombinations) display little correlation and are often simply not present in the other bleed, while higher-frequency sequences match exceedingly well. This might be explained by lower frequency clones being less likely to be captured by sequential sampling events.

3.1.5 Invariant TCR detection

There exist various subpopulations of phenotypically distinct $\alpha\beta$ T-cells that make use of so-called 'invariant' TCRs, which are composed of near-identical α chains and a small pool of β V genes. Currently, we know of three such cell types, which use these restricted diversity receptors to recognise non-canonical ligands bound to various MHC-like molecules. Discovered first, invariant natural killer T-cells (NKT) respond to glycolipid antigens bound to CD1d, using a TRAV10 - TRAJ18 containing TCR (132, 133, 134). Mucosal-associated invariant T-cells (MAIT) are MR-1 restricted and recognise microbial vitamin metabolites with TCRs using TRAV1-2 primarily in combination with TRAJ33, although variants using TRAJ12 and TRAJ20 have recently been described (135, 136, 137, 138). The most recently categorised population are the germline-encoded mycolyl lipid-reactive (GEM) T-cells, which use TCRs containing a TRAV1-2 - TRAJ9 α chain, and recognise mycobacterial lipids presented by CD1b (139, 140). Having noticed that two types of invariant TCR-bearing cells make use

Figure 3.7 *(preceding page)*: Non-productive TCR sharing and publicness. A: Publicness of non-productive TCR rearrangements, inferred from `Decombinator` indexes (or DCRs) that are not expected to encode a productive TCR polypeptide. P = productive, NP = non-productive. **B**: Sharing of DCRs that describe either productive or non-productive rearrangements, for either whole or subsampled repertoires. Samples were subsampled down to the lower bound of NP samples, which were 1000 for α chains and 260 for beta chains. Note that there are no subsampled P DCR results for beta, as all of these comparisons resulted in no shared clones. **C**: Frequencies of P and NP DCRs summarised in **A**. Horizontal dashes indicate group means.

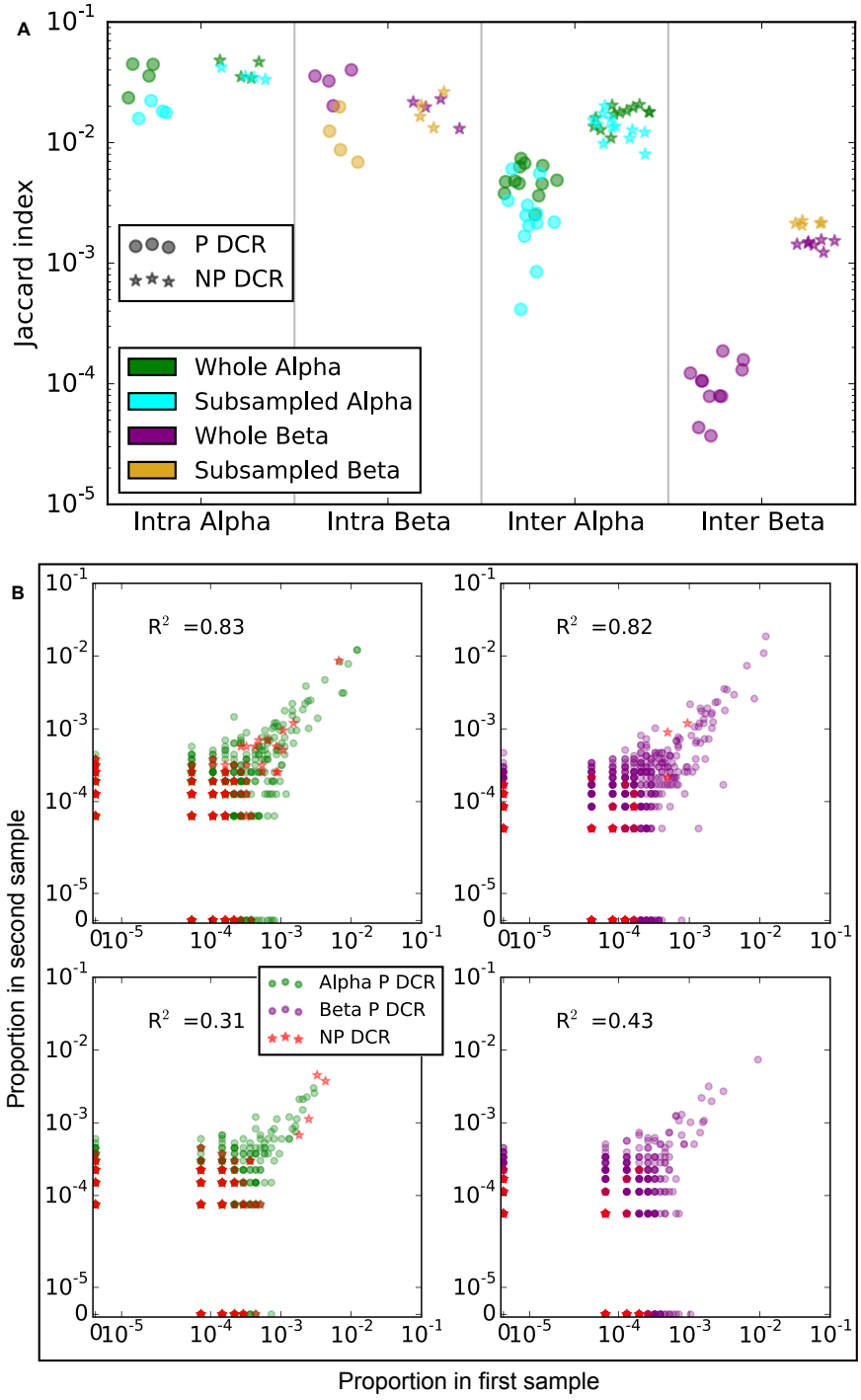


Figure 3.8: Intra-donor TCR sharing. A: Overlap in DCR repertoires of the four healty donors from whom we took two peripheral blood RNA samples, three months apart, measured by Jaccard index. Both productive and non-productive (P and NP) DCR were compared, for both complete and subsampled repertoires. 'Intra' indicates comparisons of the two bleeds from one individual, while 'inter' indicates comparison of two bleeds from different individuals (but taken at the same time). α repertoires were sampled down to 1,300 rearrangements, while β repertoires were subsampled to 300. **B**: Example plots demonstrating the intra-individual concordance of rearrangement frequencies. α and β repertoires are shown for the individuals with the highest and lowest degrees of correlation by R^2 value (HV01 and HV03; higher and lower two plots respectively). Note that the first and second sample bleeds were all taken at the same times of the year for each donor.

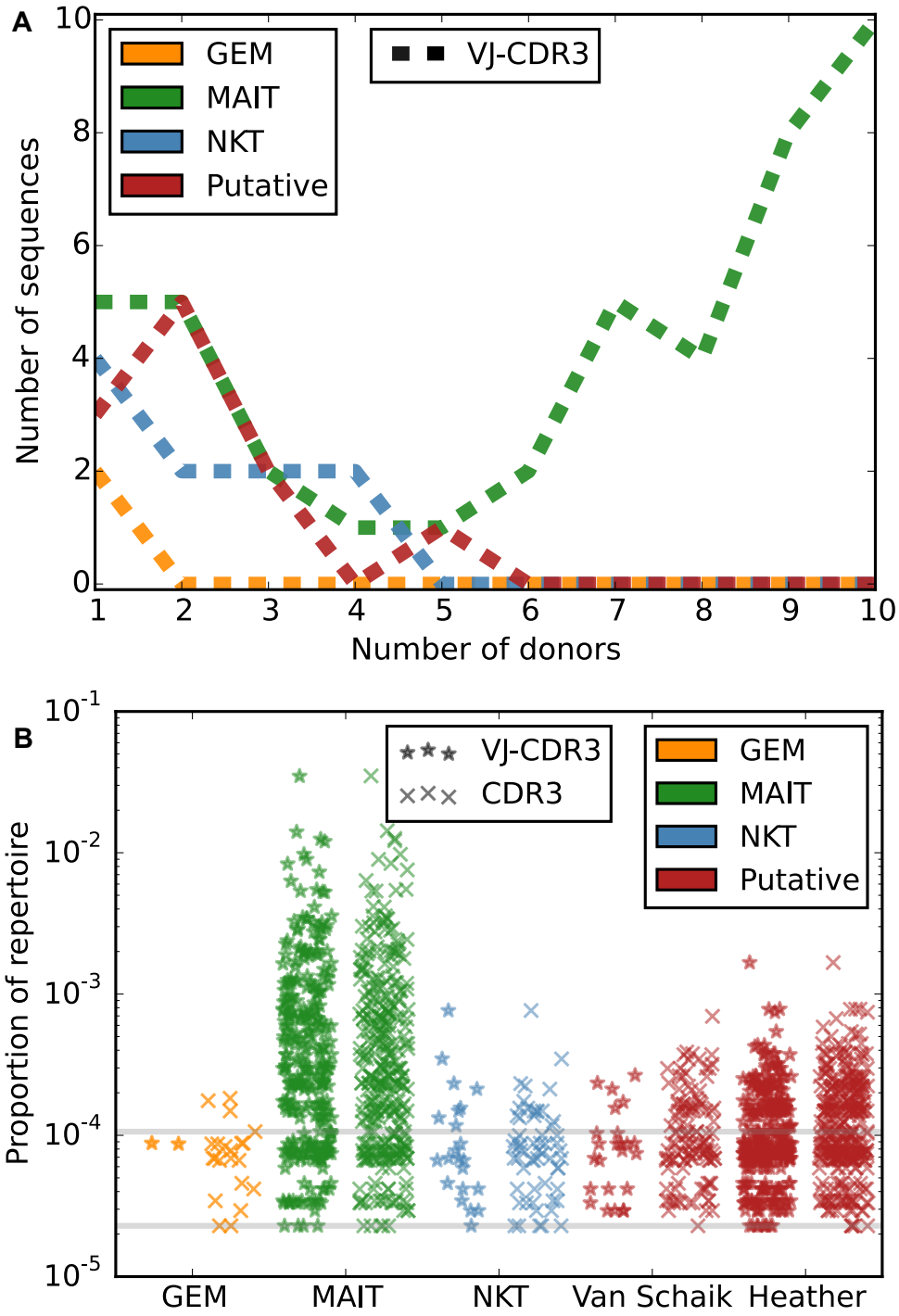


Figure 3.9: Detection of known and putative invariant α chain TCRs. A: Number of published invariant TCR VJ-CDR3s that were found in our healthy donor peripheral blood repertoires. The sources of these sequences are listed in Table A.2. **B**: The size of the published CDR3s and VJ-CDR3s searched for in our donors, as proportions of the repertoires that they were found in. Note that the 'Putative' group shown in **A** does not include the putative sequences derived from the healthy donors presented later in this chapter.

of the same α V gene (TRAV1-2), the group that discovered GEM cells have recently performed targeted deep-sequencing of α chains that specifically use this gene (141). This process has generated a number of novel sequences that are shared among multiple individuals, which they claim represent potential invariant T-cell subpopulations. I harvested published sequences for these invariant TCRs and the putative invariant shared TRAV1-2 sequences (referenced in Table A.2), and looked for them in the healthy donor repertoires in the context of the correct V and J genes.

As Figure 3.9A shows, we are able to detect invariant α chains belonging to each population at the VJ-CDR3 level to some extent. Most of these TCRs are only detectable in a small subset of donors. Only two of the five described GEM VJ-CDR3 sequences were found in this data, in different donors, with a slightly greater number of NKT and putative TRAV1-2 invariant sequences in more donors. This is likely due the prevalence of these cell types: human NKT cells account for approximately 0.2% of T-cells in blood (142), while GEM are two orders of magnitude less frequent (140). However, far more sequences derived from MAIT cells are found, with at least ten VJ-CDR3s being detected in each of the donors. While this search was repeated for β chains, this search yielded almost no results; only two MAIT β CDR3 sequences (lacking V and J gene context) were detectable, at low frequency.

Figure 3.9B shows how prevalent each detected invariant TCR sequence was in the repertoire file in which it was found. GEM sequences were only identified close to the limit of detection for this assay, with NKT and the putative sequences from Van Schaik *et al* being slightly more frequent. MAIT cell sequences are detected far more often and at higher frequency, with one sequence making up 3% of all sequences in the α repertoire it was found in. MAIT cell sequences reach such high proportions in blood that the contribution of their invariant α chain is readily apparent even by looking at V-J gene pairing frequencies, as was was shown earlier in Figure 3.3 (the most frequent pairing is TRAV1-2 - TRAJ33, the canonical MAIT combination). This is consistent with existing flow cytometry data showing that MAIT cells can be exceedingly common in blood, sometimes accounting for upwards of 10% of $CD8^+$ T-cells (143).

I also used my data to identity potential invariant TCRs based on degree of publicness. We have already observed that there are a number of α chain VJ-CDR3 sequences shared among the majority of the healthy cohort (Figure 3.6A). I took the sequences that were found in seven or more donors, and subtracted any that matched a pre-existing known or potential invariant population. This left 40 sequences (listed in Table A.3, Appendix A). One sequence was found in all ten donors, using TRAV13-1 and TRAJ6, bearing the CDR3 CAASKGGGSYIPTF. This does not appear to be any

kind of contamination, as this same sequence is generated using different nucleotide sequences in different donors. Searching for these sequences in the data produced a large number of hits, at greater frequencies than all other groups apart from MAIT (which might be expected, seeing as this list was defined by being common in this dataset).

3.1.6 Section discussion

TCR repertoires have a number of properties by which they may be characterised. Perhaps the simplest attribute to measure is the spread of frequencies of different rearrangements. The resultant distribution is characterised by an abundance of very low frequency sequences, and a small number of more common recombinations. Description of the frequency landscape of recombinations in healthy donors may function as a benchmark against which to assess samples with expected repertoire perturbations. When doing so, we must bear in mind that approaches that start with an RNA substrate (such as the one used in this thesis) do not allow one to infer a direct correlation between sequence frequency and cell counts, as different cells may express different amounts.

The presence of non-productive (NP) rearrangements detected in this RNA-derived data may initially seem surprising, given that mechanisms exist to downregulate such sequences. Various epigenetic systems mediate allelic exclusion of chromosomes that do not contain productively rearranged alleles (144, 145, 146, 147), and any transcripts that are made from such rearrangements are thought to be efficiently downregulated by nonsense-mediated decay (NMD) (148, 149). Therefore NP reads potentially just represent the 'leakiness' of the system, which seems fairly considerable in the α chain data. Allelic exclusion has been reported to be stronger at the TRB locus, a fact which can be inferred from the relative ability to find cells transcribing two productively rearranged α chains (72, 150). The majority of observed NP sequences from either chain contain premature stop codons, which are thought to be the strongest signal for downregulation by NMD (151). These two findings might seem incongruent, but likely reflect that the majority of NP reads would be expected to contain stop codons.

These data also demonstrate that some supposedly non-functional germline TCR genes contribute to relatively high frequency rearrangements. This does not necessarily correspond to their involvement in functioning TCR heterodimers: indeed some of the pseudogenes would be expected to not be able to produce any peptide at all. However I believe that there is a strong case for the official status of several of these genes to at least be reconsidered, if not reclassified. For instance, the frequently-found gene TRAJ58 is classified as an ORF as it contains an in-frame stop codon (152); however

as this resides towards the very 5' of the gene it can very easily be removed by nucleolytic nibbling during recombination. This allows it to produce perfectly theoretically functional rearrangements as appear in this data, as well as in several published reports that selected T-cells based on TCR binding (153, 154, 155). Another oft-detected ORF in our data was TRAJ35 (which is classified as such due to the presence of a cysteine in place of the phenylalanine in its FGXG motif) is found in valid rearrangements in several papers (156, 157, 158). TRDV1 particularly, which is shown to be consistently used at reasonably high frequencies in this data, has been found to be rearranged to a variety of TRAJ genes by a number of groups, some of whom detected surface expressed protein (127, 159, 160, 161, 162, 163, 164, 165, 166). The most recent paper to make note of this claims that TRDV1 is used with sufficient frequency and breadth of TRAJ genes that it merits a change of name to TRAV42/DV1 (127) (although this would not fit current IMGT numbering conventions). It even appears that IMGT is aware of TRDV1 recombining with at least one TRAJ gene (167), but does not count this as a dual TRAV/DV usage V gene (168), stating:

> *"Although the TRDV1 gene was also found rearranged to a TRAJ gene and therefore expressed with the TRAC gene ... it is considered as a TRDV gene."*

Even recombinations using genes which would prevent translation still contain information, even if they arise due to some basal transcriptional noise of the system, and thus are worthy of inclusion at the primary stages of TCR analysis. For instance, to the best of my knowledge this thesis is the first report of recombination between TRDV2 and TRDV3 with any TRAJ genes in a healthy donor's repertoire[3]. I predict that comparably high-resolution investigation of δ chain TCR repertoires would find V genes currently considered to be purely involved in α chain recombination contributing to TRDJ- and TRDC-containing δ chain rearrangements.

I have also demonstrated that β chain `Decombinator` classifier insert sequences can be searched to permit detection of TRBD genes. Due to the nature of these genes, they are often impossible to detect or assign reliably. One gains most sensitivity by using shorter match length thresholds, but this reduces the specificity of the search as shorter sequences are more readily produced by chance. This is evidenced by the presence of false positive TRBD hits in α chain sequences. Even taking these limitations

[3]The only papers I have found to describe TRDV2-TRAJ recombinations – which remarkably even included TRDD genes – did so in cells derived from leukaemia patients, and were predicted to not produce functional protein (169, 170).

into account, these results demonstrate that one is able to extract some information regarding the recombination of TRBD genes in data generated using the pipeline in this thesis. Doing so I have repeated some notable findings from other studies, such as the pairing of TRBD2 with TRBJ genes from the TRBD1 cluster, and preferential pairing between TRBV and TRBD genes. The presence of VDDJ rearrangements is particularly interesting, as among T-cells these are often thought to be the sole province of δ chains, due to the configuration of the RSS around the D genes (171, 172). This raises the possibility of there also being direct VJ β chain rearrangements existing (173). Such rearrangements would however be hard to distinguish from genuine VDJ sequences that just had heavily deleted TRBD genes. While assignation of residual D gene nucleotides is not a requirement for the majority of analyses I describe in this thesis, such efforts would be of great importance in experiments inferring the mechanisms of recombination from sequence data.

There is a growing body of work describing sharing, 'publicness' and bias in TCR repertoires. Since the first descriptions of public TCRs recognising specific influenza, lysozyme and Epstein-Barr virus antigens (174, 175, 176), such responses have been discovered in many reactions across all species studied, prompting a great deal of discussion (177, 178, 179, 180). Turner *et al* have proposed a system of classification for the different types of bias observed in a response to an antigen (177). Type I bias describes responses that deploy multiple TCRs all using the same V gene, type II bias involves TCRs that share common amino acid motifs or features in their CDR3s, and type III bias denotes the sharing of actual nucleotide sequences. Miles *et al* later proposed an additional class, type IV bias, which describes those TCRs that fall between types II and III; such sequences share V gene identity and a significant proportion of matching CDR3 sequence, but differ in a small number of residues (179). While I am not presenting any antigen-specific data in this thesis, we can still infer the existence of each of these levels of bias in this global repertoire data by merit of the shared TCR features and sequences.

One surprising result from the publicness analyses described in this chapter is the extent to which non-productive (NP) rearrangements are shared between people. While infrequently detected (for reasons discussed in Section 3.1.1), on a per-TCR basis NP sequences were more likely to be shared between people than those that were productively rearranged. I propose that these data are consistent with the hypothesis that TCR sharing occurs as a function of generation probability, i.e. these TCRs are for some reason more likely to be made in multiple individuals (131). In this model, TCRs are produced at different efficiencies, therefore those that are generated most easily or most

often are those most likely to be shared. Productively rearranged receptors, regardless of potential publicness, will be subject to thymic selection, thus a portion of these will not be detected in the periphery. Non-productively rearranged receptors however do not contribute to ligand recognition and therefore cannot be selected against, therefore it is reasonable to assume they will exist in different individuals at some rate more proportional to their production. NP sequences are therefore more shared than productive as they have not had some potential public sequences removed. This higher level of sharing is only manifest between individuals; NP TCRs were not more likely to be found in two bleeds of the same person taken several months apart. This is consistent with the idea that we are observing NP TCRs at some rate proportional to their production; those we see at relatively high-frequencies being retained have likely been generated in thymocytes whose functional TCR chains have promoted their expansion.

These data do not however provide information as to the mechanism of unequal TCR production frequencies. One of the leading models for explaining public TCR responses is that of convergent recombination, which postulates that such sequences arise by merit of multiple recombination events producing the same nucleotide sequence (178, 181, 182). However one group has applied various statistical analyses to a selection of NP TCR sequences (obtained by sequencing from genomic DNA) and found that convergent recombination does not explain sharing, lending weight to the possibility that biases in the recombination machinery are instead responsible for production of public sequences (183, 184).

Related to measuring publicness is the detection of sequences known to exist in the non-pMHC-restricted T-cell subsets that express relatively invariant TCRs[4]. In this study, working from a fraction of RNA extracted from a 2-3ml sample of peripheral blood, the sequences we can most readily detect are those belonging to MAIT cells, which are highly prevalent cells in human blood[5]. However if other populations were a focus of research one could increase their detection through sequencing RNA either from larger volumes of blood, sorting cells, or by isolating RNA from other tissues in which these cells reside.

One exciting possibility offered by repertoire sequencing is the identification of potential unknown invariant populations. Van Schaik *et al* recently reported just such an exercise, looking for sequences that were highly shared in TRAV1-2$^+$ cells from three

[4]However seeing as each so-called 'invariant' TCR is actually a pool of several similar but non-identical sequences, the field may at some point wish to address this choice of word.

[5]Note that NKT cells are more prevalent in mice than in humans (185), while MAIT cells are more prevalent in humans than in mice (135), so this finding is likely species-specific.

or more donors (141). They found a number of sequences, some of which were identifiable in this data, with more regularity than seen for iNKT or GEM cell sequences. An analogous search for highly shared sequences in my data, found a variety of highly shared sequences, one of which was present in all ten donors. Without any evidence of a phenotypically unique cell subset possessed of non-pMHC restriction we cannot differentiate these sequences from just highly shared clonotypes, perhaps representing public responses against common antigens. The line between invariance and type IV TCR bias will be a technologically difficult one to draw. Characterisation of cells using single-cell RNA sequencing data, permitting both identification of TCR chains and expression of lineage-specific markers, could represent a sequencing based solution to more confidently assert the existance of such cells. As for presentation, there are a great many MHC and MHC-like genes that could be responsible for displaying antigens to such hypothetical subsets. The original draft of the human MHC predicted 128 different protein-coding genes (186), while later work – which counted a larger portion of chromosome 6 – raised this estimate to 252, of which roughly 70 are predicted to have immune functions (187). A number of class I-like genes additionally exist outside of the MHC (188) which might serve in such a role. Indeed, all of the molecules that we currently know to display antigens to invariant T-cell populations – CD1 and MR1 – fall into this category, with their genes located on chromosome 1 (189, 190).

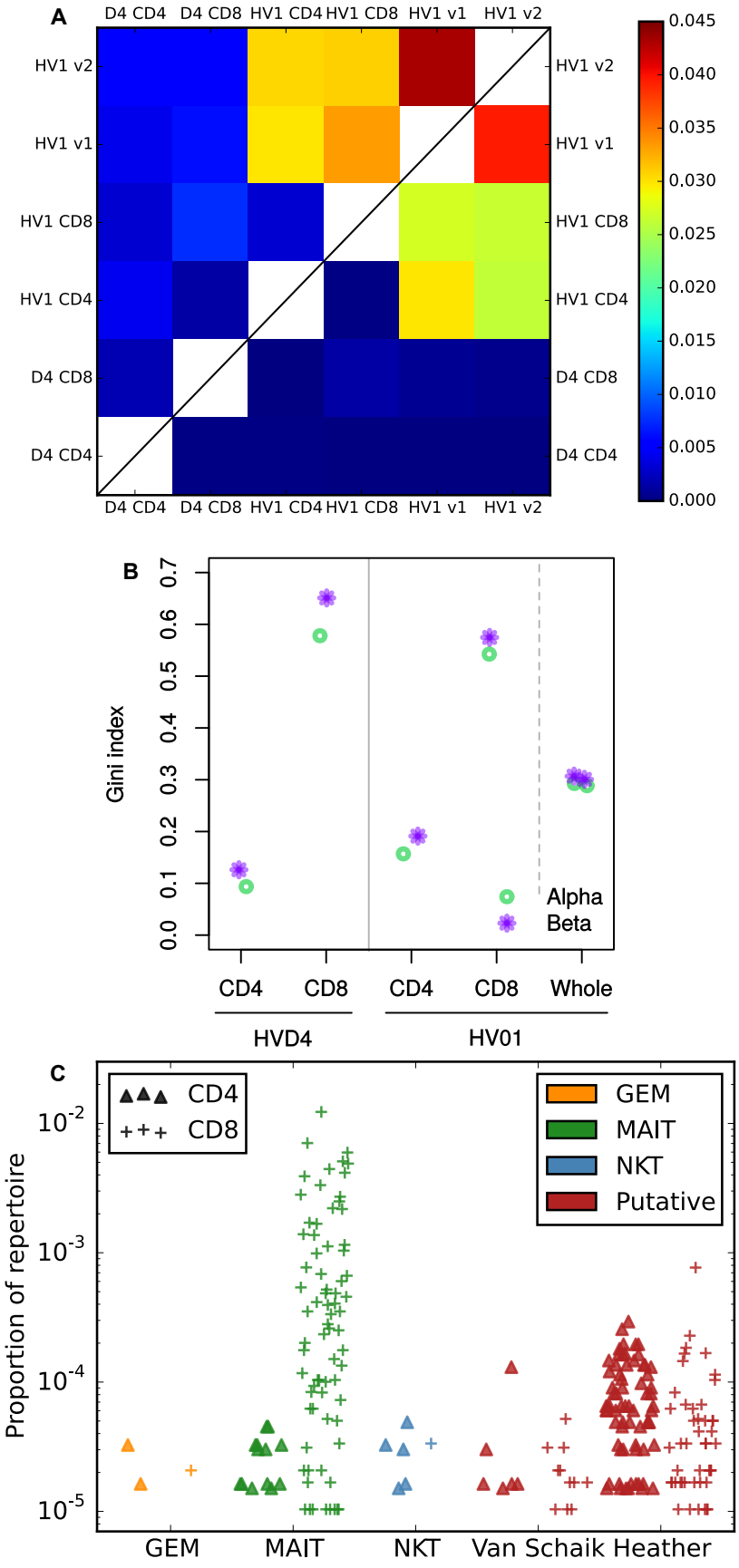
High-throughput sequencing is therefore able to extract and quantitate a number of diverse inter- and intra-individual TCR repertoire features, working from a small sample of blood.

3.2 Subset-specific TCR repertoire findings

In addition to investigating the global repertoire of peripheral blood, I also undertook a number of trial experiments to assess features of TCR repertoires of particular cell subsets. Herein I additionally explore data produced using different protocols, which only amplify rearrangements using a limited selection of V genes.

3.2.1 CD4 and CD8 TCR repertoires

T-cells can be divided into two major functional subsets – T helper cells or cytotoxic T-lymphocytes (CTLs) – based on the expression of either the CD8 or CD4 glycoprotein TCR co-receptor. As these molecules associate with T-cells that recognise peptides bound to different families of MHC molecules – class I or II – we hypothesised that their TCR repertoires would differ substantially, reflecting the significant structural


A
D4 CD4
D4 CD8
HV1 CD4
HV1 CD8
HV1 v1
HV1 v2
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
B
Gini index
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Alpha
Beta
CD4
CD8
CD4
CD8
Whole
HVD4
HV01
C
CD4
CD8
GEM
MAIT
NKT
Putative
Proportion of repertoire
10^{-2}
10^{-3}
10^{-4}
10^{-5}
GEM
MAIT
NKT
Van Schaik
Heather

differences in ligands. In collaboration with Dr Theres Oakes, PBMC from two healthy donors were sorted into $CD4^+$ and $CD8^+$ populations and their TCR repertoires were sequenced. Cell purities are displayed in Appendix A in Figure A.2.

Overlap analysis revealed that there is very little sharing between the CD4 and CD8 compartments within a donor (Figure 3.10A). The CD4 and CD8 compartments of the two different individuals actually share more sequences with one another than do the separate populations within each individual. One of the healthy volunteers who donated blood for sorting was also one of the donors who provided two samples for previous analyses (HV01, see Figure 3.8). The blood to be sorted was drawn approximately two years after the two previous whole blood samples, yet still shares a large number of sequences with both CD4 and CD8 populations. The overlap between the two whole blood samples of this sample taken three months apart was included in Figure 3.10A for reference.

I also inspected the inequality of the sorted repertoires. CD8 repertoires display a far more unequal distribution than CD4, representative of a more oligoclonal distribution (Figure 3.10B). The whole blood samples from the same donor display intermediate equalities, being a combination of both populations. This difference in inequality is apparent even when viewing the VJ gene pairing frequencies of each population (Figure A.3 to A.6, Appendix A). CD4 populations in both chains clearly display a more even spread of gene usage compared to CD8, which is typified by a combination of hotspots and many infrequently used combinations.

Another feature that was apparent from viewing the VJ pairing of one of our sorted repertoires (HV01) was the segregation of large amounts of TRAJ1-2 - TRAJ33 rearrangements (the gene pair that contributes to canonical MAIT cell invariant TCRs) to the CD8 repertoire (Figure A.3). Donor HVD4's CD8 repertoire does not prominently display this pairing (Figure A.4, although there are indeed a number of MAIT

Figure 3.10 *(preceding page)*: Clonal composition of the sorted CD4 and CD8 TCR repertoires. A: Heatmap showing the overlap between sorted CD4 and CD8 subsets of two donors, HV01 and HVD4 (D4), by Jaccard index. Sorted samples from HV01 were taken approximately two years after the two HV01 bleeds (v1 and v2) that were used in Section 3.1. Upper-left hand triangle shows comparisons of α chain rearrangements, lower-right triangle shows beta. **B**: Clonal inequality of CD4 and CD8 compartments by Gini index. Unsorted whole blood samples from HV01 from two years before were included for reference. Plots **A** and **B** were both generated using subsampled repertoires to control for different sample sizes; 10,000 DCRs were used for alpha, while 20,000 were used for beta. **C**: Proportional frequencies of the VJ-CDR3 sequences of invariant α chain TCRs (discussed in Section 3.1.5) as found in either file of CD4 and CD8 sorted TCR data.

cell sequences present). Instead this population is dominated by large expansions of clonotypes made using TRAV12-2 and TRAJ18, which account for 16.4% of all rearrangements. Within this VJ combination, 88.5% produce CDR3s that fit the pattern CAVXXDRGSTLGRLYF (where X represents any amino acid). This raised the possibility that we might be able to confirm and/or determine the CD4/CD8 expression patterns of different invariant TCR clonotypes. I searched for the presence of known and potential invariant α chain sequences in the sorted repertoire data (Figure 3.10C). As with previous analyses, very few GEM and iNKT sequences were detected, however those that were found were predominantly in the CD4 compartment. MAIT sequences were present in both populations, but only reached the large populations sizes observed previously in the CD8 subset. The putative rearrangements from Van Schaik's and my own data are found distributed approximately evenly between both populations. Of particular note is the sequence that was earlier found to be present in all ten healthy donor whole blood samples (TRAV13-1 - TRAJ6, CAASKGGSYIPTF). Not only was it found exclusively within the $CD4^+$ compartment of both sorted samples, but it was highly prevalent in each, being the most frequent non-MAIT invariant sequence found in the CD4 data (from HVD4). The CD8 VJ-CDR3 sequence with the highest frequency was TRAV12-2 - TRAJ5, CAVNNTGRRALTF (also in HVD4), one of the sequences that was earlier identified in eight of ten of our whole blood healthy samples, although this sequence was found at low frequency in the CD4 population of both donors. Of those 40 sequences, the majority were found in both CD4 and CD8 repertoires; ten were found exclusively in CD4, while only one was found exclusively in CD8 (Table A.3).

3.2.2 Clonality distribution differences using targeted V amplification protocols

In the process of developing the error-correcting pan-TCR amplification protocol used in the rest of this thesis (the process of which is covered in Chapter 5), I tested various intermediate protocols which specifically target a limited number of V genes. These intermediate protocols were used to compare the repertoires of two pairs of cell subsets: naïve versus memory, and cord blood versus adult peripheral blood.

3.2.2.1 $CD4^+$ $CD45RA^+$ versus $CD45RO^+$ clonal distributions

The first such experiment targeted rearrangements using the α and β V genes TRAV8-4 and TRBV12-3 (see Section 5.2.2 for details). Three healthy donors were bled and PBMC extracted. One donor's PBMC were processed and sequenced, while the other

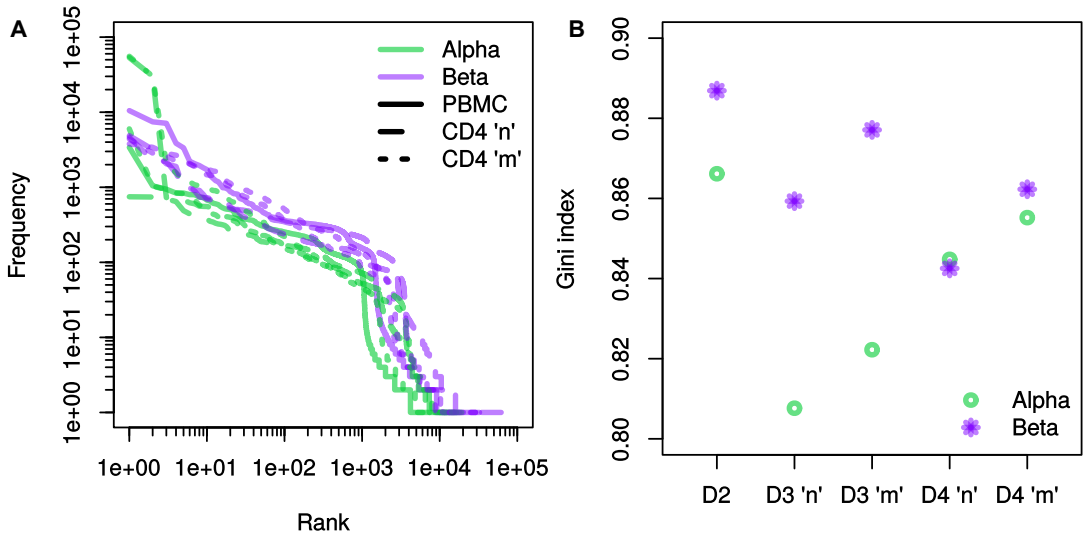


Figure 3.11: Clonal composition of CD4$^+$ CD45RA$^+$ or CD45RO$^+$ TRAV8-4$^+$ and TRBV12-3$^+$ repertoires. **A**: Logged rearrangement rank versus logged frequency for the five samples sequenced using the simple V amplification protocol, targeting specific V genes. Thus α results are exclusively those that are shown by `Decombinator` to be using TRAV8-2 or TRAV8-4 genes (which are highly similar in their 3' regions, and are thus indistinguishable). β samples shown use TRBV12-3. D2 samples were whole, unsorted PBMC. D3 and D4 were sorted into CD4$^+$CD45RA$^+$ ('naïve' or 'n') and CD4$^+$CD45RO$^+$ ('memory' or 'm'). **B**: Gini indexes for the repertoires shown in **A**.

two donors' cells were sorted into $CD4^+$ populations expressing either CD45RA or CD45RO, using magnetic assisted cell sorting (MACS), before sequencing. These populations have been used to enrich for naïve and memory populations respectively (191, 192). It must however be noted that using differential CD45 isoform expression does not provide true naïve-memory separation, as there is a subset of antigen-experienced memory cells that lose CD45RO and re-express CD45RA; such highly differentiated cells are called T_{EMRA} (T effector memory RA^+) (193), and will be included in our 'naïve' sorted population.

Figure 3.11A shows that the pattern of rearrangement frequencies by rank are broadly analogous between the groups. This implies that there are both expanded clones in the naïve population and a large number of memory clonotypes that are not expanded. Examination of the Gini indexes of these populations (Figure 3.11B) does however reveal that within a donor, the $CD4^+CD45RO^+$ population does display greater inequality, indicative of a greater proportion of the memory repertoire being occupied by relatively expanded clones.

3.2.2.2 Cord blood versus adult whole blood clonal distributions

The next subset experiment used a more refined experiment: while still priming from specific V genes in the PCR, the requisite sequences were introduced to allow error- and frequency-correction (described later in Sections 5.2.5 and 5.3.2), analogous to the approach used to correct the majority of the data in this thesis. This 'barcoded V' (BV) protocol targeted two additional V genes to the previous V-primed strategy – TRAV21 and TRBV20-1 – which had consistently appeared at high frequencies in our preliminary pan-V experiments. This protocol was used to explore the clonotypic structure of TCR rearrangements in cord blood relative to whole peripheral adult blood, focusing on this small number of genes.

Analysis of the specific-V data revealed that both adult and cord blood display the typical linear log-rank pattern, in which each rearrangement is a certain proportion bigger than the previous one (Figure 3.12A and B). Thus both populations contain a small number of relatively expanded and a large number of very-low frequency sequences; however sequences in adult blood reach greater frequencies than cord blood. This feature is recapitulated in their diversity indexes: cord blood repertoires display lower Gini indexes and greater Shannon entropies, indicative of more evenly distributed diverse repertoires (Figure 3.12C and D).

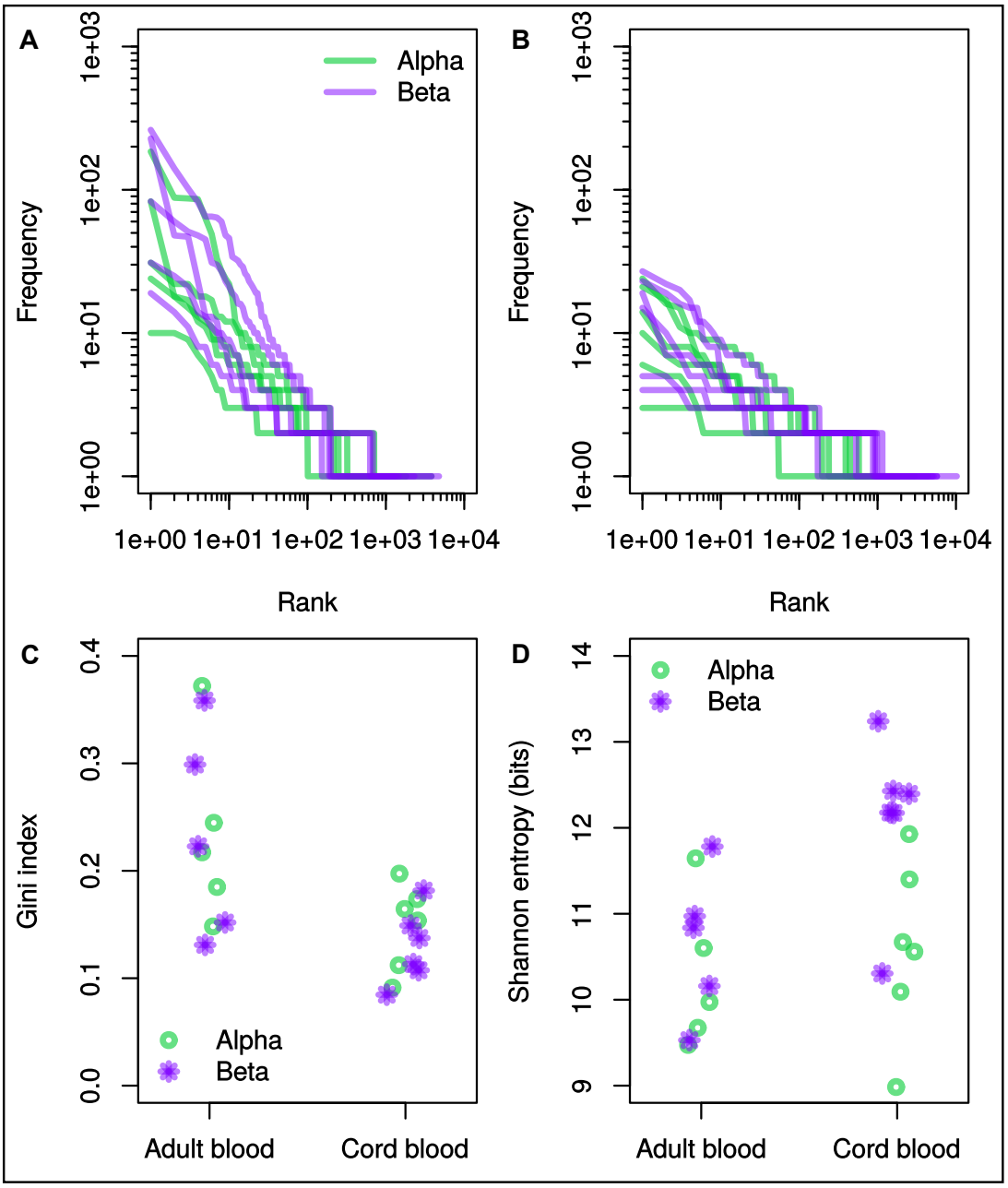
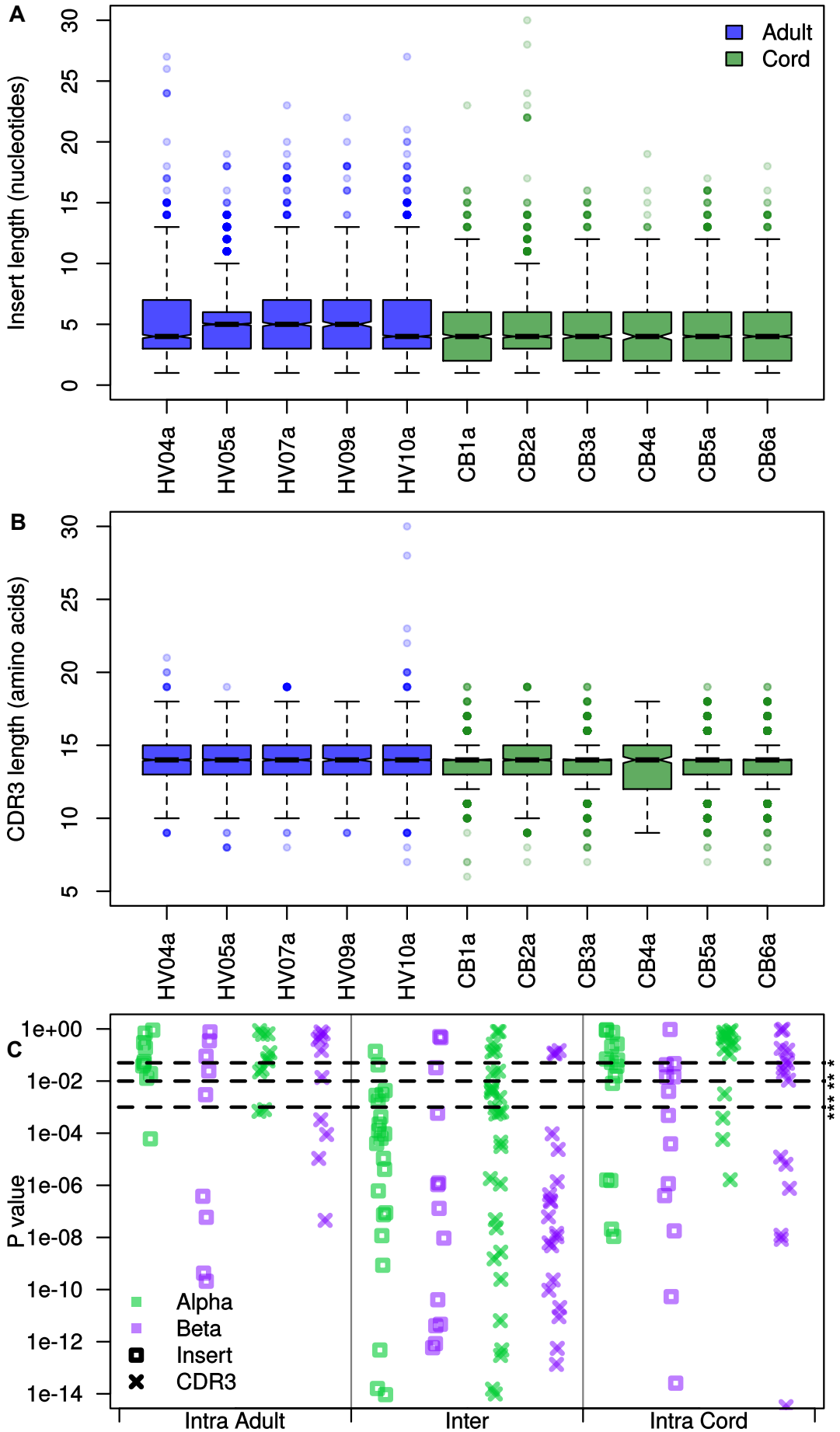


Figure 3.12: Clonal distributions and diversity indices of adult and cord blood V-specific repertoires. Logged rank-frequency distributions for whole peripheral adult blood (**A**) and neonatal cord blood (**B**). **C**: Gini indexes for adult and cord blood samples. **D**: Shannon entropies for adult and cord blood samples. Cord blood samples displayed significantly lower Gini indexes ($p = 5.746e\text{-}05$) and higher Shannon entropies ($p = 0.03458$), by one-sided Wilcoxon rank sum tests. All data in this section was produced using the BV protocol, and therefore only contains rearrangements using TRAV8-2/TRAV8-4, TRAV21, TRBV12-3 and TRBV20-1.


A
Insert length (nucleotides)
0
5
10
15
20
25
30
Adult
Cord
HV04a
HV05a
HV07a
HV09a
HV10a
CB1a
CB2a
CB3a
CB4a
CB5a
CB6a
B
CDR3 length (amino acids)
5
10
15
20
25
30
HV04a
HV05a
HV07a
HV09a
HV10a
CB1a
CB2a
CB3a
CB4a
CB5a
CB6a
C
P value
1e+00
1e-02
1e-04
1e-06
1e-08
1e-10
1e-12
1e-14
*** *** ***
Alpha
Beta
Insert
CDR3
Intra Adult
Inter
Intra Cord

I investigated the distributions of the lengths of the inserts[6] and CDR3s of the two groups. Interestingly, despite adult donors' rearrangements tending to contain a greater number of non-templated nucleotides than cord blood (Figure 3.13A), CDR3s from both groups contain a median of 14 amino acids (Figure 3.13B). Despite the confounding presence of residual TRBD gene nucleotides this pattern was also observed in β chains, producing the same median length CDR3s (Figure A.7A and B). Pairwise comparisons of either length distribution between all donors reveals that inter-group pairs are more significantly different than intra-group pairs (Figure 3.13C), i.e. any given repertoire from one group is likely to have a length distribution more similar to another member of that group than the other. Within groups CDR3 length profiles also differ less than insert length profiles. We can infer two main findings from these data. First, that there exists a greater level of heterogeneity in how many inserted nucleotides one finds in TCR repertoires than in the lengths of the resultant CDR3s of those rearrangements. The second finding is that that both these length distributions differ significantly between neonatal cord and peripheral adult blood.

Sharing analyses that were first performed on the pan-V healthy data (presented earlier in Section 3.1.4) were applied to the BV data of this Section. While only considering a limited number of V genes, there are sequences from both chains that are present in the majority of donors (Figure 3.14A). These groups also share similar numbers of sequences with one another, although the variation is greater among adult samples (Figure 3.14B). This data thus provides another resource to look for highly public α chain sequences. Table A.4 in Appendix A details those VJ-CDR3s that were present in seven or more of the BV-processed samples. One of the V genes targeted – TRAV21 – was also found to be used by some of the highly public sequences discovered

[6]The fifth field of the `Decombinator` classifier that contains the nucleotide sequence between the recognisable ends of the V and J genes.

Figure 3.13 *(preceding page)*: Adult and cord blood α chain TCR length distributions. **A**: Box plot of α chain rearrangement insert lengths, i.e. the number of nucleotides between the recognisable ends of the V and J genes, for adult (HV) and cord blood (CB) samples. **B**: Box plot of α chain CDR3 lengths (running inclusively from the conserved cysteine contributed by the V gene to the conserved phenylalanine in the J gene). **C**: P-values obtained from running two-sided Wilcoxon rank sum tests on all the insert and CDR3 length distributions of pairs of donors shown in **A**, **B** and Figure A.7 in Appendix A. 'Intra' indicates comparisons within a group (e.g. cord blood versus cord blood) while 'Inter' indicates one distribution from each group. Dashed horizontal lines indicate levels of significance: 0.05 (*), 0.01 (**) and 0.001 (***). Note that despite a similar nomenclature, the adult healthy volunteers were not the same as those used in the pan-V gene results in Section 3.1.

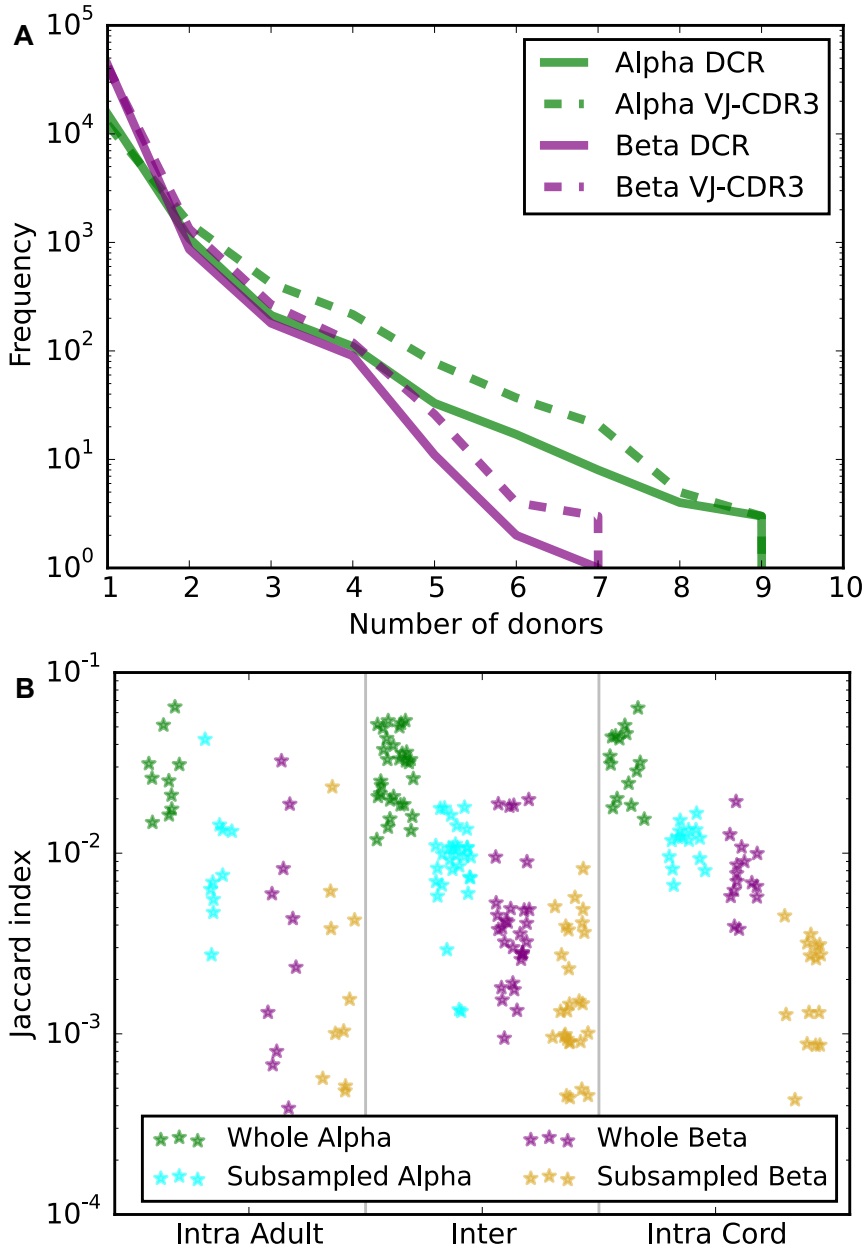


Figure 3.14: TCR publicness and sharing in the BV data. A: Publicness of nucleotide (DCR) and VJ-CDR3 sequences in the data produced using the barcoded V protocol, i.e. how many sequences were there found in a given number of donors. Both healthy adult and cord blood samples were pooled. Note that CDR3 sequences alone are not plotted as they exactly matched the VJ-CDR3 data, possibly as a result of targeting specific V genes in the amplification process. **B**: Sharing between the different groups amplified with the BV protocol, measured by Jaccard index. α repertoires were subsampled to 400 sequences, β repertoires to 1,200.

previously in data produced with the BLR protocol (Table A.3); cross-referencing these results showed that a number of sequences are found in both analyses. Thus we find sequences that are shared among up to fifteen different donors in our data alone.

3.2.3 Section discussion

In this section I report results from preliminary analyses made using combinations of specific cell populations and V gene repertoires. Most notable are the differences between the TCR repertoires of $CD4^+$ and $CD8^+$ T-cell populations. The most striking difference is the disparity of evenness of repertoires between the two groups, in agreement with previous findings (194), including deep-sequencing reports (99). This separation of polyclonality between $CD4^+$ and $CD8^+$ might be explained by differences in their proliferative programs; upon stimulation, CD8 cells undergo a greater amount of division, achieving larger burst sizes, that are more persistent than clonal exansions in CD4 cells (69, 195). Serendipitously, one of the donors bled for cell sorting was also one of the donors for whom there were data for the two whole blood samples (discussed in Section 3.1, which were taken approximately two years apart from the sorted samples), revealing a slightly higher long-term retention of sequences in the CD8 compartment.

The other notable finding is one replicating an observation from a previous TCR repertoire paper, that an individual's $CD4^+$ and $CD8^+$ populations do not show much overlap in TCR usage (83). In fact the CD4 and CD8 populations of the different donors shared more sequences than did the two populations within the same donor, highlighting the relationship between TCR identity and T-cell subtype. This makes intuitive sense, as CD4 and CD8 cells are recognising different pools of peptides (or 'peptidomes') in the context of different classes of MHC molecules, therefore one might expect a separation in their TCR repertoires. It was shown as early as the late 1980s that TCR sequences were in some way deterministic of the fate of that cell, as T-cells in mice expressing transgenic TCR sequences derived from a CD8 cell line were found to preferentially develop into $CD8^+$ cells (196). According to a current model of lineage commitment, this repertoire property is likely enforced by the signal strength of the TCR-pMHC reaction during development: stronger, longer signals produce CD8 cells, and short, weaker signals lead to CD4, while any cells assigned the incorrect fate are later deleted (40). Note that this if effectively a combination of the 'instructive' and 'stochastic' models originally proposed (197).

Data produced from MACS sorted cells underwent the initial pilot TCR amplification and library preparation protocol featured in this thesis. As this protocol lacks the

ability to apply barcode-based error correction, the data are somewhat rudimentary and are likely to be less robust quantitatively. Comparison of Figure 3.11 with similar data (e.g. Figure 3.1B and C) suggests the effect of PCR amplification and errors are skewing the results. The broad findings however – that low frequency clones occur in memory populations, and expanded clones could well exist in naïve populations – agree with those found in previous deep-sequencing efforts (83, 198).

A more refined protocol that also only targets specific V genes for amplification was applied to cord and adult whole blood samples. Cord blood is of particular medical interest due to the combination of large numbers of haematopoietic stem cells and relatively naïve T-cell populations possessed of lesser cytotoxic abilities, making it an attractive source of material for stem cell transplants (199). It also presents an interesting research opportunity to immunologists, as it allows inspection of T-cells arguably at their most naïve, having been shielded from microbial antigens whilst *in utero*. This likely explains the clonal distribution we observe in the cord blood, which is far more evenly distributed than adult blood, without so many clonal expansions, confirming previous spectratyping results (200). However there are *some* expanded sequences, which may reflect sequences that were produced frequently, perhaps as a result of convergent recombination or recombination bias as discussed above. I also demonstrate that despite producing the same median length CDR3s, adults' repertoires tend use TCRs with more inserted nucleotides (either N or P nucleotides) to encode them. Again, this supports previous published reports that describe foetal rearrangements tending towards shorter CDR3s than infants' (201), or adult blood tending towards longer CDR3s than cord blood (202). We know that in human embryos, Tdt – the enzyme responsible for the non-templated addition of nucleotides during recombination – is expressed before the 22nd week of gestation[7] (204). Therefore the difference between the degree of non-templated nucleotide addition between cord and adult blood cannot be explained by lacking the enzyme. Instead, I hypothesise this difference is a result of two features; the lack of expanded clones in cord blood and the probability distribution of producing TCRs with different numbers of insertions. Probabilistically speaking, the formation of TCRs with increasing numbers of nucleotides is less likely with each addition, therefore the longer the insert sequence the rarer that sequence can be expected to be produced by recombination. Adult repertoires have had both more chances to make such sequences and to increase their frequency by cell division

[7]Note that is not the case in mice, as they do not express Tdt until three to five days after birth (203).

(as some cells bearing many insertions can be expected to expand in response to their cognate antigen), thus such sequences should be more frequent in adults.

3.3 Intra-gene splicing of TCR transcripts

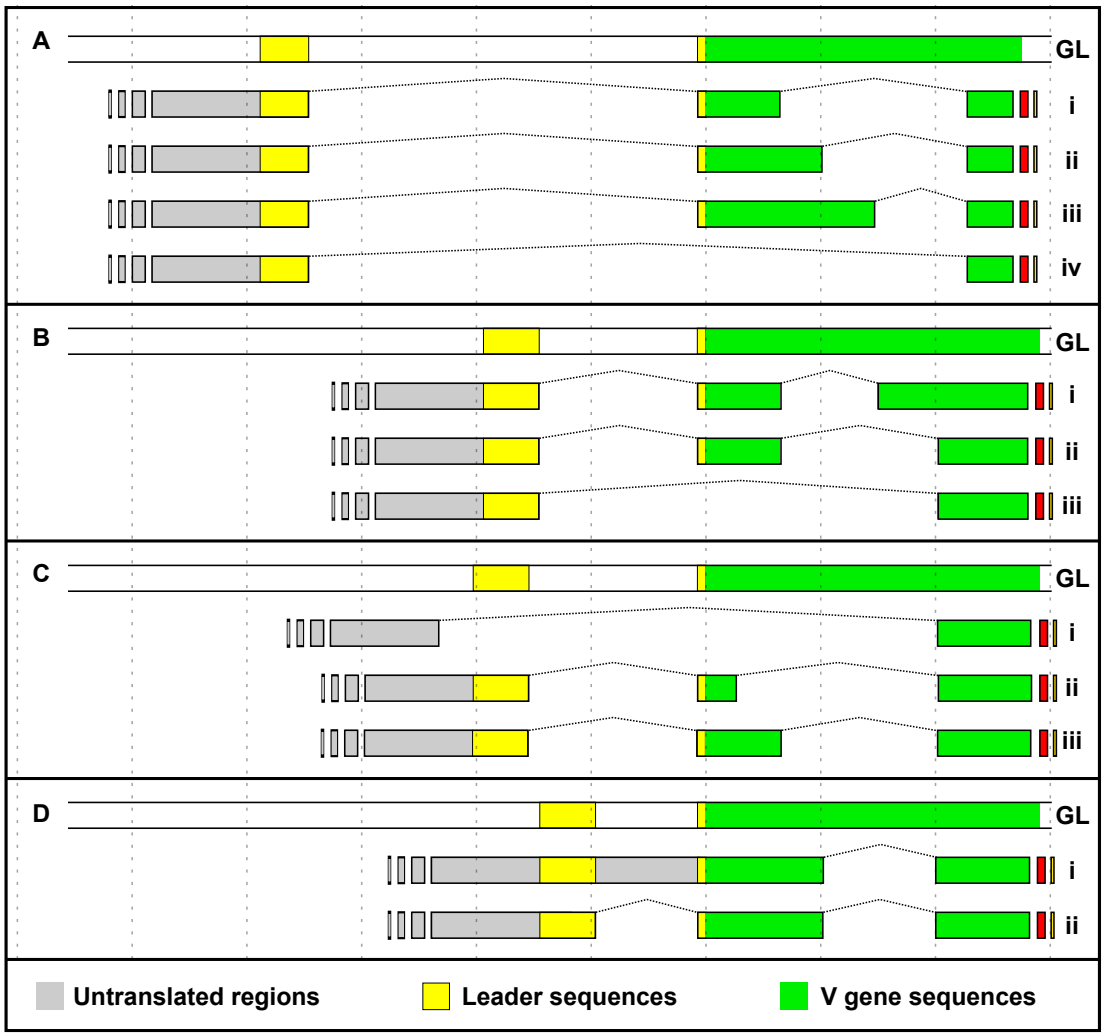


Figure 3.15: Examples of intra-V gene splicing. Examples of genes that display non-canonical splice events that interrupt V gene sequences, for TRAV1-1 (**A**), TRBV7-8 (**B**), TRBV7-3 (**C**) and TRBV2 (**D**). Chromosomal germline (GL) sequences for each sequence are included for comparison. Schematics are to scale: vertical dashed lines indicate 100 bp intervals.

Detecting rearrangements using non-functional TCR genes (as was shown in Figure 3.2) required the design a new set of V and J gene tag sequences, which are used to bioinformatically determine which specific genes are used in a TCR rearrangement[8].

[8]Discussed later in Section 5.3.1.1

During the course of testing this new 'extended' tag set, the frequency of one particular TRAV gene dropped significantly relative to its use in the original tag set. By inspecting the nucleotide sequences of reads that were assigned recombinations using the original tag set but not with the new, it was revealed that a number of transcripts using TRAV1-1 were subject to alternative splicing, to an splice acceptor site that lies within the new tag sequence.

In order to more systemically assess the extent of splicing within V genes I first needed to obtain nucleotide sequences that spanned the complete V gene. The sequencing protocol used in this thesis produces paired-end reads (i.e. one read from both ends of the template molecule), each 250 base pairs long[9]. By using the program `FLASH` (117) I was able to merge all of the reads that overlap in their 3's for one donor to obtain sequences which should contain whole transcript sequences. I then made use of the molecular barcodes introduced uniquely to each cDNA during the amplification process to error-correct our data. This differs to the error-correction employed in assigning the recombined TCR sequence: this splicing analysis requires accurate base calls across the entire length of the sequence, whereas the standard analyses are more tolerant of inaccuracies that would not affect calling the identity of a rearrangement. Accordingly more stringent filtering criteria were applied: in short, only those amplicons which were covered by more than three high-quality sequencing reads were considered, and then the most common base at each position was calculated to obtain a consensus for each transcript. These error-corrected transcripts were then processed using `Decombinator`, giving us rearrangement information. Recombined sequences up to the 3' of the V gene were then mapped onto the appropriate germline locus using the RNA alignment software `STAR` (114), and inspected for alternative splicing (AS).

A number of V genes in both chains were found to be subject to alternative splicing at some frequency: 30 distinct splice events were recorded divided across the 17 genes listed in Table A.5. Figure 3.15 shows scaled schematic representations of four example genes that between them demonstrate the breadth of different splice events taking place. These alternative spliceoforms often employ the standard splice event required to add the first exon (which contains the majority of the leader sequence, termed L-PART1 under IMGT nomenclature (19)) but then also splice out a downstream section of V sequence (e.g. Figure 3.15A i to iii). Some transcripts do however display other splicing profiles, such as splicing the L-PART1 sequence directly into the middle of the V gene (Figure 3.15A iv), making use of a splice donor upstream of L-PART1 (3.15C i), or failing to remove the leader sequence intron (Figure 3.15D i).

[9]This process is detailed later in Section 5.2.6.

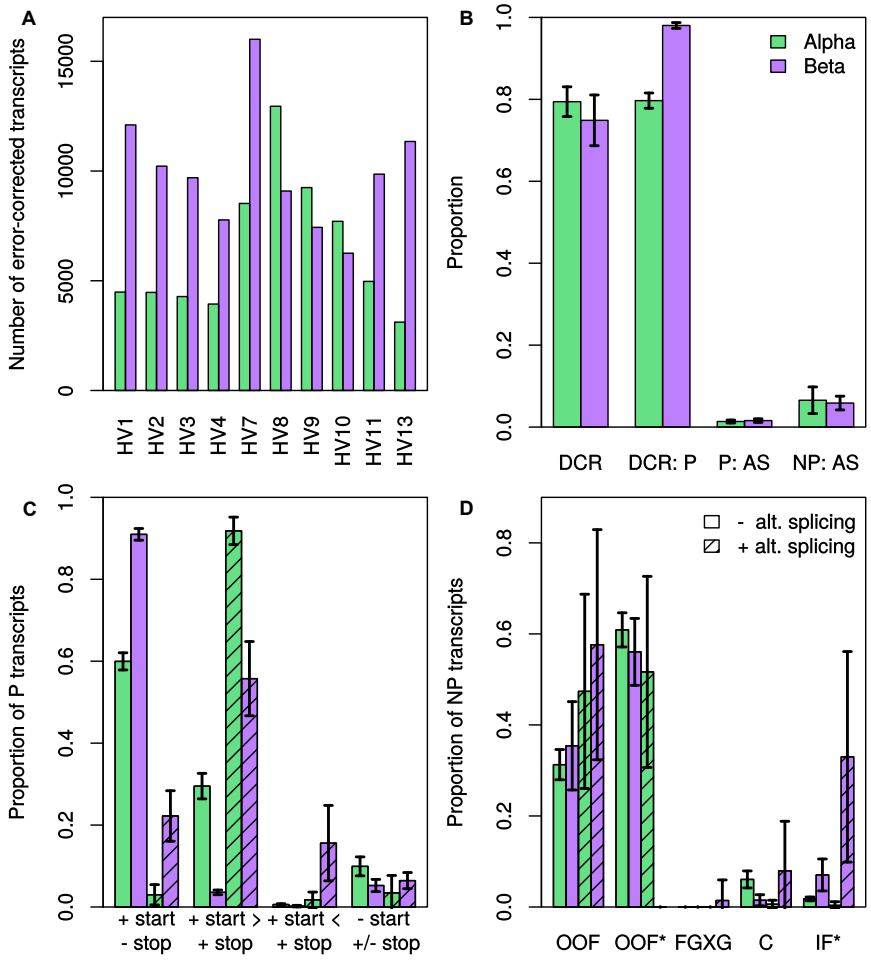


Figure 3.16: Extent of intra-V gene splicing. **A**: Number of merged, paired-end individual transcript consensus sequences that were successfully produced per healthy volunteer (HV). **B**: Proportions of transcripts that made it through each step of the analysis. DCR = those transcripts in which `Decombinator` successfully assigned a rearranged TCR, using the original tag set. DCR:P = of those assigned a rearrangement, how many appear to contain a valid productive rearrangement based solely on the `Decombinator` classifier, performed as in Figure 3.1. P/NP: AS = Of the reads determined to be productive (P) or non-productive (NP) based on their DCR, those that were shown to contain one of the alternative splicing (AS) events listed in Table A.5. **C**: Assessment of what proportion of reads predicted to be productive based on their `Decombinator` classifiers stand to theoretically translate based on the presence of start and stop codons in their complete transcript sequence. Categories are: those transcripts that contain only start codons (+ start - stop); those that contain both with the last start codon being downstream of the last stop codon (and thus might be translated, + start > + stop); those with both but stop codons after the last start codon (+ start < + stop), and those that contain no start codons at all (-start +/- stop). **D**: The proportions of different criteria found among the NP transcripts preventing their productivity based on `Decombinator` classifiers. OOF = out-of-frame but lacking a stop codon (at least upstream of where the constant region would start); OOF* = out-of-frame, also containing a stop codon; IF* = in-frame but containing a stop codon; FGXG = lacking the conserved motif at the end of the CDR3; C = lacking the conserved cysteine residue at the start of the CDR3. α chain results are shown in green, β chain in purple, with any transcripts found to contain splice events (as listed in Table A.5) hatched. All error bars indicate standard deviation.

I then quantified how frequently these splice events occurred in each of the ten healthy donors sequenced, after equivalent processing. Figure 3.16A shows how many overlapped and error-corrected transcripts were obtained using the high stringency pipeline. Of those transcripts produced, the majority are successfully assigned rearrangements and seem to encode productive polypeptides (Figure 3.16B). By looking for the splice sites identified above in either the productive (P) or non-productive (NP) we can see that splicing is proportionally enriched among the NP. In order to assess what effect AS might be having on seemingly-productively rearranged receptors, such transcripts were searched for the presence of start and stop codons (Figure 3.16C). In order to feasibly be productive, a transcript must contain a start codon that is followed by no stop codons (as these transcripts were reverse transcribed upstream of their proper stop codon). We can see that AS-transcripts show a large reduction in the proportion of transcripts that contain a start codon without a stop codon present. Instead they possess a far larger proportion of transcripts that contain stop codons but might still yet be productive as the last stop codon precedes the 3'-most start codon. Interestingly there is an increased tendency among α chain transcripts to fall in this category relative to β chains, even in unspliced transcripts. There does not appear to be any considerable difference between unspliced and AS transcripts in terms of the profiles of transcripts that would not be expected to translate, i.e. those that lack a start codon or have one followed by a stop codon. Note that this analysis is limited in two respects. One, we can only obtain full-length sequences for those transcript that are short enough to be overlapped, and two, these sequences may not actually correspond to the actual 5' end of mRNA sequences, rather just where the reverse transcriptase molecule dissociated. I also looked to see whether there was a difference in the reasons why different unspliced and AS transcripts were classified as NP (Figure 3.16D). As there are so few sequences that are NP, of which only a fraction are spliced (with averages of 60 and 8 NP/AS transcripts per donor) this analysis is limited. However there does appear to be at the very least an increase in the frequency of in-frame sequences that contain stop codons (i.e. as a result of non-templated nucleotide deletion or addition) in AS-transcripts.

3.3.1 Section discussion

There are a number of insights that can be drawn from this analysis. The first is that alternative splicing at the TCR loci appears to consist of a far more diverse series of splice events than has previously been recogised. In addition to describing a variety of novel cryptic splice donor sites, I believe this to be the first report of human TCR pre-mRNAs undergoing alternative splicing to cryptic acceptor sites within the

sequence of the V gene itself. There is one report that I can find of an equivalent phenomenon, in which a non-productively rearranged α chain in a clonal murine hybridoma was shown to have undergone inter-V gene splicing (205). This analysis finds over 10,000 transcripts that undergo this intra-V gene AS, allowing one to quantitatively and qualitatively probe the properties of such sequences. These splice events should theoretically be able to dramatically regulate a transcript's expression. That they are enriched among TCRs that should not be able to produce a valid polypeptide based on their rearrangement ('recombinatorially non-productive') suggests that it might be employed as a mechanism in these cases to further regulate message that will not be contributing towards surface receptors. The different profiles in criteria for non-productivity between spliced and unspliced transcripts might reflect AS preferentially being employed against particular subsets of transcripts. For instance it known that AS-mediated incorporation of stop codons can lead to degradation of message by nonsense-mediated decay (206). However the possibility remains that these transcripts just represent some low-level errors of the transcription and splicing machinery, and the tremendous depth of high-throughput sequencing has brought them above the level of detection; functional experiments would be required to settle the matter.

Perhaps more interesting is the small fraction of sequences which we would determine to be productive, based on their recombined hypervariable region, but which contain upstream intra-V splicing events. It is probable that this would either prevent translation entirely, or at least alter the protein structure so much that the product would no longer be able to contribute to a valid surface-expressed receptor ('transcriptionally non-productive'). This raises the possibility that AS might be able to function to suppress the expression of valid TCR rearrangements, perhaps to prevent the expression of autoreactive receptors.

These findings could also impact on the design of repertoire sequencing experiments. All protocols of which I am aware (including the one I developed and employed in this thesis) involve targeted sequencing around the recombination junction to extract clonotypic information. Many even use V-gene specific primers to amplify just this region. While it is not important for this work, it is feasible other groups may be trying to extract specific rearrangements to use functionally downstream; researchers should remain aware of the fact that just because the cDNA describing the CDR3 appears productive, alternative splicing events upstream could be preventing that rearrangement from being expressed as working protein. Sequencing technologies permitting longer reads than are currently available would be required to ensure transcripts are functional based on their entire lengths. Similarly, construction of TCR sequences for

gene-therapy purposes might benefit from modulation of cryptic splice sites to prevent unwanted alternative splicing of transgenes.

3.4 Chapter discussion

High-throughput sequencing of TCR repertoires effectively consists of the application of novel disruptive technologies to a relatively new field. It is clearly a very powerful tool for research, yet as a field we are still in the early stages of exploring what we might accomplish using these methods. In this chapter, I have displayed an overview of some of the kinds of experiments and analyses one can employ in repertoire sequencing efforts, indicating the range of research questions that we might begin to address. Working from small samples of peripheral blood from healthy individuals, I have shown it is possible to extract a great deal of information about various aspects of TCR biology. At the most simple level, I have described the basic clonotypic and germline gene distributions in my data. These findings recapitulate those of the earliest TCR repertoire sequencing efforts, particularly with respect to the striking nature of the relative abundances of the lowest and highest frequency sequences and the differential usage of different V and J genes (75, 76). This provides us an impression of the 'shape' of the repertoire, which allows us the chance to use this as a reference point when looking for perturbations to the repertoire, such as in disease. An important related aspect of this appreciation of repertoire configuration is how to score and interpret diversity.

I have also looked in my data for the presence of shared or public sequences, as well as those belonging to the so-called 'invariant' TCR T-cell subpopulations. Even without focusing on antigen-specific cell populations I can readily show the presence of a large number of sequences that are shared between multiple donors. Some of these sequences are even found in all donors: while these mostly consist of sequences known to be found in MAIT cells, at least one other sequence was found to be equally public, with several others not far behind. As more donors and cell compartments are sequenced, I anticipate that we will find a far larger number of shared sequences, blurring the lines between what we currently consider to be public and private, a prediction supported by the observation of yet more low-frequency but highly shared sequences even when only priming off a small number of V genes. I also expect that in line with current findings (138), cells matching the phenotypic profiles of the non-pMHC restricted TCR populations but with increasing deviation from the exact canonical sequences will be found, which might eventually necessitate the need to reassess usage of the word 'invariant'. I also demonstrated how by sorting cells into subpopulations prior to sequencing one can not only gain information about those sorted samples,

but apply sequence-matching analyses to infer properties about unsorted populations. Finally I have shown how we might use repertoire sequencing data to learn about aspects of TCR biology unrelated to the clonal composition of the repertoire in revealing the presence of multiple novel splice events that occur within the TCR V gene sequences, which might represent an under-appreciated level of TCR regulation.

Several themes that emerge from this chapter are worthy of particular note. The first is how these experiments speak to the richness of the data produced by repertoire sequencing. However – due to the age of the technique – there is still little consensus about what information is most important and how best to present it, which presents a challenge for both analysis and interpretation. Indeed, as is the case for all HTS, with decreasing costs and greater access to sequencers we could be heading towards a scenario in which data analysis instead of generation becomes the limiting step (207).

Another important aspect of this data is the frequency with which we find uncommon phenomena (such as alternative-splicing or use of unexpected V-genes) that might not be detected or be more readily overlooked from studies using lower-throughput or -resolution repertoire methods. It is probable that as the field investigates increasing numbers of large, multi-dimensional datasets more thoroughly we will discover increasing numbers of such novel events or contradictions to existing rules, and we should be prepared to obtain and assimilate such information. For example, the IMGT databases contain a great deal of well-annotated and easily accessible information regarding variable lymphocyte receptors from a number of species, and their classification and numbering systems allow sensible interpretation of data between studies (19, 101, 102). This has made it in invaluable resource for repertoire studies, demonstrable by the fact that almost every high-throughput repertoire paper either cites the resource or makes use of its nomenclature. However, as we have seen in this chapter, it certainly contains incorrect classification of some genes, therefore – at least in some small way – experiments which are based exclusively upon their annotations are flawed from the outset. One could imagine a scenario in which looking for a particular specific TCR clone could be important, such as tracking minimum residual disease in a leukaemia patient: if this clone happens to be using a V gene that was not included in the multiplex primer panel or rearrangement analysis software then it unlikely to feature or be detected in the data. This chapter serves as a reminder that we must remain open-minded in our approach, and that even the most well-curated databases are still subject to errors. I imagine that IMGT must soon adopt a system whereby the findings of such repertoire sequencing experiments can be incorporated. Another area where this might be useful

is the identification of new alleles of different germline genes, which has already been reported in one RepSeq analysis (183).

There are a number of caveats and limitations both to repertoire sequencing generally and the analyses presented here specifically. One of the major issues is that of sampling: in any one experiment we can only observe a fraction of total cells taken typically from one tissue, usually blood[10]. These tissues are themselves subject to sampling processes as cells enter and leave, divide and die. This will affect the nature of the questions that can be answered using repertoire sequencing. One is only able to reliably study cells which appear frequently appear above the rate of detection of the experiment: one would not be able to study say GEM cells at the depth of sampling used in this study without great cost. Given the conservation between samples, it is however probable that the general TCR repertoire features and population structures obtained from a small but random sample are indeed likely to be representative of the whole tissue. In a similar vein, Eugster *et al* were able to detect antigen-specific T-cell responses from sequential small blood samples, however the exact clones used were often different between samples (208). It must also be stressed that in this thesis I was only making use of a fraction of RNA derived from just two to three millilitres of blood, yet was still able to generate this wealth of data: if greater depth is required one could easily take more blood, process more RNA or sequence more of the library.

Sampling leads on to another important consideration for repertoire sequencing, that of diversity. Diversity is a key parameter of the T-cell response – arguably even the 'purpose' of TCRs – as it corresponds on some level to the number and nature of antigens that can be responded to (209). However we cannot compare sample diversity in its purest sense (i.e. number of different sequences, or richness) due to the sampling effect. Instead, in this thesis I have adopted two diversity indices to measure different aspects of TCR repertoire diversity, which can then be compared between samples without knowing the true size of the repertoire. Being able to measure repertoire diversity in healthy individuals provides a baseline against which to compare repertoires from patients who we might expect to display altered T-cell populations.

The final major consideration is that of expression. With the exception of Harlan Robins' group, and anyone who makes use of data generated either in collaboration with them or through the company he helped found, the majority of groups currently using deep-sequencing to study TCR repertoires use RNA as their starting material. This has a number of technical benefits, such as the downregulation of non-productive sequences

[10]This problem affects mouse studies to a lesser extent, as more T-cells can be sampled, i.e. a greater proportion of a given mouse's total T-cells can be sampled by splenectomy than can be obtained from a human by phlebotomy.

and the ability to perform unbiased pan-TCR amplification (discussed in detail in Section 5.1.2 in Chapter 5), yet also brings some caveats, the foremost being that mRNA expression levels are mutable. TCR expression has a further complication, as not only can mRNA levels change through modulation of transcription or turnover as with any transcript, but such processes could theoretically differentially affect TCRs as each V gene brings its own promoter and regulatory regions to a rearrangement (210, 211). Therefore we must not confuse sequence frequencies with cell frequencies. However, the two are related, if not linearly. While not extensively studied, TCR expression has been shown in several *in vitro* model systems to increase upon T-cell stimulation prior to cell division (212, 213, 214, 215), even as TCR molecules are internalised from the cell surface (216, 217). This implies that there are two forces driving the frequency of TCR mRNA molecules in a blood sample (and thus the probability of their detection by sequencing): the number of cells of that clonotype, and their activation status, which could be advantageous in certain experiments. Using gDNA should allow description of relative abundance of cells, assuming an unbiased amplification protocol. However, if one requires absolute quantitation of cells from repertoire sequencing data, it would require using a barcoded amplification approach analogous to the one described in this thesis, but using gDNA as a substrate, a strategy that to my knowledge has not yet been employed.

In spite of these limitations, in this chapter I have demonstrated how one can apply TCR repertoire sequencing to even small samples of peripheral blood and extract a great deal of diverse information relating to T-cell biology.

4

Perturbations to the T-cell receptor repertoire during HIV infection

4.1 Introduction

4.1.1 HIV and AIDS

June 1981 saw the beginnings of the Western outbreak of acquired immunodeficiency syndrome (AIDS) in the United States of America, which made it to the United Kingdom by the end of that year (218, 219). The syndrome consists of increased likelihood of pathology from opportunistic infections, often from typically non-pathogenic microbes, as well as certain uncommon malignancies and autoimmune conditions, in patients with no history of immunodeficiency (220). Pneumonia from *Pneumocystis jiroveci* (then called *Pneumocystis carinii*) and Kaposi's sarcoma (later found to be caused by the eponymous herpesvirus (221, 222)) were particularly common forms of AIDS associated disease (223, 224). Moreover the syndrome was initially enriched among certain demographics, particularly men that have sex with other men, intravenous drug users, haemophiliacs and Haitians (220). In the following years, several groups – notably those headed by Luc Montagnier and Robert Gallo – isolated cytopathic T-cell tropic retroviruses from patients with AIDS (225, 226, 227). These were later discovered to all be the same virus and the causative agent of AIDS, which was eventually renamed human immunodeficiency virus (HIV) (228, 229, 230).

HIV has been the subject of a great deal of research, thus much is now known about it and its life cycle. It belongs to the lentivirus family of retroviruses (231). Primary infection can often be asymptomatic or produce mononucleosis symptoms, and initial peak viraemia will typically be contained as the cellular and humoral responses to the virus develops (232, 233). Patients usually then enter a phase of seeming good health with minimal viraemia, which can last from months to years, although the virus remains highly active in lymphoid tissues (234, 235). As the disease progresses, patients suffer from a gradual loss of $CD4^+$ T-cells, which prompted research that identified CD4 itself

as HIV's primary receptor for cell entry (236, 237). This CD4 cell death was previously thought to be the result of apoptosis of uninfected bystander T-cells, likely as a result of activation-induced cell death or by action of some viral component (238). However recent evidence suggests that the bulk of CD4 cell loss in lymphoid tissues may be due to unsuccessfully infected cells undergoing pyroptosis in response to aborted reverse transcription products, creating an inflammatory focus that in turn could recruit more T-cells to the site of infection (239, 240). This loss of CD4 cells and the chronic systemic inflammation seen during HIV are thus considered the driving forces of the immunodeficiency that causes AIDS, making CD4 count a useful clinical marker of disease progression (241).

HIV also perturbs the $CD8^+$ compartment, both through driving antigen-specific clonal expansion as well as the general bystander activation and exhaustion observed during infection (242, 243). $CD8^+$ cytotoxic T-cells (CTL) are important mediators of anti-HIV immunity. Strong CD8 responses correlate with control of initial viraemia, while subsequent loss of or escape from viral-specific CTLs associates with increasing viral titres and clinical progression (244, 245, 246, 247). Qualitative differences in the CD8 T-cells of individuals who achieve good virological control correlate with a delay in progression to AIDS (248, 249, 250). Furthermore certain HLA class haplotypes associate with either greater or lesser ability to control disease progression (251).

3'-azido-3'-deoxythymidine (AZT), a thymidine analogue that inhibits HIV reverse transcriptase (RT) by chain termination, was shown to possess antiviral activity both in vitro and in vivo, providing the first antiretroviral therapy (ART) for HIV (252, 253). Inspection of viral sequences from treated patients with decreasing sensitivity to treatment however revealed that the RT gene could acquire resistance mutations (254). The development of additional drugs, and particularly new classes of drugs that target different stages of the HIV life cycle, allowed the development of combination ART (cART, also called highly active antiretroviral therapy, HAART) drug regimens (255). HAART reduces viral loads, improves CD4 cell levels and reduces the risk of opportunistic infections, while reducing the risk of resistance as the virus would need to mutate multiple residues simultaneously (256). Accordingly cART has become the standard of care for patients with HIV (257).

Current therapeutic options are however unable to completely remove virus from an individual as HIV integrates its genome into that of the host, allowing it to persist in the face of treatment, necessitating constant medicating (233). Moreover patients on HAART still remain at greater risk of certain non-AIDS related diseases, which are typically those often associated with aging, displaying increased rates of cardiovascular,

kidney, liver and bone disease, cancer and various neurological syndromes (reviewed in (241)). Despite increasing life expectancy of HIV^+ individuals substantially, these treated patients still have lower life expectancies than HIV^- people (258). This suggests that HIV is able to cause some physiological damage (either before or during therapy) that treatment is unable to fully rectify, or that the drugs themselves have off-target effects. At the end of 2013 there were 35 million people estimated to be infected with HIV globally, of whom approximately 12.9 million were receiving ART (259), therefore the burden both of the disease and any therapy-associated pathology remain important subjects of research.

4.1.2 HIV TCR repertoire analysis

There have been a number of studies using various techniques investigating TCRs in HIV-infected individuals. Flow cytometry, DNA hybridisation and quantitative PCR have shown decreased expression of certain V genes, with certain families being depleted to variable extents in different individuals (260, 261, 262, 263). Furthermore spectratyping analysis revealed a skewing of CDR3 length distributions in certain V families, indicative of clonal expansions (264, 265, 266). A decrease in TCR diversity associated with HIV infection has also been measured using a number of experimental approaches (267, 268, 269), although such observations have typically been limited to either antigen- or T-cell subset-specific populations. The overall impact on the clonal structure of the T-cell population of the virus depleting $CD4^+$ T-cells while simultaneously being targeted by (and eventually escaping from) CD8 cells remains incompletely understood.

As discussed in Chapter 3, the introduction of high-throughput DNA sequencing (HTS) allows analysis of the TCR repertoire at much greater depth than was previously possible. The wide ranging T-cell effects and large global disease burden make HIV-infection an eminently suitable testing ground for such analysis. Therefore in this chapter I describe the sequencing of TCR repertoires from a cohort of ART-naïve HIV patients. RNA was extracted from the peripheral blood of 16 adult HIV-positive patients due to start ART at two time-points, both immediately prior to ('v1') and three months after commencing therapy ('v2'). Patients were excluded if they were currently showing symptoms for AIDS related illness, or if they had been recently vaccinated. Each RNA sample was processed using the error-correcting barcoded-cDNA protocol that was used in Chapter 3 (detailed later in Section 5.2.6). The patient data was compared to the healthy repertoires of the ten healthy donors explored in Chapter 3, four of whom were also sampled twice after a three month interval.

4.2 HIV infection associates with a dramatic loss of TCR diversity

Given the variety of reports of reduced diversity in the CDR3 repertoires of HIV patients I used the indexes employed in Chapter 3 to assess the diversity in the pre-treatment patient samples. Figure 4.1A and B display a significant difference between the Gini index and Shannon entropy of healthy and HIV-positive repertoires respectively. The reduced diversity reflected by the difference in these indexes was manifest in both α and β chains, although the complexity of α chain repertoires was less than those of β (Figure 4.1B), consistent with the lack of a diversity region and consequently less junctional diversity (78). The high inequality and low complexity of repertoires from HIV patients could be considered equivalent to a relatively oligoclonal population in comparison to healthy individuals. In order to assess whether this was the case I measured the average proportional frequency of the 50 largest CDR3s in each donor: in patients these common sequences accounted for approximately ten times the proportion of their respective donors than in healthy donors (Figure 4.1C). I performed a number of linear regressions to see whether any clinical scores correlated with the repertoire diversity metrics. Surprisingly, the most significant correlation was between the CD8 cell count and the Gini index (Figure 4.1D), suggesting that particularly unevenly distributed repertoires might be produced in part by large clonal expansions in the CD8 pool. Despite the low CD4 T-cell counts in these patients, CD4 values did not correlate with Gini index (see Figure B.1A in Appendix B). Multiple regressions of both indexes against various clinical scores (performed by my colleague Katharine Best) confirmed that CD8 cell count and Gini index were the two parameters that correlated best (Figure B.1B).

These findings seemed to run in opposition to the simplest potential explanation, that the decrease of diversity was solely attributable to the loss of CD4 T-cells as a result of viral infection. I wished to further explore the hypothesis of CD8 expansions being a pressure driving down diversity, however we did not have access to fresh or sorted blood samples from these patients. In order to overcome this limitation, an alternative assay was devised that requires neither access to patient samples nor sorting of HIV-positive cells. TCR sequences were randomly sampled from the $CD4^+$ and $CD8^+$ repertoires sorted from healthy donors (discussed earlier in Section 3.2.1) *in silico*, and combined in different amounts to simulate mixed repertoires from donors with different CD4:CD8 ratios. The diversity indexes of these simulated repertoires are shown in Figure 4.2. As the proportion of CD8 cells increases, so does the repertoire

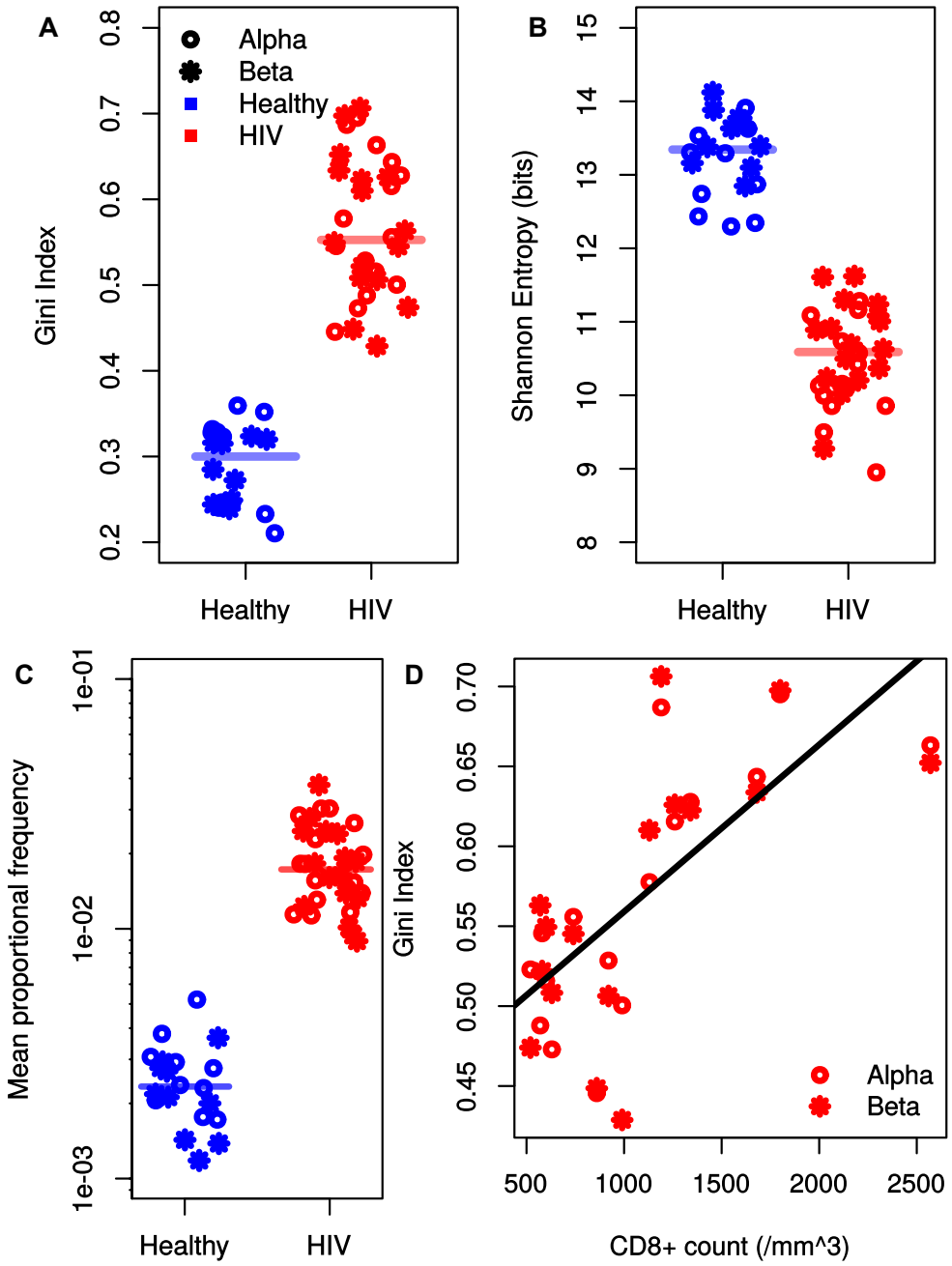


Figure 4.1: TCR-repertoires of HIV-infected patients are typified by lower clonal diversity and higher inequality than healthy donors. **A**: Evenness of healthy and HIV-infected whole, unsampled CDR3 repertoires as measured by Gini index. HIV^+ Gini indexes were significantly greater than healthy donors' ($p = 7.937e\text{-}15$, the lower bound for this test in R, the statistical software used). **B**: Information content (as a diversity index) measured by Shannon entropy of CDR3 repertoires. HIV^+ Shannon entropies were significantly lower than healthy donors' ($p = 7.937e\text{-}15$). Between both healthy donors and HIV patients, alpha repertoires have significantly lower Shannon entropies ($p = 2.582e\text{-}05$). **C**: Mean proportional frequencies of the 50 most common CDR3s for healthy (blue) and HIV-infected (red) donors: HIV^+ mean frequencies are significantly higher than healthy donors ($p = 7.937e\text{-}15$). P-values in **A**, **B** and **C** are all Wilcoxon rank sum tests. **D**: Linear regression of CD8 T-cell count during the v1 bleed versus Gini index ($R^2 = 0.51$, $p = 4.30e\text{-}6$).

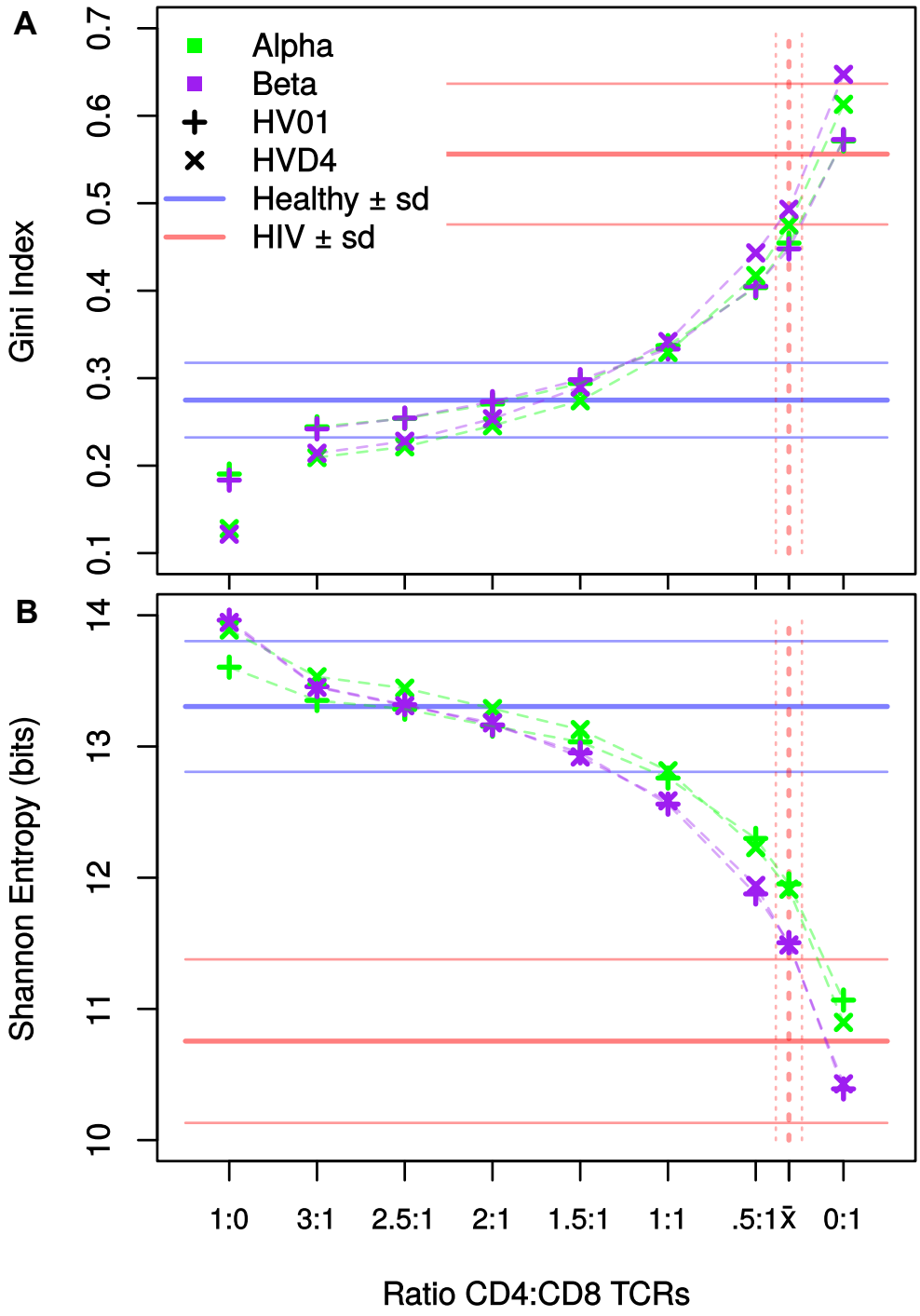


Figure 4.2: Simulating TCR repertoires with different CD4:CD8 ratios. $CD4^+$ and $CD8^+$ T-cells were sorted from PBMC extracted from two healthy donors (HV01 and HVD4) and sequenced. Fixed numbers of TCR classifiers were randomly sampled from each population and mixed in different ratios to simulate populations with different ratios, and then Gini index (**A**) and Shannon entropy (**B**) were calculated. Horizontal lines indicate the mean index values of the healthy (blue) and HIV-infected (red) repertoires in our dataset ± standard deviation. The mean ratio observed in our patient clinical data (0.31:1) is denoted by $\bar{x}$ (vertical red line, ± standard deviation)

inequality (Figure 4.2A), while the complexity falls (Figure 4.2B). The diversity indices of the simulated repertoires intersect the mean values observed in the sequenced data of the healthy repertoires at ratios of 2.5:1 to 2:1, within typical expected ranges for a healthy population (270). However, when we reconstitute simulated repertoires with a ratio of 0.31 – the average CD4:CD8 ratio seen in the HIV patient's clinical data – we see that the reconstituted repertoires are still yet more diverse, showing both lower Gini indexes and higher Shannon entropies. These findings demonstrate that the low diversity observed in the HIV^+ samples does not appear to solely be due to a decreasing CD4:CD8 ratio; instead, CD8 expansions as well as CD4 cell loss must be driving the inequality and low complexity in HIV^+ patients.

4.3 TCR repertoire population structures are perturbed in HIV patients

In spite of the stochastic nature of V(D)J recombination, there are a number of repertoire features that are reported shared between individuals, potentially reflecting constraints of the recombination machinery (77, 179, 271). These properties are characteristic of, and potentially contributory towards, what we consider to be 'healthy' TCR repertoires. Therefore I wished to establish whether these features were altered during HIV infection.

One such property is the sharing of TCR clonotypes across multiple individuals (discussed above in Section 3.1.4), possibly as a result of convergent or biased recombination processes (130, 131, 181, 184, 271). Given the dramatic loss of diversity observed in HIV patients, I hypothesised that we should expect to see less sharing. I calculated the publicness of those CDR3s that were only found either in healthy or HIV-infected patients; those that were limited to HIV patients were far less public, occurring less often and in fewer donors for both α (Figure 4.3A) and β chains (Figure B.2A). Calculation of the Jaccard index – a normalised measure of sharing – between pairs of different donors revealed that any two HIV patients share fewer CDR3s than do any two uninfected controls (Figure 4.3B). I also observed that the largest CDR3s in the healthy donors are also encoded by a greater number of nucleotide rearrangements than those in HIV-infected donors (Figures 4.3C and B.2B). This loss in translational convergence occurs even though the most common CDR3s in HIV patients account for a far greater proportion of their respective repertoires, reflecting the extreme expansions of individual rearrangements.

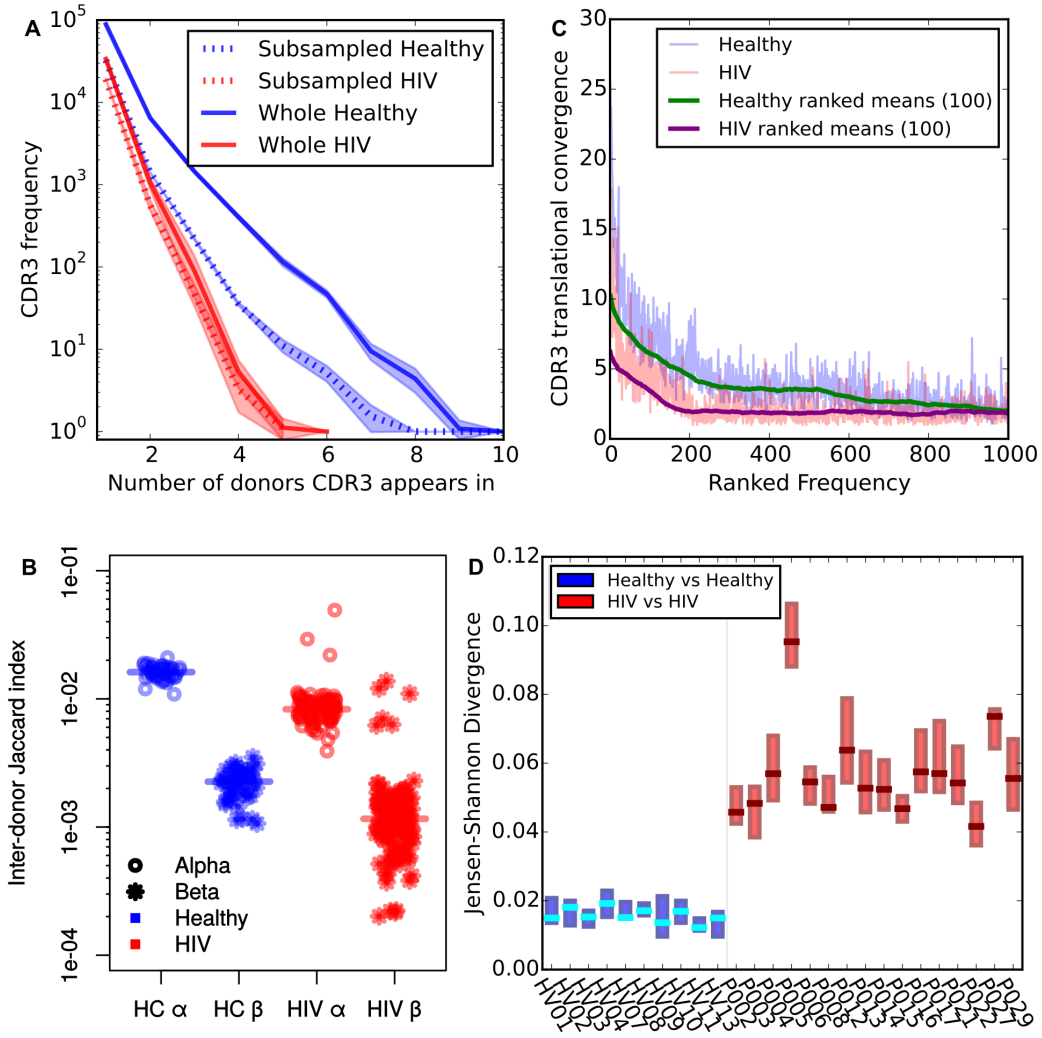


Figure 4.3: HIV$^+$ TCR repertories are less shared, public, and translationally convergent than healthy controls, and are highly divergent in their V and J gene use. A: α chain CDR3s are less public in HIV-infected than healthy repertoires. The number of CDR3s shared by different numbers of individuals is shown, considering only those CDR3s that were found exclusively in either healthy (blue) or HIV$^+$ (red) donor repertoires. This analysis was applied to whole samples (solid lines) and a random subsample of 5000 CDR3s (dotted lines). To account for the greater number of patients than controls, ten HIV$^+$ samples were randomly selected and publicness measured: this was repeated 100 times, shaded areas represent standard deviation. **B**: Inter-donor Jaccard index measuring sharing between pairwise combinations of healthy control (HC) or HIV-infected CDR3 repertoires, for 5000 randomly chosen CDR3s. Horizontal lines indicate medians. Inter-donor Jaccard scores were significantly higher in healthy donors than HIV$^+$ patients for both chains (p = 2.2e-16 and 1.147e-13 respectively, unpaired Wilcoxon rank sum test). **C**: α CDR3s display a lower mean translational convergence in HIV$^+$ donors than healthy, i.e. were encoded by fewer nucleotide sequences, determined by counting the number of unique `Decombinator` identifiers that produce the same CDR3. Blue and red lines indicate the mean translational convergence for each group (healthy/HIV respectively) for each CDR3 rank; green and purple lines show a sliding window averaging across 100 ranks. **D**: Barplot showing the difference in TRAV usage, by Jensen-Shannon Divergence (JSD). Donor-specific proportional gene usage matrices, before each pair of healthy or HIV$^+$ donors was compared. Only interquartile ranges and medians are shown.

I also assessed the differences between the V and J gene expression profiles of the two groups, by comparing frequency distributions of pairs of donors using the Jensen-Shannon measure of divergence (JSD). Values for α V genes are shown in Figure 4.3C, while the remaining genes' divergence values are in Appendix B, Figure B.3. Healthy donors' repertoires diverge from one another by a relatively small, consistent amount, while any two HIV$^+$ patients differ by greater, more variable amounts. Comparing healthy donor to HIV-patient distributions reveals an intermediate level of divergence (data not shown for clarity). These data suggest a scenario in which there exists a healthy core of gene usage, where any two individuals can differ from one another by a small amount. HIV infection then drives an individual's usage profiles away from this healthy core, but it does so in a different direction in different people, producing idiosyncratic profiles that differ somewhat from typical values, but greater from other patients. I also note that there is no difference between infected and uninfected donor profiles for TRBJ (Figure B.3C), perhaps due to the smaller number of genes or the confounding effect of the D region. Examination of the raw gene usage profiles themselves (Figure B.4) reveals that most genes on average are used an equivalent amount, just with greater variance among patients; a small number of genes in each locus, however, show a consistent difference in frequency between healthy controls and HIV-infected individuals (e.g. TRAJ33 and TRBV6-1, Figure B.4B and C).

4.4 Effect of antiretroviral therapy on TCR repertoires of HIV$^+$ individuals

By comparing the treatment-naïve samples with those taken after an average of 14 weeks of therapy I was able to assess the impact of short-term ART upon the TCR repertoires of HIV patients. The four healthy donors who were also sampled after a comparable time period act as controls of typical repertoire variation over time and between samples.

By inspecting the clinical parameters we can infer that the ART had the expected effects: viral loads drop in all patients, typically beneath the limits of detection, while absolute CD4 cell levels and CD4:CD8 ratios increase significantly (Figure B.5). The only major lymphocytic parameter to not significantly change over the course of the treatment is the CD8 T-cell count (Figure B.5B), which could indicate that the CD8 expansions hypothesised above are persistent in the face of treatment. This analysis also revealed an interesting caveat of our patient data: the length of time elapsed before

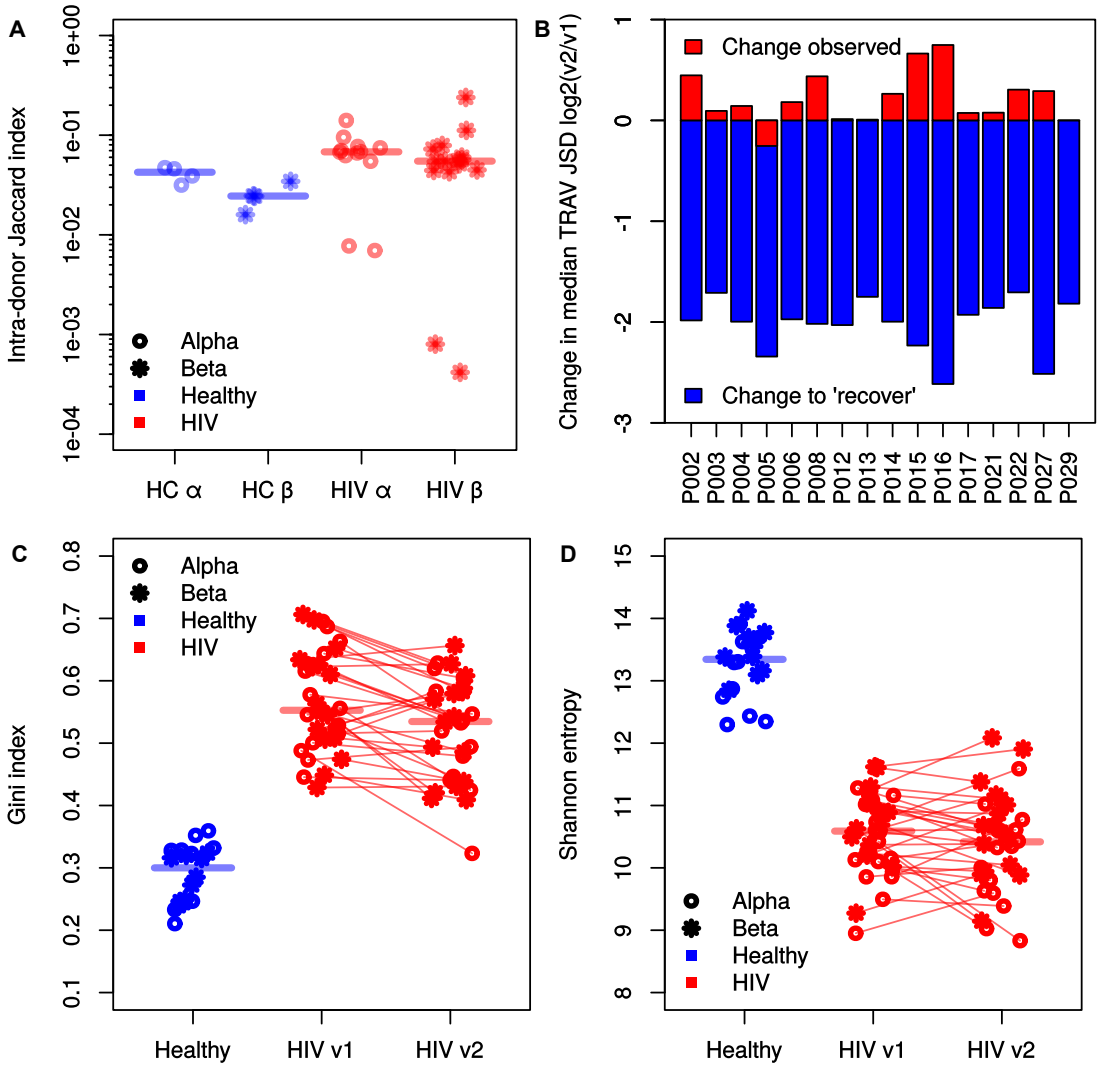


Figure 4.4: ART partially reduces repertoire inequality, but general diversity and idiosyncratic gene usage do not recover. A: Intra-donor Jaccard indices, describing overlap between v1 and v2 bleeds for both healthy control (HC) and HIV$^+$ donors for 5000 CDR3s subsampled from whole repertoires. Horizontal lines indicate medians. Healthy controls have significantly lower α and β chain intra-donor sharing by Jaccard ($p = 0.02912$ and 0.01367 respectively, one-sided unpaired Wilcoxon signed rank test). **B**: Log2 fold change in median TRAV Jensen-Shannon Divergence (JSD) from v1 to v2 (red). Blue bars indicate the log2 change in median JSD that would be required for that patient to reach the mean value of the median JSDs for the ten healthy controls, i.e. return them to the average level of divergence healthy controls display in Figure 4.3D. **C**: Gini indexes of whole CDR3 repertoires of healthy donors, and both v1 and v2 bleeds of HIV$^+$ patient repertoires. Horizontal lines indicate medians. Pre-ART bleeds have significantly higher Gini indexes than post-ART ($p = 0.0001673$, paired one-sided Wilcoxon test). **D**: Shannon entropies of whole CDR3 repertoires of healthy donors, and both v1 and v2 bleeds of HIV$^+$ patients. Horizontal lines indicate medians. There is no significant difference between pre- and post-ART Shannon entropies ($p = 0.1609$, paired one-sided Wilcoxon test). Fold changes are shown in Figure B.8.

each patient began therapy correlated with their CD4:CD8 cell ratio at time of the pre-treatment bleed (Figure B.6). This might be explained by differences in viral control influencing the likelihood of patients seeking rapid treatment.

In contrast to the inter-donor Jaccard indexes shown in Figure 4.3B, calculation of intra-donor CDR3 repertoires revealed that the overlap between two samples from the same donor is greater for the HIV patients than the healthy donors (Figure 4.4A), i.e. more sequenced are retained within an HIV-patient over the same time period. Given the various levels of subsampling in such sequencing efforts – including cells sampled during phlebotomy, mRNAs processed during amplification and amplicons recorded in the sequencer – it is only ever possible to obtain data for a fraction of the entire TCR repertoire. That these HIV patients have far fewer unique CDR3s at both time points but still retain more sequences over therapy suggests that clonal expansions are persistent, even in the face of therapy. However there are two patients that have exceedingly low Jaccard indexes in both chains, indicating that they have undergone dramatic repertoire remodelling events. Post-treatment gene usage divergence values were calculated as before: not only did divergence not return to healthy ranges, in the majority of patients it actually increased (Figures 4.4B and B.7). The majority of patient repertoires (11 out of 16) did observe a decrease in Gini index on ART, indicating a partial recovery in repertoire equality (Figures 4.4C and B.8A). However, the information content as measured by Shannon entropy did not significantly change over the course of treatment (Figures 4.4D and B.8B). Thus three months of ART is sufficient to reduce viral load to undetectable levels and begin to replenish CD4 cell numbers, but does not restore the properties of the TCR repertoire to the ranges observed in healthy donors.

4.5 Tracking sequence-specific CDR3s on ART

It is not yet possible to infer the quaternary structure of TCRs from primary sequence, much less their binding specificities. Moreover most high-throughput experiments, including this one, do not yet routinely obtain paired α and β chain sequence and MHC haplotype information. However there are a number of reports that CDR3s from T-cells reactive against the same antigens often share sequence identity or features, and some TCRs can recognise the same peptide in different HLA backgrounds (272, 273, 274, 275). There are also reports from several groups of public TCR sequences reactive against HIV antigens (276, 277, 278), sometimes even finding the same CDR3 in combination with different V and J genes (249). This lead me to question whether our data contains any CDR3 sequences of known specificity, as a proxy for such specificities

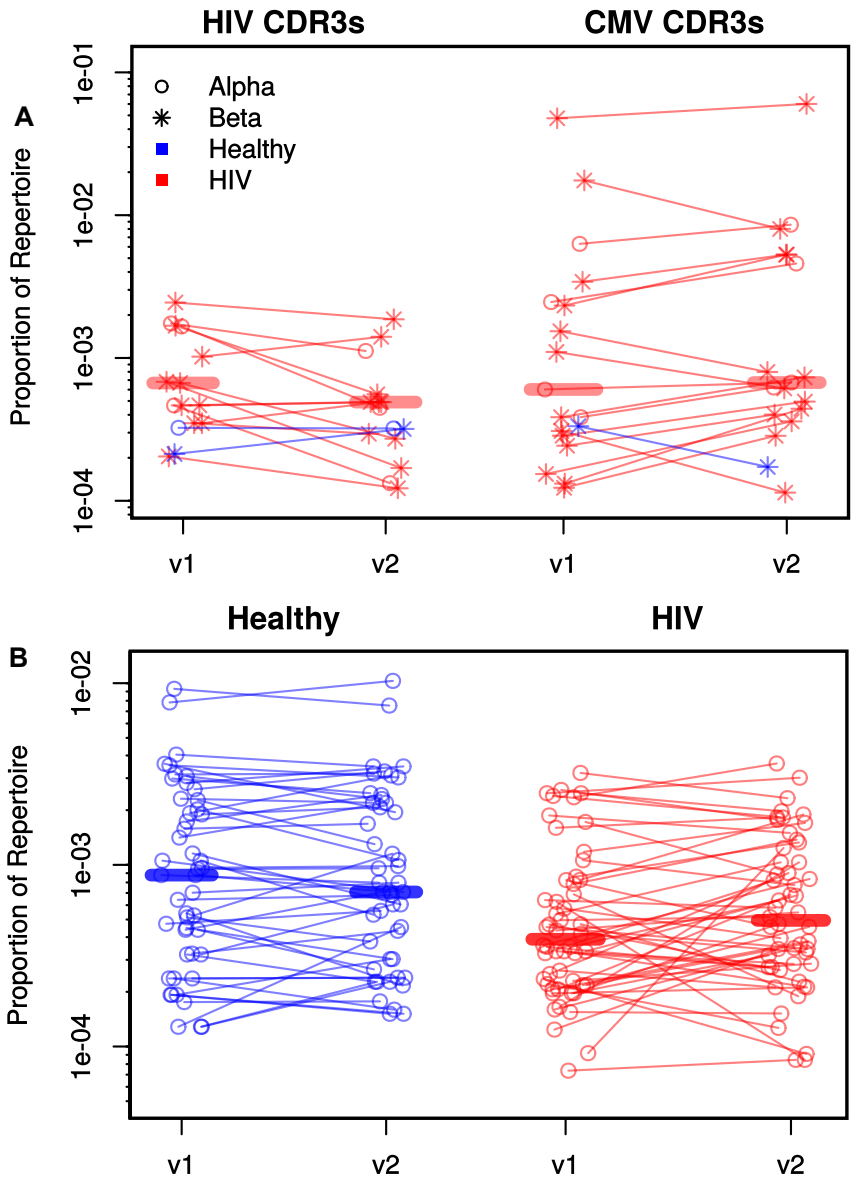


Figure 4.5: Dynamics of viral-associated and invariant MAIT cell CDR3s. A: Proportional frequencies of the published, antigen specific CDR3s that were also found in the whole CDR3 repertoires pre- and post-treatment. CDR3s were only included if they appeared twice or more in both of a patient's bleeds. Horizontal lines indicate medians. HIV-related CDR3s significantly decreased after three month of therapy, while CMV CDR3s increased ($p = 0.01799$ and 0.04919 respectively, paired Wilcoxon signed rank test). HIV-associated CDR3s were only found in six of the 16 HIV$^+$ patients, with the majority (7/13) all found in one donor (P017). The two HIV-associated CDR3s found in healthy volunteers were from the same donor (HV04). CMV-associated CDR3s were found across nine HIV$^+$ donors (with four, the most, being contributed by P027), while the single healthy response was found in HV03. **B**: Proportional frequencies of α chain sequences that match published MAIT cell VJ-CDR3 sequences, for healthy control (left) and HIV donors (right). Horizontal lines indicate medians. Healthy donors' median MAIT CDR3 frequencies are significantly higher than those of HIV patients at both time points ($p = 0.006366$ and 0.03558, unpaired one-sided Wilcoxon signed rank test). There is no significant change in frequency for either healthy or HIV$^+$ results from bleed v1 to v2 ($p = 0.2929$ and 0.1403 respectively, paired one-sided Wilcoxon signed rank test).

potentially existing in the patients. I therefore harvested the amino acid sequences of published CDR3s known to be specific to HIV antigens from the literature (referenced in Appendix Table B.1). These were typically obtained through isolating HIV-specific pMHC-tetramer positive T-cells from infected patients. A comparable set of sequences were also collected for CMV- and EBV-specific CDR3s was also collected, as these are other well-studied, widespread chronic viruses known to drive large oligoclonal T-cell expansions (81, 279, 280).

The relative change and raw frequency values for those published viral-associated CDR3s that were found in our data (present at least twice in both of an individual's bleeds) are shown in Figures 4.5A and B.9. Thirteen HIV-associated CDR3s were present in the HIV patients' sequence data, coming from six of the 16 individuals sampled. These sequences included three α chain CDR3s, one of which was present in two donors while the third differed by only a single amino acid. Seventeen CMV-associated CDR3s were detected from nine patients, while six EBV-associated sequences were found across five patients. Interestingly, the frequency of HIV-associated CDR3s decreased in frequency over the course of therapy, while the majority of CMV- and EBV-associated sequences grew more frequent. These viral-associated sequences were almost completely absent from the healthy controls, with only one or two low-frequency responses identified per virus (although this will partially reflect the greater number of HIV patients sequenced). If we assume that these viral-associated sequences may indeed represent T-cells playing a role in specific antiviral responses, this data would suggest that during ART anti-HIV clones are able to contract as viral load falls, which would leave 'space' in the repertoire into which clones of other specificities could expand. The nature of this 'space' would likely be decreased competition for some shared or limiting resource. The extent of the herpesvirus response is particularly noteworthy. The rate of return for this analysis was much greater than for HIV – despite finding more sequences, I only searched for 265 published CMV CDR3s compared with 871 specific for HIV – and the resultant CDR3s account for greater proportions of their repertoires (Figure B.9A). This is consistent with observations that CMV- and EBV-specific TCRs often belong to very large and highly shared clones (81, 275).

I also looked for the presence of the invariant α chain TCRs of the non-pMHC restricted T-cell subsets discussed in Section 3.1.5. MAIT cell VJ-CDR3 sequences had been highly prevalent in the healthy controls (see Figure 3.9), and are still often detectable in the HIV-patient repertoires (Figure 4.5B). However there was a significant reduction in the proportional frequency of MAIT cell TCR sequences detected in untreated HIV patients, that did not significantly change after three months of treatment.

This depletion of MAIT cell sequences is so profound that we can see it at the gene level: the V and J genes for the canonical MAIT alpha chain (TRAV1-2 and TRAJ33) as well as the V genes known to contribute to their beta chains (TRBV6-1, TRBV6-4 and TRBV20-1 (281)) were among the minority of genes that were used significantly less in the HIV patient data (Figure B.4A-C). Similarly, despite there being no published iNKT VJ-CDR3 sequences observed in this analysis, the genes encoding their invariant TCRs (TRAV10 and TRAJ18 (282)) were also less common in the repertoires from HIV-infected samples. NKT cells are known to be reduced in HIV patients (283, 284, 285); it is possible that a far wider range of iNKT CDR3s exists than is currently known and some are being depleted in these patients, which would explain a decrease in the associated V genes. No GEM sequences were found using these critera, consistent with their low detection even in healthy individuals.

4.6 Discussion

Through application of the unbiased, error-correcting TCR amplification and sequencing protocol developed in the course of my PhD, I have explored the effect of HIV infection and subsequent antiretroviral treatment upon TCR repertoires. HIV patients' repertoires differ from those of healthy individuals in every regard examined. They are typified by marked clonal inequality, wherein a relatively small number of highly expanded, private rearrangements occupy a large proportion of the repertoire. Furthermore this inequality correlated with the patients' $CD8^+$ cell counts. However neither Gini index values nor CD8 counts correlated with CD4 counts, suggesting that the two compartments are being altered by different processes. Cell counts of both populations failed to correlate with the length of time individuals remained infected but untreated, which varied widely in this cohort. These correlations, combined with the fact that *in silico* simulation of repertoires of decreasing CD4:CD8 ratios failed to reproduce the low inequality observed in the patients, speak to a model in which $CD8^+$ clonal expansions help drive the loss of repertoire evenness, independent of CD4 cell death. These two components of T-cell perturbation drive reductions in repertoire diversity, translational convergence and inter-donor TCR sharing, endowing patients with highly idiosyncratic gene usage profiles that lie far outside the ranges observed in healthy uninfected individuals.

A current weakness of TCR sequencing approaches is the inability to assign specificity, as predicting protein structures and behaviours of large multi-protein complexes remains extremely challenging in general, and particularly difficult for TCR-pMHC interactions (reviewed in (286)). Even were there mechanisms to infer TCR antigens

from sequence, this current study would not be geared towards this end as we obtained neither paired $\alpha\beta$ chain information nor patient HLA haplotypes. I nevertheless identified a small number of CDR3s in the HIV patients' repertoires that have previously been reported as belonging to viral-specific T-cell clones. The frequency of these CDR3s undergo changes that we might expect if they were reactive against their associated viruses: HIV-associated CDR3s got rarer while CMV- and EBV-associated CDR3s became more prevalent. This matches a variety of reports that found a decrease in HIV-reactive clone frequency during ART while CMV-reactive clones increased (268, 287, 288), which is in turn reinforced by the frequent loss of CMV viraemia during treatment for HIV (289, 290). These findings lend weight to the possibility that the viral-associated CDR3s in this data might actually be derived from TCRs which share those specificities, and the changes observed in Figure 4.5A represent the contraction of HIV-reactive clones as ART depletes the source of antigen, allowing re-emergence of protective responses and pathogen control. If so, it is possible these shared viral-associated CDR3s could represent members of a subset of receptors in which antigenic specificity is concentrated in one of the two chains (291, 292).

Questions as to the nature of the expanded sequences remain. These could represent $CD8^+$ clones that are yet to contract in response to diminished antigen levels, or still be responding to residual infection. These need not consist exclusively or even primarily of HIV-reactive clones. For example, of the viral-associated CDR3s I found, those associated with CMV were found at much higher frequencies than those associated with HIV. Indeed, Betts *et al* reported that only an average 6.31% of total CD8 cells from treatment-naïve HIV-patients were specific for HIV antigens (293). The rest of the expanded clones may be driven to proliferate in an antigen-independent manner, or by exposure to other non-HIV pathogens as a result of increasing immunosuppression (294, 295). For instance it has been shown that HIV induces gastrointestinal damage that leads to increased translation of gut microbial antigens to the circulation (296, 297), which stands to act as a substantial source of persistent antigens for T-cells to react against. Another potential explanation for such prevalent, persistent clones is the more recent discovery that the integration of HIV at different genomic sites associates with clonal expansion and retention (298), even when the integration consists of defective provirus (299), presumably as a result of interacting with the expression of neighbouring host genes. However due to the lack of correlation between CD4 count and either diversity index it seems unlikely that this phenomena causes widespread effects in this data.

Not only does this data describe a profound perturbation to the normal TCR repertoire in the context of chronic HIV infection, these abnormal population structures persist despite otherwise effective ART. Three months of ART is often associated with a reversal of AIDS-associated pathology, clinical improvement and decreased risk of morbidity, even though $CD4^+$ T-cell counts may remain well below normal levels (300, 301). Nevertheless, it is apparent that the recovery of the $CD8^+$ T-cell repertoire remains incomplete despite effective treatment. Large $CD8^+$ clonal expansions are often found to persist after years of therapy (reviewed in (302)), and significant 'holes' in the repertoire may remain. Insufficient epitope coverage as a result of profound CD8 clonal inequality may also explain why HIV patients with low CD4:CD8 ratios are at a greater risk of developing non-HIV related morbidity and mortality (270, 303), even when viral replication is controlled.

In this data I have described that the invariant alpha chain TCRs of MAIT cells – which can comprise in excess of 10% of the circulating CD8 T-cell pool under normal circumstances (304) – are profoundly depleted in this HIV^+ cohort, and remain so over the course of treatment. This is in agreement with similar findings from recent efforts that described the same phenotype using multiparameter flow cytometry (143, 304, 305). It is unclear whether the under-representation of these cells in the blood occurs as a result of increased cell death or trafficking to another tissue, although they do not appear to be preferentially targeted by the virus (143), consistent with their primarily $CD8^+$ or double negative profile (306). However, given their contribution to bacterial ligand recognition at mucosal sites, their absence in even treated HIV^+ patients may represent gaps in the immune protection from certain microbes. Invariant TCR chain NKT cells have also been reported to be selectively depleted early during HIV infection (284, 285, 307), which may similarly lead to gaps in viral protection, although these sequences remain below the thresholds of consistent detection in this study. The loss of these non-pMHC restricted T-cell subsets are thus prime examples of potential holes in the repertoires of HIV patients.

Further longitudinal samples from the patients in this and other cohorts at later time-points after starting therapy will be instructive in assessing the extent and repertoire correlates of long-term immunological recovery. Sequencing RNA from a larger blood sample would also increase the range of detection in these assays. Sorting of patient blood into different populations – particularly $CD4^+$ and $CD8^+$ – before sequencing would allow much greater resolution in the descriptive and predictive power of the experiment, although handling and cell sorting of blood from HIV patients presents its own technical and safety challenges. In this vein, an ideal experiment would likely

involve high-throughput single-cell whole transcriptome sequencing, from which one could extract both paired TCR information and expression of T-cell subset markers. There is obviously more to T-cell immunity than just which clonotypes are present at what frequencies. In the case of HIV, particular clonotypes may be present but exhausted as a result of chronic stimulation (243), and remain unable to offer further protection. Thus transcriptomic profiling of cells with known TCRs would allow qualitative descriptions from which we might estimate phenotype.

In conclusion, my data contribute to the field by describing in detail how HIV infection profoundly alters the TCR repertoire. Partial reversal of inequality may be rapid and lead to improved immune function, coinciding with the first more rapid phase of T-cell recovery (308). Long-term damage to the repertoire may be much more difficult to repair, particularly in adults where naive peripheral T-cell pools are maintained by homeostatic cell division as opposed to thymic production of new cells, which may contribute to long-term morbidity (309, 310). TCR sequencing and similar approaches offer a chance to gain new insight into such processes, which in larger studies may provide useful biomarkers for patient stratification.

5

Protocol development

5.1 Introduction

High-throughput DNA sequencing (HTS) of variable antigen receptor repertoires (or 'Rep-Seq' (74)) represents a tremendous opportunity to dissect immunological phenomena with the utmost breadth and accuracy. However, while the transition to such platforms offers great throughput and resolution, it also introduces a number of technical hurdles which must be overcome in order to produce valid, good quality data. This chapter deals with the nature of these hurdles, and how I have addressed them during my PhD. Herein I detail the amplification procedures and sequence-based metrics that were used to judge efficacy and appropriateness of different strategies. The biological interpretation of the data output from these protocols was covered above in Chapters 3 and 4.

5.1.1 DNA sequencing technologies

In order to fully appreciate the development of protocols for sequencing TCR repertoires, we must first understand how sequencing technologies work, and how high-throughput sequencers are different to earlier generation machines.

5.1.1.1 First-generation DNA sequencing

Despite Watson and Crick solving the three-dimensional structure of DNA in 1953 (311)[1], the ability to 'read' or sequence DNA did not follow for some time. Initial efforts focused on sequencing the genomes of single-stranded RNA bacteriophages and bacterial ribosomal RNA subunits – which were the most readily available pure populations of single RNA species at the time – through what we would now consider to be extremely arduous techniques, involving a great deal of radioactive labelling and two-dimensional fractionation (313, 314, 315). It was by using such techniques that Walter Fiers' laboratory was able to produce the first complete sequence of a gene encoding a protein, that of the coat protein of bacteriophage MS2 (316), followed four years

[1]Notably working from crystallographic data produced by Rosalind Franklin and Maurice Wilkins (312).

later by the complete genome (317). Around this time various researchers[2] began to develop methods for sequencing DNA, again making use of radiolabelled nucleotides. Original observations that one could make use of DNA polymerase to incorporate nucleotides at the ends of bacteriophage genomes that possessed 5' overhangs (318) lead researchers to create their own partially double-stranded DNA molecules through the addition of synthetic oligonucleotides, enabling sequencing by incubating with mixtures of variably-labelled nucleotides (319, 320, 321, 322). A notable alternative technique involved using a number of different chemical techniques to cleave DNA preferentially at different bases, allowing inference of the original sequence by comparison of the resultant molecule lengths (323). However the major breakthrough that forever altered the progress of DNA sequencing technology came in 1977, with the development of Sanger's 'chain-termination' or dideoxy technique, in the same year that his group sequenced the first DNA genome, that of bacteriophage ϕX174 (or 'PhiX') (324, 325). The chain-termination technique makes use of chemical analogues of the deoxyribonucleotides (dNTPs) that are the monomers of DNA strands; dideoxynucleotides (ddNTPs) lack the 3' hydroxyl group that is required for extension of DNA chains (as it forms the bond with the 5' phosphate of the next dNTP) (326). Radiolabelled ddNTPs mixed into a DNA extension reaction at a fraction of the concentration of standard dNTPs results in DNA strands of each possible length being produced, as the dideoxy nucleotides get randomly incorporated as the strand extends, halting further progression. By performing four parallel reactions containing individual ddNTP bases and running the results on four lanes of a polyacrylamide gel, one is able to use autoradiography to infer what the nucleotide sequence in the original template was, as there will a radioactive band in the corresponding lane at that position of the gel. While working on the same principle as other techniques (that of producing all possible incremental length sequences and labelling the ultimate nucleotide), the accuracy, robustness and ease of use led to the dideoxy chain-termination method – or simply, Sanger sequencing – to become the most common technology used to sequence DNA for years to come.

A number of improvements were made to Sanger sequencing in the following years, which primarily involved the replacement of phospho- or tritrium-radiolabelling with fluorometric based detection (allowing the reaction to occur in one vessel instead of four) and improved detection through capillary based electrophoresis, both of which contributing to the development of increasingly automated DNA sequencing machines (327, 328, 329, 330, 331, 332, 333). These sequential alterations produced the first

[2]Principally Ray Wu, Walter Gilbert and Fred Sanger, the latter of whom had been very instrumental in the earlier development of RNA sequencing techniques.

crop of commercial DNA sequencing machines (334) which were used to sequence the genomes of increasingly complex species. These 'first-generation' DNA sequencing machines produce reads slightly less than one kilobase (kb) in length: in order to analyse longer fragments researchers made use of techniques such as 'shotgun sequencing' where overlapping DNA fragments were cloned and sequenced separately, and then stitched together *in silico* (335, 336). The development of techniques such as polymerase chain reaction (PCR) (61, 62) and recombinant DNA technologies (337, 338) further aided the genomics revolution by providing means of generating the high concentrations of pure DNA species required for sequencing. These advances also aided sequencing by less direct routes. For instance, the Klenow fragment DNA polymerase[3] had originally been used for sequencing due to its ability to incorporate ddNTPs efficiently; however more sequenced genomes and tools for genetic manipulation provided the resources to find polymerases that were better able to cope with the additional chemical moeities of the increasingly modified dNTPs used for sequencing (340). Eventually, newer dideoxy sequencers – such as the ABI PRISM range developed from Leroy Hood's research, produced by Applied Biosystems (341), which allowed simultaneous sequencing of hundreds of samples (342) – came to be used in the Human Genome Project, helping to produce the first draft of that mammoth undertaking years ahead of schedule (343, 344).

5.1.1.2 Second-generation DNA sequencing

Concurrent to the development of large-scale dideoxy sequencing efforts, another technique appeared that set the stage for the first wave in the next generation of DNA sequencers. This method markedly differed from existing methods in that it did not infer nucleotide identity through using radio- or fluorescently-labelled dNTPs or oligonucleotides before visualising with electrophoresis. Instead researchers utilised a recently discovered luminescent method for measuring pyrophosphate synthesis[4] to infer sequence by measuring dNTP incorporation when each nucleotide is washed through the system in turn, over the template DNA which was affixed to a solid phase (346). Note that despite the differences, both Sanger's dideoxy and this pyrosequencing method are 'sequence-by-synthesis' (SBS) techniques, as they both require the direct action of DNA polymerase to produce the observable output (in contrast to say the Maxam-Gilbert technique, which uses chemical degradation to produce different sized radi-

[3] A fragment of the *E. coli* DNA polymerase that lacks 5' to 3' exonuclease activity, produced through protease digestion of the native enzyme (339).

[4] This consisted of a two-enzyme process in which ATP sulphurylase is used to convert pyrophosphate into ATP, which is then used as the substrate for luciferase, thus producing light proportional to the amount of pyrophosphate (345).

olabelled fragments). This pyrosequencing technique, pioneered by Pål Nyrén and colleagues, possessed a number of features that were considered beneficial: it could be performed using natural nucleotides (instead of the heavily-modified dNTPs used in the chain-termination protocols), and observed in real time (instead of requiring lengthy electrophoreses) (347, 348, 349). Later improvements included attaching the DNA to paramagnetic beads, and enzymatically degrading unincorporated dNTPs to remove the need for lengthy washing steps. The major difficulty posed by this technique is finding out how many of the same nucleotide there are in a row at a given position: the intensity of light released corresponds to the length of the homopolymer, but noise produced a non-linear readout above four or five identical nucleotides (349). Pyrosequencing was later licensed to 454 Life Sciences, a biotechnology company founded by Jonathan Rothburg, where it evolved into the first major successful 'next-generation sequencing' (NGS) technology.

The sequencing machines produced by 454 (which was later purchased by Roche) were a paradigm shift in that they allowed the mass parallelisation of sequencing reactions, greatly increasing the amount of DNA that can be sequenced in any one run (350). Libraries of DNA molecules are first attached to beads via adapter sequences, which then undergo a water-in-oil emulsion PCR (emPCR) (351) to coat each bead in a clonal DNA population (where ideally on average one DNA molecule ends up on one bead which amplifies in its own droplet in the emulsion). These DNA-coated beads are then washed over a plate with picoliter reaction wells that only fit one bead each; pyrosequencing then occurs as smaller bead-linked enzymes and dNTPs are washed over the plate, and pyrophosphate release is measured using a charged couple device (CCD) sensor beneath the wells. This set up was capable of producing reads around 400-500 base pairs (bp) long, for up to the million or so wells that contained suitably clonally-coated beads (350). This parallelisation increased the yield of sequencing efforts by orders of magnitudes, for instance allowing researchers to completely sequence a single human's genome – that belonging to DNA structure pioneer, James Watson – far quicker and cheaper than a similar effort by DNA-sequencing entrepreneur Craig Venter's team using Sanger sequencing the preceding year (352, 353). The original 454 machine (and first HTS option widely available to consumers), the GS 20, was later superceded by the 454 GS FLX, which offered a greater number of reads (by having more wells in the 'picotiter' plate) as well as better quality data (354). This principle of performing huge numbers of parallel sequencing reactions on a micrometer scale is what came to define the second-generation of DNA sequencing, often made possible as a result of improvements in microfabrication and high-resolution imaging (355).

A number of technologies sprung up following the success of 454, however the most important among them is arguably the Solexa method of sequencing, later acquired by Illumina (354). Instead of parallelising by performing bead-based emPCR, adapter-bracketed DNA molecules are passed over a lawn of complementary oligonucleotides bound to a flowcell; a following solid phase PCR produces neighbouring clusters of clonal populations from each of the individual original flow-cell binding DNA strands (356, 357). This process has been dubbed 'bridge amplification', due to replicating DNA strands having to arch over to prime the subsequent round of polymerisation off neighbouring surface-bound oligonucleotides (354). Sequencing itself is achieved in a SBS manner using fluorescent 'reversible-terminator' dNTPs, which cannot immediately bind further nucleotides as the fluorophore occupies the 3' hydroxyl position; this must be cleaved away before polymerisation can continue (358). These modified dNTPs and DNA polymerase are washed over the primed, single-stranded flow-cell bound clusters in cycles; each cluster incorporates the reverse complement of the base at the position of the cycle number – which can be monitored with a CCD by exciting the fluorophores with appropriate lasers – before enzymatic removal of the blocking fluorescent moieties and continuation to the next cycle and hence next position in the sequencing template. While the first Genome Analyzer machines were initially only capable of producing very short reads (up to 35 bp long) they had an advantage in that they could produce paired-end (PE) data, in which the orientation of the molecules could be reversed, allowing observation of the sequence at both ends of each DNA cluster. The standard Genome Analyzer version (GAIIx) was later followed by the HiSeq, a machine capable of even greater read length and depth, and then the MiSeq, which was a lower-throughput (but lower cost) machine with faster turnaround (359, 360).

A number of other sequencing companies, each hosting their own novel methodologies, have also appeared and had variable impacts upon both what experiments are feasible and the market at large. In the early years of second-generation sequencing perhaps the third major option (alongside 454 and Solexa/Illumina sequencing) (361) was the sequencing by oligonucleotide ligation and detection (SOLiD) system from Applied Biosystems (which became Life Technologies following a merger with Invitrogen) (362). As its name suggests, SOLiD sequenced not by synthesis (i.e. catalysed with a polymerase), but by ligation, using a DNA ligase, building on principles established previously with the open-source 'polony' sequencing developed in George Church's group (363). In brief, following a bead emPCR, clonal DNA populations are interrogated using cycles of hybridising and ligating short degenerate, labelled oligonucleotides: each fluorescent colour corresponds to a particular dinucleotide at a given position in the

template, which can be observed (giving '2 base encoding') before cleaving the fluorescent moeity off and continuing the ligation cycle. Capable of producing paired-end reads – highly accurately, due to measuring pairs of nucleotides rather than single monomers, meaning an error would have to occur on both observations – the SOLiD platform is not able to produce the read length and depth of Illumina machines (364), although it remains competitive on a cost per base basis (365). Another notable technology based on sequence-by-ligation was Complete Genomic's 'DNA nanoballs' technique, where sequences are obtained similarly from probe-ligation but the clonal DNA population generation is novel: instead of a bead or bridge amplification, rolling circle amplification is used to generate long DNA chains consisting of repeating units of the template sequence bordered by adapters, which then self assemble into nanoballs, which are affixed to a slide to be sequenced (366). The last remarkable second-generation sequencing platform is that developed by Jonathan Rothburg after leaving 454. Ion Torrent (another Life Technologies product) is the first 'post-light sequencing' technology, as it uses neither fluorescence nor luminescence (367). In a manner analogous to 454 sequencing, beads bearing clonal populations of DNA fragments (produced via an emPCR) are washed over a picowell plate, followed by each nucleotide in turn; however nucleotide incorporation is measured not by pyrophosphate release, but the difference in pH caused by the release of protons (H^+ ions) during polymerisation, made possible using the complementary metal-oxide-semiconductor (CMOS) technology used in the manufacture of microprocessor chips (367). This technology allows for very rapid sequencing during the actual detection phase (365), although as with 454 (and all other pyrosequencing technologies) it is less able to readily interpret homopolymer sequences due to the loss of signal as multiple matching dNTPs incorporate (368).

This oft-described 'genomics revolution' has drastically altered the cost and ease associated with DNA sequencing. This growth occurs in excess of that seen in the computing revolution described by Moore's law – that the complexity of microchips (measured by number of transitors per unit cost) doubles approximately every two years – as sequencing capabilities between 2004 and 2010 doubled every five months (369). The various offshoot technologies are diverse in their chemistries, capabilities and specifications, providing researchers with a diverse toolbox from which to design experiments. However in recent years the Illumina sequencing platform has been the most successful, to the point of near monopoly (370)[5] and thus can probably considered to have made the greatest contribution to the second-generation of DNA sequencers.

[5]In 2012 the board of Illumina claimed that approximately 90% of the world's total sequencing output is produced using their machines (371).

5.1.1.3 Third-generation DNA sequencing

There is some discussion about what defines the generations of DNA sequencing technology, particularly regarding the division from second to third (372, 373, 374, 375). Arguments are made that single molecule sequencing (SMS), real-time sequencing, and simple divergence (from what are currently accepted as second-generation technologies) should be the defining characteristics of third-generation; it is also feasible that a particular technology might straddle the boundary. For the purposes of this discussion, I am taking the most consistently agreed upon characteristic, in which third-generation technologies are those capable of sequencing single molecules, negating the need for the clonal amplification techniques that typify the previous generation. The first SMS technology was developed in the lab of Stephen Quake (376, 377), later commercialised by Helicos BioSciences, and worked broadly in the same manner that Illumina does, but without any bridge amplification; DNA templates become attached to a planar surface, and then propriety fluorescent reversible terminator dNTPs (so-called 'virtual terminators' (378)) are washed over one base a time and imaged, before cleavage and cycling the next base over. While relatively slow and expensive (and producing relatively short reads), this was the first technology to allow sequencing of non-amplified DNA, thus avoiding all associated biases and errors (372, 374). As Helicos filed for bankruptcy early in 2012 (379) other companies took up the third-generation baton.

At the time of writing, the most widely used third-generation technology is the single molecule real time platform (SMRT) from Pacific Biosciences (380), available on the PacBio range of machines. During SMRT runs DNA polymerisation occurs in arrays of microfabricated nanostructures called zero-mode waveguides (ZMWs) – essentially tiny holes in a metallic film over a chip – which exploit the decay properties of light passing through very small apertures to allow visualisation of single fluorophores (due to the zone of laser excitation being so small), even over the background of neighbouring molecules in solution (381). By depositing single DNA polymerase molecules inside the ZMWs and washing over the DNA library of interest along with fluorescent dNTPs, the extension of DNA chains by single nucleotides can be monitored in real time; after incorporation the fluorescent dye is cleaved away, ending the signal for that position (382). In addition to the benefits of sequencing single molecules in a very short amount of time (additionally producing kinetic data, allowing for detection of modified bases (383)), PacBio machines are capable of producing incredibly long reads, up to and exceeding 10 kb in length, which are ideal for de novo genome assemblies (372, 380).

Perhaps the most anticipated arena for third-generation DNA sequencing development is the promise of nanopore sequencing, an offshoot of the larger field of using

nanopores for the detection and quantification of all manner of chemicals and molecules, beyond the scope of just biology (384). The potential for nanopore sequencing was first established even before second-generation sequencing had emerged, when researchers demonstrated that single-stranded RNA or DNA could be driven across a lipid bi-layer through α-hemolysin molecules[6] by electrophoresis; moreover, passage through the channel blocks ion flow, decreasing the current for a length of time proportional to the length of the nucleic acid (385). There is also the potential to use non-biological, solid-state technology to generate suitable nanopores, which might also provide the ability to sequence double stranded DNA molecules (386, 387). Oxford Nanopore Technologies, the first company offering nanopore sequencers, has generated a great deal of excitement over their nanopore platforms GridION and MinION (388, 389), the latter of which is a small (mobile phone sized) USB device which was released to end users in an early access trial in 2014 (390). Despite the admittedly poor quality profiles currently observed, it is hoped that such sequencers represent a genuinely disruptive technology in the DNA sequencing field, producing incredibly long read (non-amplified) sequence data far cheaper and faster than we ever could before (384, 389, 391).

5.1.2 Considerations for repertoire sequencing

Sequencing variable immune receptor repertoires presents different technological and scientific challenges to more typical sequencing efforts, such as genome sequencing or RNA-seq, which must be addressed.

The region of interest and target of the sequencing will most often be that encoding the complementarity determining region 3 (CDR3), due to its hypervariability and its role in antigen determination. By targeting the CDR3 and its flanking sequences one is able to obtain information about both the non-templated sequence and germline genes used in a particular rearrangement. While there may indeed be biological information located outside of this region – such as polymorphisms further towards the gene termini – this is usually less sought after, and would require a different sequencing approach.

Having selected the target, there are two further considerations: choice of nucleic acid substrate, and method of amplification. The latter consideration is required as the current generation of commercial DNA sequencing platforms (discussed in Section 5.1.1) require higher concentrations of nucleic acid than is easily obtained from appropriate biological samples.

[6]Large ion channels from *Staphylococcus aureus*, that can remain open for extended periods.

TCR sequences differ markedly between its genomic DNA (gDNA) and messenger RNA (mRNA) forms (Figure 5.1A). Primarily, gDNA contains introns, the most significant of which resides between the splice donor site at the end of the J gene and the splice donor at the start of the constant region: depending on the TCR chain and genes involved, this sequence spans anything from 2.5 to 70 kb (inferred from IMGT LocusView (104)). After removal of introns by splicing during transcript maturation, the constant region forms a contiguous sequence with the end of the rearranged J gene. There will also be differences in copy number; the gDNA version will be present once per cell, while the RNA could be present at greater amounts (or even be absent). Indeed a number of papers have reported both temporary up- and down-regulation of TCR mRNA in T-cell cultures following treatment with different strength activation signals[7] (212, 213, 392, 393). As V(D)J recombination is predicted to only produce in-frame sequences one third of the time (due to the random addition and removal of nucleotides), sequencing gDNA would be expected to sequence more non-productive rearrangements than sequencing RNA, since mechanisms exist to prevent or downregulate expression of such failed rearrangements (discussed in Section 3.1.1). Therefore we would expect strategies that target mRNA as opposed to gDNA to be relatively enriched in productive TCR rearrangements. There are additional practical considerations to the choice of nucleic acid, such as the relative increased stability of DNA compared to RNA, or the availability of banked biological and clinical samples.

The challenge with amplifying TCRs lies in the fact that the different V genes can differ substantially in their sequence, thus providing no shared region from which to prime. There are two typical solutions. The first is to generate panels of different primers targeting each of the V genes in one direction, and either all of the J genes (Figure 5.1B, arrows) or the constant region (Figure 5.1B and C) in the other direction. These primers would then be pooled to run a multiplex PCR, in which multiple oligonucleotide pairs prime multiple targets in one reaction. The alternative option is to add a known adapter sequence to the 5' end of all cDNA molecules (with respect to the TCR coding sequence). This allows amplification of all rearrangements with a single primer pair complementary to both the adapter and the constant region (Figure 5.1C). The addition of a known adapter sequence can be achieved using a number of molecular biochemical techniques; such reactions are generally termed rapid amplification of cDNA ends, or RACE (394). Both techniques have merits and disadvantages. By targeting the J genes instead of the constant region in a multiplex PCR one is able

[7] Although it is difficult to know what impact changes in cell size and number have had on these data.

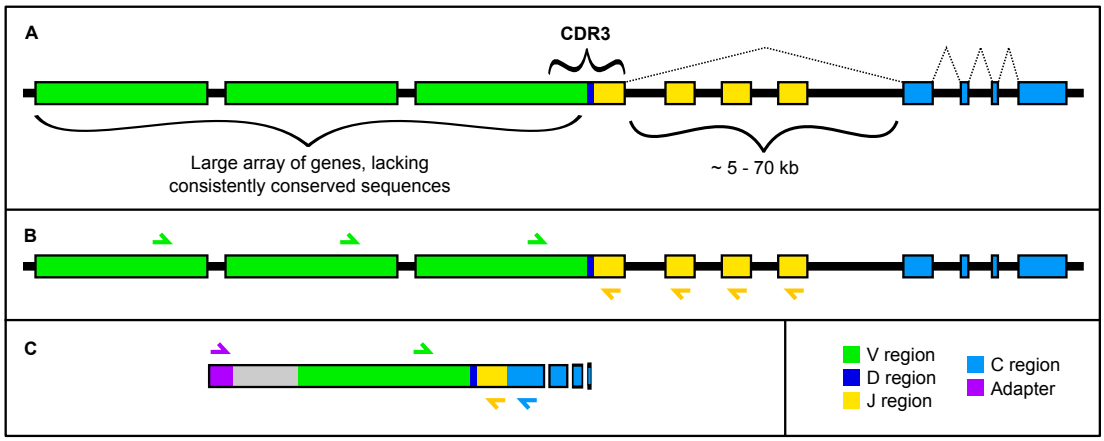


Figure 5.1: How the choice of nucleic acid impacts on choice of TCR amplification protocol. **A**: Schematic of a rearranged TCR genetic locus, indicating the panels of V and J genes, the site of interest (CDR3) and introns (dotted lines indicate splicing events), including the large J-C intron, which can reach up to 68 kb (the distance from the furthest functional alpha chain J gene (TRAJ57) splice donor site to the splice acceptor site of the constant region (TRAC). **B**: Schematic of a multiplex PCR reaction to amplify rearranged TCR chains from gDNA, using a panel of V gene and J gene specific primers (green and yellow arrows respectively). **C**: Schematic of amplifying TCRs from cDNA (reverse transcribed either from constant region specific or poly-T primers), either in a multiplex PCR from V primers to J or constant region primers (green, yellow and blue primers) or using a single primer pair targeting the constant region and a RACE adapter added to the 5' end of all rearrangements (blue and purple arrows).

to amplify directly from gDNA (avoiding the large J-C intron). A multiplex PCR also allows more control over how long PCR products (or 'amplicons') will be, which is important as different sequencing technologies have different sequence length requirements and capabilities. However when mixtures of primers are used there is potential for 'primer bias', wherein different primer pairs can amplify with different efficiencies, which would lead to bias in the output data (395, 396). There can be no primer bias in the a RACE approach, as each polymerisation event uses the same primers. Multiplex PCR approaches could additionally fail to capture the full spectrum of rearrangements of a given chain depending on whether or not all V genes have been included in the primer panel.

5.1.3 Published TCR methodologies

In the few years in which adaptive immune repertoires have been explored using HTS both approaches have been applied to TCRs. Interestingly, despite the capability of multiplex PCRs to start with gDNA templates, a large number of groups with published protocols use RNA (83, 126, 130, 198, 271, 397). There is one notable exception to this, the protocol developed in the laboratory of Dr Harlan Robins, whose TCR repertoire sequencing papers using gDNA templates make up a substantial proportion of the literature in the field to date (398). They have additionally performed a number of experiments in a model TCRγ multiplex system using synthetic DNA libraries covering all possible V-J combinations, allowing measurement and minimisation of the effect of primer bias by altering primer concentrations (399). As for 5' RACE techniques, the vast majority of published techniques make use of the 'template switching' effect (75, 77, 88, 97, 99, 127, 400). A template switch reaction exploits the terminal deoxynucleotidyl transferase activity of certain reverse transciptase (RT) enzymes, which under certain conditions will add short non-templated homopolymers to the end of the nascent cDNA chain (401, 402). Addition of oligonucleotides that contains a complementary 3' homopolymer sequence can then bind act as a primer, allowing the RT to swap strands and produce a DNA strand in the same orientation as the original mRNA[8] (404).

When I began developing the TCR sequencing protocols used in this thesis there were very few published TCR repertoire papers, with even fewer publicly accessible datasets. The few publications that provided both protocol and data have been extremely useful, both as illustrations of how different groups have addressed the chal-

[8]This is analogous to how the retrovirus from which the RT is derived replicates its own RNA genome (403).

lenges of repertoire sequencing, and as a source of data to use in benchmarking my own protocols. Accordingly I have described the process by which three such datasets were produced in Section C.1 of Appendix C.

5.1.3.1 Our protocol; initial decisions

Our group elected to work towards developing a RACE protocol, i.e. extracting RNA from some source of T-cells, adding a known sequence to the V-proximal end of TCR cDNAs and then amplifying between that adapter and the constant region. The main justification for this was a desire to minimise all potential sources of technical bias. Additionally, as we are interested in the effects of infection upon the repertoire – rather than say studying the process of V(D)J recombination itself – we do not expect to need the non-productively rearranged sequences that will not be contributing to binding.

There were a number of other criteria for our eventual protocol to fulfil. It should be sensitive, in that it permits the detection of low-frequency clonotypes. It should be robust, and able to avoid or overcome error, which will always be produced at some level in a high-throughput system; an early paper that sequenced transgenic mice bearing single TCRs detected numerous sequences produced from through the accumulation of errors (405). Furthermore, if such a technology is ever to be of clinical use, it should also be as rapid and cost-effective as possible.

5.1.4 A note on FASTQ quality scores

The analyses presented in this chapter deal with investigating and comparing different sets of high-throughput DNA sequencing data. A necessary component of such comparisons involves considering sequence quality scores. These are values that give an estimate as to the veracity of the individual DNA base calls. Although different sequencing platforms calculate their quality scores differently, the values produced are theoretically variants on the same principle established by the program `Phred`[9]. In this scheme, the quality (Q) of a given base call is a value logarithmically related to the probability (P) of that base call being correct, so that $Q = -10 \times log_{10}(P)$ (407). In this framework, a quality score of 10 (Q10) means the probability of that base call being incorrect is 1 in 10, while Q20 corresponds to 1 in 100 and so on, allowing predicted accuracies to be calculated from quality scores. While different platforms still

[9] `Phred` was developed alongside automated base-calling during the Human Genome Project, before even the invention of second-generation sequencing platforms (406, 407).

use this basic premise, due to the differences in error profiles calculation by the respective base-calling softwares, quality scores are not always directly comparable between technologies.

5.2 Sequencing development

In the course of my PhD I performed a number of iterations of sequencing protocol development, which informed and culminated in the protocol I used to produce the data described in Chapters 3 and 4.

5.2.1 Comparison results from published datasets

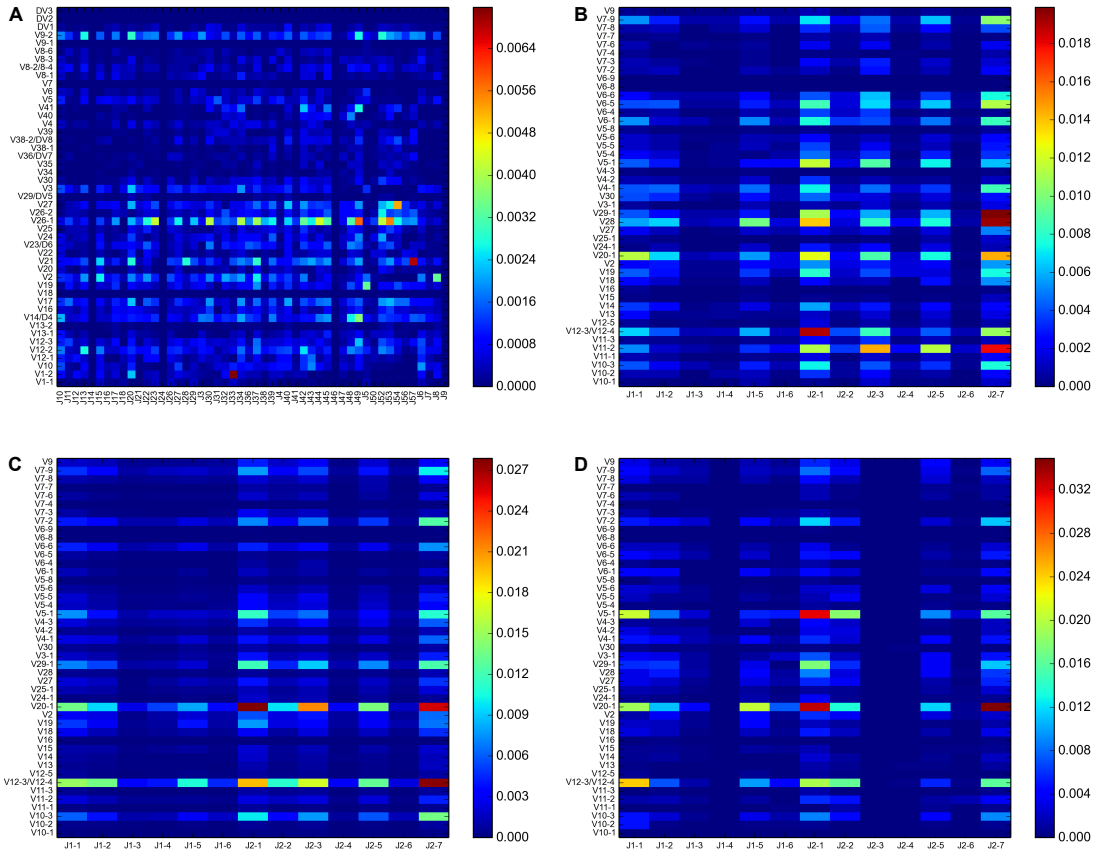


Figure 5.2: Example V-J usage plots for the previously published datasets Produced by `PlotDCR` (a modified version of the plotting modules from `Decombinator` v1.4). **A** and **B**: Alpha and beta distributions from the Wang *et al* paper (SRR030416). **C** and **D**: Beta VJ distributions from the Warren *et al* (SRR060699) and Bolotin *et al* (SRR494403) papers respectively.

I analysed three existing, published TCR-HTS datasets (described in Section C.1) using the bioinformatic pipeline I applied to the data I generated, in order to provide

controls. Table C.1 in Section C.2 of Appendix C shows the raw read and analysis numbers for three FASTQ files downloaded from each dataset, and example V-J usage information is shown in Figure 5.2.

5.2.2 Simple V amplification

5.2.2.1 Protocol

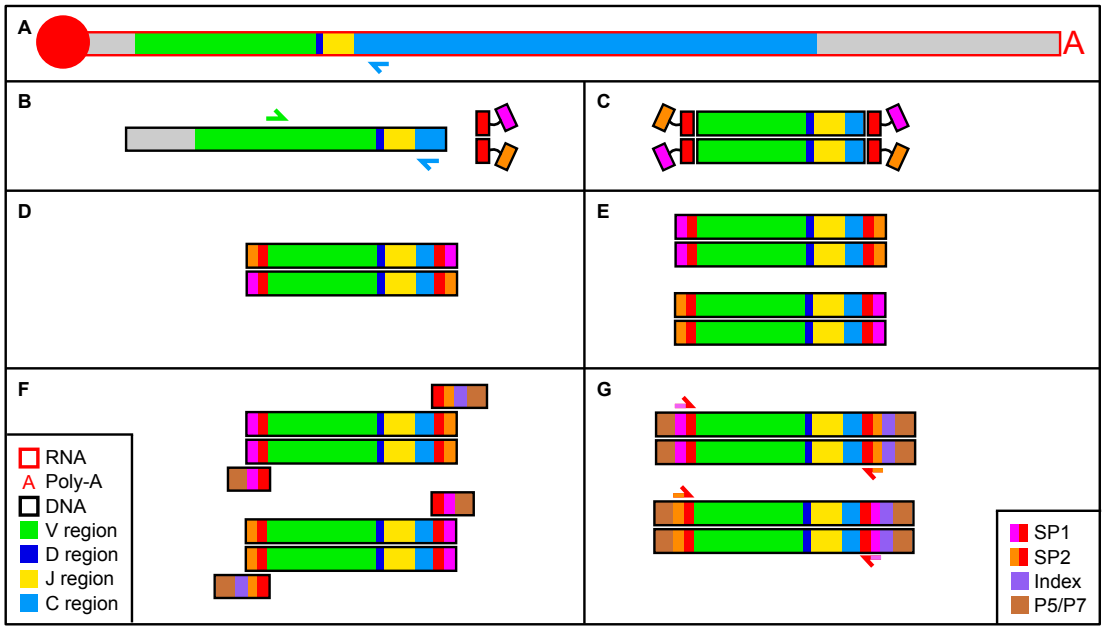


Figure 5.3: Schematic detailing the amplification and library preparation steps involved in our initial simple V amplification experiment. A: TCR mRNA molecule schematic, with different regions approximately to scale. Rearrangements are reverse-transcribed using oligonucleotides directed against the J-region proximal (5') end of the constant region (blue arrow). **B**: Rearrangements that make us of particular V regions are specifically targeted with specific V gene primers (green arrow) and relatively nested constant region primers (blue arrow). Partially double-stranded 'Y-adapters' for adapter-ligation library preparation also shown (right): each consists of two oligonucleotides which are only complementary at one end (red), which makes up the ligation interface. **C** to **G**: Standard adapter-ligation library preparation. Y-adapters are ligated to both ends of the now double-stranded amplicons (**C** and **D**), which results in different adapter sequences being present at either end of the resultant molecules (**E**), permitting paired-end sequencing from either end. Indexing and flow-cell binding adapter sequences can then be added by PCR (**F**), before paired-end sequencing of final libraries (arrows, **G**).

My initial HTS experiment was kept relatively simple. One TCRα and one TCRβ V gene – TRAV8-4 and TRBV12-3 – were selected, and amplified with specific V-region primers. These genes were chosen as they are found in the TCR chains of the Jurkat

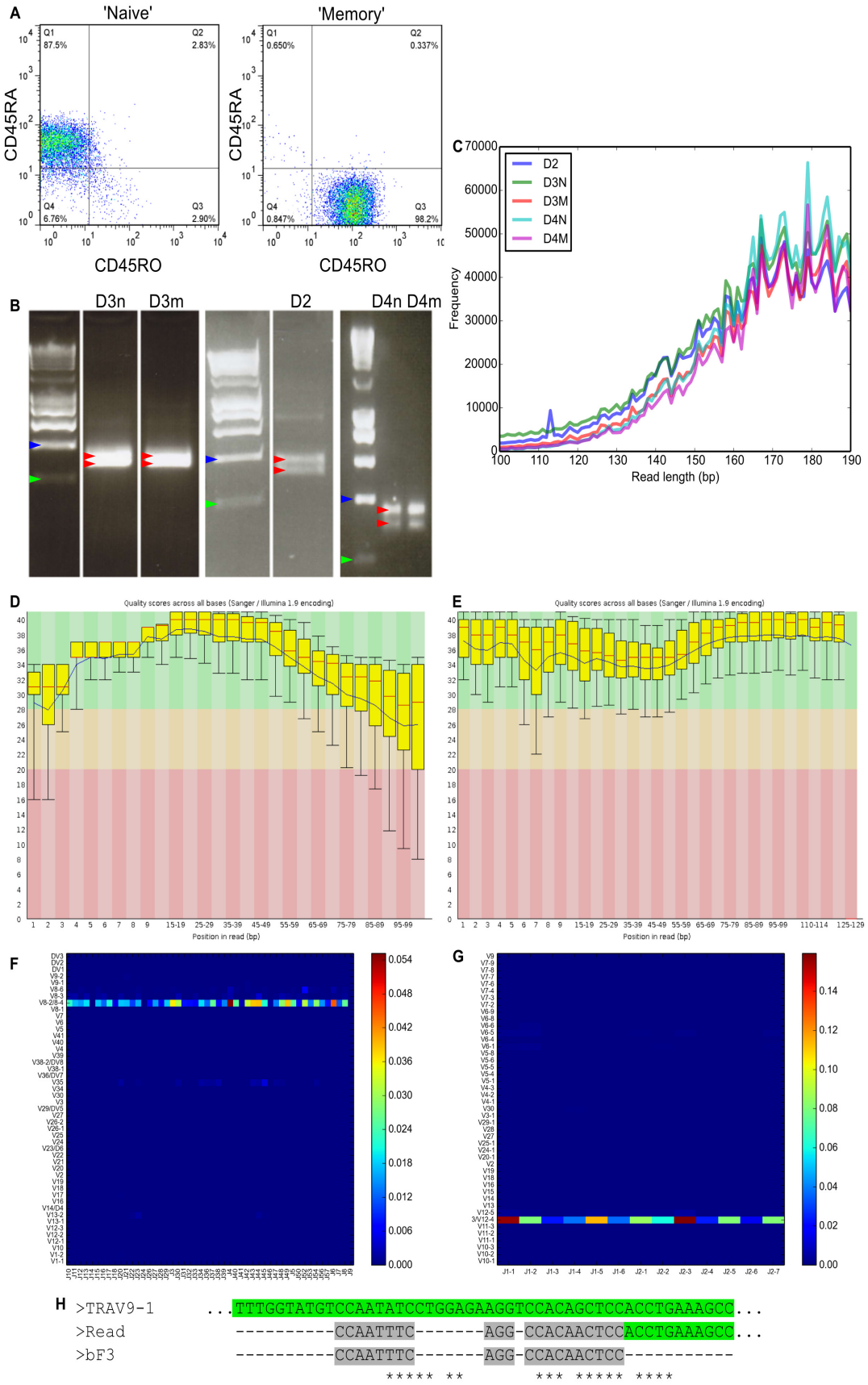
cell line[10]. Jurkat is a widely used, well-studied and easily grown leukaemic cell-line (410, 411), providing a ready supply of clonal TCRs with which I could validate the developing protocol. I therefore designed a multiplex PCR that would amplify all TCR rearrangements that use the Jurkat V regions (Figure 5.3A and B). After amplification, purified TCR amplicons underwent standard adapter-ligation library preparation (Figure 5.3B to G).

I also trialled using magnetic assisted cell sorting (MACS) to isolate different T-cell subsets. This technique relies on attaching paramagnetic beads to specific cell types via antibodies specific for particular cell surface markers, allowing retention of known cell types in a magnetic field (412, 413). I performed negative selection MACS on peripheral blood mononuclear cells (PBMC), whereby all cell types aside from those of interest are bound by antibodies and retained in the magnetic field, providing the sorted cell population in the eluate. Two donors' PBMC were sorted into $CD4^+CD45RA^+$ and $CD4^+CDR45R0^+$ populations, broadly representative of naïve and memory phenotypes (see Section 3.2.2.1). Two rounds of magnetic sorting were required to achieve the enrichment shown in Figure 5.4A, wherein $CD45RA^+$ cells were sorted to 87.5% purity, and $CD45RO^+$ cells to 98.2%. In addition to the two sorted donors, an additional donor's unsorted PBMC underwent the multiplex PCR. Figure 5.4B shows the products of the specific-V RT-PCR by agarose gel electrophoresis. Amplified bands were excised, purified, and sent to Professor Paul Kellam's group at the Wellcome Trust Sanger Institute, where the adapter-ligation was performed. Fully processed libraries were then sequenced on the Illumina HiSeq platform, producing two 100bp paired end (PE) reads.

5.2.2.2 Sequencing results

Sequences produced were tested for the presence of recombined TCRs (explained later in Section 5.3). As Table C.2 shows, this protocol produced a wealth of data, averaging almost 13 million reads per sample, or approximately 1.4 gigabases (gB) of data per sample. I was however only able to detect rearranged TCRs in a very small proportion of those reads: on average only 0.32% of TCRα and 0.77% of TCRβ input reads seemed to contain TCRs, far below the observed assignment rate for the existing published data sets (Table C.1). Closer examination of the reads revealed that the read length of the data (100bp) is simply too short to detect the majority of rearrangements. In order to try and salvage as many reads as possible, I ran the data through `FLASH`, a program that merges overlapping reads in paired-end FASTQ data (117). This should allow

[10]GenBank accession numbers X02592 (408) and K02885 (409) respectively.


A
'Naive'
'Memory'
CD45RA
CD45RO
Q1 87.5%
Q2 2.83%
Q4 6.76%
Q3 2.90%
Q1 0.650%
Q2 0.337%
Q4 0.847%
Q3 98.2%
B
D3n
D3m
D2
D4n D4m
C
D2
D3N
D3M
D4N
D4M
Frequency
Read length (bp)
D
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
E
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
F
G
H
>TRAV9-1 ...TTTGGTATGTCCAATATCCTGGAGAAGGTCCACAGCTCCACCTGAAAGCC...
>Read ----------CCAATTTC-------AGG-CCACAACTCCACCTGAAAGCC...
>bF3 ----------CCAATTTC-------AGG-CCACAACTCC----------
***** ** *** ***** ****

TCR assignation for those clones whose V region could not be identified in 100bp, but whose total amplicon length is short enough to overlap enough to merge, giving enough sequence to make the assignation. While the rates of reads that merged was not large (approximately 15% on average, see Table C.2), they were greater than expected. These libraries were produced from gel-purified amplicons whose migration through agarose suggested sizes of approximately 300bp (Figure 5.4B). Given that the sequencing produced PE reads of 100bp each (and factoring in the 10bp minimum overlap that `FLASH` requires by default) the upper bound read length we should expect to see should be 190bp, significantly shorter than the apparent size of our bands and outside the area of the gel-purification. Plotting the distribution of the lengths in the overlapped reads reveals that 190bp is indeed the maximum length observed (Figure 5.4C). Due to the nature of this distribution, I would conjecture that either some nucleolytic process is occurring in between gel purification and library preparation, truncating amplicons at some low rate, or a fraction of shorter amplicons are migrating at faster than expected rates so as to be included in the gel purification, allowing them to be included in this analysis.

Whatever the cause of the short sequences, the rate of TCR assignation in the merged reads is greater than those reads that did not overlap, by at least two orders

Figure 5.4 *(preceding page)*: Results of simple V amplification and HiSeq sequencing experiment. **A**: Example purities of 'naïve' and 'memory' MACS-sorted T-cell populations, by flow cytometry. Cell populations were first gated on CD3 and CD4. **B**: Agarose gels showing the PCR products that were excised and purified for sequencing. D = donor; N = naïve, M = memory. Left-hand lane in each gel shows Hyperladder I (Bioline): blue and green triangles indicate 600bp and 400bp bands respectively. Red arrows indicate the α and β amplicons that were gel excised for purification (based on Jurkat sequences, approximate expected band sizes were 360bp and 300bp for α and β respectively). **C**: Length distributions of extended reads merged by `FLASH`. **D**: Representative quality score barplot for the D2 raw FASTQ file, as produced by `FastQC`. **E**: Representative quality score barplot for the D2 α inter-tag FASTQ file, i.e. the FASTQ data for the reads that contribute to the `Decombinator`-assigned rearrangement, delimited by the outermost bases of the V and J tag sequences. `FastQC` plot features are: central red line = median, blue line = mean, upper and lower whiskers = 10% and 90% points, yellow box = 25-75% inter-quartile range. Green, amber and red backgrounds represent typical values deemed as 'good', 'passable' or 'poor' quality. **F** and **G**: Frequency plots produced by `PlotDCR` (a modified version of the plotting modules of `Decombinator`) showing the V-J pairing frequency for α and β chain rearrangements respectively for a representative donor (D2). **H**: Example read indicating mis-priming (off-target amplification as a result of the bF3 primer binding a non-targeted site), resulting in presence of additional V genes in the HiSeq data. Green indicates V region identity, grey indicates primer sequence identity and asterisks indicate exact matches.

of magnitude (Table C.3). This corroborates the hypothesis that poor assignation in this HiSeq data is a result of the short read length, as longer reads contained more detectable rearrangements. I also observed that the overlapped reads displayed a lower average predicted frequency of errors, when considering the quality of the basecalls used to identify the hypervariable region (p-value = 0.0012, one-sided Wilcoxon rank sum test)[11]. This will partially be due to the nature of the `FLASH` merging algorithm, as lower quality scores can be discarded in favour of higher ones at positions covered by both reads (117). Overall read statistics for combined samples are shown in Table C.4. In order to assess the distribution of potential errors across the reads I visualised the quality scores using `FastQC` (108), a widely used FASTQ data checking tool. Both raw sequences and 'inter-tag' region FASTQ files – those that contain the sequence used by `Decombinator` to assign a TCR – were processed (Figure 5.4D and E). The inter-tag sequences show much higher and more evenly distributed quality scores. Illumina read quality is known to fall towards the 3' of the read (414), from a combination of loss of phasing of molecules within the cluster, loss of fluorescent signal intensity over time and changes in the cross-talk observed between fluorophores at later cycles (415, 416). The increasing quality at the start of the read is due to the Illumina base-calling software using the information from the first few cycles to calibrate the quality-determination for later cycles. Thus part of the higher and more even quality seen in the inter-tag is due to selecting from a higher quality internal subsequence, as well as the mitigation of quality loss from merging reads. Additionally, low quality 3' tails would be expected to contain an increased proportion of errors; due to the read length this is the area where the V tag for potential TCR arrangements would fall; we are thus selecting for higher quality reads as those with errors are less likely to be detected by `Decombinator`.

Figure 5.4F and G show representation frequency distribution heatmaps of α and β V-J gene usage to provide an overview of the spread of TCR clonotypes. As expected we see a polyclonal response within the V genes targeted by the primers. Low frequency clones that do not use the targeted V genes are presumed to largely represent instances of mispriming. An example of such mispriming is shownin Figure 5.4H, which was identified as a rearrangement using TRAV9-1 amplified from the primer directed against TRBV12-3 (bF3). Mis-assignment by `Decombinator` could also produce such reads.

[11] Note that there is no significant difference between the mean quality scores of these different groups, which is likely due to the fact that the Q score value is a linear representation of logged data.

5.2.2.3 Verdict

While the Illumina HiSeq was able to produce in excess of ten million reads for each of our samples, the read length prevents it from being useful for our purposes. However, the quality of the data and the rate of assignation in merged (and therefore longer) reads suggests that the Illumina technology shows promise as a suitable platform could we but attain longer reads.

5.2.3 Poly-T RACE

5.2.3.1 Protocol

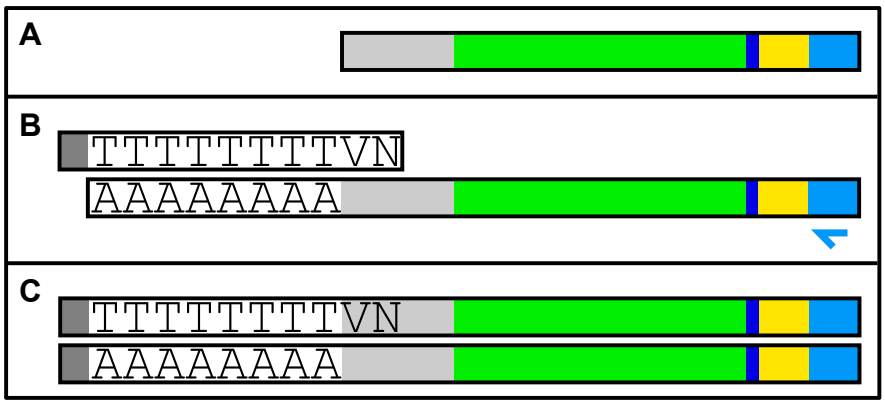


Figure 5.5: Poly-T RACE strategy schematic. **A**: TCR cDNA, produced by reverse transcription as displayed in Figure 5.3A. **B**: Tdt supplied with dATP is used to polyadenylate the 3' end of the cDNA, to which a primer bearing a 3' poly-T sequence can anneal. Amplification occurs between the poly-T RACE primer and a relatively nested constant region primer. A 'VN-hook' is used to ensure that the poly-T primer anneals at the V-region end of the poly-A tract. **C**: Amplicons produced using poly-T RACE.

Having established various protocol parameters with the simple V amplification technique, I next moved towards a protocol that would allow amplication of all possible TCR rearrangements in a non-biased manner. This involved switching to a rapid amplification of cDNA ends (RACE) protocol, which allows amplification of a cDNA molecule in which only a single priming site (e.g. a TCR constant region) needs to be known through addition of another priming site to the appropriate terminus of the cDNA (394). In the first RACE trial I performed this adapter was added through treatment with Tdt[12], supplied with a single deoxyribose nucleoside triphosphate (dNTP), which results in the homopolymer 'tailing' of the 3' end of the cDNA (and thus the 5' end of the original transcript) (25, 417). All such labelled transcripts can then be

[12]This enzyme is coincidentally the same one required for non-templated addition of nucleotides during the TCR recombination process.

amplified using one primer against the original known sequence (the constant region) and another that contains in its 3' a homopolymer of the complement nucleotide to that which was used in the tailing reaction, which in my case was dATP (Figure 5.5).

Initial attempts to translate the reaction conditions of the single V PCR to the poly-T RACE PCR were unsuccessful, so I made a number of alterations. These included purifying DNA products in between various stages of the reaction, and addition of a 'VN' dinucleotide 'hook' to the 3 end of the poly-T primer – where V is any nucleotide apart from T, and N is any nucleotide – encouraging the annealing of the poly-T primer at the 3' end of the poly-T tail. The most significant improvement was seen in the switch to and subsequent optimisation of a reaction using Phusion DNA polymerase (New England Biolabs), replacing the Taq polymerase variant used previously. Phusion is a 'second generation' recombinant polymerase, engineered to increase processivity and speed through addition of a DNA binding protein domain (418). In addition to this increased processivity (i.e. how many nucleotides a given polymerase molecule incorporates before dissociation from the DNA) Phusion reportedly has higher fidelity, making fewer mistakes (419), which should reduce sequence errors. Phusion has even been reported to produce libraries which retain higher post-homopolymer quality sequences (by Sanger sequencing) (420), making it a credible suitable polymerase.

5.2.3.1.1 Poly-T RACE cloning

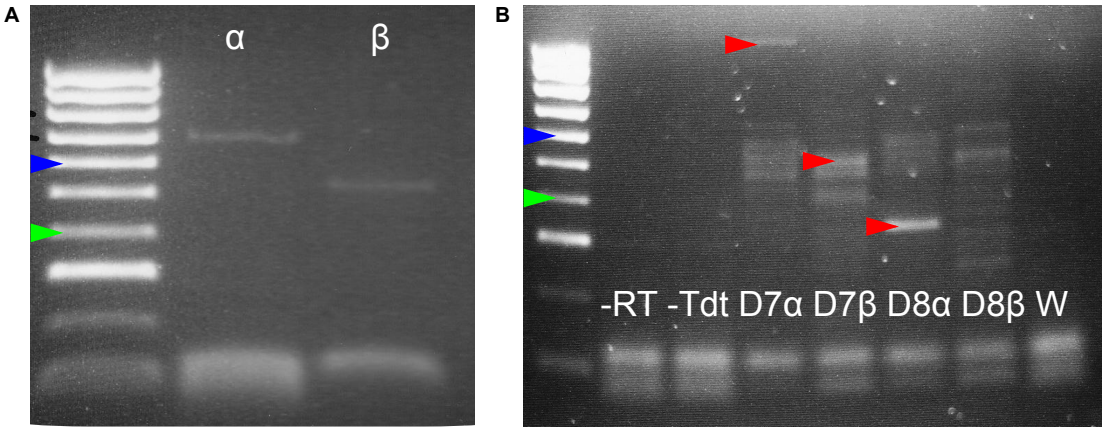


Figure 5.6: Agarose gel scans showing amplicons produced using the preliminary poly-T RACE protocol. Original genetic substrate taken from either lysed Jurkat cells (**A**) or two healthy donors' PBMC (**B**). Left-hand lane in each gel shows 100bp Hyperladder IV (Bioline); blue and green triangles indicate 600bp and 400bp bands respectively. Red arrows in B indicate those bands that were excised and re-amplified before cloning into pUC19 for Sanger sequencing (see Appendix Figure C.1). Control lanes were: −RT (minus reverse transcriptase); −Tdt (minus homopolymer-tailing); W (PCR water control).

In order to test the efficacy of the optimised Phusion protocol I first amplified the TCR from RNA extracted from Jurkat cells. As Figure 5.6A shows, I obtained a single band for each chain, representing the alpha and beta rearrangements: Sanger sequencing of these PCR products corresponded to the published sequences for these chains. Application of this protocol to PBMC samples from two healthy donors revealed a ladder of bands, representing clonotypes of differing mRNA length and frequency (Figure 5.6B). In order to test whether this protocol was reliably amplifying genuine rearrangements, I excised bands from the gel shown in Figure 5.6B (red arrows), purified the DNA, re-amplified and cloned into plasmids to allow Sanger sequencing.

All normally sized TCRβ chains cloned from donor 7 (D7) were legitimate, productive TCR sequences (Appendix Figure C.1A and B). However, the D7 and D8 alpha chains (which had been chosen due to their aberrant band sizes) proved less typical. D7α (Figure C.1C) produced the longest band from the poly-T RACE PCR, which turned out to be due to the V region being used by this rearrangement possessing a unexpectedly long 5 untranslated region (UTR)[13]. This sequence does not appear to represent gDNA contamination, as the leader intron immediately upstream of the V region has been correctly spliced. Conversely, the D8α amplicon (Figure C.1D) was the smallest distinct band, which yielded another peculiarly primed but correctly rearranged receptor. In this instance a molecule of the TRAC primer aRC1 seems to have lost the five 3'-most nucleotides (likely through the strong 3' to 5' exonucleolytic activity of Phusion), allowing amplification between two aRC1 moleculs (Figure C.1E). These results indicate that this protocol is capable of selectively amplifying TCR sequences. However they also indicate both the presence of unexpectedly long mRNA species and erroneous priming events. I also observed that two of the cloned sequences – Figure C.1B and another, non-productive rearranged clone (not shown) – contain more thymidine residues in the RACE homopolymer than were present in the primer. This is likely caused by polymerase slippage, a form of errors in which repetitive sequences cause the enzyme to dissociate, allowing the newly produced strand to shift in place before re-association, forming insertions and deletions (421).

5.2.3.2 Sequencing results

5.2.3.2.1 454

[13]Note that this sequence belongs to a canonical MAIT cell sequence, underlining how frequent these cells are, as discussed in Section 3.1.5. Furthermore, this cDNA could potentially have been even longer; it contains a poly-A tract 668bp upstream of the start of its V region to which the our poly-T primer could bind, relatively nested to a potential more 5' Tdt-added adenine tract.

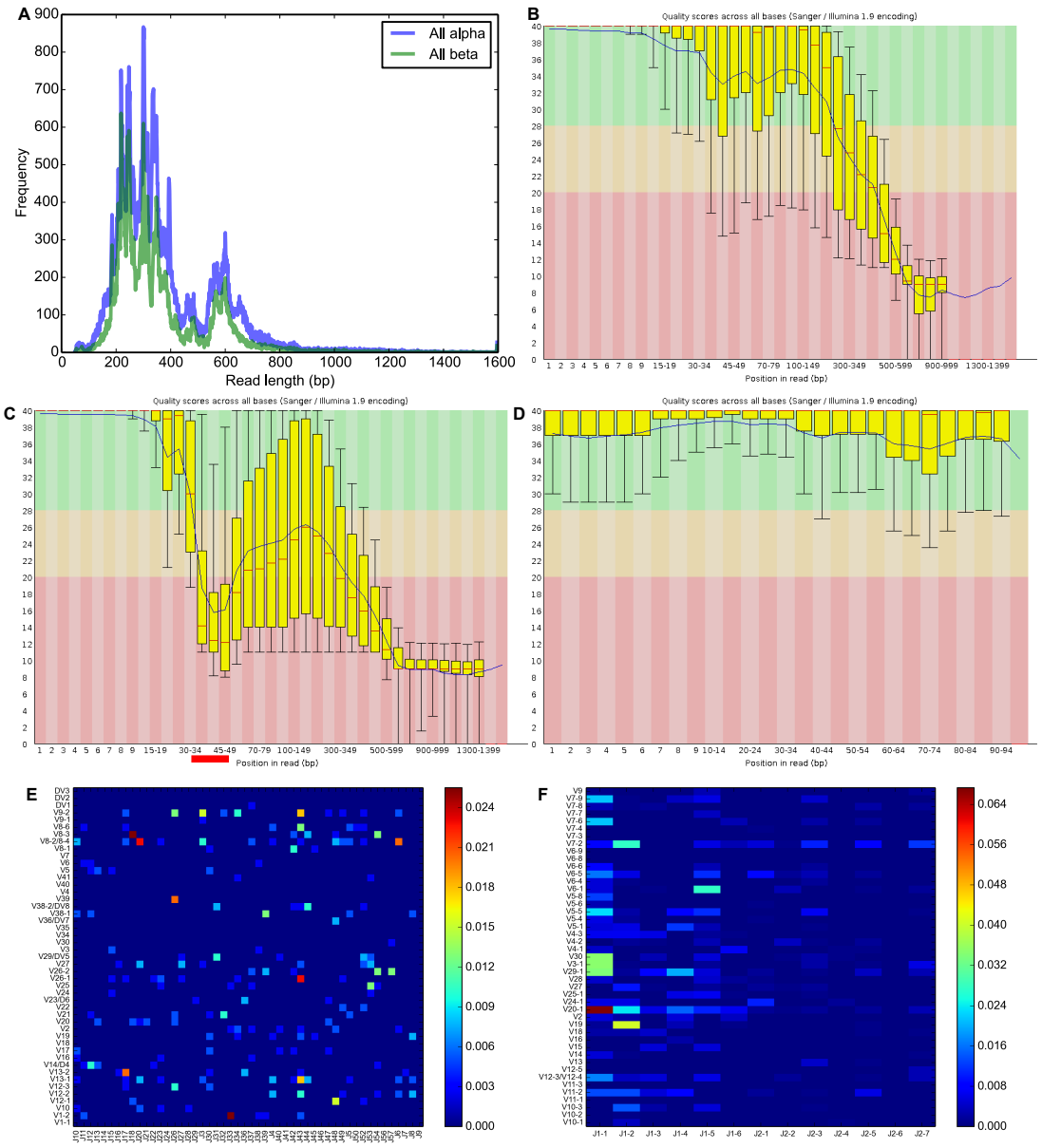


Figure 5.7: Summary data for the poly-T RACE amplification products as sequenced on the 454 GS FLX. A: Length distributions of the 454 data, for the α and β read files pooled from all donors. The upper limit read length of 1600bp is the number of different 'flows' performed by the machine (where each nucleotide is washed over the wells of the plate in turn plate) (422). It is unlikely these rare sequences represent genuine TCR transcripts. **B**: Quality score barplot for the raw FASTQ file of a representative donor (D10, α file), produced by `FastQC` (see Figure 5.4 legend). **C**: Quality score for the reads (sorted from all files) that contain the RACE homopolymer at the start of the read, showing the impact of the homopolymer on the quality of the read. Red line indicates approximate position of the homopolymer in the read (which is variable, due to the indels the homopolymer causes). **D**: Quality score for the inter-tag FASTQ sequence for D10 α. **E** and **F**: Frequency plots produced by `PlotDCR` showing the V-J pairing frequency for α and β chain rearrangements respectively for D10.

After the low rate of TCR assignment observed using the Illumina HiSeq, I next opted to try Roche's 454 sequencing platform, which has been used in a number of early TCR repertoire sequencing strategies (83, 97, 129, 198). This technology is capable of much longer read lengths (350). Purified amplicons were produced from four healthy donors and again sent to collaborators at the Wellcome Trust Sanger Institute for sequencing. Amplicons pools were randomly fragmentated by sonication and subjected to standard library preparation (analogous to that illustrated in Figure 5.3B-G).

Read statistics from this run are displayed in Table C.5. The samples contained an average 19,000 reads each, below the 25-30,000 expected given the capacity of the plate[14] and how many samples we multiplexed. Of these reads, an average 41% were assigned TCRs by `Decombinator`. The 454 produces reads of differing lengths; the length distributions are displayed in Figure 5.7A. The quality score distribution in 454 reads (Figure 5.7B) mirrors that of the HiSeq data (Figure 5.4D). 454 qualities are also known to decrease towards the 3' of the read, and for similar reasons (accumulation of noise and loss of signal) (416, 423).

Due to the mechanism of the 454 base-calling software, its output is particularly prone to insertion and deletion errors ('indels') following homopolymer tracts (361, 368), therefore I assessed the impact of these sequences in this data. I pooled the reads from all donors and divided them into either orientation, i.e. those that contain the RACE homopolymer either in their 5' or 3' end. Despite theoretically having been produced in equal proportions during the PCR, reads that contained a 5' RACE homopolymer only accounted for 19% of the total. These reads showed a distinctly poorer quality profile than whole reads (Figure 5.7C), with quality scores declining precipitously after the approximate location of the poly-T. As in the HiSeq data, the quality distribution for the inter-tag FASTQ sequence is similarly improved relative to the whole read scores (Figure 5.7D), again suggesting that `Decombinator` does indeed select out higher quality reads. Figure 5.7E and F show the α and β VJ gene usages for a representative donor; despite a spread of clonalities these distributions were different between donors, and did not resemble those observed in published datasets (Figure 5.2).

5.2.3.2.2 MiSeq – adapter-ligation

In parallel to the 454, I also explored the then-newly available Illumina MiSeq platform. This makes use of the same Illumina reversible-terminator chemistry as the HiSeq, but with longer reads and a faster turnaround (424). TCRs were amplified with

[14]Each run can produce up to 1×10^6 reads per run, of which we were allocated a quarter.

the poly-T RACE amplification (Figure 5.5) followed by NEBNext adapter-ligation (Figure 5.3B-G) and run on the MiSeq.

Table C.6 shows the summary data for these original MiSeq test samples. Initial trials spiking small amounts of TCR library into runs again produced both fewer reads than expected and a lower rate of `Decombinator` assignation than in the 454 data (31% versus 41%). The read quality was also lower than the reported capabilities of the machine, lower even than unrelated samples sequenced in the same run. This appears to be due to the RACE homopolymer having a markedly deleterious effect on the quality of a given read; post-RACE-homopolymer sequences rapidly fall to the lower limit (Figure 5.8A)[15]. The quality of reads that originate from the constant region of the amplicon show greater quality scores (Figure 5.8B), but still below the reported accuracy of the machine. The distribution of homopolymer tract length (Figure 5.8D) further indicates polymerase slippage. V-J pairing frequency (Figure 5.8E and F) again revealed a preference for the Jβ1 cluster, in contrast to previous published datasets (Figure 5.2B-D). This lead to an alteration of the beta constant region primers: instead of the one original primer (βRC1), which used two degenerate bases to cover the sequences of both of the TRBC genes[16], two primers exactly specific for each TRBC were designed and mixed in equal concentrations.

5.2.3.2.3 MiSeq – PCR library preparation

Neither 454 nor Illumina platforms appear able to efficiently sequence libraries containing the homopolymer introduced during the poly-T RACE protocol. However I observed that reads starting from the constant region end of the amplicons – those that contain the V(D)J information – tended to not suffer from this poor quality. Therefore we theorised that by sequencing the libraries unidirectionally from the constant region one could obtain all the clonal information without wasting reads by sequencing through the RACE homopolymer. I modified the library preparation protocol to add the Illumina sequencing and flowcell-binding elements specifically to both ends of TCR cDNA via a stepwise PCR protocol (as opposed to the random addition of sequences through ligation, as depicted in Figure 5.3B-G). This also offered the opportunity to use the constant region primer itself as the sequencing primer, permitting sequencing

[15]The lower limit of Illumina quality scores is Q2, although this does not refer to a specific measured error rate. Instead it is used if successive base calls towards the end of a read are consistently low quality (e.g. Q15 or lower) to indicate that such sequences are probably not suitable for downstream analysis (425).

[16]Accession numbers M12887 and M12888.

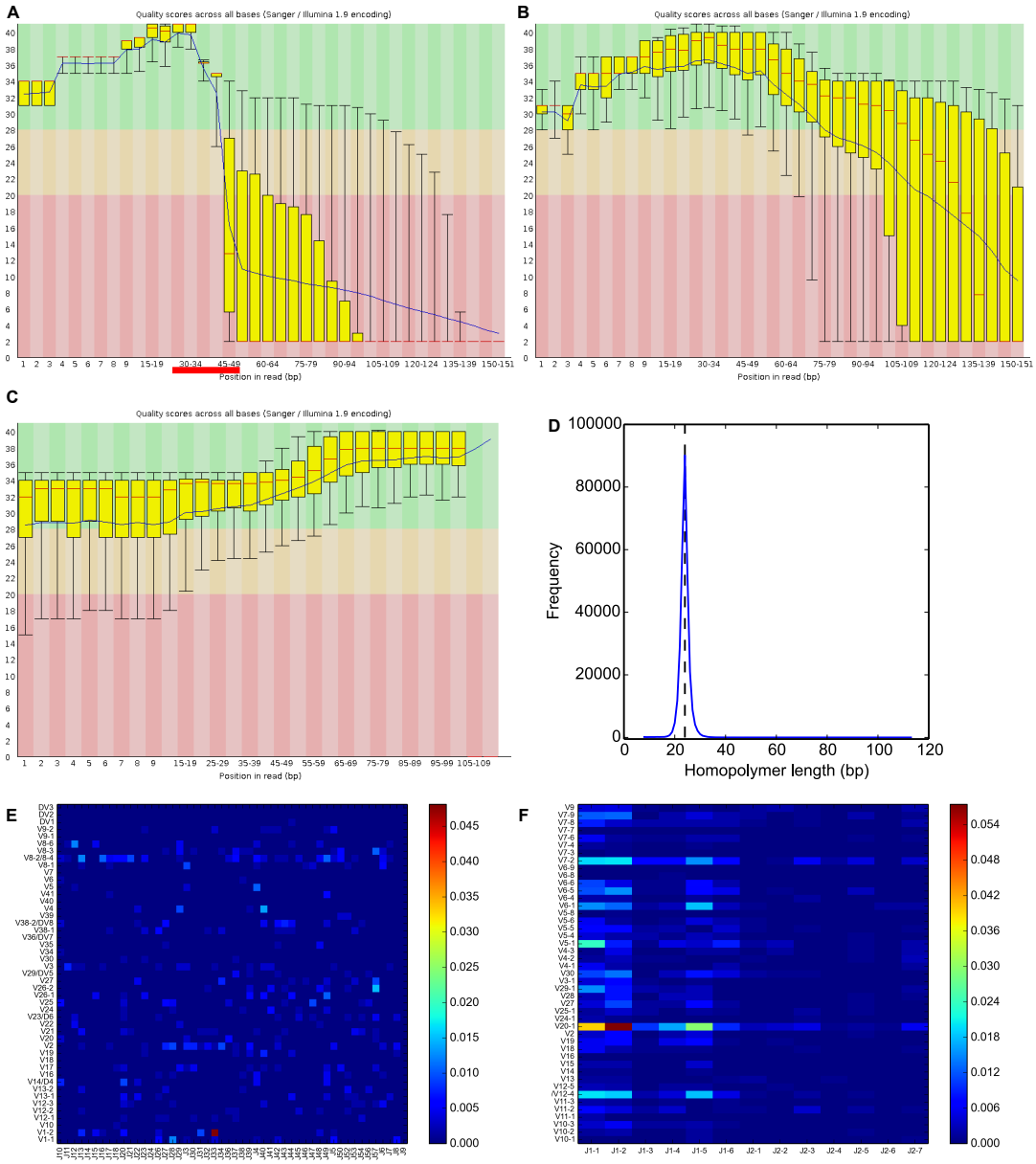


Figure 5.8: Results for the first poly-T RACE amplicons being sequenced on the MiSeq, having undergone NEBNext adapter-ligation library preparation. A: Quality barplots of sorted reads that were found to contain the poly-T RACE homopolymer, produced by `FastQC` (see Figure 5.4 legend). Red line indicates approximate location of homopolymer. **B**: Quality distribution of the sorted reads that did not contain the homopolymer. **C**: Quality barplots for the inter-tag FASTQ sequence, i.e. the FASTQ data that contributed to `Decombinator`-assigned TCRs. The quality of the inter-tag data increases because the `Decombinator`-delimited TCR FASTQ data has been converted to the correct orientation for transcription (i.e. V→ (D)→ J→ C). **D**: Frequency of different length homopolymers of the sorted reads that made up the data summarised in **A**. The first 70 nucleotides of each read were searched for a 'seed' homopolymer consisting of eight matching nucleotides (setting the lower bound), and then any additional identical bases were counted. 90113 out of the total 292096 reads (31%) contained homopolymers of equal length to that of the primer. **E** and **F**: Frequency plots produced by `PlotDCR` showing the V-J pairing frequency for alpha and beta chain rearrangements.

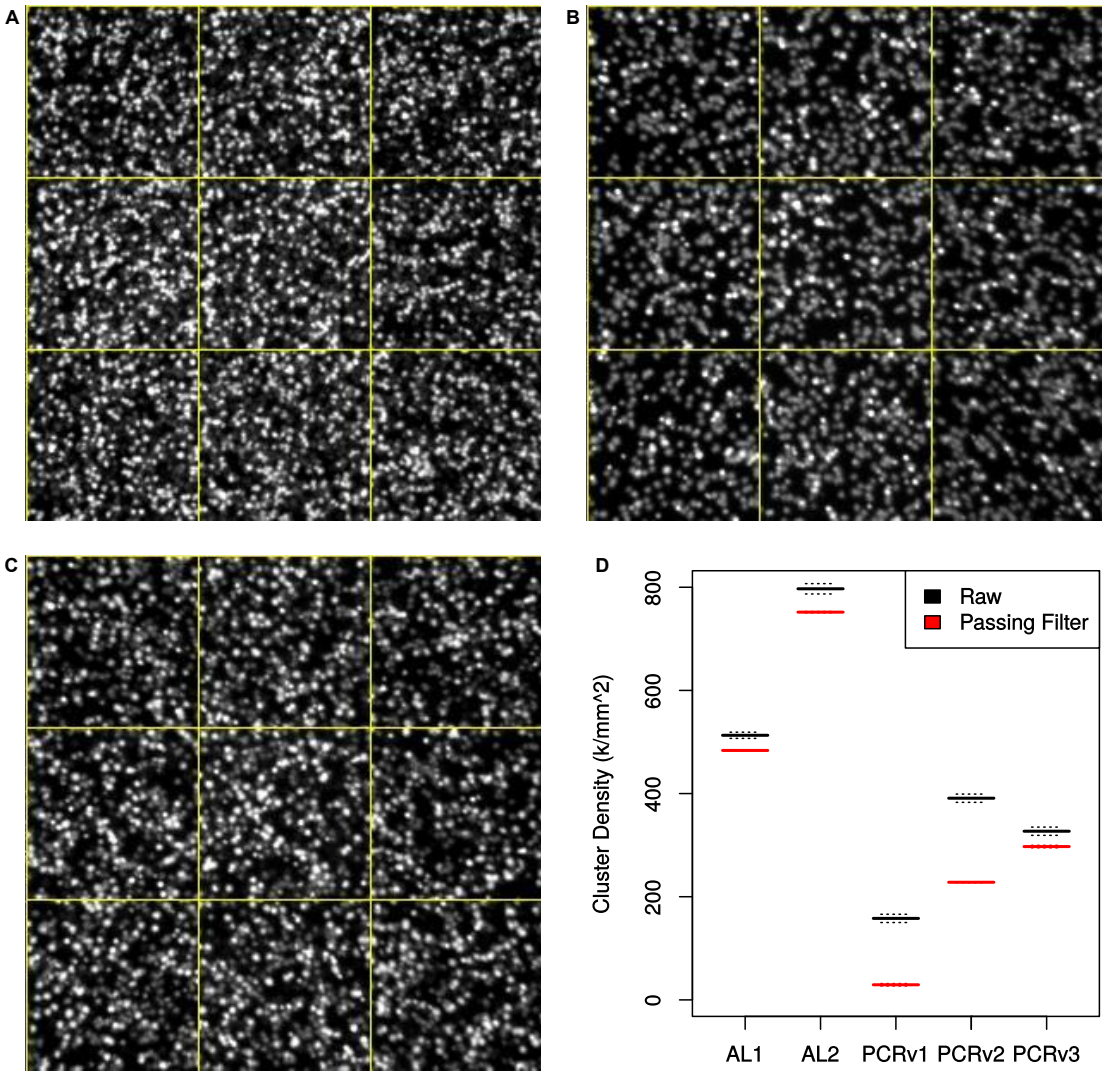


Figure 5.9: Clustering on the MiSeq flowcells. **A** to **C**: Thumbnail images of DNA clusters produced by bridge amplification during different MiSeq runs. Images are taken from the adenosine channel for the first cycle of each respective run, but are representative of all channels. Images are produced by the MiSeq software, and sample different sections of the tile for an overview of cluster formation. **A**: An example of a successful run performed using the same versions kits on the same MiSeq machine, with a mean raw cluster density of 964k/mm^2. **B**: The clustering of the first MiSeq run in which stepwise PCR was used to add the required sequencing elements (158k/mm^2). These clusters are diffuse and low-intensity, which likely contributed to low levels of cluster detection. **C**: The second run had PhiX genome spiked in to provide additional diversity produced greater density (391k/mm^2). Given the poor rate of TCR assignment many of these clusters likely correspond to PhiX reads. **D**: Average raw cluster densities (black) and clusters that pass Illumina filters (e.g. chastity and phasing filters, red) for each iteration of initial MiSeq runs. 'AL' = adapter-ligated, i.e. the runs presented in Table C.6. TCR libraries constituted a minor proportion of these AL runs, so the majority of the clusters should be formed from the unrelated libraries (which were randomly sheared viral genomes). 'PCR' runs had Illumina adapter sequences introduced by stepwise PCR. 'PCRv1' was the initial attempt (sequencing off the TCR constant regions with custom primers), 'PCRv2' introduced PhiX spike-ins (and reverted to standard SP1 oligos) and 'PCRv3' introduced random nucleotides between SP1 and TCR constant regions, to augment initial nucleotide diversity for every cluster. The image shown in A was kindly provided by Anthony Brooks at UCL Genomics.

further into the V gene. Initial attempts suffered from low cluster density, and only a small proportion of those clusters passed all requisite filters. However visual inspection of images taken by the sequencer did not appear to show loss of clusters. Instead there was a relative increase in low-intensity, more diffuse clusters (Figure 5.9A and B), which might be less readily accepted by the detection software. Illumina base-calling software calibrates cluster detection based on the results of the first few cycles of imaging, assuming that there should be approximately equivalent proportions of each nucleotide at each step. This has lead to reports of problems when sequencing so-called 'low-diversity' libraries (426). In order to overcome such low nucleotide diversity, six random nucleotides were added between the constant region and sequencing primers, and highly diverse 'PhiX' libraries[17] were spiked in. Cluster densities improved following these alterations (Figure 5.9C and D), as did quality scores and rates of TCR assignation, although all values remained far below the reported capabilities of the machine. Binding sites for the primary Illumina sequencing primer (SP1) were also reintroduced, so as to be able to sequence under the most controlled and tested conditions. However even as number of raw reads per run increased, the proportion of reads that Decombined remained low; large amounts of the data was made up of reads that mapped to the PhiX genome, or were seemingly noise.

One hypothesis that could explain the lack of cluster generation is that the sequences required for bridge PCR and Illumina sequencing were not being correctly incorporated during library preparation. Individual amplicons were cloned from final sequenced libraries, revealing that all requisite sequences were present in the appropriate locations and orientations (Appendix Figure C.2). Therefore this method of library preparation does not appear to be responsible for the low cluster density.

Consultation with Illumina sequencing experts lead us to the conclusion that the most probable cause of the consistently low cluster density numbers (and hence poor read depths) was the RACE homopolymer. One possible explanation posited is that the RACE homopolymer might interact with the poly-T oligonucleotide sequences that are used to attach the P5 and P7 binding elements to the flowcell. This issue is seemingly separate from the poor quality in sequences downstream of homopolymers, and is independent of library diversity, as the problem was maintained even after adding leading random nucleotides and spiking in PhiX DNA. Thus the inclusion of such long homopolymers in the libraries likely precludes one from achieving optimal output.

[17] Genomic libraries of the bacteriophage ϕX 174 – a common control library in the sequencing industry – produced through random shearing and adapter ligation.

5.2.3.2.4 Effect of native TCR homopolymers

Even without considering the RACE poly-T tract, we would expect there to be a certain number of homopolymer sequences occurring within TCR transcripts. Therefore I also assessed the extent of the these 'native' homopolymers and whether they affect the quality of our reads, for both the 454 and MiSeq data. I counted the number of homopolymers of four bases or longer in the prototypical alleles of the IMGT-described functional TCR genes, revealing that homopolymers occur in the majority of germline genes: over 350 were found spread across 132 of the 155 germline genes (Figure 5.10A). A similar search applied to a published, rearranged $\alpha\beta$ TCR data set that did not use homopolymer RACE – which would contain any additional homopolymers produced during recombination – revealed that 88% of reads contained homopolymers of four bases or longer, and that those reads typically contained at least two such sequences (Figure 5.10B). This indicates that even without considering the effect of the RACE homopolymer the majority of reads in a TCR library can be expected to contain homopolymer sequences. Any TCR sequencing approach must therefore be able to cope with them. In order to assess whether these native homopolymers impacted upon quality, I first combined all of the reads for all samples of a given run. For the 454 data, reads containing the RACE homopolymer in their 5' were filtered out. A 200bp window downstream of the sequencing primer – that should contain any V(D)J information without reaching the poorest quality at the 3' of reads – was then extracted and searched for homopolymers of five nucleotides or longer. The last five nucleotides of any such homopolymer, and the next 15 bases, were then extracted. Random sequences of comparable length were also extracted from the same window of the reads for comparison. The median quality of 454 reads during a homopolymer are much lower compared to random sequences, requiring almost ten subsequent base calls to return to baseline (Figure 5.10C). In contrast, the median quality of the MiSeq data within and following homopolymers is unperturbed relative to the random sequence control.

5.2.3.3 Verdict

The convenience of homopolymer-tailing 5' RACE in studying TCRs was realised shortly after its development (427), and was still being used for lower-throughput analyses at the time of our adopting it as a possible protocol for high-throughput TCR repertoire sequencing (428). Therefore in my PhD I trialled sequencing TCR libraries amplified using this approach on both Roche's 454 GS FLX platform, and Illumina's MiSeq platform.

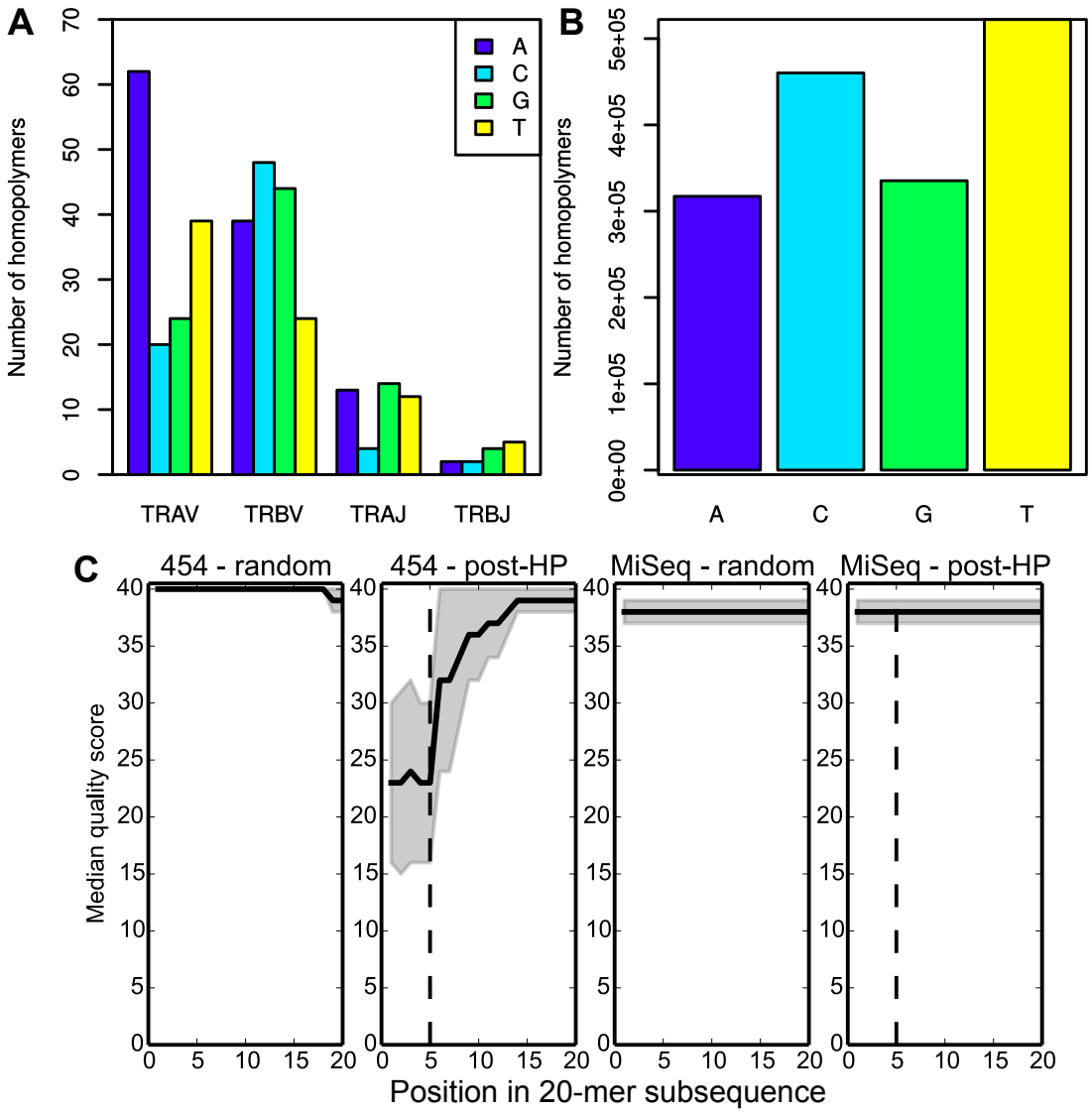


Figure 5.10: The extent and effect of 'native' homopolymers (HP) in TCR repertoire sequence data. A: The number of homopolymers of four nucleotides or longer in the germ-line V and J genes of the human α and β chain TCR loci. Only the prototypical alleles for functional genes have been included. 356 total homopolymers were detected, across 132 out of 155 germ-line genes. **B**: The number of homopolymers of four nucleotides or longer in the pooled data from Wang *et al* (83) (see Section C.1), which was not produced using a homopolymer approach, and sequenced using 454 technology. 1,634,849 homopolymers were detected, across 747,829 of the total 847,610 reads. **C**: The effect of native TCR homopolymers on the quality of the data. Homopolymers of five nucleotides or longer were detected, and the last five nucleotides plus the next 15 bases were harvested into separate FASTQ files, for both the pooled 454 data and that of the third, most productive MiSeq run (PCRv3). Due to having more reads (and homopolymers), the post-homopolymer subsequences from the MiSeq data were subsampled to match the number of detected 454 subsequences (31285 reads). An equal number of reads were then randomly subsampled from the initial pooled FASTQ files, and then random 20-mers were subsampled from these (from between positions 40 and 240, thus only sampling the higher-quality section of the read that should contain any V(D)J information while skipping the initial primer regions). Solid black lines represent median quality scores for each position, grey shaded areas represent median absolute diversity, dotted black lines indicate the end of the homopolymer sequences.

454 seems to be an inappropriate technology for the sequencing of TCR libraries, primarily due to a propensity for quality scores to decline following homopolymer sequences – indicative of a lower probability of base calls being correct. Given the regularity with which homopolymers occur natively within the TCR transcriptome, this represents a substantial source of potential error, theoretically producing a less accurate assessment of a sample repertoire. While producing reads that are long enough for our purposes, there are a number of other technical and practical reasons that make 454 sequencing a less viable option. 454 sequencing requires more DNA per library preparation than Illumina methods (several μgs as opposed to less than one μg), necessitating either more PCR cycles (increasing risk of error) or more substrate (which might be limiting in clinical samples). It is also consistently more expensive than Illumina sequencing (355) and has a slower turnaround. In controlled comparisons, Illumina based technologies typically produce fewer errors than 454, partially as a function of the greater depth afforded by Illumina[18] (360, 361, 368). The 454 propensity towards indels makes it additionally less suitable for repertoire sequencing. Such errors are less readily detected by `Decombinator` (which has mechanisms for correcting substitution errors located in V or J tag sites) and are more likely to turn a functional TCR sequence into a non-functional one (appearing as frameshift mutations), as opposed to just a different TCR sequence (which might then still translate to the same CDR3 sequence) (361, 368). This relative abundance of indel errors in raw TCR sequence data from the 454 platform (as well as the Ion Torrent platform, which operates in a similar pyrosequencing manner) has been noted previously, further suggesting Illumina as a more appropriate technology (126). Finally, even had the 454 quality profile been better, the choice to not pursue this technology in our sequencing program proved serendipitous, as in 2013 Roche revealed that it planned to stop selling and ultimately discontinue the technology, instead focusing on its later acquisition PacBio (429).

Sequence data produced using the MiSeq lacked the post-homopolymer decline in quality observed in the 454 results. It also possessed a number of features that are highly suited to our needs: it produces reads that are long enough for us to capture both V and J tags (permitting `Decombinator` assignation), doing so on runs that are both very quick to run and relatively cost-effective. However reads originating from the end of the amplicons that contain the RACE homopolymer – as in the 454 data – were of very poor quality, and therefore of little practical use. Efforts to rectify this

[18]It is worth noting that the quality scores of the two technologies are not directly comparable due to differences in how the respective softwares calculate error: each is designed to detect their major source of errors, which is indels for 454 and substitutions for Illumina (354, 355, 416).

problem through unidirectional sequencing from the constant region failed to produce the volume of data that the machine should be capable of. In the case of the MiSeq, this problem was not due to low diversity libraries, as it persisted even with the addition of PhiX spike-ins and random nucleotides downstream of the sequencing primer.

Despite its ability to amplify a polyclonal population of sequences, TCR amplification through homopolymer-RACE consistently produced libraries refractory to both 454 and Illumina sequencing, and thus is not a suitable amplification process for high throughput repertoire analysis. This finding has implications for those groups who may have employed such strategies in the past for lower throughput cDNA analyses and are considering the transition to NGS.

5.2.4 Ligation RACE

5.2.4.1 Protocol

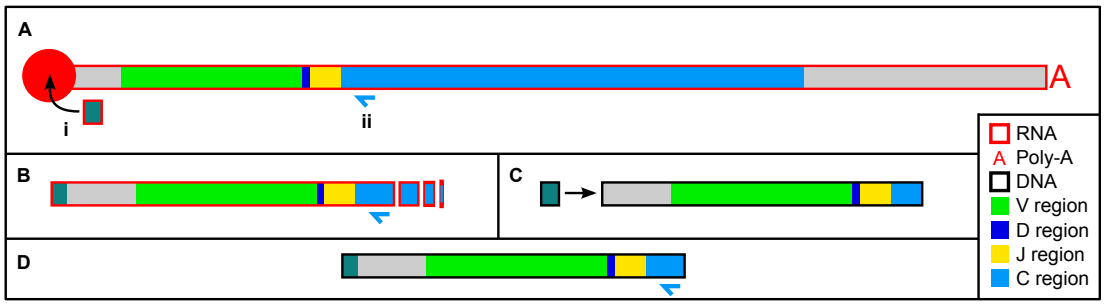


Figure 5.11: Ligation RACE strategies. Starting with TCR mRNA (**A**), we explored a number of ligation RACE strategies, including ligation of adapters to mRNA 5' caps directly (**Ai** and **B**), or reverse-transcription followed by adapter-ligation to cDNA 3's (**Aii** and **C**). Both produce equivalent adapter-labelled cDNA molecules (**D**).

In order to retain the ability of the poly-T PCR to amplify a wide range of V-bearing TCR rearrangements whilst improving on the efficiency of clustering in the bridge amplification, alternative 5' RACE strategies were tested. A number of 5' RACE strategies employ some form of oligonucleotide ligation to add a known sequence, offering an alternative to homopolymer tailing (430). The first such strategy we tried used RNA-ligation, wherein RNA oligonucleotides bearing the PCR priming site are ligated specifically onto fully processed mRNAs using an RNA ligase (431). Treatment of whole RNA from a sample of lysed cells with phosphatase removes the 5' phosphate group from any RNA molecules that lack a 7-methylguanosine (m7G) cap. Tobacco acid pyrophosphatase (TAP) is then used to 'decap' the fully-processed mRNAs, removing the m7G while leaving a 5' phosphate group, a requirement for ligation. Then

addition of an RNA oligonucleotide adapter that contains a 3' hydroxyl group (but itself lacking a 5' phosphate) in the presence of an RNA ligase results in the attachment of the adapter to the 5' end of previously-capped mRNAs. Concurrent with the testing of the RNA-ligation RACE, various members of the laboratory – Prof. Chain, Dr Oakes and myself – also trialled several DNA ligation techniques, the most successful of which involved using an RNA ligase to attach a DNA adapter to the 5' UTR of our cDNA molecules (432, 433). Both of the two major tested ligation-RACE strategies are illustrated in Figure 5.11. Amplicons then underwent a comparable PCR-based library preparation technique to the initial MiSeq attempts.

5.2.4.2 Sequencing results

Table C.7 shows the read statistics of the two major ligation protocols tested. Despite very high rates of TCR assignation using the RNA ligase protocol, quality remained relatively low. Moreover, the TCRs assigned displayed skewed V and J gene usage and pairing, particularly in the RL1 sample (Figure 5.12A and B). In contrast, the reads from the DNA ligation protocol are of better quality, with a median that remains above Q30 well past the 150 or so nucleotides typically required for `Decombinator` assignation (Figure 5.12C and D), with a V-J usage pattern far more in keeping both with our own and others' previous findings (c.f. Figures 5.8 and 5.2). We also note that the percentage of reads filtered out from the Decombined subset (i.e. those that contained ambiguous base-calls in their inter-tag region, or whose tags were too far apart to encode a genuine receptor) has dropped relative to that seen in earlier runs, again confirming a decline in the occurrence of errors.

5.2.4.3 Verdict

Initial trials with RNA and DNA ligation RACE approaches seemed to remedy the poor clustering and low quality reads that characterised the poly-T libraries. Despite giving slightly lower rates of TCR assignation by `Decombinator` than RNA ligation, DNA ligation appears to offer better quality data, with a gene usage profile that better fits those of previous methods and other laboratories' published data[19]. Therefore the DNA ligation RACE approach appeared to be the best candidate for further development.

[19]It should also be noted that concurrent with testing these ligation protocols, Illumina modified both the sequencing kits and base-calling software to the effect that the technology copes better with low-diversity libraries (434).

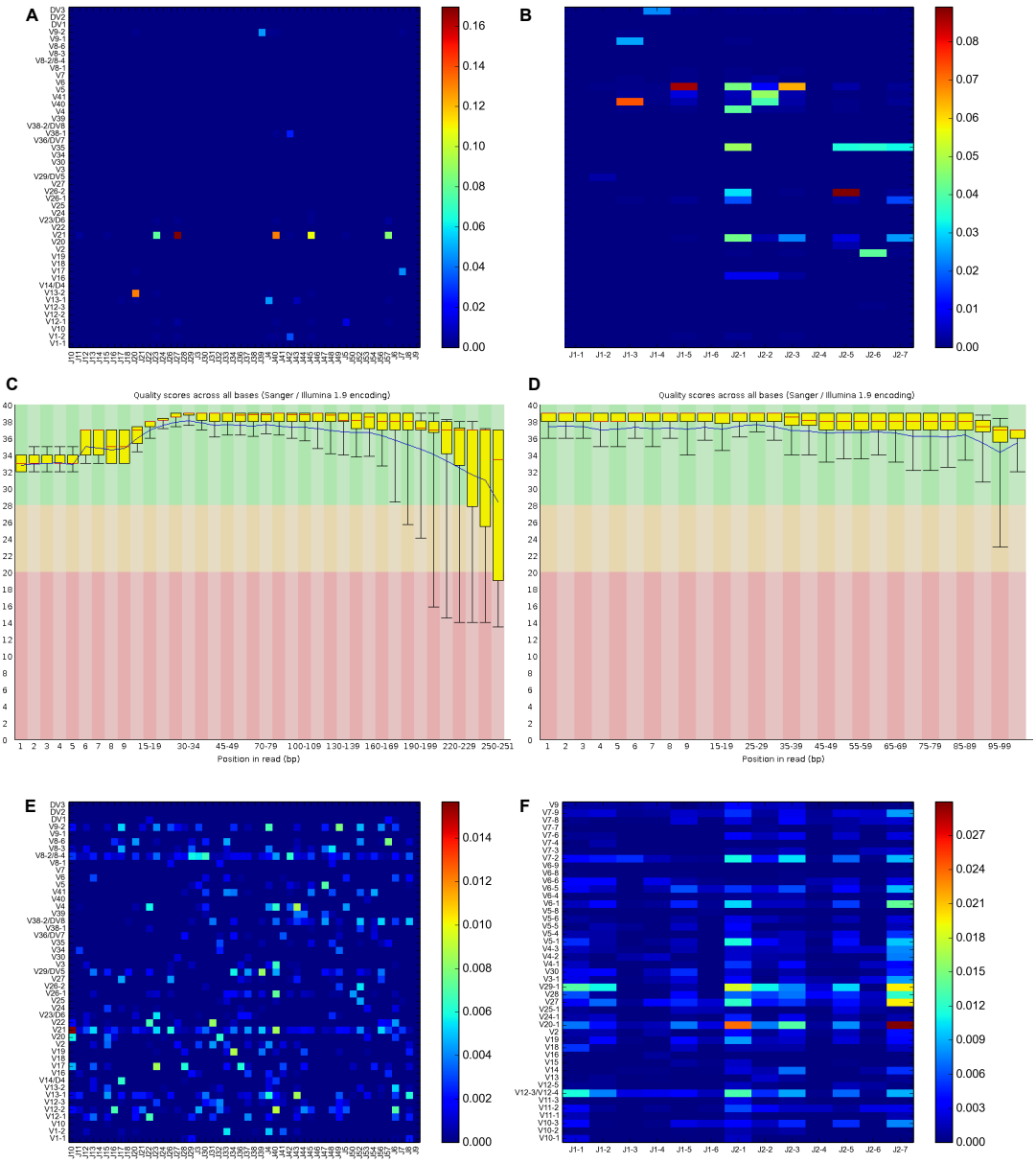


Figure 5.12: Performance of the basic RNA and DNA ligation RACE protocols. **A** and **B**: Proportional V-J gene usage in the TRA and TRB data for sample RL1, prepared using the RNA ligation strategy. **C** and **D**: The whole read and inter-tag quality scores for the α FASTQ files of sample DL2, prepared with the DNA ligation protocol. **E** and **F**: Proportional V-J gene usage in the α and β data for DL2, produced by `PlotDCR`. The RNA ligation strategy was performed by Prof. Chain, while the DNA ligation sequencing was performed by Prof. Chain and Dr. Theres Oakes.

5.2.5 Randomly-barcoded V amplification (BV)

5.2.5.1 Protocol

I next addressed the twin issues of sequence error and PCR duplication. No matter how effective an amplification procedure, all raw sequencing data will retain some errors (substitutions, insertions or deletions) that differ from the sequence of the original input molecules. While the nature of `Decombinator` analysis partially mitigates some errors (discussed later in Section 5.3), errors could still remain within the inter-tag region. PCR duplicates are an even harder problem to address: the heavy sampling intrinsic to deep-sequencing libraries followed by logarithmic amplification makes it very hard to know whether and how the frequency of reads belonging to a given TCR measured after sequencing reflect the true frequency of that TCR in the sample. Some groups have addressed this issue by using algorithms which cluster TCR sequences simply based on their similarity (76, 126). While this will reduce the number of artifactual TCRs, it could also combine genuine TCRs that just happen to differ by a small amount, thus removing actual diversity.

Beyond merely increasing diversity at the start of our reads, the introduction of random nucleotides to our libraries afforded an opportunity to identify and remove errors that accumulate during PCR and sequencing. The addition of random sequences prior to amplification can effectively label each initial input molecule, provided there are enough random sequences. After sequencing, these random sequences can be used as molecular barcodes: differences between TCRs sharing a barcode represent accumulated errors, while the number of different barcodes associated with a given TCR reflects the actual number of different TCR cDNA molecules that had daughter amplicons detected in the sequencing run. Such approaches have begun to be explored across fields that utilise HTS technology (435, 436, 437). The first paper to use the current generation of deep-sequencing platforms to sequence variable antigen receptor repertoires actually made use of such a technique (438), but the practice has not been widely adopted in subsequent protocols.

In order to test the feasibility of including such a technique in our protocols I again reverted to a simpler amplification, selecting a restricted number of V genes from which to prime, so that we could assess the results of 'barcoding' our libraries in a restricted dataset (see Figure 5.13). I selected the two V regions targeted in our initial simple V amplification protocol (TRAV8-4 and TRBV12-3), as well as TRAV21 and TRBV20-1, the genes which appear to often be used most frequently in the repertoires yielded from earlier experiments. Not only would this simplify the amplification procedure,

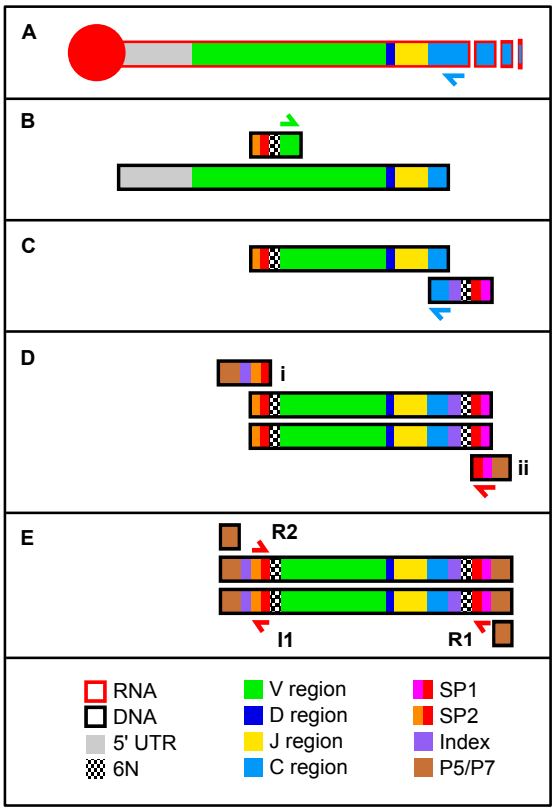


Figure 5.13: Schematic of the barcoded V (BV) amplification protocol. **A**: As in previous protocols, TCR RNA is reverse transcribed using oligonucleotides directed against the 5' of the α and β constant regions (blue arrow). Red circle represents mRNA 5' cap. **B**: Second strand synthesis is achieved using oligonucleotides complementary to one of four specific V genes. Sequences 5' to the V-primer sequence add a random hexamer (6N) and an Illumina sequencing primer (SP2). **C**: A 'third strand' single round extension reaction incorporates another random hexamer at the other end of the amplicons (as well as SP1, the other sequencing primer, and an index for demultiplexing), completing the 12-mer barcoding of TCR cDNA. **D**: A four-cycle PCR is used to add: another index (that which is sequenced in the dedicated Illumina indexing read) and P7 at one end (**i**) and P5 at the other (**ii**). P5 and P7 are the elements required for cluster generation, as they bind to the oligonucleotides that coat the flow-cell, permitting bridge amplification. **E**: A final PCR directed against the P5 and P7 elements (for 22 cycles) amplifies full-length amplicons to sufficient concentrations for sequencing. Amplicons are finally purified, quantified, sized and normalised before sequencing on the MiSeq. Three sequencing reads are obtained per library: the first (R1) from SP1, the second (I1) from the reverse complement of SP2 (producing the standard Illumina index read) and finally from SP2 itself (R3). R1 contains the V(D)J information, and therefore is Decombined. R2 contains information pertaining to the other end of the amplicons, which includes 5' UTR (untranslated region) information. The concatenated first six bases of R1 and R2 form the random dodecamer barcode which is used to error-correct the sequence. R1 and I1 both contain index sequences, which allow multiplexing of multiple samples per run.

but would allow us deeper insights into the clonal hierarchy within a small number of V families.

Briefly, following reverse transcription a second strand extension reaction was performed, which selectively adds Illumina sequencing primer 2 (SP2) followed by a random hexamer to the targeted V regions. After removing unbound primers, a novel 'third strand' reaction is performed, wherein a single cycle extension is performed using an oligonucleotide bearing Illumina sequencing primer 1 (SP1) followed by a random hexamer and the J-region proximal constant region sequence, followed by removal of unbound primers. In this manner each molecule of cDNA (bearing the correct V regions) is labelled with a dodecamer of random nucleotides. There are approximately 16.7 million different combinations of random nucleotides in a 12-mer ($4^6 \times 4^6$). Therefore when analysing RNA from small populations of T-cells[20] the probability of any two TCRs being labelled with the same barcode should be very small. Following 'barcoding' a short PCR introduces the final required sequencing elements and amplifies.

5.2.5.2 Sequencing results

This protocol produced a large number of reads, with good quality scores and high rates of `Decombinator` assignation (Table C.8); an average 61% of reads from each FASTQ file contained a rearranged TCR. The quality remains high well into the read (Figure 5.14A) and throughout the inter-tag region (Figure 5.14B). As with the original simple V amplification experiment that we ran on the HiSeq (Figure 5.4), I specifically observed the targeted V regions in the output data, albeit with trace amounts of other, non-primed Vs (Figure 5.14C and D).

5.2.5.3 Verdict

Only a small number of V genes were amplified and analysed in this manner so as to simplify the protocol, both to allow testing of the barcoding and gain a deeper insight into clonal structure within specific V genes. By efficiency of `Decombinator` assignation alone, barcoded single V amplification – combined with PCR library preparation and MiSeq sequencing – could be regarded as the most efficient protocol discussed yet, producing a large number of good quality, identifiably TCR reads. The major improvement however appears to be the error- and frequency-correction capabilities provided by the unique incorporation of random barcodes, which is discussed below in Section 5.3.

[20]The few ml of blood from which RNA is extracted in these experiments likely contains fewer than five million T-cells, and only a fraction of the total RNA is used for sequencing.

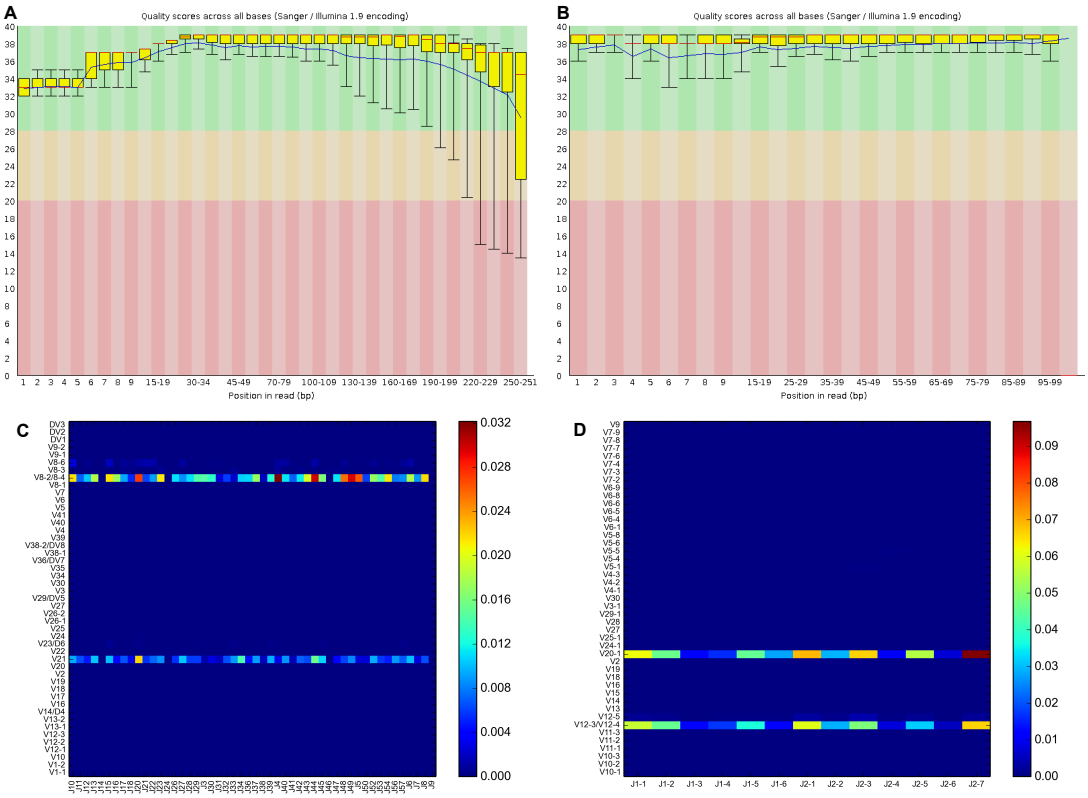


Figure 5.14: Overview of barcoded simple V amplication protocol results. A and **B**: Representative `FastQC` quality score barplots for whole and inter-tag sequence FASTQs for a representative sample (HV4), sequenced on the Illumina MiSeq. **C** and **D**: V-J proportional usage heatmaps for sample HV4. V genes targeted in the PCR were TRAV8-4, TRAV21, TRBV12-3 and TRBV20-1. The effect of adding and 'collapsing' our sequences based on these molecular barcodes is discussed later in Section 5.3.

5.2.6 Barcoded Ligation RACE (BLR)

5.2.6.1 Protocol

The final step in producing a TCR sequencing pipeline that fulfilled the initial goals was to add the error-correcting functionality of barcoding to the pan-TCR amplifying DNA-ligation RACE protocol, combining the best aspects of the more successful of the previous protocols tested.

This new barcoded-ligation RACE (BLR) strategy consists of reverse transcribing mRNA, ligating on a barcoded-SP2 oligonucleotide, performing second and third strand syntheses (adding SP1 and a second random hexamer) and then proceeding with the second and final PCR steps as in the randomly-barcoded specific-V protocol. This protocol is shown in detail in Figure 5.15.

5.2.6.2 Sequencing results

The BLR amplification procedure proved highly effective, producing a large number of reads per sample, a large percentage of which contained TCRs (Table C.9). Furthermore, these reads are of very good quality, with a sequencing error predicated to occur on average only once per 3000 base calls, within the inter-tag region.

The quality profiles for representative files again show high quality data for the entirety of the portion of the reads that contain the V(D)J sequence (Figure 5.16A and B). Plotting the VJ gene usage reveals polyclonal repertoire usage (Figure 5.16C and D), comparable to that of samples produced by other laboratories (Figure 5.2). Moreover, these libraries are uniquely barcoded, permitting further error-correction and frequency estimation by 'collapsing' (see Section 5.3), which will further improve upon the quality and accuracy of the data.

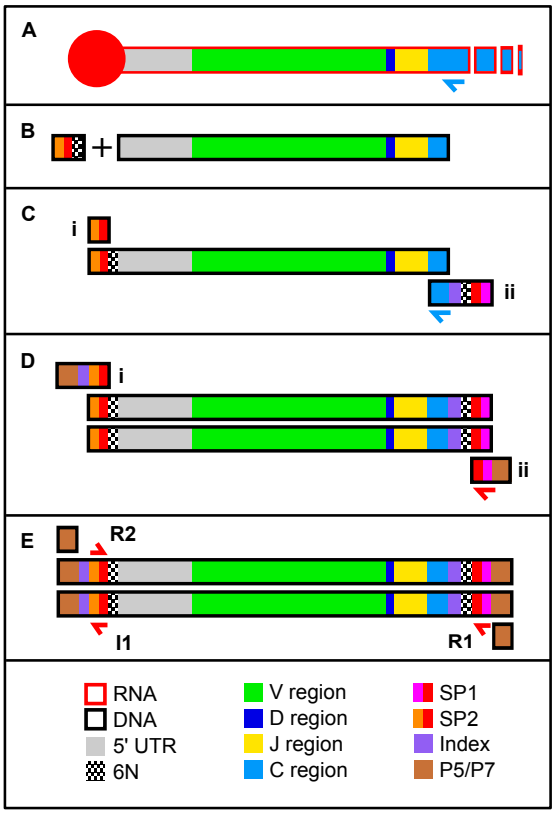


Figure 5.15: Schematic of the barcoded ligation RACE (BLR) protocol. A: As in previous protocols, TCR RNA is reverse transcribed using oligonucleotides directed against the 5' of the α and β constant regions (blue arrow). **B**: RACE is achieved through ligation of DNA adapter to the 3' end of the cDNA; adapter consists of Illumina sequencing primer SP2 and a hexamer of random nucleotides (6N). **C**: Separate single round second (**i**) and third strand (**ii**) reactions allow incorporation of another random hexamer at the other end of the amplicons (as well as SP1, the other sequencing primer, and an index for demultiplexing), completing the random 12-mer barcoding of TCR cDNA. **D**: A four-cycle PCR is used to add: another index (that which is sequenced in the dedicated Illumina indexing read) and P7 at one end (**i**) and P5 at the other (**ii**). **E**: A final PCR directed against the P5 and P7 elements (for 23 cycles) amplifies full-length amplicons to sufficient concentrations for sequencing. Amplicons are finally purified, quantified, sized and normalised before sequencing on the MiSeq. Three sequencing reads are obtained per library: the first (R1) from SP1, the second (I1) from the reverse complement of SP2 (producing the standard Illumina index read) and finally from SP2 itself (R3). R1 contains the V(D)J information, and therefore is Decombined. R2 contains information pertaining to the other end of the amplicons, which includes 5' UTR (untranslated region) information. The concatenated first six bases of R1 and R2 form the random dodecamer barcode which is used to error-correct the sequence. R1 and I1 both contain index sequences, which allow multiplexing of multiple samples per run. This is the protocol used to amplify and sequence the majority of the samples processed in this thesis.

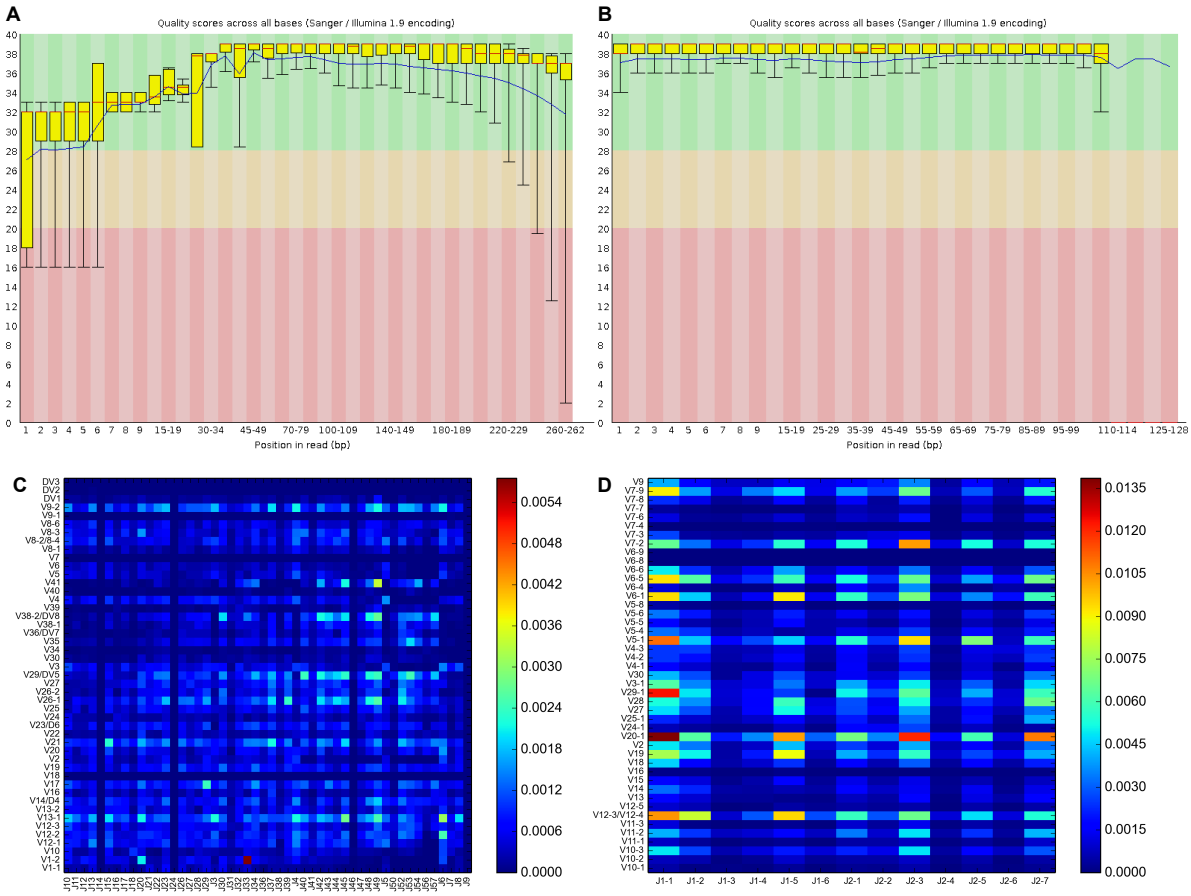


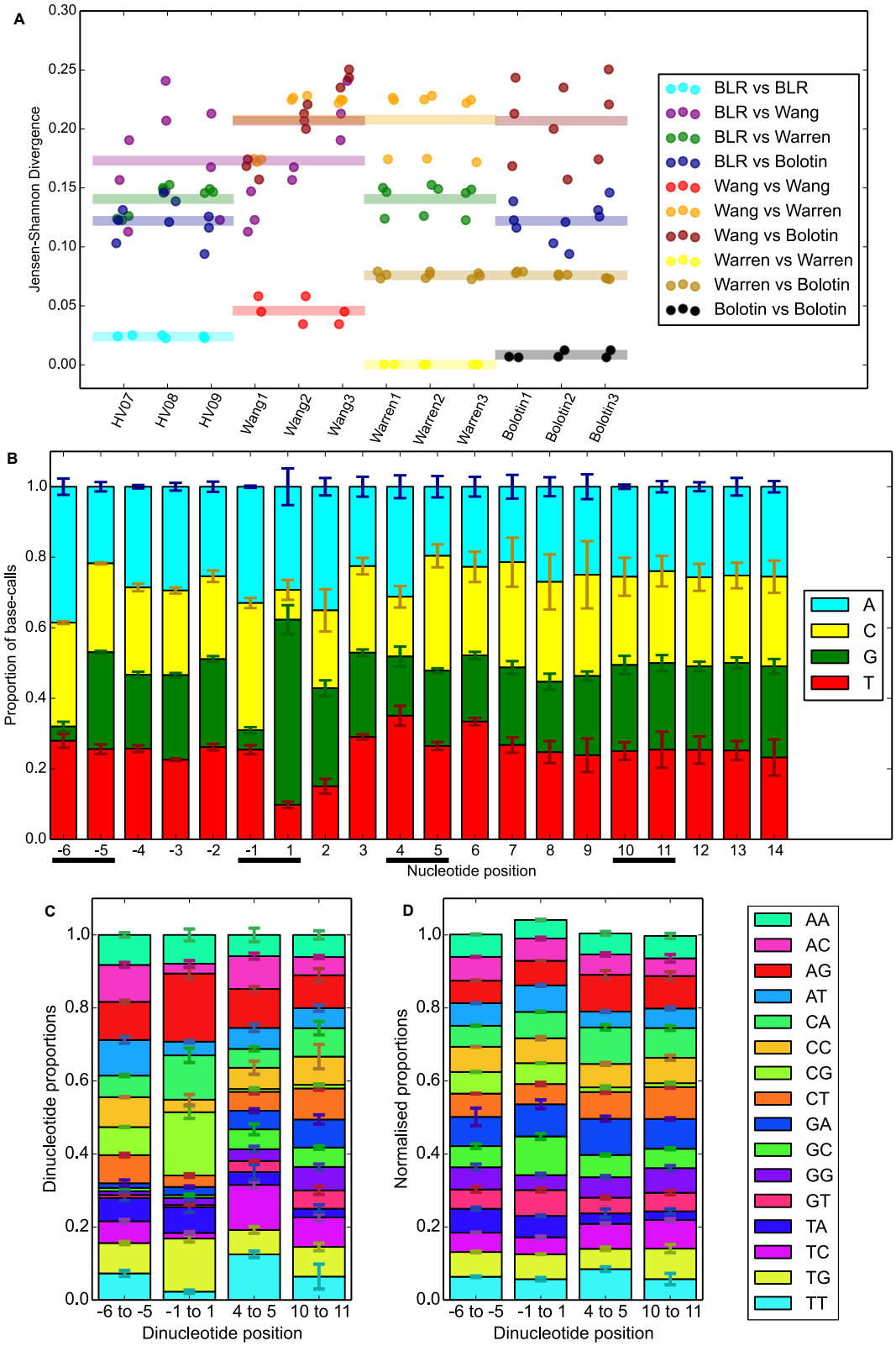
Figure 5.16: Quality score boxplots and VJ gene usage heatmaps for a representative whole blood healthy donor sample (HV07), produced using the BLR strategy. A: Quality score distribution for the raw, complete FASTQ file. **B**: Quality score distribution for the inter-tag region (i.e. from the start of the V tag to the end of the J tag). **C** and **D**: VJ proportional frequencies for α and β chain rearrangements.

5.2.6.2.1 Assess bias

While this protocol should be exempt from primer bias – as all rearrangements are amplified using the same primers – it is possible that bias gets introduced during the ligation step, perhaps dependent on terminal nucleotide or V gene identity. In order to address whether this was the case I first compared V gene usage profiles between three healthy donor repertoires produced using the BLR strategy and those of previous publications (summarised earlier in Table C.1 and Figure 5.2). The proportional usage of each TRBV gene was calculated for each repertoire sample, and these distributions were compared against all other samples both within and between different protocols, using the Jensen-Shannon measure of divergence (JSD, Figure 5.17A). Intra-protocol comparisons (i.e. comparing the results of different samples produced with the same method) produce low and consistent levels of divergence. Of the intra-protocol comparisons, the Warren and Bolotin samples are the most internally conserved (yellow and black), the Wang samples are the most divergent (red), while the BLR samples are intermediate (cyan). We would expect the BLR samples to be more divergent as each sample came from a different donor, while all of the data from the previously published papers represent repeats from the same bleed of one donor; the Wang data therefore appears to be the most inconsistent with regards to V gene usage. These findings are reflected in the inter-protocol comparisons: the Warren and Bolotin samples are the least divergent from one another (gold), either of these compared against the BLR data are slightly more divergent (green and navy), and then any comparison against the Wang data is yet still more divergent (purple, orange and brown). Therefore we can deduce that in terms of usage profiles of the different TRBV genes, my data appears to be within the divergence range of existing published data sets.

In order to look specifically at whether the DNA ligation process engineered a specific bias, I made use of the data produced from the second of the paired-end Illumina reads[21] (see Figure 5.15E). As the ligation oligonucleotide consists of the binding site for the R2 sequencing primer (SP2) followed by six random nucleotides, the first six bases of the actual reads will correspond to the ligation-introduced barcode, while the seventh base onwards will represent the 5' end of that TCR transcript. I measured the proportions with which each base is used at each position across the start of R2 (Figure 5.17B). I observed a noted bias in the prevalence of different bases in the random hexamer, particularly in the first and last positions (-6 and -1). There seems to be very little usage of guanosine, which corresponds to cytosine being used infrequently

[21] The 'R2' (read 2) data.


A
Jensen-Shannon Divergence
0.30
0.25
0.20
0.15
0.10
0.05
0.00
HV07
HV08
HV09
Wang1
Wang2
Wang3
Warren1
Warren2
Warren3
Bolotin1
Bolotin2
Bolotin3
BLR vs BLR
BLR vs Wang
BLR vs Warren
BLR vs Bolotin
Wang vs Wang
Wang vs Warren
Wang vs Bolotin
Warren vs Warren
Warren vs Bolotin
Bolotin vs Bolotin
B
Proportion of base-calls
1.0
0.8
0.6
0.4
0.2
0.0
-6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Nucleotide position
A
C
G
T
C
Dinucleotide proportions
1.0
0.8
0.6
0.4
0.2
0.0
-6 to -5
-1 to 1
4 to 5
10 to 11
Dinucleotide position
D
Normalised proportions
1.0
0.8
0.6
0.4
0.2
0.0
-6 to -5
-1 to 1
4 to 5
10 to 11
Dinucleotide position
AA
AC
AG
AT
CA
CC
CG
CT
GA
GC
GG
GT
TA
TC
TG
TT

in the adapter molecules at these positions. Similarly there exist large preferences in the bases at the start of the actual transcript read, particularly in the first position – which contains a guanosine in over half of the reads – which normalises around position seven or eight. By measuring the proportion with which different dinucleotide pairs occurred across the ligation junction, I could assess which bases were actually being ligated together. Comparing these pairings to dinucleotides at other sites either at the start of the random hexamer or within the early transcript read revealed a markedly different pattern of dinucleotide usage across the different positions, with a particular preference at the junction site for dinucleotides ending with a G residue (Figure 5.17C). This was likely explained by the exceeding abundance of Gs at the +1 position observed in Figure 5.17B. To control for this, the frequency of dinucleotide pairs was normalised by dividing by the frequencies of the individual bases at each position, giving a truer representation of how likely a base is to pair with another. Figure 5.17D reveals that at the start of the random barcode (-6 to -5), there is practically no preference in base pairing, and thus the prominent inequality of bases at the -6 position observed in Figure 5.17B is likely some artefact of the oligonucleotide generation. At the ligation junction (-1 to +1) there is a slight loss of uniformity in the base pairing observed for dinucleotides that begin with a guanosine, although as Gs make up less than 10% of bases at the last position of the adapter this finding could reflect a sampling artefact. The two positions from later in the read show a similar pattern of dinucleotide usage

Figure 5.17 *(preceding page)*: Assessment of bias in the BLR protocol. A: Inference of potential bias by comparing differences in V gene usage distributions, as measured by Jensen-Shannon Divergence (JSD). Samples used were three healthy volunteer (HV) samples produced using the BLR protocol, as well as three samples each from three previously published data sets (see Section C.1). All possible pairwise sample combinations were compared and JSD calculated (with a greater JSD meaning a greater difference in V gene usage). Horizontal lines indicate group means. **B**: Proportional usage of each nucleotide position in the read 2 (R2) data (see Figure 5.15E), for those paired-reads that successfully assigned a TCR in their R1 reads (for the same HV samples used in **A**). Note that the first six base-calls of every read correspond to the random hexamer barcode sequence incorporated at that end of the molecule (-6 to -1), which would then be followed by the 5' UTR sequence of the TCR (1+). **C**: Raw proportions of all the different possible dinucleotides, at the four sites underlined with black bars in **B**, corresponding to the start of the random hexamer (-6 to -5), the interface between the ligation oligo and the cDNA molecule (-1 to +1) and two sites within the 5' of the transcript (4 to 5 and 10 to 11). **D**: Normalised proportions of the dinucleotide proportions displayed in **C**, where the value of each dinucleotide is divided by the proportion of each of the individual nucleotides at their respective positions (and divided again by 16, the number of dinucleotide combinations, in order to provide an equivalent scale to **C**). All error bars indicate standard deviation.

to each other that is different from the other two positions, likely reflecting the innate biases and motifs of the human transcriptome. For example we observe a very low frequency of CG dinucleotides, which is a well-documented feature of animal genomes due to the propensity of methylated thymines to deaminate to thymidine residues (439). Therefore at the site of ligation we do not appear to detect a particular bias that should preclude or influence our protocol from detecting any particular TCR transcript based on its 5' primary sequence.

5.2.6.3 Verdict

The barcoded ligation RACE strategy fulfils the required criteria: it produces high quality data (in the expected ranges of the machine), from which one can obtain a large number of polyclonal TCRs. Furthermore, as each cDNA molecule gets uniquely labelled with random barcodes one is able to pass output `Decombinator` data through an error-correction process, further improving the qualitative and quantitative aspects of the repertoire data.

I have demonstrated that the ligation occurs without an obvious bias. The observed imbalance in the non-templated nucleotides in the ligation adapter does not necessarily reflect on the likelihood of bias in our system, as the production of 'random' sequences in synthetic oligonucleotides is not truly random (440), a feature that is often observed in transcriptome sequencing efforts that use random primers for reverse-transcription (441). Similarly the differential base preferences we observe here at the start of TCR transcripts, particularly in the first position – where there is an extreme preference for guanosine – are not themselves indicative of bias. Despite a great deal of heterogeneity, there are a number of biased motifs that appear in or around transcriptional start sites (TSSs), which often result in an increase production of G residues in the first bases transcribed (442, 443, 444). There is also an additional methodological explanation for the pronounced G bias in the first base of our transcripts, as a result of using a Moloney Murine Leukemia Virus (MMLV) derived reverse transcriptase (RT) enzyme. MMLV derived RTs possess terminal deoxynucleotidyl transferase activity, which results in the addition of several non-templated bases after reverse transcribing an mRNA; however this activity has a strong bias for the inclusion of cytosine residues[22] (401, 402). Even in reaction conditions such as these which aren't optimised for such extension there is often a single or double C residue added (401). As the sequences measured in Figure

[22]Incidentally this enzyme activity provides the basis for template-switching amplification protocols used in other TCR repertoire RACE studies, as a poly-G containing oligonucleotide can anneal to this tail and prime the second strand reaction.

5.17 are in the complementary strand with respect to the cDNA, any C residues that the RT incorporates to the cDNA would contribute to the G bias we see in the +1 and +2 positions. When I normalise for the frequency of each nucleotide either at the end of the ligation adapter or the start of the transcript we observe that there is very little bias in nucleotide pairing (and no pairing that does not occur at all). This is supported by a report which detailed that T4 RNA ligase does not show preferential bias when considering the primary structure of cDNA adapters 3'-ligated to RNA molecules (a process which was significantly improved through the addition of random nucleotides at the end of the adapter) (445). Therefore all and any TCR cDNA species could theoretically be labelled without preference using this protocol.

Due to the success of this protocol, I applied it to the major sample sets of this thesis, those investigating the global TCR repertoires of healthy individuals and HIV patients, as described in Chapters 3 and 4. However, the protocol has subsequently been further refined within the group. Improvements include combining certain steps to reduce the number of bead purifications that are required, which should reduce the number of low-frequency TCR cDNAs lost throughout the protocol. We are additionally exploring the use of quantitative, real-time PCRs in the amplification protocol, to allow us to halt the reaction after an optimal number of samples. In this way we can reduce the number of cycles to the absolute minimum required to produce sufficient concentrations for sequencing: the fewer cycles of PCR, the less chance there is for any errors to occur.

5.3 Bioinformatic development

"Screening peripheral blood T cells for the expression of up to 20 different Vβ families ... will provide an extensive characterisation of TCR Vβ expression. However, this analysis would also require large numbers of cells ... and generate massive amounts of data, which could quickly make the analysis impossible to perform."

Jeff Faint *et al*

From 'Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression', Journal of Immunological Methods, 1999 (56).

5.3.1 TCR repertoire analysis with `Decombinator`

At the outset of this project, there existed very few published descriptions of TCR analyses, and even fewer publicly available programs. Where described, most methodologies primarily made use of pairwise alignment algorithms to detect the presence of

particular V, J (and sometime D) regions, such as: DNAPlot (as used in IMGT/HIGHV-QUEST) (446, 447), BLAT (198, 448), BLAST, (97, 449) and SWA (83, 271, 450). Whilst effective, these methodologies suffer from a relative lack of speed when compared to certain other techniques which might be applied, limiting the rate of analysis of large data sets. To overcome this hurdle, then-PhD student Dr. Niclas Thomas in our group developed a novel TCR analysis suite named `Decombinator` (111), which has been used for all of the TCR rearrangement analyses in this thesis. This makes use of an extremely efficient string matching algorithm described by Aho and Corasick (451), which is widely used across both bioinformatics and numerous other computing-related fields (452).

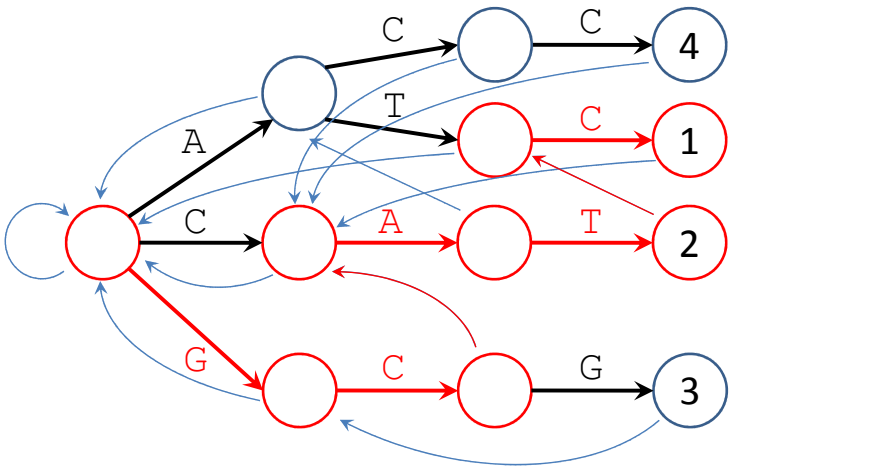


Figure 5.18: Schematic detailing an Aho-Corasick string matching automaton. The trie is generated prior to searching, based on the four keyword patterns (or tags) to search for. The trie is then applied to the search string (or read). The traversal of the trie is illustrated in red. Starting at the leftmost node (the root), the branch to navigate to is decided based on matching the next character in the string. If the criteria to continue along the branch is not met, the automaton instead navigates along the failure functions (leftwards arrows), using the longest suffix so far (e.g. the GC on the keyword 3) to find the longest prefix match (the C at the start of pattern 2). In this example, one pass of the search string will yield matches for patterns 2 and 1, in that order.

In our implementation, `Decombinator` uses the Aho-Corasick (AC) finite state automaton to rapidly search a given sequencing read for the presence of one of a list of 'tag' sequences, the presence of which should uniquely identify a particular V or J gene. Finding both a V and J tag allows full description of a TCR rearrangment using a five-part classifier consisting of: which V and J genes were used; how many nucleotides were exonucleolytically removed from each of these genes, and what the intervening 'insert' DNA sequence is (from the end of the V to the beginning of the J). From these five characteristics the entire DNA sequence of the rearranged chain can be recapitulated. Note that we do not explicitly assign a D region (for β or δ chains), as these are short sequences that are often typically too deleted to unambiguously assign (75); any remaining D region sequences will be contained within the insert sequence. Compared to raw sequencing data, the classifiers produced by `Decombinator` are easily stored, transferred and manipulated in downstream processes.

The speed of AC string matching lies in how it finds sequences: before traversing a string, it uses the panel of keyword patterns (i.e. the V or J tags) to generate a trie (Figure 5.18). When passing through the string to be searched (the sequencing read) the finite state automaton navigates the tree node by node, depending on which nucleotide is present at the next position. The AC algorithm simultaneously looks for each subsequence in one pass of the read, and in the event of a mismatch uses the longest suffix matched at any one time to skip to the longest possible prefix match in another branch of the trie. These features combine to produce the incredible rate of V/J gene assignment we see with `Decombinator`, orders of magnitude above that seen using pairwise alignment (111). As a direct pattern-matching algorithm, an error within the tag sequence itself would ordinarily preclude assignation. To overcome this limitation, if complete V or J tags are not found, `Decombinator` splits each tag into two half-tag sequences and searches for these. Successfully finding a half-tag is followed by checking the neighbouring sequence where the rest of the full tag would ordinarily lie against the relevant partner half-tag: if the comparison of these two sequences produces a Hamming distance of one (i.e. a one nucleotide mismatch) the V or J can be assigned.

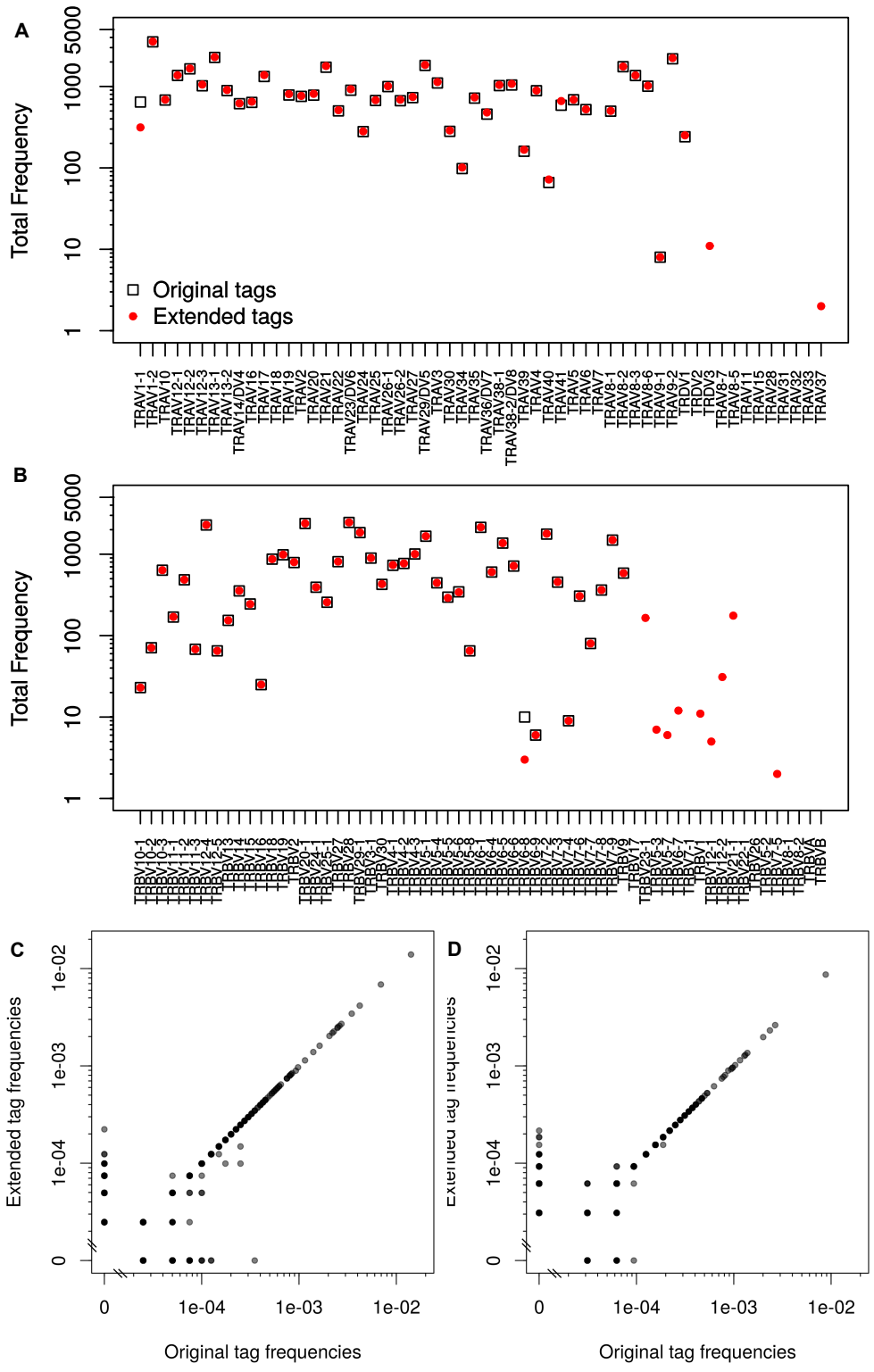
5.3.1.1 Generation of new `Decombinator` tag sequences

In the course of this project I made several contributions to `Decombinator`, including the generation of human V and J α chain tag sequences. α tags were designed in a similar method to that used by Dr. Thomas for generation of the β tags (and later γ/δ tags): complete sequences for the prototypic allele of all functional TRAV and

TRAJ genes were downloaded from IMGT/GENE-DB (19), and then potential 20-mer windows located 20 nucleotides away from the recombining edge (to allow room for deletion) were assessed for uniqueness. If potential tags for multiple genes were identical, the tag window was shifted and uniqueness re-assessed. Developing tags for the α chain is simpler than for β in some respects, as the V genes share less identity, demonstrable as fewer TRAV genes fall into subgroups[23]: there are only seven subgroups containing 20 members in the TRA locus compared to 16 subgroups totalling 60 members in TRB (19, 453). Therefore in the tag set originally published with `Decombinator`, only two TRAV genes – TRAV8-2 and TRAV8-4 – are included under one tag. I also produced tags for the three δ chain V genes that reportedly do not recombine with α J genes, for reasons discussed in Section 3.1.2.

Tags were originally designed to cover only the prototypical alleles (i.e. the first one discovered) for all TCRs with 'functional' gene status in IMGT, thereby ignoring those that were not thought to contribute. However, when investigating sequences that failed to be assigned by `Decombinator`, I observed that a proportion contained rearranged TCRs that make use of these other genes (as described above in Section 3.1.2). I therefore designed a new set of tags that covered the prototypical alleles of every α and β V or J gene for which IMGT has an entry, regardless of functionality. As shown earlier in Figure 3.2, a number of these are indeed used in multiple donors, and are thus worthy of inclusion for recombination level analyses at least. In the design of these extra tags, I also strove to maintain at least a minimum distance of 20 nucleotides between the start of the tag and the end of the gene (where possible) to minimise the possibility of tags being altered during recombination. By being able to move further into the V region (thanks to longer read lengths from the MiSeq) I was also able to design tags with increased resolution between related genes, lowering the chance that errors could result in misassignment. In order to check whether this new extended tag set was still able to recognise the same original clones, I ran `Decombinator` on sample healthy individuals and compared the frequency at which I found each gene. As Figures 5.19A and B and C.4 show, the extended tags are typically able to find either equivalent or greater numbers of sequences. There are some exceptions, such as TRAV1-1, the loss of which was found to be due to intra-V gene splicing to a cryptic acceptor site (see Section 3.3). Comparison of the frequency of individual clones in each tag set is remarkably conserved (Figure 5.19C and D). The new tag set therefore largely recapitulates the findings of the original tags, save with increased resolution in

[23]Subgroups are classified as families of germ-line genes which posses greater than 75% nucleotide identity.


A
Total Frequency
1
10
100
1000
5000
Original tags
Extended tags
B
Total Frequency
C
Extended tag frequencies
0
1e-04
1e-03
1e-02
Original tag frequencies
D
Original tag frequencies

a small number of genes while permitting detection of genes which may be involved in recombinations, regardless of their official IMGT status or expression at the cell surface.

5.3.1.2 Limitations of standard `Decombinator` analysis

As with any analysis tool, `Decombinator` suffers from a number of small limitations which must be taken into consideration when reviewing data produced using it. The foremost is one of accuracy. Whilst Dr. Thomas demonstrated a 98.8% accuracy rate when using *in silico* sequences allowed a 0.1% error rate (111) – a value on the order of that which we might expect in an Illumina run (361) – errors are not equally distributed among reads (405). Therefore it is probable that certain reads accrue enough errors in the tag regions so as to be assigned incorrect V or J genes. Alternatively erroneous production of a tag sequence at an inappropriate location in the read might result in a sequence that contains a genuine, error-free recombination elsewhere in the read being overlooked. Moreover, an error in the insert sequence cannot be corrected by any comparison to germline, as this region contains non-germline-encoded information. A 2013 publication by Bolotin *et al* showed that in their hands `Decombinator` produced a number of erroneous additional CDR3 sequences upon translation, despite scoring very accurately on V and J assignation, implying that errors within the insert sequence are artificially inflating diversity with false sequences (454), although this analysis made use of an early version that lacks more recently added error-filters.

Tag-based gene assignation results in specific limitations based on the design of the tags. With the inclusion of the extended tag set described in this thesis, `Decombinator` is now able to detect rearrangements using a larger panel of genes, regardless of functionality. However many TCR genes have multiple known alleles. If single nucleotide polymorphisms (SNPs) occur within tag regions certain alleles may be less likely to be detected. Conversely, alleles might differ in their ability to recognise antigen (recorded for at least one TRAV and one TRBV gene (455, 456)) but remain conserved in their tag region, which might affect correct data interpretation. Tag placement presents another problem: despite being typically at least 20 nucleotides away from the recombining

Figure 5.19 *(preceding page)*: Comparison of original and extended `Decombinator` tags. Plots of a representative healthy donor (HV09) showing the comparative performance of `Decombinator` in combination with the original and extended tag sets. **A** and **B**: Frequency of `Decombinator` rearrangements using particular TRAV and TRBV genes, when using either the original or new extended tag sets. **C** and **D**: Correlation of recombination frequency for α and β chain rearrangements.

edge of the germline gene, it is possible that a rearrangement could still have a greater number of deletions and remain productive, yet be undetectable to `Decombinator` currently. A related issue is that some V genes are simply too homologous in their 3' sequence, and thus one is unable to assign tag regions that are able to distinguish them.

These problems are largely soluble through the production of different panels of tags. For instance, if one wished to distinguish between different allelic forms of V and J genes, tags could be generated for each allele. Similarly, recognition of hyper-deleted genes could be achieved through selection of tags further away from the recombining edges. Both of these approaches would however require longer reads of higher quality than are routinely achievable. Other problems remain, such as how to deal with potential sequence errors remaining in the insert sequences. I have addressed this in this work through the unique labelling of transcript molecules, as described in Section 5.3.2.

5.3.2 'Collapsing' TCRs: error-correction and improved frequency estimation

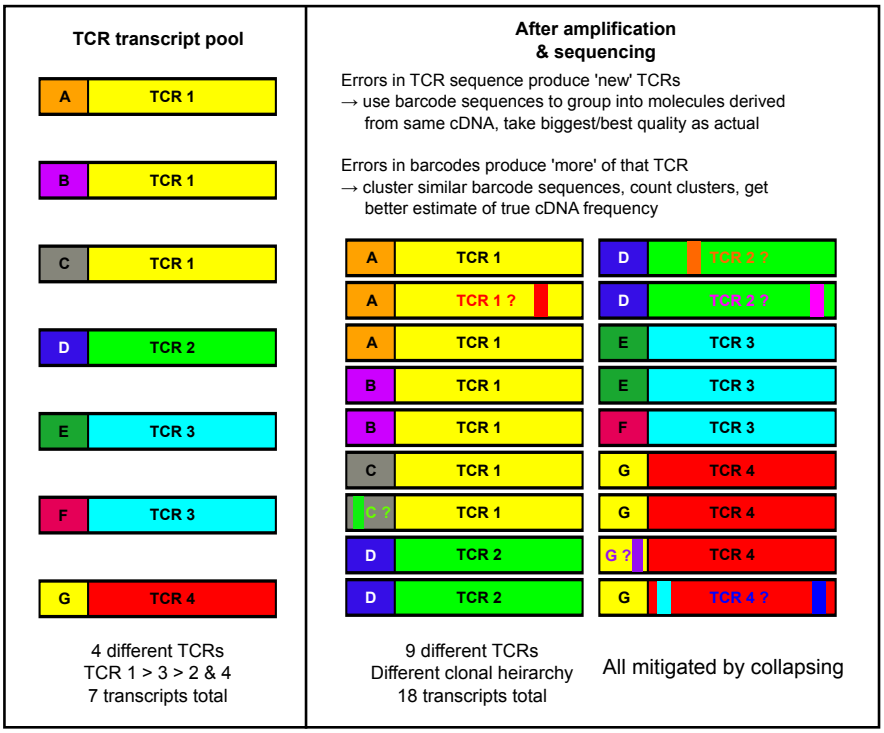


Figure 5.20: Schematic detailing the need for error-correction. Following processing, a labelled transcript pool (left) accrues PCR and sequencing errors (right), and differential amplification or sampling may result in changes to the frequency hierarchy.

Both PCR and deep-sequencing (which invariably requires additional PCR steps) produce sequence errors that may artificially inflate the diversity and disrupt the heirar-

chy of a repertoire sample (Figure 5.20). As discussed in Sections 5.2.5 and 5.2.6 above, the addition of random molecular barcodes to cDNA molecules prior to PCR amplification provides us the ability to mitigate for PCR and sequencing errors and duplications. In order to take advantage of this feature, I developed an alternative TCR analysis pipeline. The first step involves concatenating the two random hexamer sequences from the start of both sequencing reads (see Figure 5.15E) to form the 12-mer barcode. Next, a modified version of `Decombinator` analyses each read, outputting the sequence and the quality information of this barcode as well as the inter-tag region, in addition to the standard five fields. The product of this 'verbose' `Decombinator` (`vDCR`) can then be fed into the final script in the pipeline that performs the error-correction and frequency-estimation. 'Collapsing' in this context refers to the reduction both in frequency and number of clones, by both mitigating PCR duplication and removing clones produced by PCR or sequencing errors. This is achieved in two steps. First, all assigned TCRs that share the same barcode are examined in turn; the most frequent nucleotide sequence is presumed to be that which most likely represents the genuine transcript. Any reads that share the same barcode sequence yet do not match this prototypical sequence are thus likely to be errors and can be ignored. The second step is to count the number of different barcodes that associate with a given rearrangement – after performing some clustering to account for errors in the barcodes themselves – thus providing a far more accurate estimation of the true frequency of that transcript than the read count. A flowchart detailing these steps is shown in Appendix C, in Figure C.3.

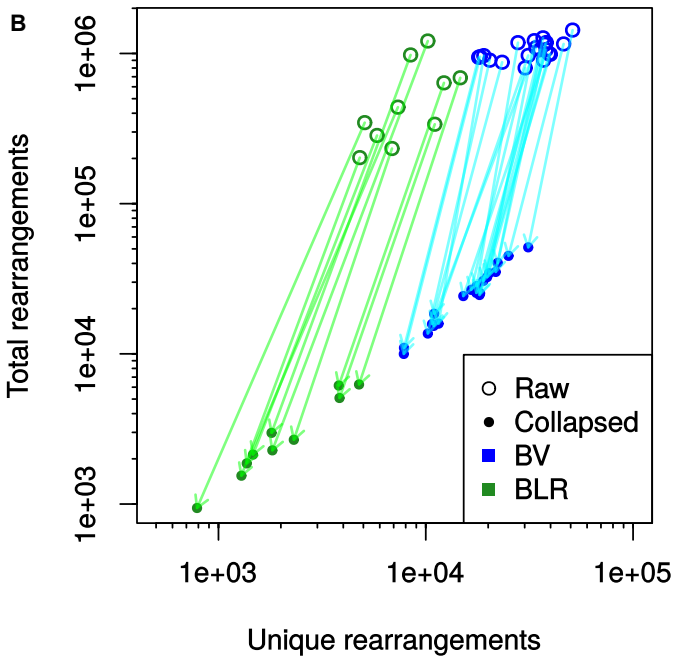
With the addition of the error-correction software – named `CollapseTCR` – the entire bioinformatic pipeline as used in this thesis is summarised in Figure 5.21A. Collapsing repertoire data leads to approximately twofold and fourfold reductions in the the number of unique TCR sequences for the BLR and BV protocols respectively due to removal of erroneous TCRs, and a 40-fold and 170-fold decrease in total number of rearrangements, due to mitigated PCR duplication (Figure 5.21B). An example of the effect this might have for a given rearrangement is illustrated in Figure 5.21C, where I have shown all of the unique `Decombinator` assignations that were associated with a particular random barcode sequence before collapsing. Due to the the large number of possible barcodes (which is far in excess of the number of TCR cDNA molecules we would expect in samples of this size) the expectation would be that all of these classifiers therefore represent the PCR progeny of one original founder molecule. In the pattern typically displayed in all samples, there is one rearrangement that is far more frequent than all others, with a number of lower frequency but highly similar

A

i	`@HWI-M01520:79:000000000-A7TUF:1:1101:15622:1335`
	`CCCTCCTACCTCCTTGTAGGAACTGGTGGGAACACGTTTTTCAGGTCCTCTGTGACCGT...`
	`+`
	`1>>AA13>>>AFAFFFFFCCCBBFGGGGGCGGGGHFGHHGGHHEFHHHHHHHHGFGHCG...`
		Amplified libraries sequenced on Illumina MiSeq, FASTQ output

ii	`6, 12, 1, 5, GTGGGGGGGCCG, [ID], [TCR-seq], [TCR-qual], [barcode-seq], [barcode-qual]`	TCRs assigned using verbose DCR, which additionally ouputs TCR and barcode (6N+6N) sequence information
iii	`6, 12, 1, 5, GTGGGGGGGCCG, 560`	Assigned TCRs error-corrected/frequency-estimated based on their barcodes, using CollapseTCR
iv	`CASSLGGGADEQYFGPG`	Error-corrected TCRs can further be translated and CDR3 regions extracted

B



C

```
69x:  3,  33,  8,  7,  CCAGGCGGGGGGGGG
 2x:  3,  33,  8,  9,  CCAGGCGGGGGGGGGCA
 2x:  3,  33,  8,  7,  CCGGGCGGGGGGGGG
 2x:  3,  33,  8,  7,  CCAGGGGGGGGGGGG
 1x:  4,  33,  9,  7,  TCCAGGGGGGGGGGG
 1x:  3,  33,  9,  7,  GCCAGGGGGGGGGGG
 1x:  3,  33,  9,  7,  CCCAGGCGGGGGGGG
 1x:  3,  33,  8,  7,  CCAGGGGGGGGGGG
 1x:  3,  33,  8,  7,  CCAGGGGGGGGGG
 1x:  3,  33,  8,  7,  CCAGGCGGGGGGG
 1x:  3,  33,  8,  7,  CAAGGCGGGGGGGG
 1x:  3,  33,  11, 7,  GGTCCAGGCGGGGGGGG
```

rearrangements. In this particular example, there were 83 reads all associated with the same barcode, which between them comprised 20 different inter-tag DNA sequences producing 12 unique `Decombinator` classifiers. Therefore instead of the 12 different clonotypes we would naïvely interpret as having a cumulative frequency of 83, collapsing allows us to disregard the sequences that have emerged through PCR amplification and error, revealing the single, accurate TCR representation.

Plotting the relationship between raw and collapsed frequency – at least for those genuine sequences that survive collapsing – reveals that the overall correlation is strong (Figure 5.22). However, variance is much greater for lower frequency clones, so that one is less able to reliably infer the collapsed frequency of a TCR if its raw frequency is below a certain threshold. Depending on the sample, this threshold seems to lie somewhere between 100 and 1000 copies, which likely reflects the various sampling processes involved. The frequency collapsing data can also be expressed as a duplication rate, calculated by dividing the raw observed frequency of a TCR chain by its eventual error-corrected, collapsed frequency, telling us to what extent that sequence was multiplied. I observed a similar pattern of duplication rates across the different samples; a representative example is shown in Figure 5.23A. Low frequency recombinations exhibit a wide variance in duplication rates, as might be expected given the large spread of low frequency clones illustrated in Figure 5.22. Moreover these low frequency data fall into one of a number of gradients, which is determined by their collapsed frequency (the denominator of the duplication rate). However, similar to previous findings, above a certain size (an approximate raw frequency between 500 and 1000, or a collapsed

Figure 5.21 *(preceding page)*: The global effect of collapsing TCR repertoires. **A**: Schematic of the various steps of the full bioinformatic pipeline used in this thesis. **i**: FASTQ data as produced by sequencing our amplified libraries on the Illumina MiSeq, which contains both nucleotide sequence and base-call quality information. **ii**: TCRs are assigned by `vDCR`, containing the typical `Decombinator` five-part classifier in addition to the sequence and quality information for both the inter-tag region as well as the barcode region (a concatenation of the two random hexamer sequences that were uniquely incorporated prior to amplification). **iii**: `CollapseTCR` uses barcode information to remove erroneous TCR classifiers and produce a better estimate of true cDNA frequency in sample. **iv**: Collapsed repertoires can then be translated and CDR3 regions extracted for further analysis, using the **CDR3ulator** script. **B**: Overview effect of collapsing by showing the decrease in frequency of unique and total TCRs detected from the raw, unprocessed samples of the healthy peripheral blood controls from both the barcoded V (BV) and barcoded ligation race (BLR) protocols. **C**: All of the `Decombinator` classifiers that were associated with a particular unique molecular barcode (`ACCGCCGGGGAC`) in the α TCR `vDCR` output file for healthy peripheral blood sample HV01v1.

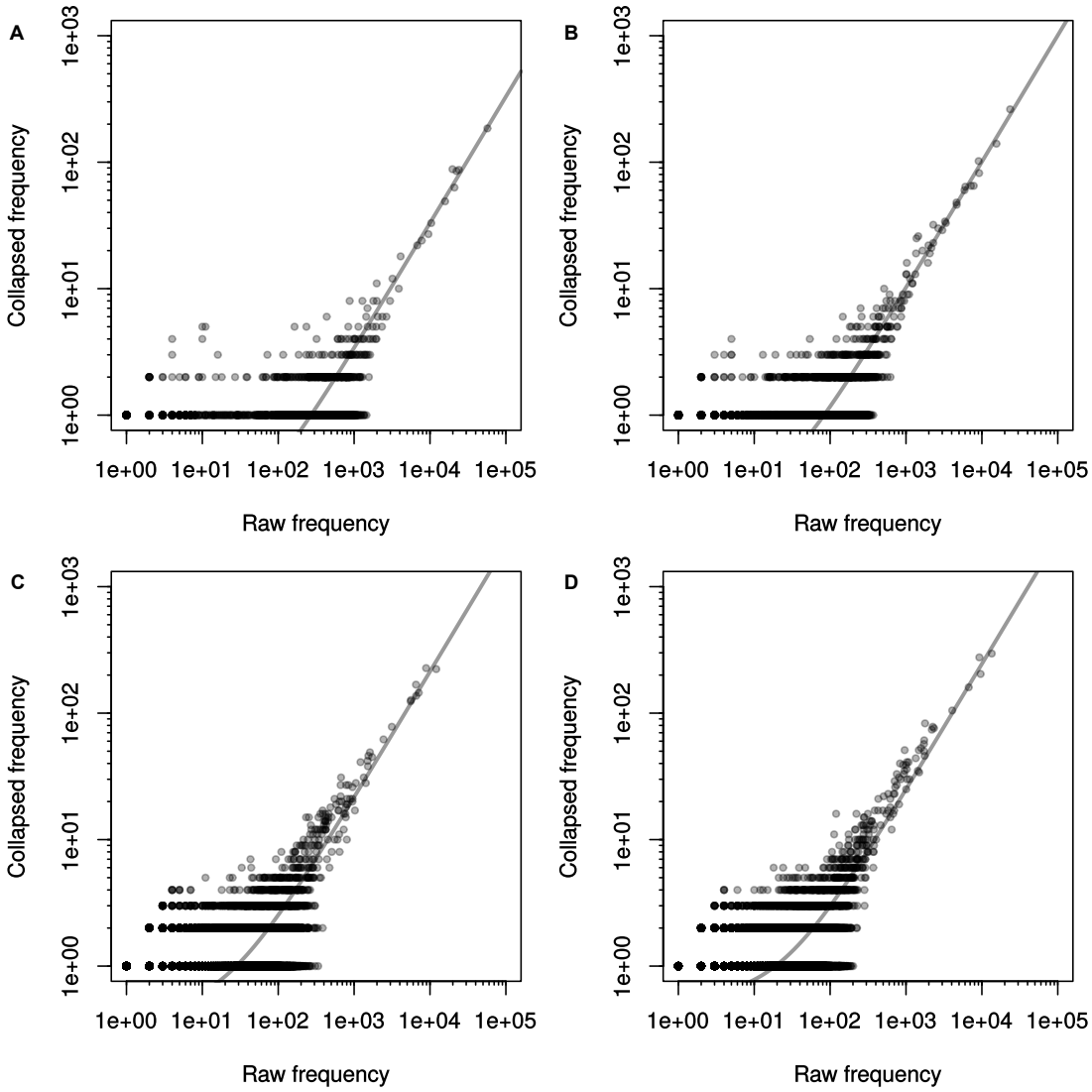
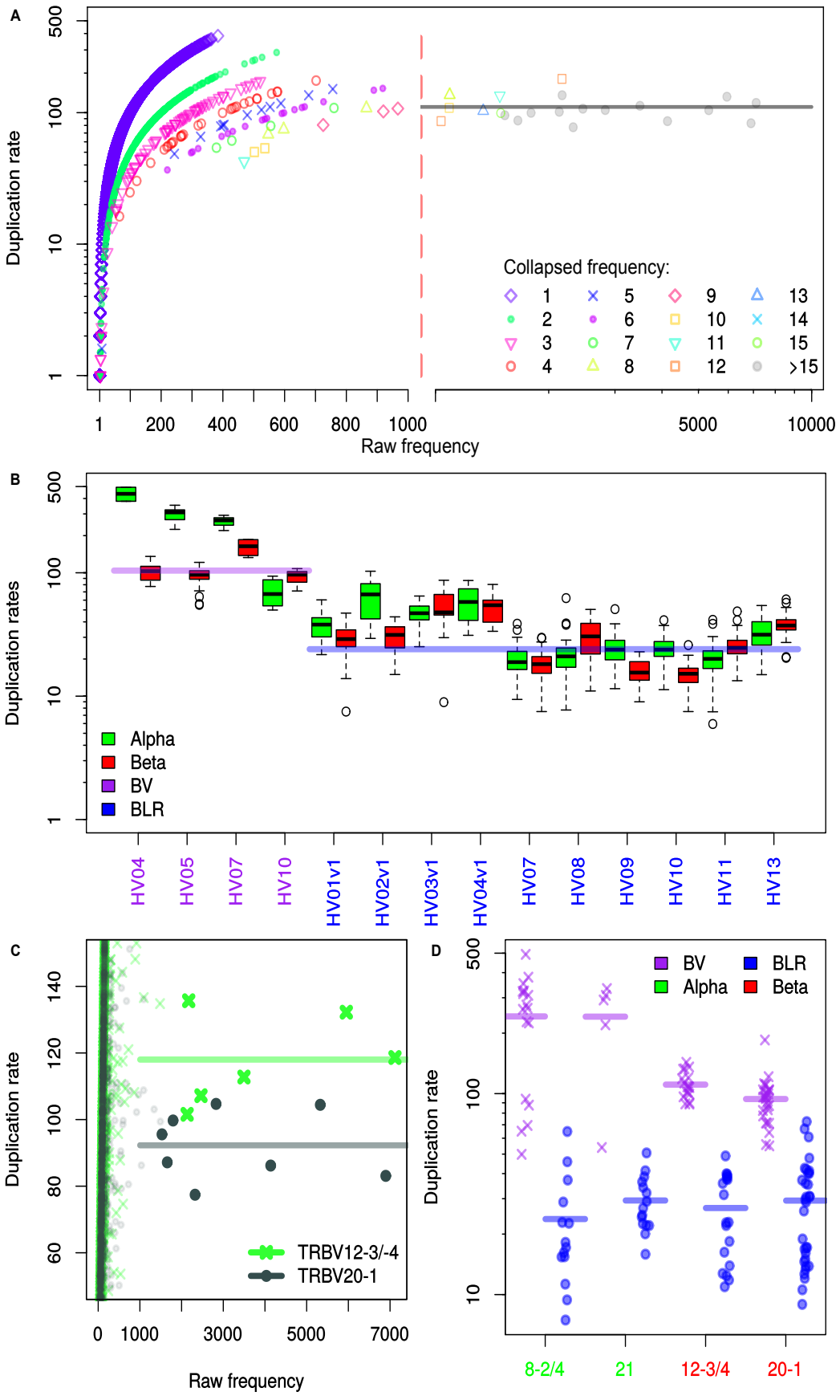


Figure 5.22: Correlation between raw and collapsed TCR frequencies. A and **B**: α and β TCR samples produced using the BV protocol (sample HV5). **C** and **D**: α and β samples produced using the BLR protocol (sample HV01v1). R^2 values for all samples in order: 0.96, 0.96, 0.92, 0.92 . P-values were all lower than 2.2e-16, the smallest value calculable using the R statistical analysis software linear regression module.


A
Duplication rate
500
100
10
1
1
200
400
600
800
1000
5000
10000
Raw frequency
Collapsed frequency:
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
>15
B
Duplication rates
500
100
10
1
Alpha
Beta
BV
BLR
HV04
HV05
HV07
HV10
HV01v1
HV02v1
HV03v1
HV04v1
HV07
HV08
HV09
HV10
HV11
HV13
C
Duplication rate
140
120
100
80
60
0
1000
3000
5000
7000
Raw frequency
TRBV12-3/-4
TRBV20-1
D
Duplication rate
500
100
10
BV
BLR
Alpha
Beta
8-2/4
21
12-3/4
20-1

frequency above 15) duplication rates began to normalise. Therefore clones above a certain size are amplified at a rate proportional to their prevalence (explaining the good correlation of high frequency clones in Figure 5.22). I also observed that the different barcoded protocols produce different duplication rates, with BV amplification showing a fourfold median increase relative to BLR (Figure 5.23B). Close examination revealed that the different V genes targeted in the BV protocol do not always display the same average duplication rates, while the BLR protocol, which amplifies all rearrangements using the same primers, does (Figure 5.23C and D). This difference is particularly obvious when comparing the difference in TRAV and TRBV gene duplication rates in the BV protocol (where α genes are duplicated approximately 17% more), but holds true when comparing the two different specific β genes targeted (TRBV12-3 and TRBV20-1, Figure 5.23D and Table C.10). Some of the increased variance in the BLR duplication rates may be due to these data coming a greater number of batches than the BV data: BLR results are the product of four preparation batches sequenced across five different sequencing runs, while all BV data is the product of two experiments.

Figure 5.23 *(preceding page)*: Use of unique molecular identifiers allows measurement of duplication rates of different TCR rearrangements. Duplication rate = raw rearrangement frequency / collapsed frequency. **A**: Representative plot illustrating the typical distribution of duplication rates within a given TCR sample (sample HV04b, BV protocol). Clonotypes are coloured by their collapsed frequency (red dotted line hightlights broken x axis at 1000, which switches to a logged axis). Lower frequency clones display highly variant duplication rates along paths determined by their actual collapsed frequency (as this value is the denominator in the calculation of the rate), while larger clones (those with collapsed frequencies greater than 15, or raw values greater than 1000, approximately) tend to show approximately equivalent duplication rates. Grey horizontal bar indicates the mean frequency of the rearrangements with collapsed frequencies greater than 15. **B**: Duplication rates for all the healthy control barcoded sample rearrangements with collapsed frequencies greater than 15 (apart from BV HV09, as this had a very small sample with a very 'flat' repertoire, with very little duplication observed). Median group values are shown as horizontal lines (104.15 for BV, 24.0 for BLR). **C**: Example plot section showing the difference in duplication rate between the two different V regions primed for in the BV β amplification protocol. Analysis limited to those rearrangements that have a collapsed frequency greater than 15 (shown with larger and darker plot points); TRBV12-3/4 duplication rate is significantly higher than that of TRBV20-1 (p-value = 0.002, unpaired one-sided Welch's t-test). **D**: Duplication rates of all rearrangements with collapsed frequencies greater than 15 for each of the different V genes targeted in the BV protocol. Horizontal lines indicate group means.

5.3.2.1 vDCR/Collapsing verdict

The pipeline described in this thesis employs a modified version of `Decombinator` to detect and classify TCR rearrangements, followed by error- and frequency-correction made possible by the introduction of random molecular barcodes to transcript sequences prior to amplification. The modifications made to `Decombinator` itself largely fall into two categories: those that increase the breadth and specificity of TCRs which it can detect, through the design of new sets of V and J tag sequences, and those that remove specific kinds of errors, through the addition of simple filters. The other major change made was to make `Decombinator` output some additional fields, containing sequence and quality information for the TCR and barcode sequences, which allows us to collapse the data and obtain a more accurate representation of what the actual repertoire in our sample was.

Collapsing is probably the single most important contribution to the robustness of the protocol. In developing this methodology I have demonstrated that using raw high-throughput TCR sequence data is flawed both qualitatively and quantitatively. Over half of the unique TCRs measured are actually the product of sequencing or PCR errors. Moreover the raw frequency of these TCRs only correlates well with the the collapsed frequency above a given frequency: this stochasticity of low-frequency clonotypes (which make up the majority each sample) would correspond to an inability to correctly assign the clonal hierarchy or perform any number of quantitative assessments accurately. These lower frequency clones display a highly variable rate of duplication throughout our pipeline, which is perhaps not well explained by the standard paradigm of PCR amplification. It is perhaps possible that the random sampling that occurs throughout the various purifications and dilutions throughout the protocol play a role, and that larger clones simply rise above the noise.

I also observed that average duplication rates varied between the two protocols that employ cDNA-barcoding. The BV protocol rearrangements unexpectedly displayed larger average duplication rates than the BLR protocol, which seems counter-intuitive given that the BLR protocol samples underwent one more PCR cycle in total than the BV samples (27 rather than 26). A possible explanation is that the final duplication rate is related to the amount of input nucleic acid: equivalent amounts were used in both protocols, however BV specifically only targets a fraction of the overall repertoire, and thus will produce a smaller amount of cDNA from a given blood RNA sample. It is feasible that this lesser amount of input material has therefore more 'room' during the PCR, and can attain higher duplication rates than protocols in which there might be more competition in the reaction. Other members of the group have performed

quantitative PCR of subseqeuent versions of the BLR protocol and demonstrated that these reactions are highly efficient (in terms of the fold increase in number of molecules per cycle, giving values close to the maximum of two, data not shown), which makes it difficult to attribute the difference in duplication rates to differential amplification. In addition to these differences between the protocols, I also observed a marked difference in duplication rates between rearrangements making use of the different V genes targeted in the BV compared to the BLR protocol. This is not unexpected: primer bias – where amplicons using different primers are amplified to different extents – has been reported to occur as a result of differences in both DNA binding energies (395, 396) and certain DNA motifs being preferentially favoured for priming by different DNA polymerases (457). This is indeed the reason why we strove to develop an unbiased RACE based protocol. It is possible that with the addition of barcoding a large portion of primer bias might be mitigated.

There are two recognised limitations to the collapsing methodology I have employed. The first is that there will be a subset of erroneous reads remaining: those reads which contained incorrect base-calls in both their (inter-tag) TCR and barcode sequences will not be collapsed together, as currently it relies on a match of either one in order to recognise that clonotypes are likely derived from the same original molecule. This could be addressed in later versions of the software, however given the throughput we typically work with in our samples the number of these sequences are probably very low. The second limitation is that the clustering of barcode sequences (so as to estimate the actual frequency of the rearrangement) allows for a relatively large number of edits when considering the length of the string (all barcodes that are up to three bases different from one another are considered to represent one sequence that has accrued errors). It is for this reason that we only claim to be estimating, not exactly measuring the true frequency of that cDNA. This could be addressed by including a greater number of random bases in the barcode sequences and placing these barcodes in sections of the read from which one obtains higher quality sequence data, which would allow us to be more confident that we are both observing the barcode accurately and not artificially shrinking a rearrangement's frequency by clustering.

I have demonstrated that the current pipeline is able to accurately detect and describe polyclonal TCR repertoires in a high-throughput manner, producing robust quantitative and qualitative data, a feature that has largely been enabled through the addition of random barcodes thus allowing us to collapse our data. Furthermore this analysis has highlighted some of the flaws and biases that will presumably be present in similar repertoire sequencing efforts that have not been comparably processed; therefore

these findings should be of great interest to any group that seeks to accurately deep-sequence receptors of the adaptive immune system.

5.4 Protocol development discussion

In the course of this thesis I have developed and refined a protocol that permits unbiased amplification, effective high-throughput DNA sequencing and robust bioinformatic analysis of TCR repertoires. In the course of developing this protocol I have gleaned various findings that would aid others undergoing a similar process. In my data an unbiased RACE approach, sequenced on the Illumina platform (and avoiding the addition of homopolymers wherever possible) produced the most reliable, unbiased repertoires with the highest sequencing quality. Furthermore I have shown that a large proportion of errors unavoidably produced during the process of amplification and sequencing can be mitigated through incorporation of random molecular barcodes, which allows one to correct errors after sequencing. This pipeline therefore meets the major criteria that an effective repertoire sequencing effort should.

One might argue that the principle of barcoding could be combined with a multiplex repertoire PCR (such as I have shown using the BV protocol) and achieve comparable results to an unbiased RACE, cancelling out the effect of primer bias. I contend that this is not the case. Even if one uses barcodes to negate the effect of differential amplification rates in a multiplex PCR (meaning that the data you end up with is more accurate in terms of what TCRs are there in what frequencies), the very fact that there has been an imbalance could mean that any disfavoured target molecules are less likely to survive the various sampling events (e.g. carrying over from one PCR to the next, or sampling to the correct concentration for the sequencing machine). Such reads would therefore not be represented in the final data; collapsing can obviously only work on TCRs for which sequences are observed. Multiplex PCRs for TCR repertoires also require generation of panels of primers targeting different V (and potentially J) genes; typically this involves selecting the functional gene sections from a repository such as IMGT and selecting primer sites by some criteria. This could be expected to fail to amplify some TCR rearrangements in two ways: allelic differences at primer binding sites could result in failure to prime, which would cause an apparent reduction or loss in gene use, and inappropriately curated or inaccurate references could result in certain genes being omitted when they do in fact make up part of the repertoire (as I have demonstrated is the case with various supposedly non-functional or not-α-rearranging TRDV genes). This more philosophical failing of multiplex strategies can be extended further: in priming for each V gene (typically as close to the V(D)J

border as will still allow V gene identification) one loses a great deal of biological information which might be useful in elucidating further TCR biology (such as the investigation of alternative splicing I undertook in Section 3.3). It is for this reason that I would especially recommend a RACE based approach when sequencing either immunoglobulin genes (which due to somatic hypermutation might lose priming sites) and variable antigen receptor repertoires from other species (whose germline loci are likely to be less well annotated than say human or mouse).

We are not the first group to have made use of such a cDNA labelling approach. The first group to deep-sequence adaptive immune repertoires – measuring the antibody repertoires of zebrafish – used such an approach (438), and it has been explored for use in general RNA sequencing (435, 458). However for reasons unknown to this researcher, the field at large has not adopted the practice, with one exception: the Russian laboratory that produced the Bolotin *et al* data used for comparison throughout this chapter has since developed and published new protocols that make use of a comparable barcoded RACE technique (and similarly observe the associated improvements following normalisation) (454, 459, 460). An interesting corollary of the impact of collapsing TCR repertoire data is that existing repertoire data that does not make use of such an approach could well be intrinsically flawed with respect to its error profile and frequency distributions. Some groups reduce sequencing error with 'nearest neighbour' or similar clustering algorithms (76, 126, 183). However this potentially could mistakenly conflate genuine rearrangements that happen to be highly similar.

The barcoded ligation RACE strategy is still being further honed and developed in the group, focusing particularly on how to drive down the PCR duplication rate. Even though we can still account for said duplication by collapsing on barcodes, a lower rate would effectively free up space on the sequencing flow-cell, allowing us to theoretically sequence a greater proportion of low-frequency clones. It would also be highly beneficial to reduce the number of PCR cycles each sample receives to the bare minimum, not only to reduce duplication but the range of accumulative errors than can occur (461, 462, 463). Reducing the cycle number could be achieved by performing the final amplification in a real-time quantitative PCR in comparison to samples of known concentration; this should allow users to halt the amplification reaction at the exact cycle at which there is enough DNA to sequence. Other protocol alterations currently being tested involve incorporation of longer random stretches, solely in the ligation adapter, reducing the number of purifications required and increasing the resolution between similar barcodes.

In this study I have primarily made use of the Illumina MiSeq platform for sequencing our repertoires, as it produced a sufficient number of reads of lengths more than suitable for our purposes, with quality score distributions that are consistently found to best rival technologies. Moreover it does so with less cost than many alternatives and in a time frame that is conducive to research, and potentially clinical applications as well. These features make it the most suitable technology for most repertoire sequencing applications, which is reflected in the fact that the major commercial entities offering repertoire sequencing make use of it (464, 465). However, like all second-generation DNA sequencing technologies (which are defined by amplifying clonal populations of DNA fragments in a hugely parallel manner, followed by independent sequencing), DNA amplification is required at one if not more stages (as most machines require relatively high input DNA concentrations). This requirement can theoretically be dispensed with when working with third-generation sequencing platforms as these technologies are defined by being able to sequence single DNA molecules (372, 373, 375). Adoption of such technology (such as the SMRT platform from Pacific Biosciences (382) or the MinIon from Oxford Nanopore Technologies (388)) for sequencing un-amplified immune repertoire libraries could provide the first truly unbiased repertoire data. Furthermore these technologies are able to sequence both very long molecules (which would allow sequencing of unrearranged and genomic TCR loci in one pass) and modified nucleic acids (which might elucidate further sites of regulation of TCR expression).

In conclusion, the research that I have undertaken suggests that in order to obtain the highest quality sequence data that contains the greatest amount of biological information one should bear in mind three main points: an unbiased amplification procedure (e.g. 5' RACE) should be employed, to ensure coverage of all genuine rearrangements in the sample; an appropriate sequencing technology must be used (which in the current market will most likely involve an offering from Illumina), and final output data can be significantly improved through the inclusion of random barcode sequences prior to amplification. High-throughput DNA sequencing of variable antigen receptors represents a powerful tool in being able to dissect and understand how vertebrate immune systems operate; it is imperative that our protocols for doing so are fit enough for the task.

6

Conclusions

The T-cell receptor repertoire is an incredibly complex and dynamic biological system, and the ability to accurately measure and model it could greatly aid the understanding of a diverse range of immune processes. However this very complexity makes repertoires intrinsically difficult to measure. Technologies born out of the genomics revolution are enabling greater interrogation of the clonal make up of TCR repertoires then ever before.

In this thesis I demonstrate the development of a protocol that uses unbiased amplification, random molecular barcoding and error-correcting analysis software to sequence the TCR repertoires of RNA samples. I have used this protocol both to characterise elements of healthy donor repertoires and measure how some of these features are perturbed during HIV infection.

A number of themes emerge from the interpretation of this data and related published findings. Perhaps the most significant is the degree of sharing of TCR sequences and profiles between unrelated individuals, which might seem counter-intuitive for a system that is generally regarded to exist as a means for producing receptor diversity. Data such as these begin to redefine our conception of variable antigen receptor generation, in which the stochasticity produced by V(D)J recombination does indeed produce a tremendously diverse repertoire, but within pre-determined bounds. This seems especially true of so-called 'invariant' TCR chains, which appear in multiple individuals in phenotypically defined cell subsets. These bounds likely reflect a combination of the constraints of the recombination machinery and selection processes (both thymic and peripheral), although the contribution that each of these makes is still yet to be dissected.

The other predominant theme of this work is that of unexpected complexity. As our ability to measure TCR rearrangements increases, the limit of detection for rare and aberrant events recedes. Furthermore using an unbiased amplification reaction removes the confirmation bias associated with targeted multiplex primer approaches. These effects have manifested several times in this thesis, such as in the detection of high-frequency rearrangements using purportedly non-functional gene segments and low-frequency alternative splicing events. While the biological significance of these

phenomena remain to be established, these findings demonstrate the diverse potential applications of the technology and inform the design of future experiments.

Application of TCR RepSeq to HIV patient blood samples demonstrates the utility of the technology in understanding the pathogenesis of clinically relevant infections. The data presented complements and elaborates on the established hypotheses, producing a model in which the drastic loss of TCR diversity – as a result of both $CD4^+$ T-cell depletion and considerable expansion of a subset of $CD8^+$ T-cells – perturbs all features of the repertoire. In this data I observed significant alterations in the frequencies of CDR3s associated with antiviral responses identified in previous studies. These observations were made despite this study only sequencing a small sample of blood at relatively low depth; were we to sequence more blood or increase our read depth (and harvest or otherwise obtain more pathogen-specific TCR sequences) I predict that such analyses would prove even more fruitful.

I contend that I have demonstrated the power of deep-sequencing as a tool to investigate the repertoire, being able to both answer a variety of biological questions and generate hypotheses for future studies. I anticipate that repertoire sequencing will become an invaluable tool in a wide variety of immune studies, that will continue to grow both more commonplace and more powerful as sequencing and related technologies progress. Used in conjunction with other technologies (such as cell sorting or transcriptomics) would permit gathering of linked rearrangement and phenotypic information, increasing the dimensionality and potential information content of output data.

6.1 Future work

There remain a great number of experiments and improvements that could be made both in this work and the field of TCR receptor repertoires more generally.

With respect to the data presented here, the majority of future experiments would involve collection of both more sequences and more samples. Greater numbers of sequences per sample – by taking more blood, amplifying from more RNA or sequencing more of our libraries – would allow greater detection of rare and potentially shared sequences. Taking samples from more healthy donors and HIV patients would allow us to greater characterise the differences in TCR features both within and between groups, while taking more longitudinal samples from the same donors would provide data on the dynamics and diversity of the peripheral pool. From a clinical perspective, with enough patients and longer follow up this could be extended to measure whether and which TCR properties might show prognostic value or aid in the stratification of patients for treatment regimes. Cell sorting, particularly of $CD4^+$/$CD8^+$ populations

in the context of HIV patients, would be expected to permit greater understanding of the dynamics of the repertoire in these subsets. The major factor limiting these works is simply one of cost; given sufficient funds all of these experiments are possible today.

On a practical level, there are a number of improvements which could be made to increase the yield of information from the data produced by our (and similar) deep-sequencing protocols. Primarily, barcoding of some sort should become far more prevalent than it already is, as it is the only solution that I am aware of that counters all aspects of PCR- and sequencing-introduced errors. Modulation of primer concentrations in multiplex PCRs may well be able to effectively counteract primer bias at the gene level (399), but it is still unable to normalise read frequencies back to nucleotide frequency, and nor can it correct nucleotide sequences save by clustering, which may conflate genuinely different rearrangements that happen to have similar sequences. The barcoding protocol detailed in this thesis could itself be improved through the addition of longer barcode sequences, in order to provide better resolution between barcode sequences during the estimation of true rearrangement frequency. Improvement of sequencing quality of both TCR and barcode sequences would greatly minimise the chance of developing or retaining erroneous sequences. A related constraint is that of read length. An ideal situation might involve reading the entire variable domain of all TCR mRNA molecules in a sample (or even the whole message). This would remove all requirements for sequence inference, allowing unequivocal interpretation as to the nature and productivity of a given mRNA. This might produce enough sequence to allow development of `Decombinator` tags that would be able to detect and distinguish all TCR V gene families and alleles, and would allow far greater investigation of TCR splicing. These practical issues are largely limited by the nature of the sequencing technology available to us: improvements in read length and quality would help address these concerns.

Postscript

In 1989 Charles Janeway wrote the seminal paper 'Approaching the Asymptote? Evolution and Revolution in Immunology' (466). In addition to expounding a variety of predictions and frameworks that bore fruit across the range of immunological fields, he made the following statement:

> *"... I believe it is safe to state that our understanding of immunological recognition is approaching some sort of asymptote, where future experiments are obvious, technically difficult to perform, and aim to achieve ever higher degrees of precision rather than revolutionary changes in our understanding."*

This could arguably be applied to the practice of investigating immune repertoires, where the majority of improvements in the last two decades largely correspond to technical advances permitting analyses with greater throughput and resolution. However, I contest the notion that increasing precision and making changes in our understanding cannot go hand in hand.

Immunology is in some respects still a comparatively young field, consisting of the study of one of the most complex biological systems known (second perhaps only to that of the nervous system). Tremendous advances have been made identifying the physiological, cellular and molecular components of the immune system – only a tiny portion of which have even been touched upon in this thesis – yet large gaps in our understanding remain. Findings can be difficult to translate between models or from basic to clinical research. and the effects of interventions remain hard to predict both quantitatively and qualitatively.

However, higher-dimensional, higher-throughput technologies such as those described in this thesis can help us move towards system level understandings. By gaining greater ability to measure the constituents of the immune system, at all scales, we increase our ability to model emergent properties. A Grand Unified Theory of Immunology is still a long way off, but we can still move towards it, even if we are only increasing precision. As Jacques Miller concluded in a review article recounting his incredible contributions to the field (38):

> *"There is still a great future for immunology!"*

A

Appendix A - exploring the TCR repertoire of healthy individuals

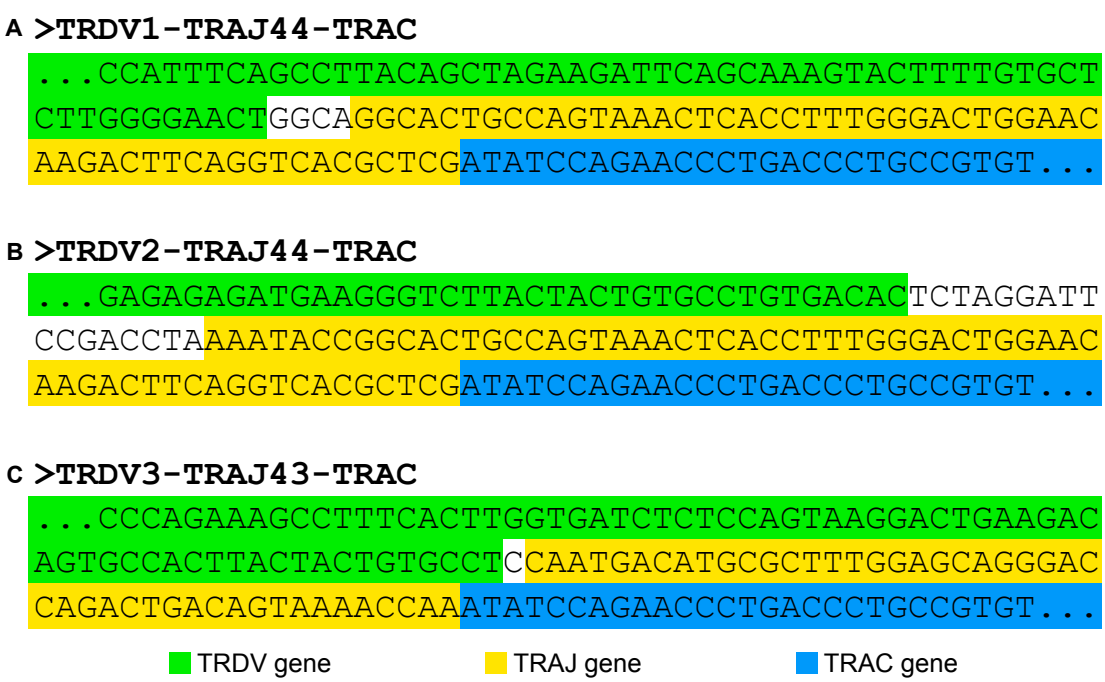


Figure A.1: Detection of TRDV usage in α chain TCR data. Identification of α chain rearrangements (i.e. containing TRAJ and TRAC) using the IMGT-defined δ chain specific V genes TRDV1 (**A**), TRDV2 (**B**) and TRDV3 (**B**), in publicly available TCR sequence data (from Wang *et al* (83), accession number SRA010149).

Gene * Allele	Sequence
TRBD1*01	GGGACAGGGGGC
TRBD2*01	GGGACTAGCGGGGGGG
TRBD2*02	GGGACTAGCGGGAGGG

Table A.1: Germline sequences of the human TRBD genes. Obtained from IMGT GENE-DB (19).

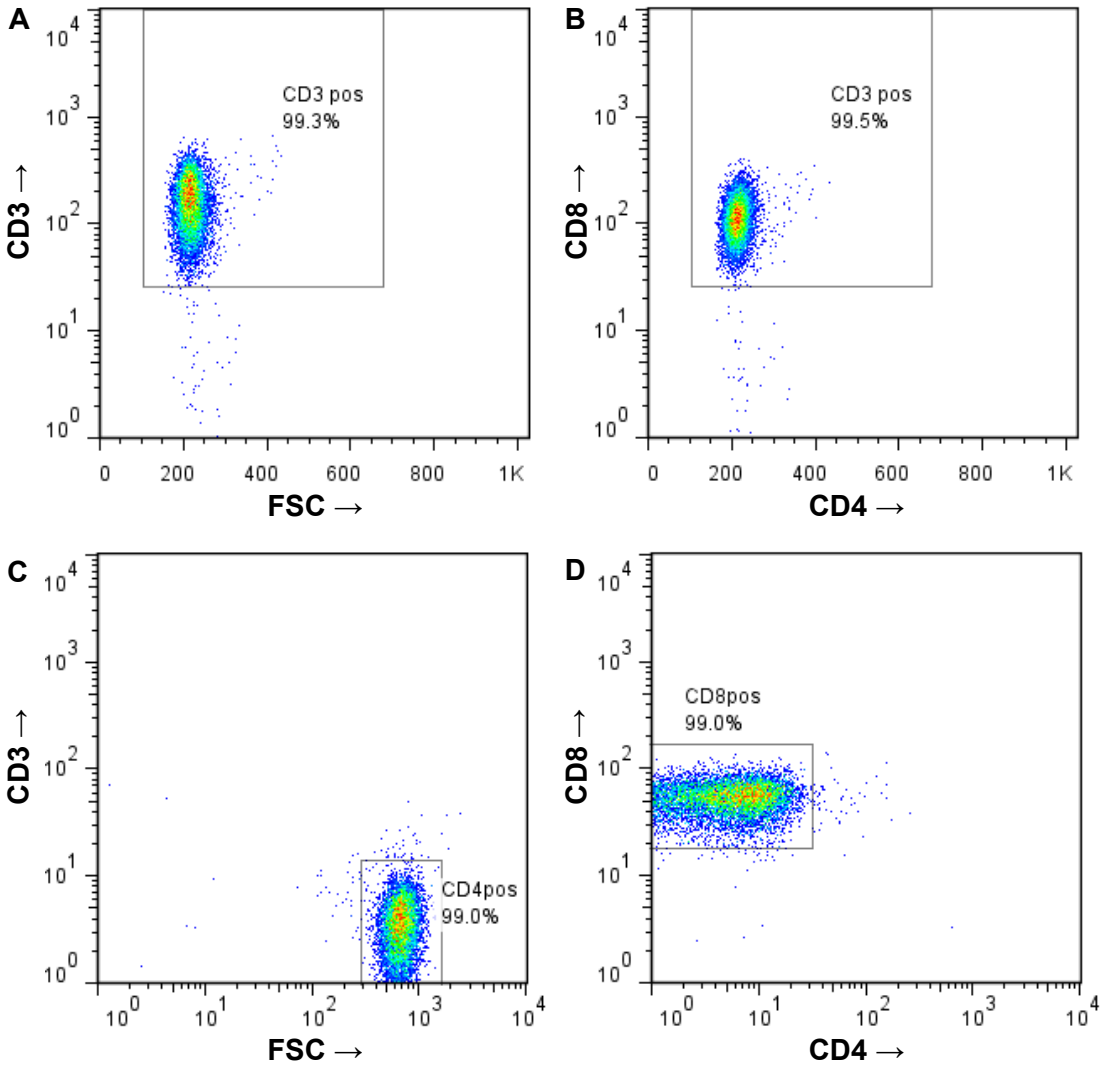


Figure A.2: Purity of CD4 and CD8 sorted T-cells. **A**: Purity of $CD4^+$ sorted cells before lysis, having sorted first for lymphocytes based on FSC/SCC (not shown), then $CD3^+$ (left) and $CD4^+CD8^-$(right). **B**: Purity of CD8+ sorted cells before lysis, having sorted first for lymphocytes based on FSC/SCC (not shown), then $CD3^+$ (left) and $CD8^+CD4^-$ (right).

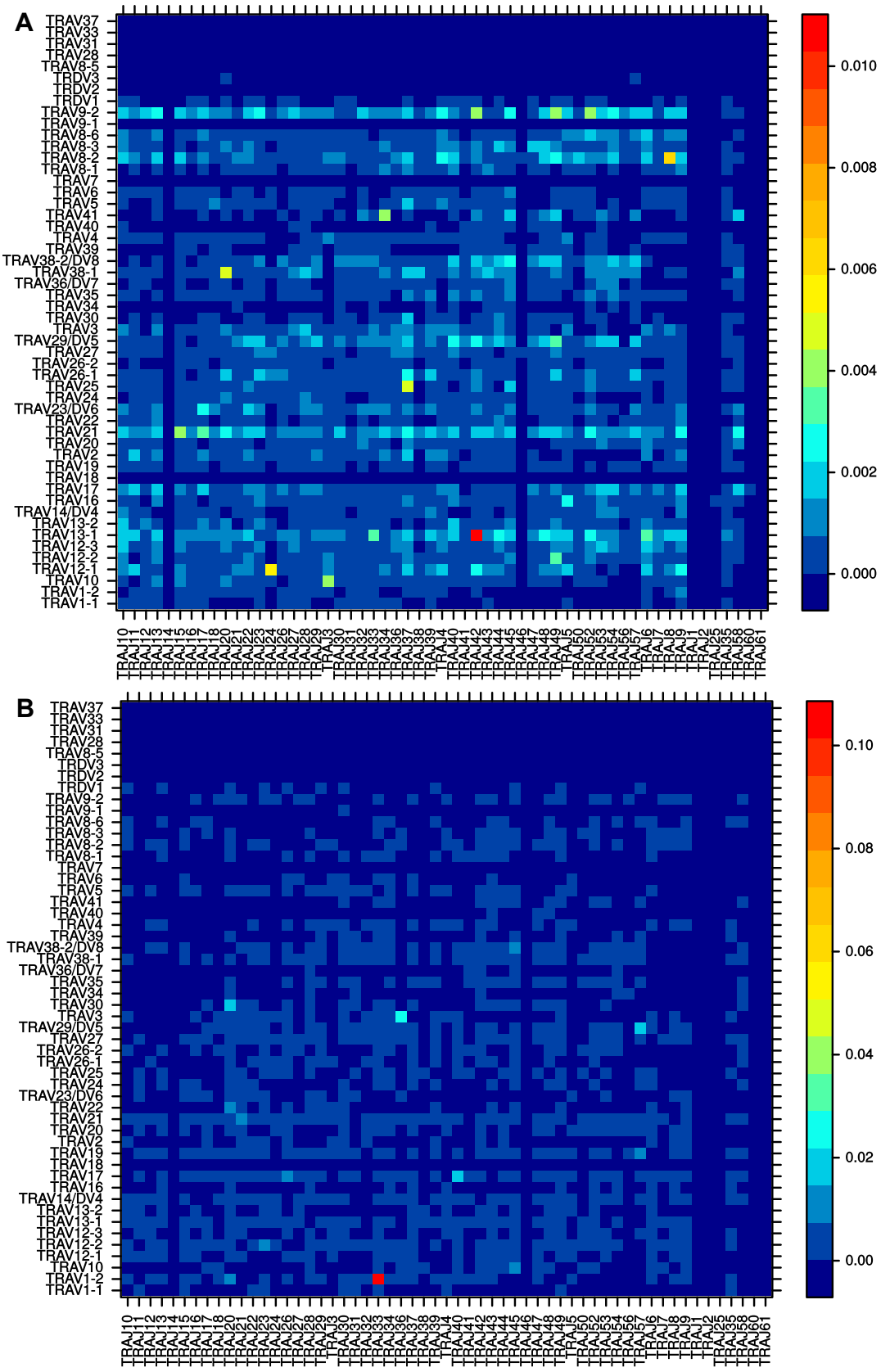


Figure A.3: Alpha chain VJ pairing frequencies for HV01. VJ pairing frequencies for HV01 VJ pairing for CD4 (**A**) and CD8 (**B**) sorted repertoires. 10,000 rearrangements were sampled from whole files.

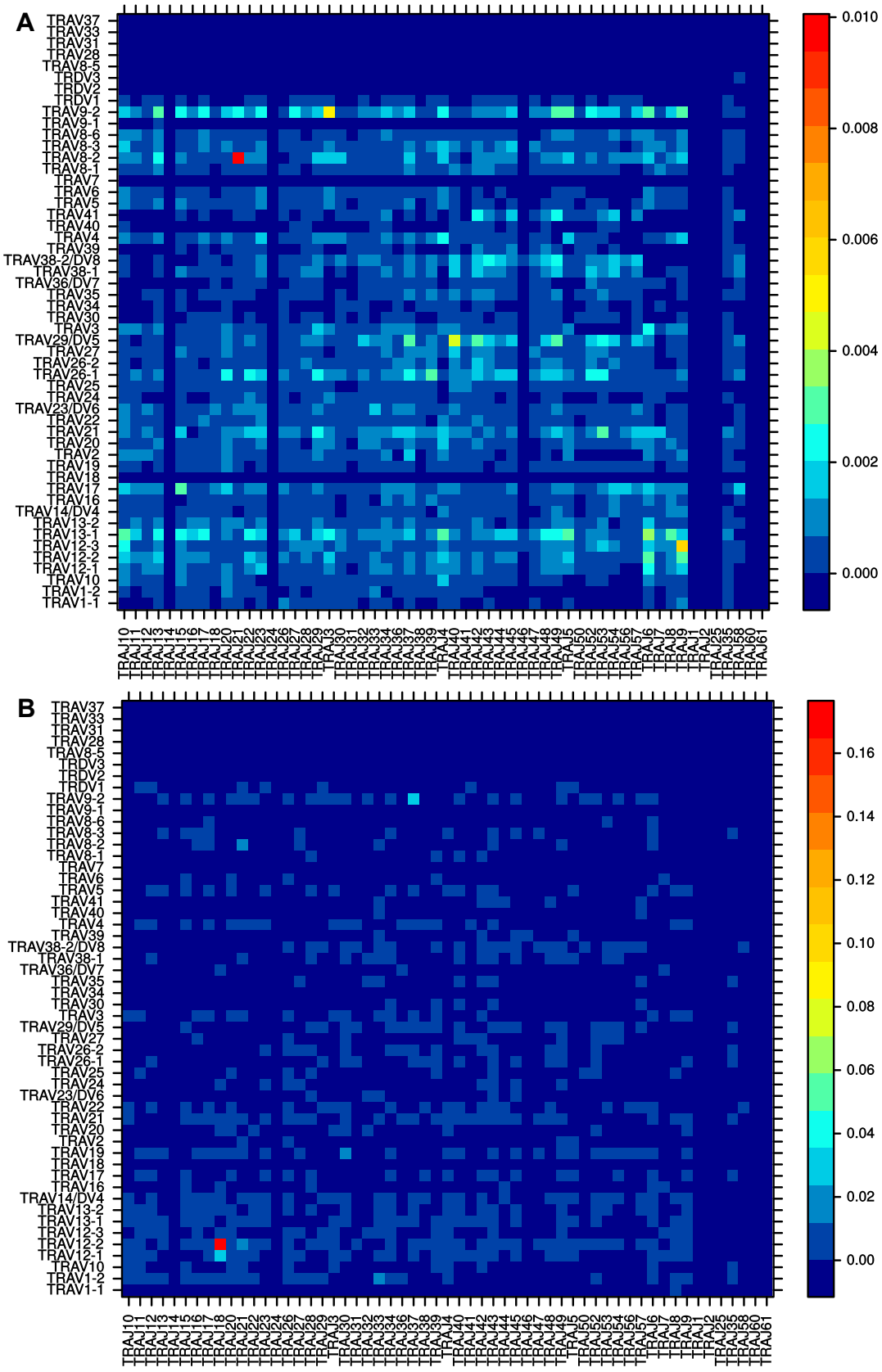


Figure A.4: Alpha chain VJ pairing frequencies for HVD4. VJ pairing frequencies for HVD4 VJ pairing for CD4 (**A**) and CD8 (**B**) sorted repertoires. 10,000 rearrangements were sampled from whole files.

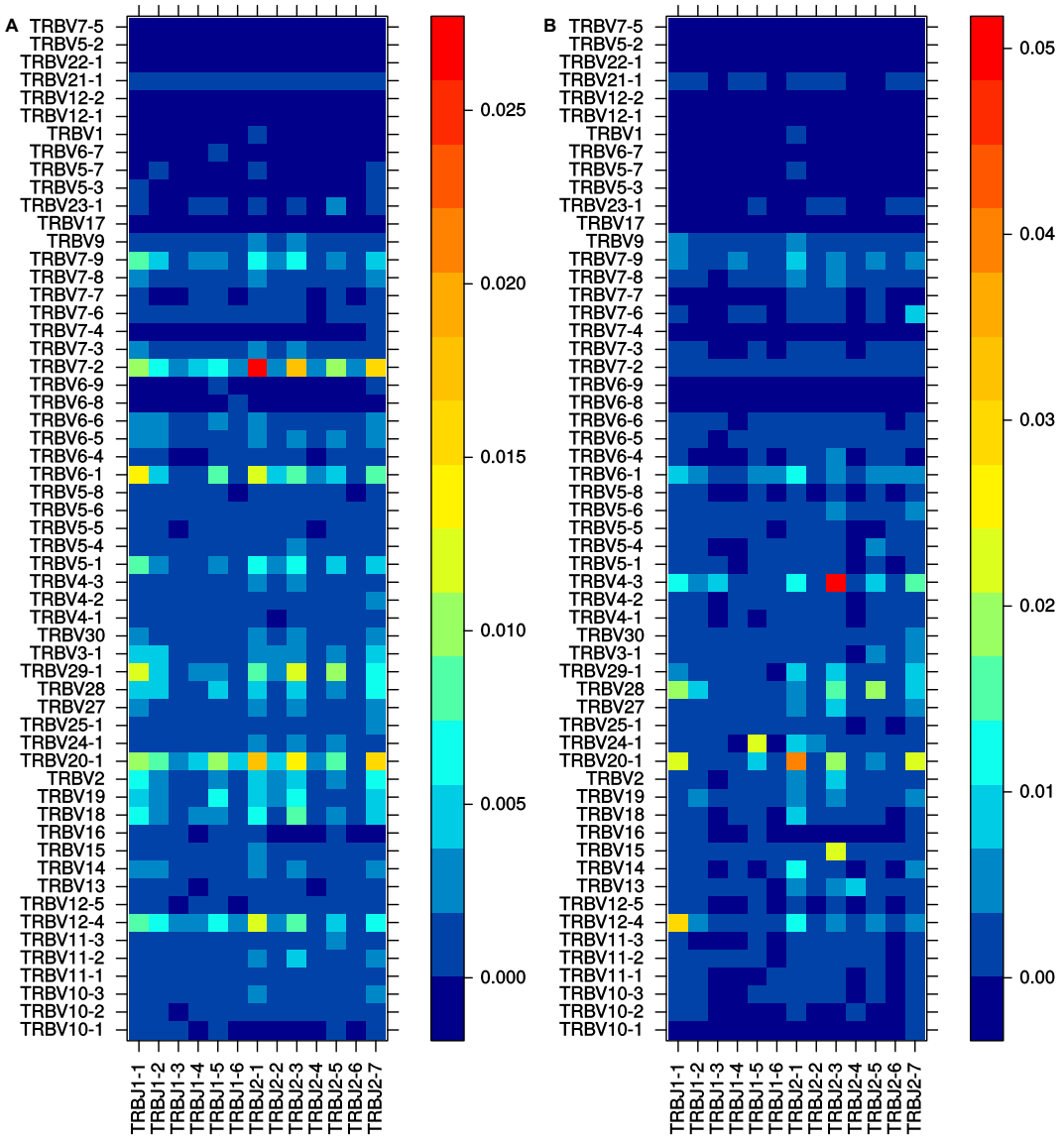


Figure A.5: Beta chain VJ pairing frequencies for HV01. VJ pairing frequencies for HV01 VJ pairing for CD4 (**A**) and CD8 (**B**) sorted repertoires. 20,000 rearrangements were sampled from whole files.

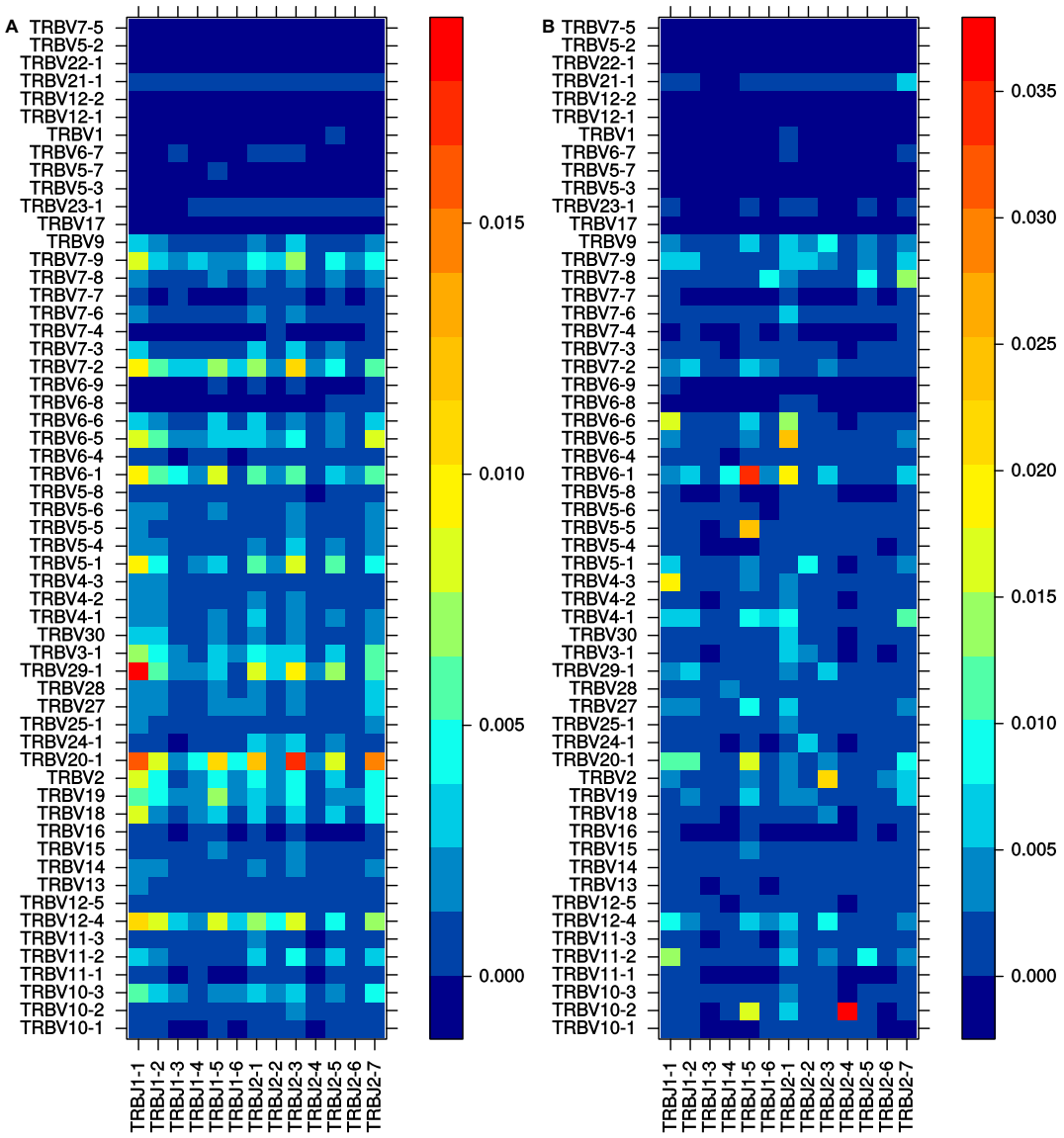


Figure A.6: Beta chain VJ pairing frequencies for HVD4. VJ pairing for CD4 (**A**) and CD8 (**B**) sorted repertoires. 20,000 rearrangements were sampled from whole files.

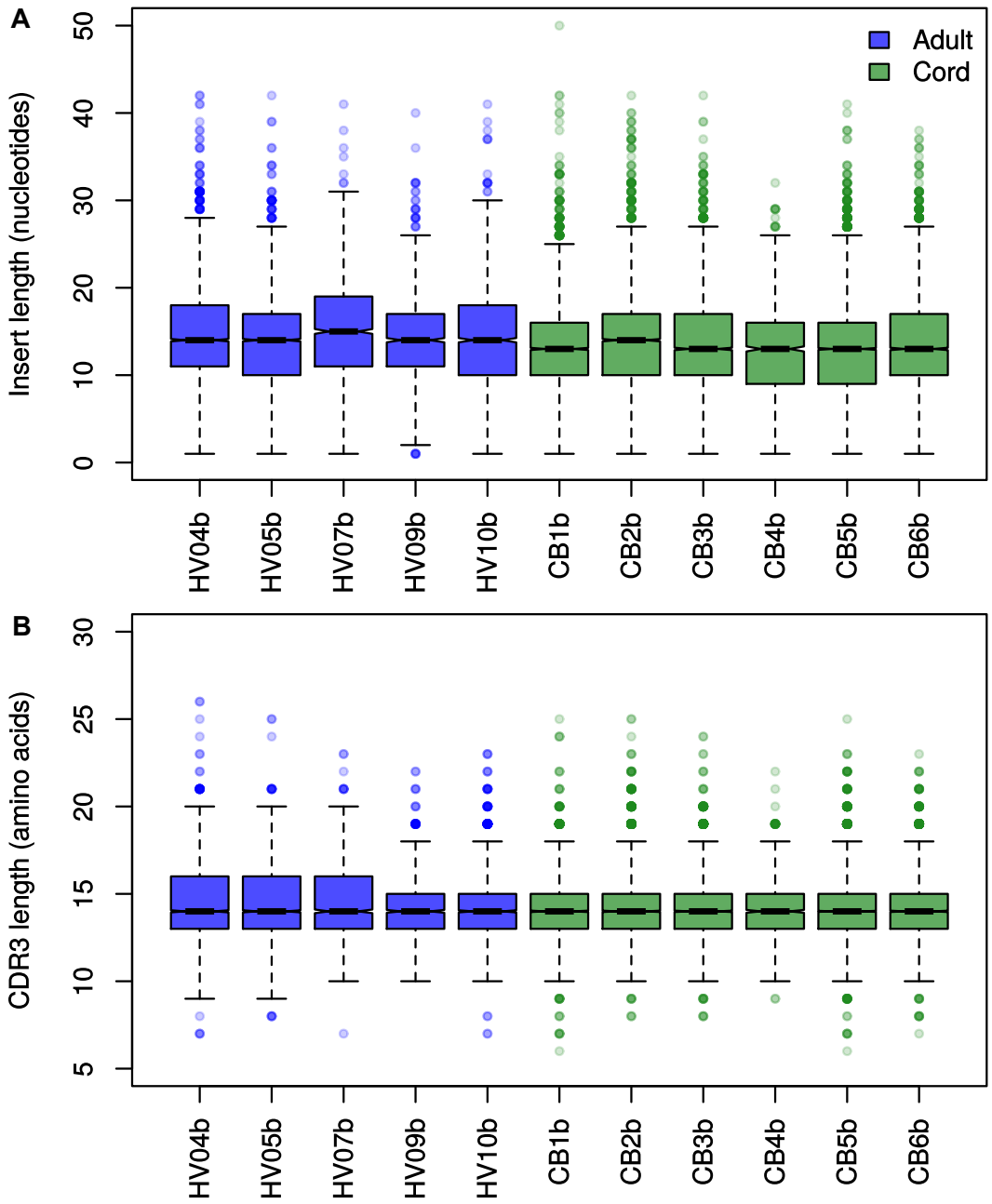


Figure A.7: Adult and cord blood β chain TCR length distributions. **A**: Box plot of β chain rearrangement insert lengths. **B**: Box plot of β chain CDR3 lengths. Data produced using the BV protocol, and therefore only targets rearrangements using TRAV8-2/TRAV8-4, TRAV21, TRBV12-3 and TRBV20-1.

Invariant Population	Reference	α	β
GEM	Van Rhijn *et al*, 2013 (140)	✓	✓
MAIT	Gold *et al*, 2010 (467)	✓	
MAIT	Greenaway *et al*, 2013 (282)	✓	
MAIT	Reantragoon *et al*, 2013 (281)	✓	
MAIT	Van Rhijn *et al*, 2013 (140)	✓	✓
MAIT	Lepore *et al*, 2014 (138)	✓	✓
MAIT	Van Schaik *et al*, 2014 (141)	✓	
NKT	Exley *et al*, 1997 (468)		✓
NKT	Han *et al*, 1999 (469)	✓	
NKT	Demoulins *et al*, 2003 (153)	✓	
NKT	Sanderson *et al*, 2012 (470)		✓
NKT	Greenaway *et al*, 2013 (282)	✓	
NKT	Van Rhijn *et al*, 2013 (140)	✓	✓
Putative	Van Schaik *et al*, 2014 (141)	✓	

Table A.2: Sources of the TCRs of T-cell subpopulations bearing invariant α chain and semi-invariant β chain rearrangements.

TRAV gene	TRAJ gene	CDR3	Number donors	Found in CD4	Found in CD8
TRAV13-1	TRAJ6	CAASKGGSYIPTF	10	✓	
TRAV12-2	TRAJ15	CAVNNQAGTALIF	9	✓	✓
TRAV13-1	TRAJ53	CAASSGGSNYKLTF	9	✓	✓
TRAV14/DV4	TRAJ3	CAMREGGYSSASKIIF	9	✓	✓
TRAV8-3	TRAJ6	CAVGAASGGSYIPTF	9	✓	✓
TRAV12-1	TRAJ15	CVVNNQAGTALIF	8	✓	✓
TRAV12-2	TRAJ5	CAVNNTGRRALTF	8	✓	✓
TRAV12-3	TRAJ8	CAMMNTGFQKLVF	8	✓	✓
TRAV13-2	TRAJ52	CAENNAGGTSYGKLTF	8	✓	✓
TRAV16	TRAJ37	CALSGGSGNTGKLIF	8	✓	✓
TRAV2	TRAJ5	CAVEEDTGRRALTF	8	✓	
TRAV21	TRAJ11	CAVMNSGYSTLTF	8	✓	✓
TRAV21	TRAJ15	CAVPNQAGTALIF	8	✓	✓
TRAV29/DV5	TRAJ52	CAASGAGGTSYGKLTF	8	✓	✓
TRAV9-2	TRAJ13	CALMNSGGYQKVTF	8	✓	
TRAV12-1	TRAJ23	CVVNNQGGKLIF	7	✓	✓
TRAV12-1	TRAJ20	CVVNNDYKLSF	7		✓
TRAV12-1	TRAJ30	CVVNMNRDDKIIF	7	✓	✓
TRAV12-2	TRAJ17	CAVMIKAAGNKLTF	7	✓	✓
TRAV12-2	TRAJ8	CAVNMNTGFQKLVF	7	✓	✓
TRAV12-2	TRAJ48	CAVNNFGNEKLTF	7	✓	✓
TRAV12-3	TRAJ17	CAMMIKAAGNKLTF	7	✓	✓
TRAV12-3	TRAJ23	CAMMIYNQGGKLIF	7	✓	✓
TRAV13-2	TRAJ53	CAENKNSGGSNYKLTF	7	✓	✓
TRAV17	TRAJ9	CATDGTGGFKTIF	7	✓	✓
TRAV2	TRAJ26	CAVEEDNYGQNFVF	7	✓	✓
TRAV21	TRAJ36	CAVKTGANNLFF	7	✓	✓
TRAV21	TRAJ17	CAVMIKAAGNKLTF	7	✓	✓
TRAV21	TRAJ8	CAVMNTGFQKLVF	7	✓	✓
TRAV21	TRAJ58	CAVKETSGSRLTF	7	✓	
TRAV21	TRAJ42	CAVMNYGGSQGNLIF	7	✓	✓
TRAV21	TRAJ50	CAVMKTSYDKVIF	7	✓	✓
TRAV23/DV6	TRAJ12	CAASRMDSSYKLIF	7	✓	
TRAV29/DV5	TRAJ45	CAASEGGGADGLTF	7	✓	✓
TRAV29/DV5	TRAJ53	CAASGSGGSNYKLTF	7	✓	
TRAV29/DV5	TRAJ43	CAASANNDMRF	7	✓	
TRAV3	TRAJ7	CAVRDDYGNNRLAF	7	✓	
TRAV3	TRAJ8	CAVRDNTGFQKLVF	7	✓	✓
TRAV8-3	TRAJ52	CAVGAAGGTSYGKLTF	7	✓	
TRAV8-3	TRAJ10	CAVGFTGGGNKLTF	7	✓	

Table A.3: Table of putative invariant α chain TCRs from the BLR data. To qualify, sequences had to be present in seven out of ten of our healthy donor samples and not belong to any other pre-existing known or putative invariant TCR group. Presence of these sequences in sorted CD4 and CD8 repertoires (see Section 3.2.1) is indicated.

V	J	CDR3	CB	Adult	BV	BLR	Sum	CD4	CD8
TRAV8-2	TRAJ42	CVVMNYGGSQGNLIF	5	4	9	6	15	✓	
TRAV8-2	TRAJ13	CVVMNSGGYQKVTF	5	4	9	6	15	✓	✓
TRAV8-2	TRAJ36	CVVSDQTGANNLFF	6	3	9	2	11	✓	✓
TRAV21	TRAJ13	CAVMNSGGYQKVTF	5	3	8	5	13	✓	✓
TRAV8-2	TRAJ23	CVVMIYNQGGKLIF	6	2	8	5	13	✓	
TRAV21	TRAJ23	CAVMIYNQGGKLIF	4	4	8	4	12	✓	✓
TRAV8-2	TRAJ32	CVVMNYGGATNKLIF	4	4	8	3	11		
TRAV21	TRAJ20	CAVNDYKLSF	5	3	8	2	10		✓
TRAV21	TRAJ11	CAVMNSGYSTLTF	5	2	7	**8**	15	✓	✓
TRAV21	TRAJ42	CAVMNYGGSQGNLIF	6	1	7	**7**	14	✓	✓
TRAV21	TRAJ36	CAVKTGANNLFF	4	3	7	**7**	14	✓	✓
TRAV21	TRAJ39	CAVMNNNAGNMLTF	5	2	7	6	13	✓	✓
TRAV8-2	TRAJ4	CVVSLSGGYNKLIF	4	3	7	6	13	✓	✓
TRAV8-2	TRAJ7	CVVSDDYGNNRLAF	4	3	7	5	12	✓	✓
TRAV8-2	TRAJ4	CVVSRSGGYNKLIF	4	3	7	5	12	✓	
TRAV8-2	TRAJ3	CVVSEGYSSASKIIF	4	3	7	4	11	✓	
TRAV8-2	TRAJ10	CVVSFTGGGNKLTF	4	3	7	3	10	✓	✓
TRAV8-2	TRAJ13	CVVTLSGGYQKVTF	5	2	7	3	10	✓	
TRAV8-2	TRAJ5	CVVSGDTGRRALTF	2	5	7	3	10	✓	
TRAV8-2	TRAJ12	CVVSRMDSSYKLIF	3	4	7	3	10	✓	
TRAV8-2	TRAJ6	CVVSGSGGSYIPTF	3	4	7	3	10	✓	
TRAV8-2	TRAJ17	CVVMIKAAGNKLTF	6	1	7	3	10	✓	
TRAV21	TRAJ32	CAVMNYGGATNKLIF	4	3	7	2	9	✓	✓
TRAV8-2	TRAJ3	CVVTGYSSASKIIF	4	3	7	2	9		
TRAV8-2	TRAJ4	CVVSFSGGYNKLIF	4	3	7	2	9	✓	
TRAV8-2	TRAJ4	CVVTFSGGYNKLIF	5	2	7	2	9	✓	✓
TRAV8-2	TRAJ11	CVVSVNSGYSTLTF	4	3	7	1	8	✓	
TRAV8-2	TRAJ15	CVVSEGQAGTALIF	4	3	7	1	8	✓	
TRAV8-2	TRAJ13	CVVSYSGGYQKVTF	5	2	7	0	7		
TRAV21	TRAJ15	CAVPNQAGTALIF	3	3	6	**8**	14	✓	✓
TRAV21	TRAJ17	CAVMIKAAGNKLTF	3	3	6	**7**	13	✓	✓
TRAV21	TRAJ8	CAVMNTGFQKLVF	5	1	6	**7**	13	✓	✓
TRAV21	TRAJ58	CAVKETSGSRLTF	2	2	4	**7**	11	✓	
TRAV21	TRAJ50	CAVMKTSYDKVIF	3	1	4	**7**	11	✓	✓

Table A.4: Table of putative invariant α chain TCRs from the BV data. Upper section: VJ-CDR3 sequences that were present in the results of seven or more donors' data produced using the barcoded and targeted V (BV) protocol (see Sections 3.2.2.2 and 5.2.5), with the number of donors given. **Lower section**: Presence of sequences identified in Table A.3, using the barcoded ligation RACE (BLR) strategy (see Sections 3.1.5 and 5.2.6) in the BV data. Column headers: CB = number of cord blood samples containing the sequence; Adult = peripheral blood healthy volunteer samples; BV = total number of samples amplified using the barcoded V protocol with the sequence; BLR = total number of barcoded ligation RACE samples with the sequence (in bold if seven or greater); Sum = total number of donors across both protocols, and CD4 and CD8 ticks designate whether that sequence was detected in the two sorted healthy donor repertoires (see Section 3.2.1).

Gene	Splice	Sequence	Number donors
TRAV1-1	D1A2	TGTTTCCATGAAGATGGGAG|GAGCTCCAGATGAAAGACTC	10
TRAV1-1	D2A2	AGCCATTGTCCAGATAAACT|GAGCTCCAGATGAAAGACTC	10
TRAV1-1	D3A2	GGGTTTTATGGGCTGTCCTG|GAGCTCCAGATGAAAGACTC	10
TRAV1-1	D4A2	CACATTTCTTTCTTACAATG|GAGCTCCAGATGAAAGACTC	8
TRAV8-1	D2A2	GGAACTGTTAATCTCTTCTG|GAAACCCTCTGTGCAGTGGA	1
TRAV8-3	D2A2	GCAACACCTTATCTCTTCTG|GAAACCCTCTGTGCATTGGA	1
TRAV13-2	D2A2	CAAGCAAGAATCTGGAAAAG|CTACTCAACCTGGAGACTCA	1
TRAV14/DV4	D2A2	CAAAGTTATGGTCTATTCTG|AAGGCAAGAAAATCCGCCAA	1
		CCAAGTTATGGTCTATTCTG|AAGGCAAGAAAATCCGCCAA	
TRAV22	D2A2	GTGAGTTGGGGCTGTCCAAG|GGACAAAACAGAATGGAAGA	1
TRAV41	D2A2	CAATCAACTGCAGTTACTCG|GAAAAGCACAGCTCCCTGCA	9
TRDV1	D2A2	GCAGTCACCCTGAACTGCCT|GGTTCTGATGAACAGAATGC	10
TRDV1	D2A3	GCAGTCACCCTGAACTGCCT|AATGCAAAAAGTGGTCGCTA	3
TRBV2	D2A2	TAATCACTTATACTTCTATT|GTTGAAAGGCCTGATGGATCA	10
TRBV7-2	D2A2	GGGAAAGGATGTAGAGCTCA|GGCAACAGTGCACCAGACAAA	9
TRBV7-2	D2A3	GGGAAAGGATGTAGAGCTCA|GTGCACCAGACAAATCAGGGC	9
TRBV7-2	D2A4	GGGAAAGGATGTAGAGCTCA|GAGAGGACTGGGGGATCCGTC	3
TRBV7-3	D2A2	GGGAAAATATGTAGAGCTCA|GGCACGGGTGCGGCAGATGAC	7
TRBV7-3	D2A3	GGGAAAATATGTAGAGCTCA|GTCAGGCCTGAGGGATCCGTC	9
TRBV7-6	D-1A2	GCTCACAGTGACACTGATCT|GAGAGGCCTGAGGGATCCATC	3
TRBV7-6	D1A2	TCCTGGGTTTCCTAGGGACA|GAGAGGCCTGAGGGATCCATC	4
TRBV7-6	D2A2	TGGAGTCTCCCAGTCTCCCA|GAGAGGCCTGAGGGATCCATC	10
TRBV7-6	D3A2	GGGACAGGATGTAGCTCTCA|GAGAGGCCTGAGGGATCCATC	10
TRBV7-8	D1A3	TCCTGGGTTTCCTAGGGACA|GAAAGGCCTGAGGGATCCGTC	8
TRBV7-8	D2A2	AGGACAGGATGTAGCTCTCA|GAATGAAGCTCAACTAGACAA	10
TRBV7-8	D2A3	AGGACAGGATGTAGCTCTCA|GAAAGGCCTGAGGGATCCGTC	9
TRBV7-9	D2A2	GGGACAGAATGTAACTTTCA|GAATGAAGCTCAACTAGAAAA	10
TRBV7-9	D2A3	GGGACAGAATGTAACTTTCA|GAGAGGCCTAAGGGATCTTTC	2
TRBV10-2	D2A2	CAAGATCACAGAGACAGGAA|GATAAAGGAGAAGTCCCCGAT	7
TRBV10-2	D3A3	GAGCCACAGCTATATGTTCT|GATCCAAGACAGAGAATTTCC	3
TRBV11-2	D1A2	CCCTCTGTCTCCTGGGAGCA|GAGAGGCTCAAAGGAGTAGAC	10
TRBV11-2	D2A3	TGGCCATGCTACCCTTTACT|GGCTCAAAGGAGTAGACTCCA	1
TRBV18	D2A2	ATCGGCAGCTCCCAGAGGAA|GAGGGCCCCAGCATCCTGAGG	10

Table A.5: List of intra-V gene splicing events. Paired-end reads for one donor were overlapped and sequences that successfully merged – thus containing whole transcripts – and contained valid rearrangements as determined by `Decombinator` were mapped against the germline loci. Alternative splicing was determined by the presence of sequences mapping to non-canonical TCR exon boundaries, i.e. anything that was not joining L-PART1 to L-PART2 to complete the leader sequence. Splice terminology: the splice donor and acceptor sites of the leader exon were termed D1 and A1, then all novel sites were numbered relatively based on their position. '|' characters indicate splice junctions. Note that there are two TRAV14/DV4 - D2A2 entries, as this region happened to fall on a single nucleotide polymorphism.

B

Appendix B - the effect of HIV infection upon the TCR repertoire

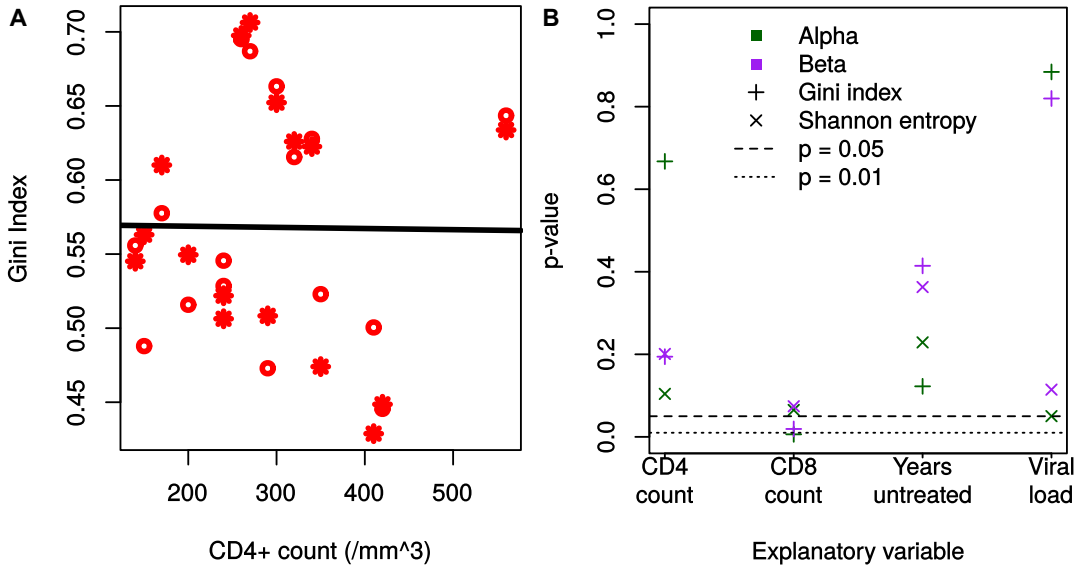


Figure B.1: Clinical score-diversity correaltions. **A**: The CD4 cell count does not correlate with the Gini index of the untreated v1 bleed in HIV patients (R^2 = -0.03317, p = 0.9458). **B**: P-values of the various possible explanatory v1 variables included in a multiple regression, showing that CD8+ lymphocyte count explains the low diversity (by either Gini index or Shannon entropy) better than other variables. Multiple regression analysis performed by group colleague Katharine Best. See Figure 4.1.

Reference	Specificity
Pantaleo et al. 1994 (471)	HIV
Wilson et al. 1998 (242)	HIV
Kolowos et al. 1999 (472)	HIV
Kou et al. 2000 (473)	HIV
Bourcier et al. 2001 (156)	HIV
Lee et al. 2002 (474)	HIV
Dong et al. 2004 (248)	HIV
Gillespie et al. 2006 (274)	HIV
Lichterfeld et al. 2006 (475)	HIV
Meyer-Olson et al. 2006 (476)	HIV
Almeida et al. 2007 (249)	HIV
Lichterfeld et al. 2007 (477)	HIV
Yu et al. 2007 (478)	HIV
Simons et al. 2008 (479)	HIV
Berger et al. 2011 (276)	HIV
Iglesias et al. 2011 (480)	HIV
Chen et al. 2012 (481)	HIV
Conrad et al. 2012b (482)	HIV
Mendoza et al. 2012 (277)	HIV
Stewart-Jones et al. 2012 (483)	HIV
Motozono et al. 2014 (278)	HIV
Sun et al. 2014 (166)	HIV
List collated by Zvyagin et al. 2014 (484)	EBV
List collated by Zvyagin et al. 2014 (484)	CMV

Table B.1: Sources of published viral- or subset-specific CDR3s. Collated CDR3 amino acid sequences are available for download (see Section 2.5) (485)

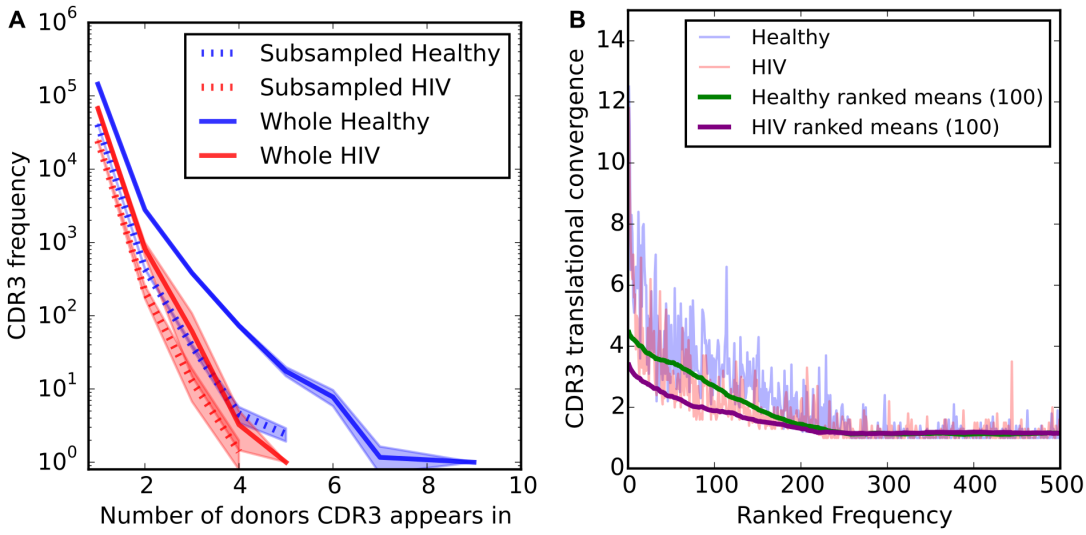


Figure B.2: HIV infection is associated with a loss of publicness and translational convergence in β chain CDR3s. **A**: β CDR3s are less public in HIV-infected than healthy comparator donors, measured by counting in how many donors each group specific (i.e. found in only HIV-infected or healthy donors) CDR3 appears. This applies for both whole samples (solid lines) and a subsample of 5000 CDR3s (dotted lines). To account for the greater number of HIV$^+$ patients than healthy controls, for any given analysis ten HIV samples were randomly selected from the total pool, and publicness measured: this was repeated 100 times, shaded areas represent standard deviation. **B**: β CDR3s display a lower mean translational convergence in HIV$^+$ donors than healthy, i.e. were encoded by fewer nucleotide sequences (and thus recombination events), as determined by counting unique `Decombinator` identifiers that produce the same CDR3. Blue and red lines indicate the mean translational convergence for each group (healthy/HIV respectively) for each CDR3 rank; green and purple lines show a sliding window averaging across 100 ranks. See Figure 4.3A and C.

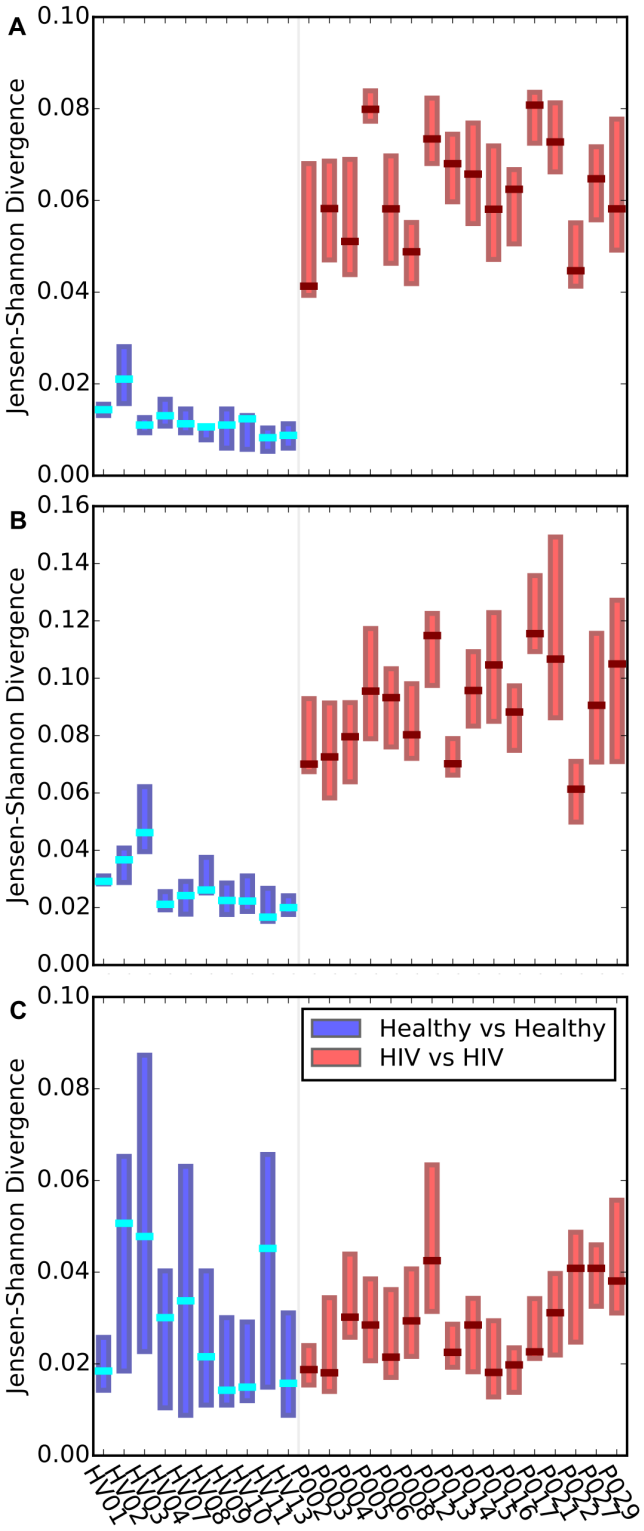


Figure B.3: Healthy donor and HIV-patient gene usage divergence. Barplots showing the difference in TRAJ (**A**), TRBV (**B**) and TRBJ (**C**) usage, by Jensen-Shannon Divergence (JSD). Donor-specific proportional gene usage matrices, before each pair of healthy or HIV$^+$ donors were compared. Only interquartile ranges and medians are shown. See Figure 4.3D.

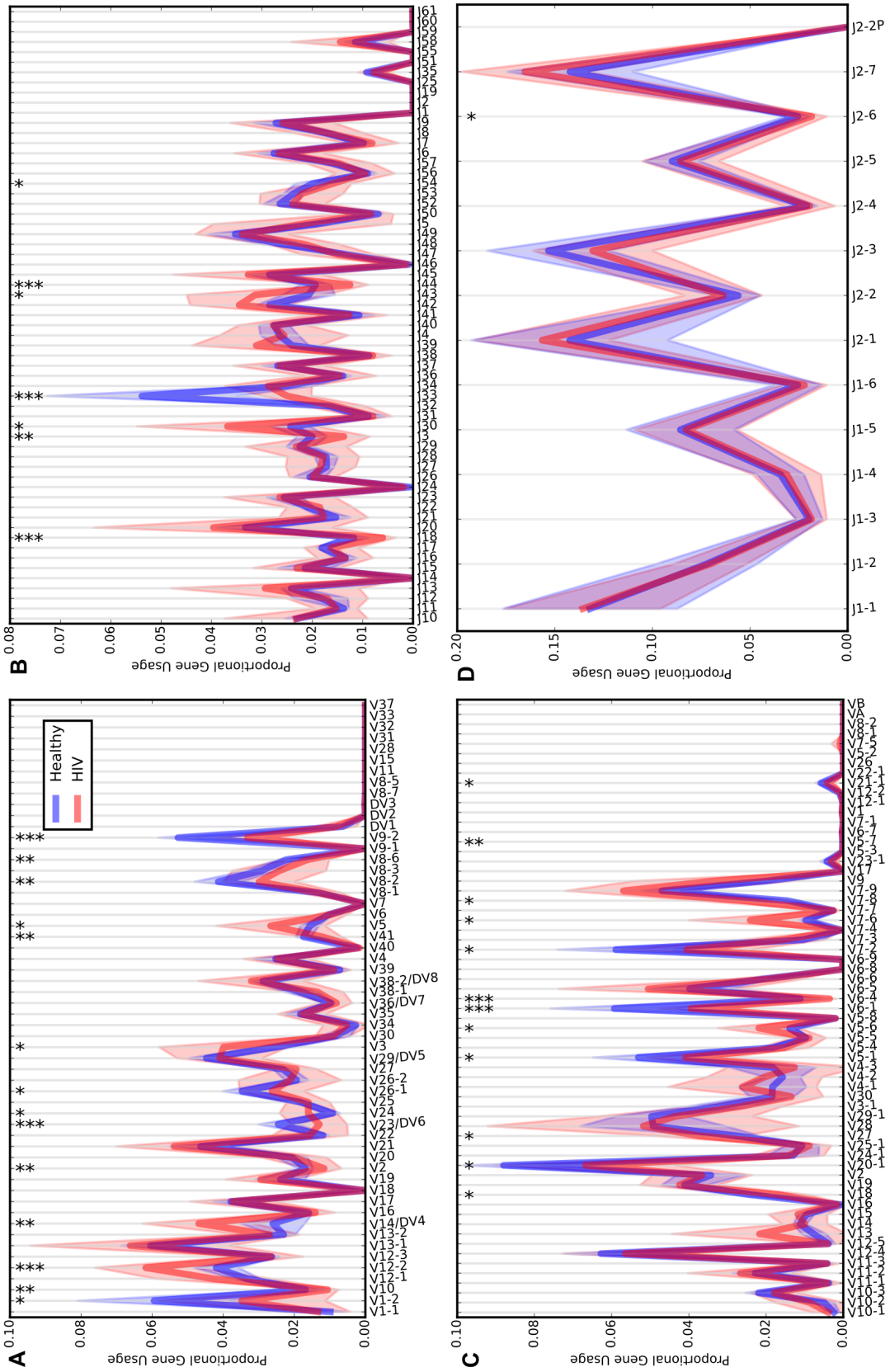


Figure B.4: Raw TCR proportional gene usage in healthy and HIV-infected samples. Raw mean (solid lines) and standard deviations (shaded areas) for proportional gene usage of TRAV (**A**), TRAJ (**B**), TRBV (**C**) and TRBJ (**D**) TCR genes. Asterisks indicate significance comparing all healthy donor values to HIV-patient values for that gene in an unpaired Wilcoxon rank sum test.

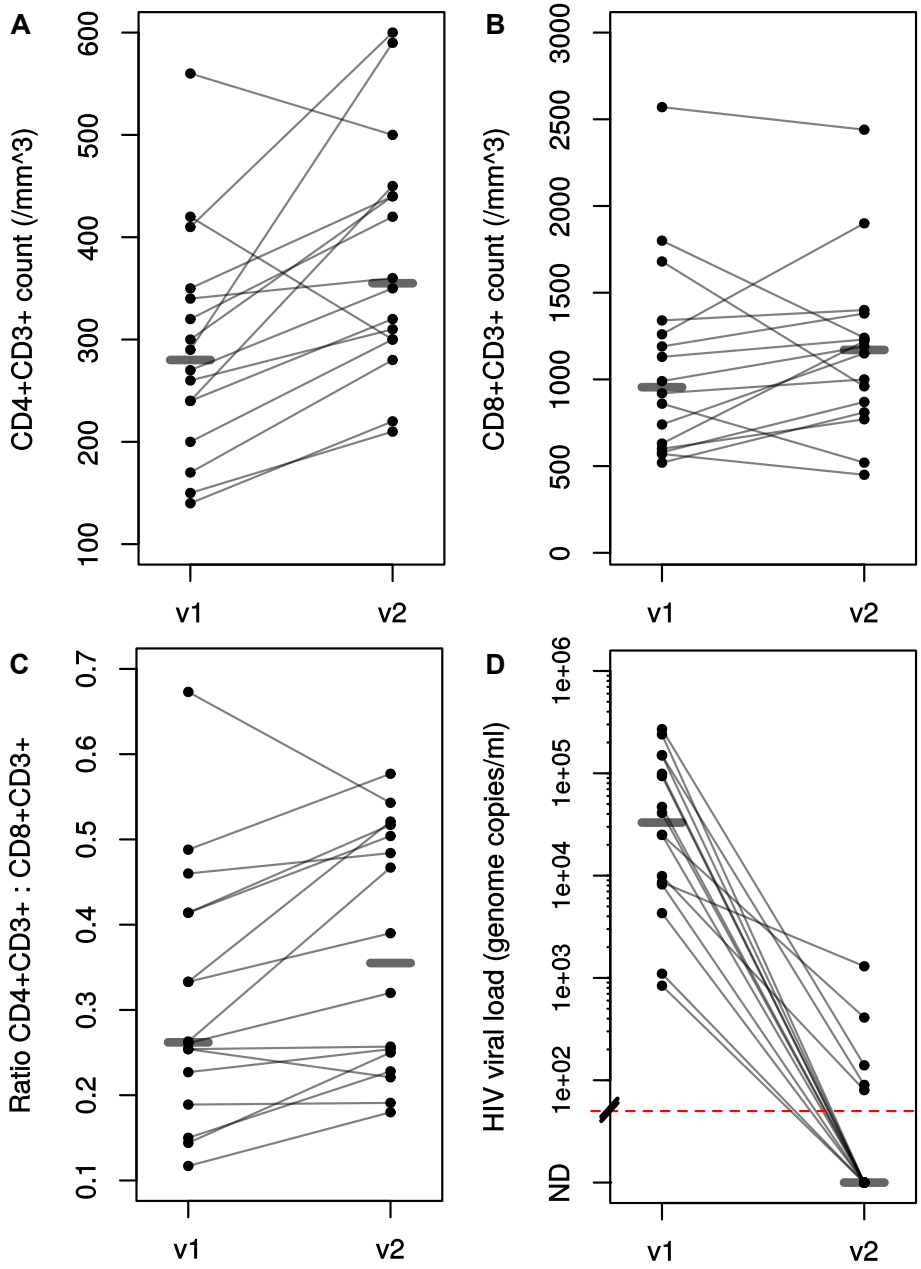


Figure B.5: Clinical features of the HIV patient samples. **A**: Change in absolute count of $CD3^+CD4^+$ lymphocytes, which significantly increased during this course of ART (p = 0.0039). **B**: Change in absolute count of $CD3^+CD8^+$ lymphocytes; no significant change observed (p = 0.17). **C**: Change in CD4:CD8 cell ratio; a significant increase was obvserved (p = 0.0046). **D**: Significant decrease in viral load on therapy, as measured by qPCR (p = 1.526e-05). ND = not detectable, which corresponds to 50 copies or lower. All p-values in this figure measured by paired, one-sided Wilcoxon signed rank tests.

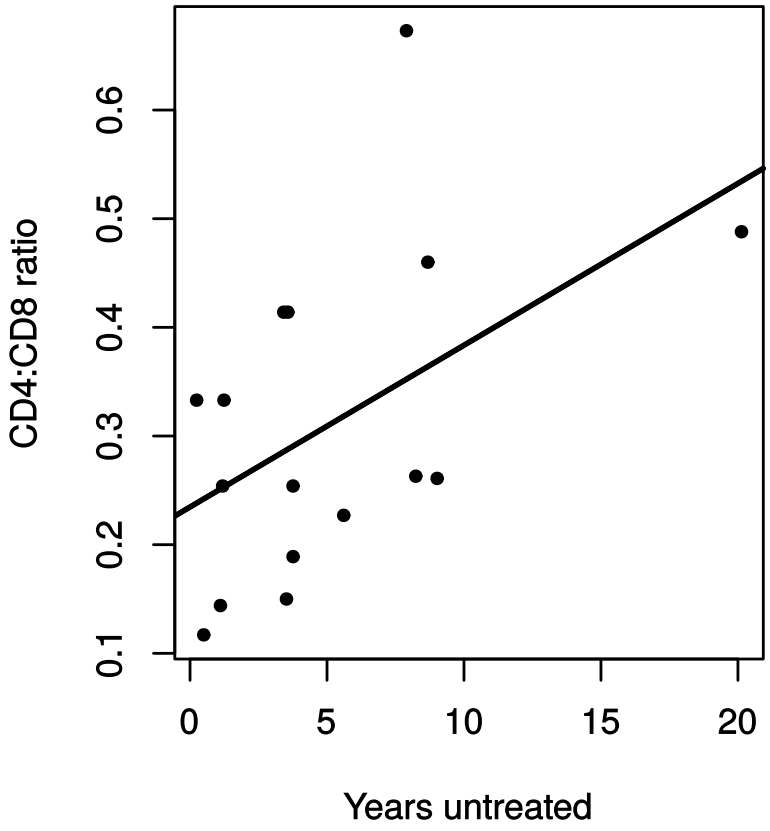


Figure B.6: Relationship between years before treatment and CD4:CD8 ratio. An unexpected feature of our patient data is that there is a correlation between the amount of time passed before starting antiretroviral treatment and the CD4:CD8 ratio at time of start therapy ($R^2 = 0.1967$, $p = 0.04844$).

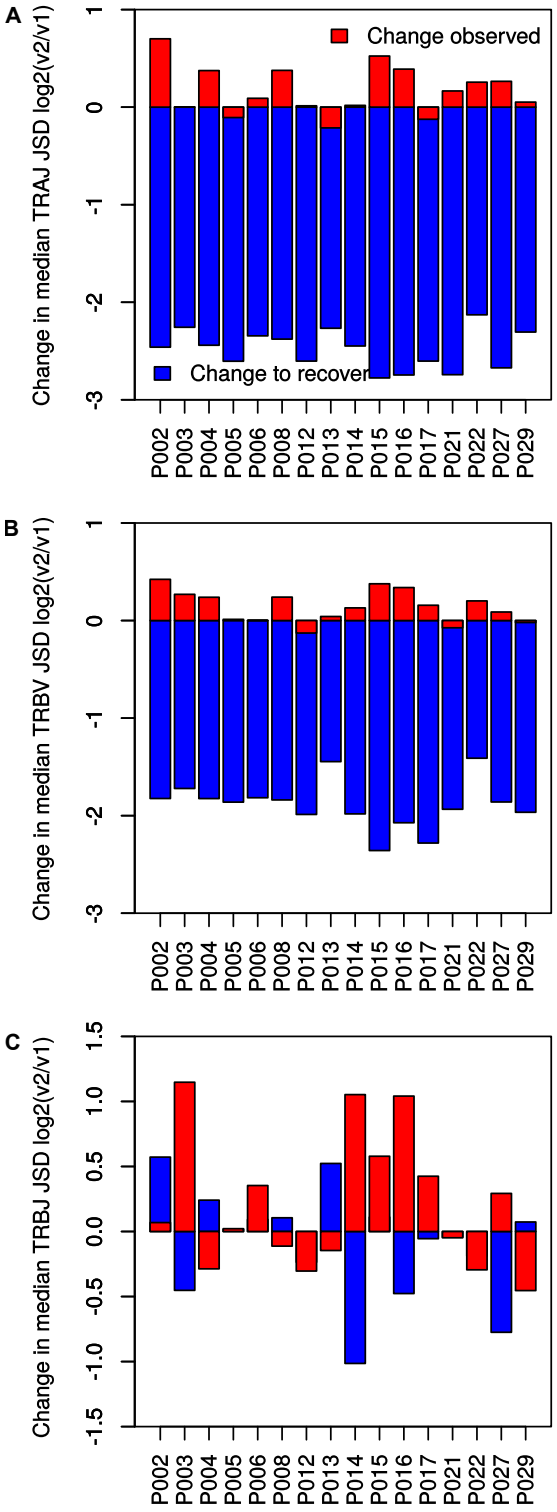


Figure B.7: Change in HIV patients' gene usage divergence in the first three months of ART. Log2 fold change in median TRAJ (**A**), TRBV (**B**) and TRBJ (**C**) Jensen-Shannon Divergence (JSD) from v1 to v2 (red). Blue bars indicate the log2 change in median JSD that would be required for that patient to reach the mean value of the median JSDs for the ten healthy controls, i.e. return them to the average level of divergence healthy controls display in Figure B.3. See Figure 4.4B.

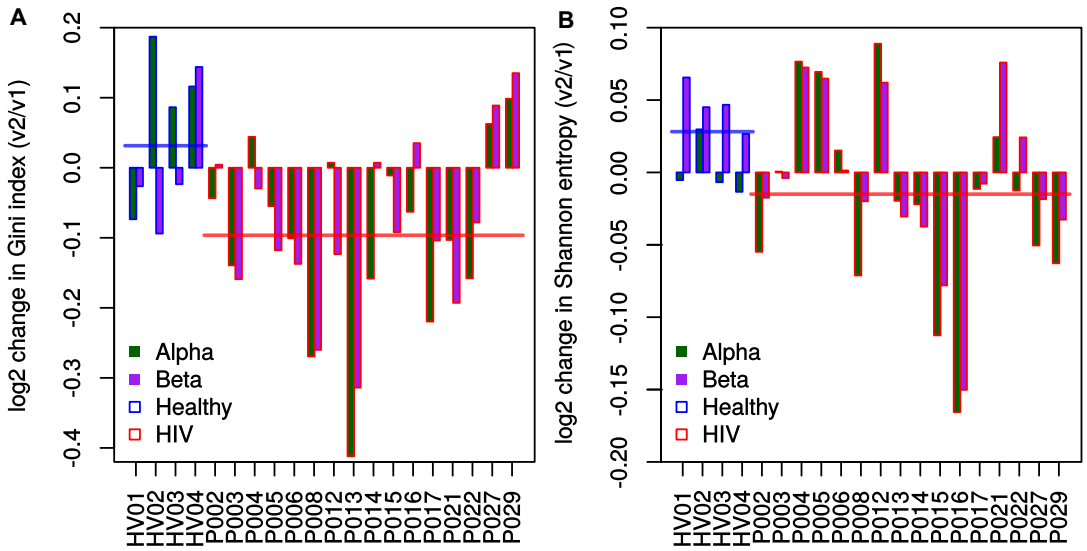


Figure B.8: HIV patients' change in diversity indexes on ART. A and **B**: Log2 fold-change in whole CDR3 repertoire Gini indexes and Shannon entropies from v1 to v2. Horizontal lines indicate group (healthy/HIV) medians. HIV$^+$ patients show a significantly lower median fold change in Gini index compared to healthy controls ($p = 0.005427$, unpaired one-sided Wilcoxon test). See Figure 4.4C and D.

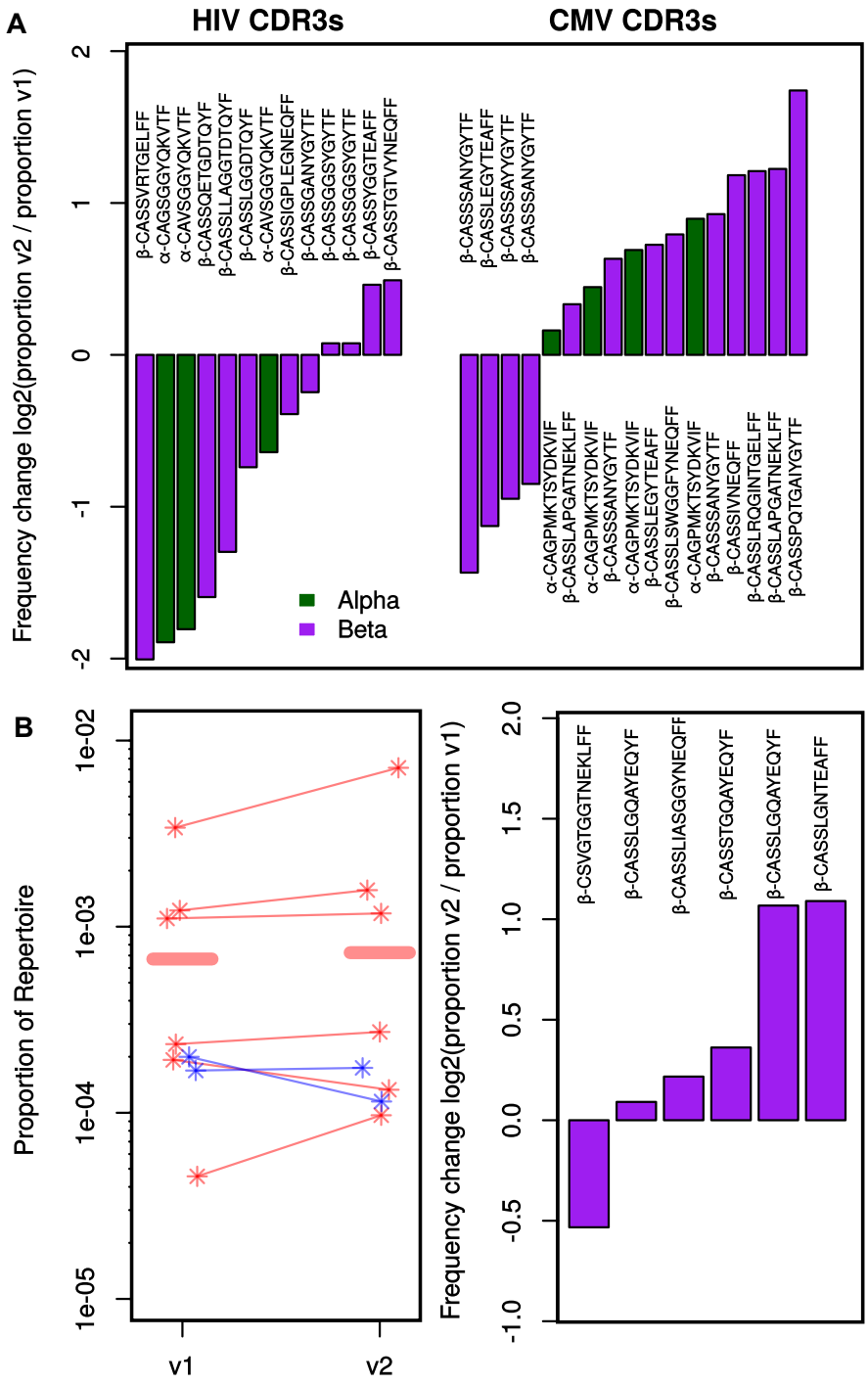


Figure B.9: HIV-associated CDR3s decrease in frequency on ART, while CMV- and EBV-associated CDR3s increase. **A**: Log2 fold-change in frequency of HIV- and CMV-associated CDR3s detected in all of the HIV-infected donors. Most HIV-related CDR3s decreased proportionally (9/13, left), while the majority of CMV-related CDR3s increased (13/17, right). CDR3s were only included in this analysis if they occurred at least twice in both v1 and v2 bleeds of a donor. **B**: Left: proportional frequencies of the published, EBV-specific CDR3s that were also found in our whole CDR3 repertoires pre- and post-treatment. Right: Log2 fold-change in frequency of EBV-implicated CDR3s. EBV-associated responses were found in five HIV$^+$ and two uninfected donors. See Figure 4.5.

C

Appendix C - developing a TCR sequencing pipeline

C.1 Previously published datasets

The earliest of the three publicly available data sets used in this thesis was produced by Wang *et al* (83). This paper sequenced the repertoires of various T-cell subsets, but I have here only analysed their 'pan T-cell' group in which $CD3^+$ peripheral blood mononuclear cells (PBMCs) were obtained using magnetic assorted cell sorting (MACS). As with all of the papers with available datasets, RNA was chosen as the input nucleic acid, and so cells were lysed, RNA extracted and mRNA reverse transcribed into cDNA. TCR rearrangements were then amplified with a so-called amplicon rescued multiplex (ARM)-PCR, which is a modified version of the Templex PCR (TEM-PCR) system developed by the same group (486). In these protocols, there is first a multiplex PCR using primers directed against the constant region (of the alpha and beta constant regions in this case) and against the different V genes. In this first PCR, there are actually two nested sets of primers for each target (V and C regions), with the inner set at a higher concentration: the lower-concentration out primers purportedly enrich the target molecules for the inner primers in the early cycles of the reaction (487). The inner primers contain one of a pair of 'universal' priming sites 5' of the V or C specific sections. Amplicons are then 'rescued' (an ambiguous term which presumably reflects a sampling, purification or dilution event) and amplified with primers directed against the universal sites, allowing amplification of all TCR rearrangements (that have been labelled with these universal sites in the previous PCR) using the same primers. The authors claim that this produces semi-quantitative results, as all rearrangements have been amplified using the same primers for the exponential phase of the PCR, however as the patent describing their technique reports that 10 to 15 cycles are used in the first PCR this statement is open to debate (488). Libraries produced by ARM-PCR were then sequenced on the Roche 454 platform.

The second dataset came from Warren *et al* (77). These authors reverse transcribe lysed PBMC RNA using TRBC specific primers, and perform a template switch reac-

tion. Next they amplified from their TRBC primer to their TS primer in a first round of PCR, and purified rearranged beta chain transcripts by purifying excised bands from agarose gel electrophoresis. A second round of PCR was performed using nested TS and TRBC primers, the latter of which contained 5' biotin moieties. The products of this PCR were then purified and randomly sheared by sonication, before using streptavidin beads to purify sequences that contain the biotinylated TRBC sequence; after another gel excision to size-select the authors have thus obtained shorter fragments (125-175bp) that contain just the TRBC proximal sequence, which contains the V(D)J information. These amplicons then underwent standard adapter-ligation library preparation (illustrated in Figure 5.3), and were sequenced on the Illumina GAIIx platform. Note that the sonication step was required to bring the amplicon lengths within the parameter range that could be sequenced on this platform at this time.

The final paper that has uploaded appropriate was from Bolotin *et al* (126). This paper actually tried several different combinations of protocols, however I have only used the data from one of these in my comparisons. In this technique PBMC RNA was reverse-transcribed and amplified using a template-switch approach as was performed by Warren *et al*. A nested PCR was then performed on the products of this PCR, in which 5' overhanging sequences added adapter regions for 454 sequencing.

There were two additional data sets that were available during the time frame in which we were developing and testing our own protocols, however these were not suitable for comparison with my data. Freeman *et al* (75) produced very short read data, which would need to be assembled into longer contigs in order for to `Decombinator` to process it, while Mamedov *et al* (97) sequenced blood from a patient with ankylosing spondylitis before and after stem cell therapy, and thus would likely not produce repertoires consistent with those of typical, healthy individuals.

C.2 Protocol development supplementary figures

A **>D7β1**

TCACACAGGAAACAGCTATGACTTTTTTTTTTTTTTTTTTTTTTTTGCATGCTCACAGAGGGCCTGGTCTAGA
ATATTCCACATCTGCTCTCACTCTGCCATGGACTCCTGGACCTTCTGCTGTGTGTCCCTTTGCATCCTGGTA
GCGAAGCATACAGATGCTGGAGTTATCCAGTCACCCCGCCATGAGGTGACAGAGATGGGACAAGAAGTGACT
CTGAGATGTAAACCAATTTCAGGCCACAACTCCCTTTTCTGGTACAGACAGACCATGATGCGGGGACTGGAG
TTGCTCATTTACTTTAACAACAACGTTCCGATAGATGATTCAGGGATGCCCGAGGATCGATTCTCAGCTAAG
ATGCCTAATGCATCATTCTCCACTCTGAAGATCCAGCCCTCAGAACCCAGGGACTCAGCTGTGTACTTCTGT
GCCAGCAGTGACGATTCTAGTCTTCGGCCCCAGCATTTTGGTGATGGGACTCGACTCTCCATCCTAGAGGAC

B **>D7β2**

GTCATAGCTGTTTCCTGTGTGATTTTTTTTTTTTTTTTTTTTTTTGATGCTCACAGAGGGCCTGGTCTAGA
ATATTCCACATCTGCTCTCACTCTGCCATGGACTCCTGGACCTTCTGCTGTGTGTCCCTTTGCATCCTGGTA
GCGAAGCATACAGATGCTGGAGTTATCCAGTCACCCCGCCATGAGGTGACAGAGATGGGACAAGAAGTGACT
CTGAGATGTAAACCAATTTCAGGCCACAACTCCCTTTTCTGGTACAGACAGACCATGATGCGGGGACTGGAG
TTGCTCATTTACTTTAACAACAACGTTCCGATAGATGATTCAGGGATGCCCGAGGATCGATTCTCAGCTAAG
ATGCCTAATGCATCATTCTCCACTCTGAAGATCCAGCCCTCAGAACCCAGGGACTCAGCTGTGTACTTCTGT
GCCAGCAGTTTAGGGGACGCCAAGGAAAAACTGTTTTTTGGCAGTGGAACCCAGCTCTCTGTCTTGGAGGAC

C **>D7α**

TCACACAGGAAACAGCTATGACTTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCACTTTGTCGCCCAG
GCTGGAGTGCAGTGGCGCGATCTCTGCTCACTGCAAACTCCGCCTCCCGGGTTCCCGCCATTCTCCTGCCTC
AGCCTCTCGAGTAGCTGGGACTACAGGCGCCCGCCACAGCGCCCAGCTAATTTTTTTGTATTTTTTGGTAGA
GACGGGGTTTCACCGTGTTAGCCAGGATGGTCTCGATCTCCTGACCTCGTGATCCGCCCACCTCGGCCTCCC
AAAGCGCTGGGATTAAAGGCGTGAGCCACCGCGCCCGGCCTACTTTGTCTTATGTTATTCCCATTTGCCGTC
TCTGTTCCTTATACATTATGCTTTTTCAACTTTACCAGAATCACTTGGATTAAAACCCGTGGATTTCTCAGT
AGGAAATGTTCATGTGAAGACACTTCTGTAGTAACAGAACCTACAGCTGCTCCTGTAGAAGGAAGTTGAAAG
TCATCCCTTCAAGAAAGGGGCTCCTCCCCTTGTAATTCTACTGGGTTTTGCATCCAGACTGAGTTTCCTTCC
CTCACCCACATGAAGTGTCTACCTTCTGCAGACTCCAATGGCTCAGGAACTGGGAATGCAGTGCCAGGCTCG
TGGTATCCTGCAGCAGATGTGGGGAGTTTTCCTTCTTTATGTTTCCATGAAGATGGGAGGCACTACAGGACA
AAACATTGACCAGCCCACTGAGATGACAGCTACGGAAGGTGCCATTGTCCAGATCAACTGCACGTACCAGAC
ATCTGGGTTCAACGGGCTGTTCTGGTACCAGCAACATGCTGGCGAAGCACCTACATTTCTGTCTTACAATGT
TCTGGATGGTTTGGAGGAGAAAGGTCGTTTTTCTTCATTCCTTAGTCGGTCTAAAGGGTACAGTTACCTCCT
TTTGAAGGAGCTCCAGATGAAAGACTCTGCCTCTTACCTCTGTGCTGCCGTGGATAGCAACTATCAGTTAAT
CTGGGGCGCTGGGACCAAGCTAATTATAAAGCCAGATATCCAGAA

D **>D8α**

ACGGCAGGGTCAGGGTTCTGCTGAGGTGCAACTATTCATCGTCTGTTCCACCATATCTCTTCTGGTATGTGC
AATACCCCAACCAAGGACTCCAGCTTCTCCTGAAGTACACATCAGCGGCCACCCTGGTTAAAGGCATCAACG
GTTTTGAGGCTGAATTTAAGAAGAGTGAAACCTCCTTCCACCTGACGAAACCCTCAGCCCATATGAGCGACG
CGGCTGAGTACTTCTGTGCTGTGAGTGATCTCGAACCGAACAGCAGTGCTTCCAAGATAATCTTTGGATCAG
GGACCAGACTCAGCATCCGGCCAAATATCCAGAACCCTGATCCTGCCCT

L region V region D region J region C region Poly-T primer

E

```
aRC1                    ACGGCAGGGTCAGGGTTCTGGATAT
D8α                     ACGGCAGGGTCAGGGTTCTGCTGAGGTGCAACTATTCAT...
GermlineTRAV8-4      ...CTGAAGGAGCCCTGGTTCTGCTGAGGTGCAACTACTCAT...
```

Figure C.1 ***(preceding page)*****: Example amplicons cloned from the DNA amplified from two healthy donors' PBMC using our poly-T RACE PCR.** Bands both expected and unexpected sizes were selected. TCR genes used in these rearrangements are **A** (D7β1): TRBV12-3 – TRBD2 – TRBJ1-5; **B** (D7β2): TRBV12-3 – TRBD1 – TRBJ1-4; **C** (D7α): TRAV1-2 – TRAJ33; **D** (D8α): TRAV8-4 – TRAJ33. Bold residues indicate mismatches relative to expected sequences. A and C contain the expected number of thymidines in their RACE homopolymer (23). B contains 24, while a clone relating to a non-productive rearrangement (not shown) was found to contain 25. E shows the overlap between the 5' end of the D8α clone, αRC1 primer and germline TRAV8-4 sequences, revealing the presumed exonucleation of the primer and subsequent mis-priming. Note that **C** is a MAIT-cell TCR sequence (see Section 3.1.5).

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
Wang1α	37,097	63.63%	0.26%	36.2	1.77×10^{-4}	99.98%
Wang1β	37,097	30.92%	1.28%	36.0	2.63×10^{-4}	99.97%
Wang2α	15,614	62.86%	0.32%	36.5	1.52×10^{-4}	99.98%
Wang2β	15,614	32.22%	0.75%	36.3	2.42×10^{-4}	99.98%
Wang3α	143,944	55.59%	0.88%	36.7	1.40×10^{-4}	99.99%
Wang3β	143,944	37.11%	4.06%	36.6	1.97×10^{-4}	99.98%
Warren1β	1,893,628	80.15%	1.61%	36.8	3.92×10^{-4}	99.96%
Warren2β	1,625,571	79.71%	0.11%	36.7	3.36×10^{-4}	99.97%
Warren3β	1,794,595	80.33%	0.02%	36.8	3.04×10^{-4}	99.97%
Bolotin1β	12,300	59.46%	1.10%	26.1	2.44×10^{-3}	99.76%
Bolotin2β	10,990	56.93%	0.78%	25.7	2.55×10^{-3}	99.74%
Bolotin3β	8,347	44.79%	1.71%	24.1	3.22×10^{-3}	99.68%

Table C.1: Overview read statistics for three sets of publicly available healthy human datasets that were available from papers published prior to completion of our own protocols. FASTQ files were downloaded from the Sequence Read Archive (SRA), with accession numbers in the displayed order: SRR030416, SRR030417 and SRR030709 (83); SRR060699, SRR060700 and SRR060701 (77); SRR494403, SRR494404 and SRR494405 (126). For each paper, the three files represent separate amplifications from different aliquots of either blood or RNA, from the same bleed of one donor. Both Wang *et al* and Bolotin *et al* FASTQ files were produced using 454 technology, while the Warren *et al* paper used the Illumina Genome Analyser (GA) IIx, the precursor machine to the HiSeq. Note that the Bolotin *et al* paper also published sequences from comparable libraries sequenced on both the GAIIx and Ion Torrent platforms, but were not included in this analysis; the reads from their GAIIx run were too short (72bp) for reliable TCR assignation by `Decombinator`, and the V gene distributions of their Ion Torrent data were inconsistent with the other platforms, suggesting a possible bias.


A >Pa1
CAAGCAGAAGACGGCATACGAGATGGCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCAATCTTTTTTTTTTTTTTTTTTTTTTGTTCT
GCAGCTGCAATGGGTGAGCAGCAAACAGGAGGTGACGCAGATTCCTGCAGCTCTGAGTGTCCCAGAAGGAGAAAACTTGGTTCTCAACTG
CAGTTTCACTGATAGCGCTATTTACAACCTCCAGTGGTTTAGGCAGGACCCTGGGAAAGGTCTCACATCTCTGTTGCTTATTCAGTCAAG
TCAGAGAGAGCAAACAAGTGGAAGACTTAATGCCTCGCTGGATAAATCATCAGGACGTAGTACTTTATACATTGCAGCTTCTCAGCCTGG
TGACTCAGCCACCTACCTCTGTGCTGTAGACGACTACAAGCTCAGCTTTGGAGCCGGAACCACAGTAACTGTAAGAGCAAATATCCAGAA
CCCTGACCCTGCCGTAGTGTAGATCTCGGTGGTCGCCGTATCATT
B >Pa2
CAAGCAGAAGACGGCATACGAGATGGCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTGCAGCTCAG
ACACAGCGAAAGAGCCTAGAACCTGGGTCCTAGTTTGCACCTAGAATATGAGGTACTTTGAGCCTTGCTAAGACCACCCAGCCCATCTCC
ATGGACTCATATGAAGGACAAGAAGTGAACATAACCTGTAGCCACAACAACATTGCTACAAATGATTATATCACGTGGTACCAACAGTTT
CCCAGCCAAGGACCACGATTTATTATTCAAGGATACAAGACAAAAGTTACAAACGAAGTGGCCTCCCTGTTTATCCCTGCCGACAGAAAG
TCCAGCACTCTGAGCCTGCCCCGGGTTTCCCTGAGCGACACTGCTGTGTACTACTGCCTCGTGGGTGACTGGGATAACTATGGTCAGAAT
TTTGTCTTTGGTCCCGGAACCAGATTGTCCGTGCTGCCCTATATCCAGAACCCTGACCCTGCCGTAGTGTAGATCTCGATGGTCGCCGTA
TCATT
C >Pa3
CAAGCAGAAGACGGCATACGAGATGGCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTGGCC
ATGGCTTTGCAGAGCACTCTGGGGGCGGTGTGGCTAGGGCTTCTCCTCAACTCTCTCTGGAAGGTTGCAGAAAGCAAGGACCAAGTGTTT
CAGCCTTCCACAGTGGCATCTTCAGAGGGAGCTGTGGTGGAAATCTTCTGTAATCACTCTGTGTCCAATGCTTACAACTTCTTCTGGTAC
CTTCACTTCCCGGGATGTGCACCAAGACTCCTTGTTAAAGGCTCAAAGCCTTCTCAGCAGGGACGATACAACATGACCTATGAACGGTTC
TCTTCATCGCTGCTCATCCTCCAGGTGCGGGAGGCAGATGCTGCTGTTTACTACTGTGCTGTGGTAGGAGGAAGCTACATACCTACATTT
GGAAGAGGAACCAGCCTTATTGTTCATCCGTATATCCAGAACCCTGACCCTGCCGTAGTGTAGATCTCGGTGGTCGCCGTATCATT
D >Pa7
CAAGCAGAAGACGGCATACGAGATGGCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTGGCA
TTTTGGCCATGGCTTTGCAGAGCACTCTGGGGGCGGTGTGGCTAGGGCTTCTCCTCAACTCTCTCTGGAAGGTTGCAGAAAGCAAGGACC
AAGTGTTTCAGCCTTCCACAGTGGCATCTTCAGAGGGAGCTGTGGTGGAAATCTTCTGTAATCACTCTGTGTCCAATGCTTACAACTTCT
TCTGGTACCTTCACTTCCCGGGATGTGCACCAAGACTCCTTGTTAAAGGCTCAAAGCCTTCTCAGCAGGGACGATACAACATGACCTATG
AACGGTTCTCTTCATCGCTGCTCATCCTCCAGGTGCGGGAGGCAGATGCTGCTGTTTACTACTGTGCTGTGGCCTCTTCTTCTGGTTCTG
CAAGGCAACTGACCTTTGGATCTGGGACACAATTGACTGTTTTACCTGATATCCAGAACCCTGACCCTGCCGTAGTGTAGATCTCGGTGG
TCGCCGTATCATT
E >Pb1
CAAGCAGAAGACGGCATACGAGATACTTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTGGGT
GTGAGGCCATCACGGAAGATGCTGCTGCTTCTGCTGCTTCTGGGGCCAGGCTCCGGGCTTGGTGCTGTCGTCTCTCAACATCCGAGCTGG
GTTATCTGTAAGAGTGGAACCTCTGTGAAGATCGAGTGCCGTTCCCTGGACTTTCAGGCCACAACTATGTTTTGGTATCGTCAGTTCCCG
AAACAGAGTCTCATGCTGATGGCAACTTCCAATGAGGGCTCCAAGGCCACATACGAGCAAGGCGTCGAGAAGGACAAGTTTCTCATCAAC
CATGCAAGCCTGACCTTGTCCACTCTGACAGTGACCAGTGCCCATCCTGAAGACAGCAGCTTCTACATCTGCAGTGCCATTCCGGGACTA
GCGGGAAGGGGCGAGCAGTACTTCGGGCCGGGCACCAGGCTCACGGTCACAGAGGACCTGAACAAGGTGTTCCCACCAGTGTAGATCTCG
GTGGTCGCCGTATCATT
F >Pb3
CAAGCAGAAGACGGCATACGAGATACTTGAGTGACTGGAGTTCGGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTGCTG
AGCACAGACACAGTGCTGCCTGCCCCTTTGTGCCATGGGCTCCAGGCTGCTCTGTTGGGTGCTGCTTTGTCTCCTGGGAGCAGGCCCAGT
AAAGGCTGGAGTCACTCAAACTCCAAGATATCTGATCAAAACGAGAGGACAGCAAGTGACACTGAGCTGCTCCCCTATCTCTGGGCATAG
GAGTGTATCCTGGTACCAACAGACCCCAGGACAGGGCCTTCAGTTCCTCTTTGAATACTTCAGTGAGACACAGAGAAACAAAGGAAACTT
CCCTGGTCGATTCTCAGGGCGCCAGTTCTCTAACTCTCGCTCTGAGATGAATGTGAGCACCTTGGAGCTGGGGGACTCGGCCCTTTATCT
TTGCGCCAGCACAAGGACAGCCGGGCAATCTGGCTACACCTTCGGTTCGGGGACCAGGTTAACCGTTGTAGAGGACCTGAACAAGGTGTT
CCCACCAGTGTAGATCTCGGTGGTCGCCGTATCATT
G >Pb6
CAAGCAGAAGACGGCATACGAGATACTTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTTTGC
CCATCCTGCCCTGACCCTGCCATGGGCACCAGCCTCCTCTGCTGGATGGCCCTGTGTCTCCTGGGGGCAGATCACGCAGATACTGGAGTC
TCCCAGGACCCCAGACACAAGATCGCTAAGAGGGGACAGAATGTAACTTTCAGGTGTGATCCAATTTCTGAACACAACCGCCTTTATTGG
TACCGACAGACCCTGGGGCAGGGCCCAGAGTTTCTGACTTACTTCCAGAATGAAGCTCAACTAGAAAAATCAAGGCTGCTCAGTGATCGG
TTCTCTGCAGAGAGGCCTAAGGGATCTTTCTCCACCTTGGAGATCCAGCGCACAGAGCAGGGGGACTCGGCCATGTATCTCTGTGCCAGC
AGCTTAGTAAGCGGGAAAGGGTTCGAGCAGTACTTCGGGCCGGGCACCAGGCTCACGGTCACAGAGGACCTGAACAAGGTGTTCCCACCA
GTGTAGATCTCGGTGGTCGCCGTATCATT
H >DD21
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTTCCCGATGAAGGTATTGTGGAAG
AAACTACACTATCGGCATTACGCGTGGCCAATGAGGTCAATGAGCTACGCATTGAATACGAATCCGCTATAAAATCCGGGTTTTCTGCCT
TTTCCACCTTTGTTAGGCATCGCCACGCCGAATGGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTAT
CATT
L region
V region
J region
C region
SP1/SP2
Index
P5/P7

Figure C.2 ***(preceding page)*****: Sanger sequenced clones produced using the stepwise PCR protocol.** Presence and orientation of sequences indicates that the stepwise PCR protocol is adding the required elements in the correct positions, and thus theoretically libraries should be able to be sequenced. TCR genes used in each rearrangement are **A**: TRAV21 – TRAJ20; **B**: TRAV4 – TRAJ26; **C**: TRAV2 – TRAJ6; **D**: TRAV2 – TRAJ22; **E**: TRBV20-1 – TRBJ2-7; **F**: TRBV5-1 – TRBJ1-2; **G**: TRBV7-9 – TRBJ2-7. **H** shows an example sequence cloned from a non-TCR library (genomic varicella-zoster virus) that sequenced with expected run output, to demonstrate matching adapter sequences. Note that SP1 sequences are absent in A-G as these clones were produced with the initial protocol, in which custom sequencing oligonucleotides were targeted against the end of the constant regions (underlined). Sequencing of clones from later libraries showed that the SP1 element was similarly correctly incorporated (analysis performed by other laboratory members, data not shown). Poly-T homopolymer lengths for clones A to H (not counting the thymidine that makes up the ultimate base of the SP2 sequence) are: 21; 17; 22; 22; 22; 22 and 24 (compared to the 23 T residues found in the primer). Base-calls that differ to reference sequences are in bold. VZV library kindly donated by Dr. Daniel Depledge.

Sample	Reads	Extended	%
D2	10,875,543	1,943,685	17.87%
D3 naïve	17,493,702	2,207,696	12.62%
D3 memory	16,410,241	1,786,562	10.89%
D4 naïve	13,068,337	1,988,492	15.22%
D5 memory	10,553,778	1,702,220	16.13%

Table C.2: Merging HiSeq reads. Number and percentages of HiSeq FASTQ reads that were successfully merged using `FLASH`, which uses overlapping portions of paired-end data to recapitulate entire target amplicons contained within the sequencing primers.

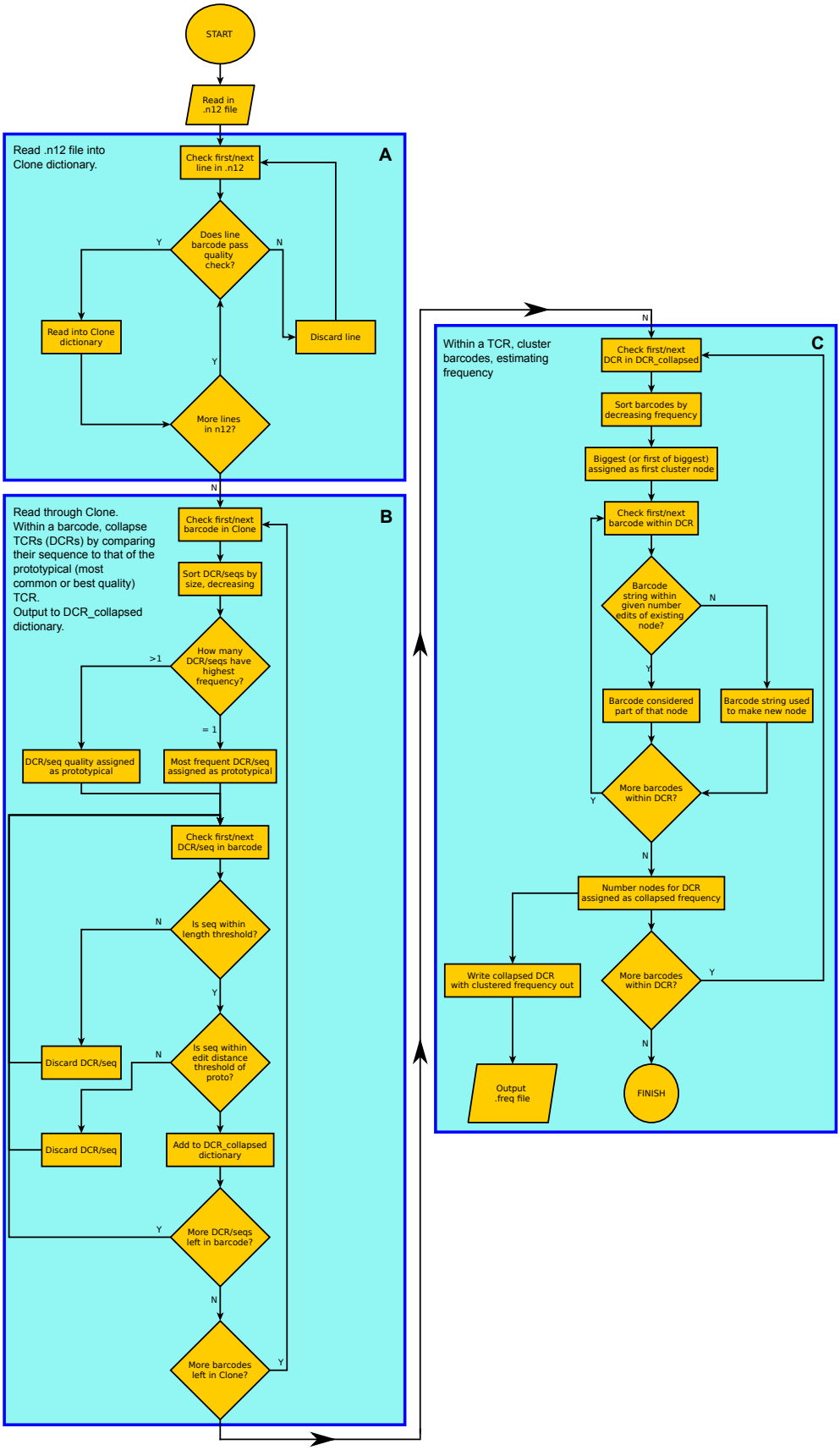

START
Read in .n12 file
A
Read .n12 file into Clone dictionary.
Check first/next line in .n12
Does line barcode pass quality check?
Y
N
Read into Clone dictionary
Discard line
More lines in n12?
B
Read through Clone. Within a barcode, collapse TCRs (DCRs) by comparing their sequence to that of the prototypical (most common or best quality) TCR. Output to DCR_collapsed dictionary.
Check first/next barcode in Clone
Sort DCR/seqs by size, decreasing
How many DCR/seqs have highest frequency?
>1
= 1
DCR/seq quality assigned as prototypical
Most frequent DCR/seq assigned as prototypical
Check first/next DCR/seq in barcode
Is seq within length threshold?
Discard DCR/seq
Is seq within edit distance threshold of proto?
Discard DCR/seq
Add to DCR_collapsed dictionary
More DCR/seqs left in barcode?
More barcodes left in Clone?
C
Within a TCR, cluster barcodes, estimating frequency
Check first/next DCR in DCR_collapsed
Sort barcodes by decreasing frequency
Biggest (or first of biggest) assigned as first cluster node
Check first/next barcode within DCR
Barcode string within given number edits of existing node?
Barcode considered part of that node
Barcode string used to make new node
More barcodes within DCR?
Number nodes for DCR assigned as collapsed frequency
Write collapsed DCR with clustered frequency out
More barcodes within DCR?
Output .freq file
FINISH

Figure C.3 ***(preceding page)*****: Flowchart overview of how the error- and frequency-correcting code of CollapseTCR works.** **A**: Files decombined with `vDCR` are read into memory; all rearrangements whose barcode is of sufficiently high FASTQ quality is stored in the Clone dictionary. **B**: The script loops through all the entries in the Clone dictionary, which looks within each barcode sequence to establish what the 'prototypical' rearrangement is, i.e. the most frequent, which is assumed to represent the true receptor chain. Also performs several other checks and assesses whether there are any instances of 'barcode clash', i.e. TCR chains that are not related detected in conjunction with the same barcode. This section is responsible for collapsing the actual `Decombinator` identifiers down, i.e. reducing the number of unique TCRs by merit of erroneous sequences being removed. **C**: Having removed erroneous sequences, the script then loops through all of the barcodes that were associated with a given rearrangement. In order to account for errors in the barcode itself, barcodes that are highly similar to one another (in this instance having an edit distance of three or less) are clustered together; the number of unique clusters are then taken as the truer estimated frequency of that TCR. Finally results are output as collapsed `Decombinator` files. Flowchart produced using yEd (yWorks).

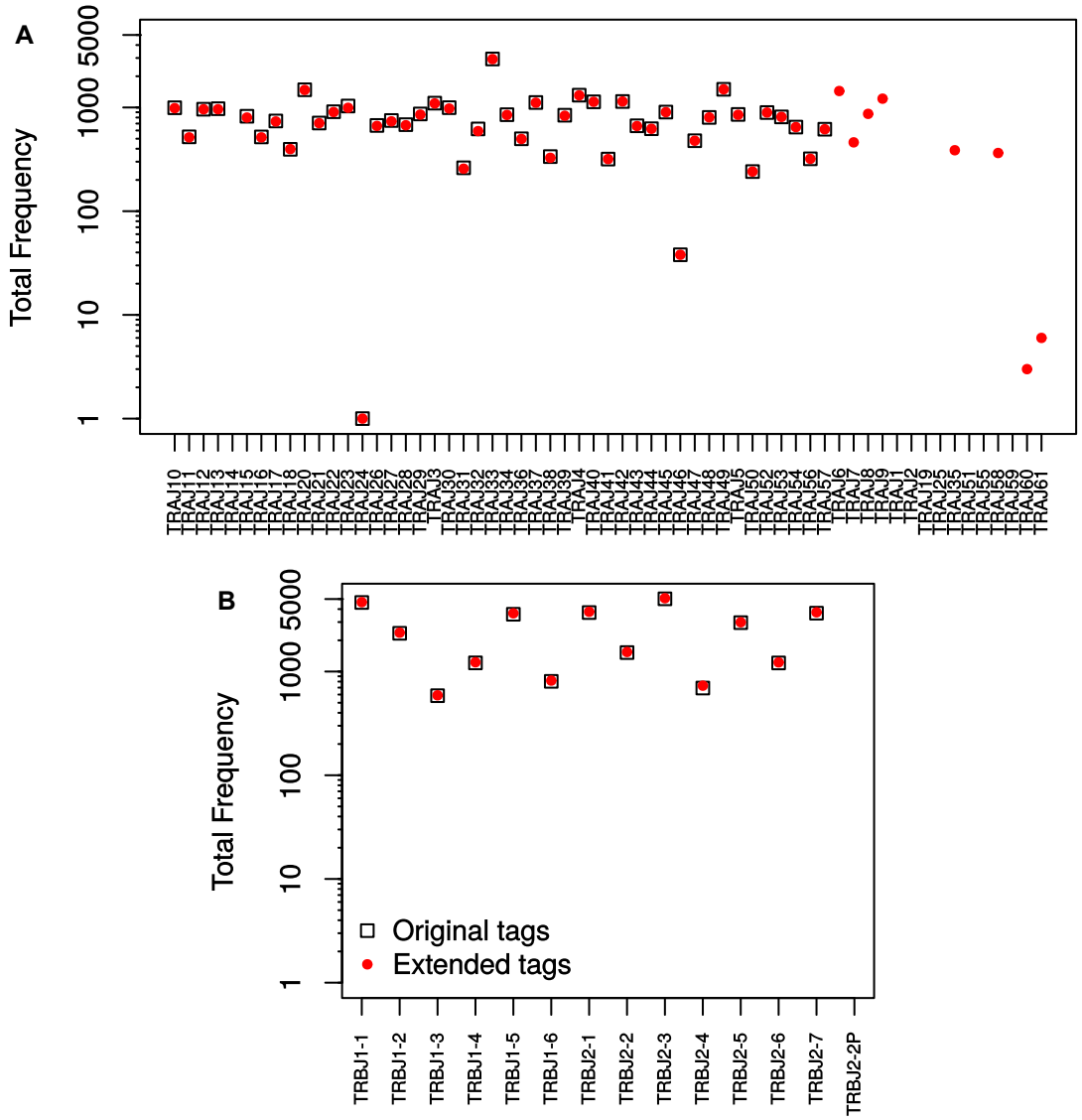


Figure C.4: TRAJ and TRBJ usage difference between original and extended `Decombinator` tag sets. Plots of a representative healthy donor (HV09) showing the comparative performance of `Decombinator` in combination with the original and extended tag sets. **A** and **B**: Frequency of `Decombinator` rearrangements using particular TRAJ and TRBJ genes, when using either the original or new extended tag sets.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
D2 extα	1,943,685	9.93%	0.16%	33.5	2.01×10^{-3}	99.80%
D2 extβ	1,943,685	21.91%	0.22%	35.3	1.41×10^{-3}	99.86%
D2 R1α	8,931,858	0.05%	0.23%	34.6	2.25×10^{-3}	99.78%
D2 R1β	8,931,858	0.14%	0.29%	34.7	1.87×10^{-3}	99.81%
D2 R2α	8,931,858	0.04%	0.89%	34.9	2.68×10^{-3}	99.73%
D2 R2β	8,931,858	0.07%	0.24%	34.8	2.55×10^{-3}	99.75%
D3 naïve extα	2,207,696	7.70%	0.16%	33.5	2.06×10^{-3}	99.79%
D3 naïve extβ	2,207,696	26.98%	0.21%	35.3	1.42×10^{-3}	99.86%
D3 naïve R1α	15,286,006	0.15%	0.24%	35.0	1.84×10^{-3}	99.82%
D3 naïve R1β	15,286,006	0.29%	0.26%	34.8	1.79×10^{-3}	99.82%
D3 naïve R2α	15,286,006	0.10%	0.97%	35.2	2.35×10^{-3}	99.77%
D3 naïve R2β	15,286,006	0.17%	0.17%	34.8	2.56×10^{-3}	99.74%
D3 memory extα	1,786,562	7.73%	0.25%	33.2	2.29×10^{-3}	99.77%
D3 memory extβ	1,786,562	24.83%	0.33%	35.1	1.53×10^{-3}	99.85%
D3 memory R1α	14,623,679	0.04%	0.42%	34.9	1.96×10^{-3}	99.80%
D3 memory R1β	14,623,679	0.06%	0.45%	34.7	1.88×10^{-3}	99.81%
D3 memory R2α	14,623,679	0.03%	1.11%	35.0	2.49×10^{-3}	99.75%
D3 memory R2β	14,623,679	0.04%	0.35%	34.8	2.55×10^{-3}	99.74%
D4 naïve extα	1,988,492	14.89%	0.20%	33.6	1.96×10^{-3}	99.80%
D4 naïve extβ	1,988,492	20.93%	0.26%	35.2	1.45×10^{-3}	99.85%
D4 naïve R1α	11,079,845	0.07%	0.33%	34.4	2.33×10^{-3}	99.77%
D4 naïve R1β	11,079,845	0.10%	0.24%	34.3	2.19×10^{-3}	99.78%
D4 naïve R2α	11,079,845	0.06%	1.03%	34.7	2.87×10^{-3}	99.71%
D4 naïve R2β	11,079,845	0.08%	0.25%	34.5	2.70×10^{-3}	99.73%
D4 memory extα	1,702,220	17.44%	0.15%	33.7	1.92×10^{-3}	99.81%
D4 memory extβ	1,702,220	20.47%	0.19%	35.2	1.42×10^{-3}	99.86%
D4 memory R1α	8,851,558	0.11%	0.10%	34.7	2.02×10^{-3}	99.80%
D4 memory R1β	8,851,558	0.09%	0.14%	34.8	1.86×10^{-3}	99.81%
D4 memory R2α	8,851,558	0.09%	0.95%	35.1	2.43×10^{-3}	99.76%
D4 memory R2β	8,851,558	0.07%	0.20%	34.9	2.50×10^{-3}	99.75%

Table C.3: Read and `Decombinator` output data for the Simple V amplification protocol, as sequenced on the Illumina HiSeq. Reads that were extended (overlapped) by `FLASH` are highlighted in grey. Records: number of FASTQ reads; number of those reads successfully assigned by `Decombinator`; how many assignations survived the N/length filter; percentage of total reads that were assigned TCRs; percentage of assigned TCRs removed by N/length filter; average quality score of inter-tag region (i.e. fastq record from the start of `Decombinator`-assigned V tag to the end of the J tag); average per base probability of an error having occurred in that inter-tag FASTQ file, and the related predicated accuracy of that sequence data.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
D2α	10,875,543	1.85%	0.18%	33.6	2.03×10^{-3}	99.80%
D2β	10,875,543	4.09%	0.23%	35.3	1.44×10^{-3}	99.86%
D3 naïveα	17,493,702	1.19%	0.23%	33.8	2.06×10^{-3}	99.79%
D3 naïveβ	17,493,702	3.81%	0.21%	35.2	1.49×10^{-3}	99.85%
D3 memoryα	16,410,241	0.90%	0.28%	33.3	2.28×10^{-3}	99.77%
D3 memoryβ	16,410,241	2.80%	0.33%	35.0	1.55×10^{-3}	99.84%
D4 naïveα	13,068,337	2.37%	0.22%	33.6	1.99×10^{-3}	99.80%
D4 naïveβ	13,068,337	3.33%	0.25%	35.1	1.49×10^{-3}	99.85%
D4 memoryα	10,553,778	2.98%	0.17%	33.7	1.94×10^{-3}	99.81%
D4 memoryβ	10,553,778	3.44%	0.19%	35.2	1.45×10^{-3}	99.86%

Table C.4: Overall read statistics for the Simple V amplification protocol Sequenced on the Illumina HiSeq. Fields are as in Table C.3.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy	Mean Length
D2 α	13,330	42.52%	0.12%	37.7	4.05×10^{-4}	99.96%	421.9
D2 β	9,714	48.45%	0.53%	36.4	6.64×10^{-4}	99.93%	480.4
D3 α	23,456	61.22%	0.33%	38.0	3.60×10^{-4}	99.96%	413.3
D3 β	25,686	28.83%	0.16%	37.2	5.88×10^{-4}	99.94%	282.5
D9 α	12,587	33.32%	0.00%	37.4	4.73×10^{-4}	99.95%	375.7
D9 β	12,432	46.45%	0.36%	35.4	8.63×10^{-4}	99.91%	507.3
D10 α	12,587	33.32%	0.00%	37.4	4.73×10^{-4}	99.95%	375.7
D10 β	38,044	33.89%	0.45%	37.0	6.19×10^{-4}	99.94%	331.6

Table C.5: Overall read statistics for the poly-T RACE protocol, sequenced on the Roche 454 GS FLX. Fields are as in Table C.3, with an additional column reporting the average read length of the original FASTQ files. Note that D9 and D10 report identical results for alpha chain files; on closer inspection, the FASTQ files were identical. One file must have been accidentally overwritten somewhere between sequencing and final processing.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
D3α	113,002	35.27%	0.00%	24.9	2.04×10^{-2}	97.96%
D3β	452,284	25.27%	0.02%	31.4	6.38×10^{-3}	99.36%
D10α	16,014	36.51%	0.03%	25.8	1.75×10^{-2}	98.25%
D10β	28,326	26.74%	0.03%	32.6	4.72×10^{-3}	99.53%

Table C.6: Overall read statistics for the poly-T RACE protocol, with NEBNext adapter-ligation library preparation. Sequenced on the Illumina MiSeq (2 x 150bp paired end (PE), version 1 chemistry kits). Fields are as in Table C.3.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
RL1α	611,664	76.88%	0.00%	37.4	4.47×10^{-4}	99.96%
RL1β	491,388	61.11%	0.01%	37.4	3.90×10^{-4}	99.96%
RL2α	827,843	62.13%	0.01%	38.1	2.47×10^{-4}	99.98%
RL3β	1,079,562	68.47%	0.00%	37.5	3.84×10^{-4}	99.96%
DL1α	3,437,933	51.84%	0.02%	37.8	3.48×10^{-4}	99.97%
DL1β	3,237,576	54.13%	0.01%	37.6	3.67×10^{-4}	99.96%
DL2α	2,477,170	31.14%	0.00%	37.9	3.15×10^{-4}	99.97%
DL2β	2,116,423	64.44%	0.00%	37.6	3.48×10^{-4}	99.97%

Table C.7: Summary statistics for the ligation RACE protocols, sequenced on the Illumina MiSeq. RL = RNA ligation, DL = DNA ligation, numbers indicate test run number per protocol. RNA and DNA ligation protocols were performed by Prof. Chain and Dr Oakes, and are included in this thesis to explain the downstream protocol choices.

Sample	Reads	% Decombined	Total %	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
HV4α	2,844,497	42.56%	66.78%	0.00%	37.6	3.80×10^{-4}	99.96%
HV4β	2,844,497	24.22%		0.00%	37.8	3.29×10^{-4}	99.97%
HV5α	2,465,640	39.65%	65.50%	0.01%	37.6	3.98×10^{-4}	99.96%
HV5β	2,465,640	25.85%		0.00%	37.7	3.37×10^{-4}	99.97%
HV7α	1,610,847	27.26%	44.92%	0.00%	33.5	1.78×10^{-3}	99.82%
HV7β	1,610,847	17.66%		0.00%	34.6	1.36×10^{-3}	99.86%
HV9α	1,533,948	22.58%	35.79%	0.00%	33.5	1.78×10^{-3}	99.82%
HV9β	1,533,948	13.21%		0.00%	34.6	1.33×10^{-3}	99.87%
HV10α	1,199,385	28.18%	47.59%	0.00%	33.4	1.80×10^{-3}	99.82%
HV10β	1,199,385	19.42%		0.00%	34.6	1.37×10^{-3}	99.86%

Table C.8: Summary statistics for the randomly barcoded simple V amplification protocol, sequenced on the Illumina MiSeq. Fields are as in Table C.3, with an additional total percentage Decombined column, as both alpha and beta reads for the same donors were sequenced under the same index, and thus were demultiplexed together into one output FASTQ file. Files were sequenced in two batches along with other samples, with ten samples multiplexed per run. Thanks go to James Opzoomer, an intern in the laboratory who performed the initial processing on the blood samples once the protocol had been sufficiently refined.

Sample	Reads	% Decombined	% Filtered	Mean Quality	Mean P Error	Predicted Accuracy
HV07α	1,530,742	77.23%	0.01%	37.5	4.18×10^{-4}	99.96%
HV07β	1,692,780	68.36%	0.00%	37.7	3.24×10^{-4}	99.97%
HV08α	1,674,080	85.50%	0.01%	37.5	4.20×10^{-4}	99.96%
HV08β	1,431,872	73.98%	0.00%	37.8	3.18×10^{-4}	99.97%
HV09α	1,549,771	82.02%	0.01%	37.7	3.68×10^{-4}	99.96%
HV09β	1,296,475	76.72%	0.00%	37.7	3.27×10^{-4}	99.97%
HV10α	1,482,059	82.15%	0.00%	37.8	3.54×10^{-4}	99.96%
HV10β	1,006,028	79.66%	0.00%	37.8	3.03×10^{-4}	99.97%
HV11α	1,377,301	85.72%	0.01%	38.0	2.90×10^{-4}	99.97%
HV11β	1,283,645	82.40%	0.00%	37.7	3.29×10^{-4}	99.97%
HV13α	1,136,590	85.15%	0.00%	38.0	2.84×10^{-4}	99.97%
HV13β	1,521,705	75.66%	0.00%	37.7	3.23×10^{-4}	99.97%

Table C.9: Summary statistics for the randomly barcoded ligation RACE (BLR) amplification protocol, sequenced on the Illumina MiSeq. Fields are as in Table C.3.

		TRAV8-24		TRAV21		TRBV12-34		TRBV20-1	
		BV	BLR	BV	BLR	BV	BLR	BV	BLR
TRAV8-24	BV								
	BLR	***							
TRAV21	BV	ns	*						
	BLR	***	ns	*					
TRBV12-34	BV	***	***	ns	***				
	BLR	***	ns	*	ns	***			
TRBV20-1	BV	***	***	*	***	**	***		
	BLR	***	ns	*	ns	***	ns	***	

Table C.10: Differential amplification observed between V genes during multiplex PCR. Levels of significance observed when comparing all of the different groups shown in Figure 5.23 with one another. All comparisons were made using Welch's t-test (unpaired Student's t-test without assumption of equal variance). P-values ¡ 0.001 = ***, ¡ 0.01 = **, ¡ 0.05 = *, ¿= 0.05 = not significant (ns). BV = barcoded V protocol, BLR = barcoded ligation RACE protocol. Purple cells indicate BV-BV comparisons, blue are BLR-BLR, while yellow indicates inter-protocol comparisons. Grey cells represent redundant combinations.

References

[1] Hyung Wook Park. **Germs, Hosts, and the Origin of Frank Macfarlane Burnets Concept of Self and Tolerance, 1936 1949**. *Journal of the History of Medicine*, **61**(4):492–534, 2006.

[2] Zeev Pancer and Max D Cooper. **The evolution of adaptive immunity.** *Annual review of immunology*, **24**:497–518, January 2006.

[3] Anne Galy, Marilyn Travis, Dazhi Cen, and Benjamin Chen. **Human T , B , Natural Killer , and Dendritic Cells Arise from a Common Bone Marrow Progenitor Cell Subset**. *Immunity*, **3**:459–473, 1995.

[4] Motonari Kondo, Irving L Weissman, and Koichi Akashi. **Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow**. *Cell*, **91**:661–672, 1997.

[5] B Y Veronika Groh, Steve Porcelli, I Marina Fabbi, Lewis L Lanier, Louis J Picker, Terri Anderson, Roger A Warnke, Atul K Bhan, Jack L Strominger, and Michael B Brenner. **Human lymphocytes bearing T cell receptor gamma/delta are phenotypically diverse and evenly distributed throughout the lymphoid system.** *The Journal of experimental medicine*, **169**(April):1277–1294, 1989.

[6] Nobumichi Hozumi and Susumu Tonegawa. **Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions**. *Proceedings of the National Academy of Sciences of the United States of America*, **73**(10):3628–3632, 1976.

[7] Maureen Gilmore-Hebert, Kathleen Herculest, Michael Komaromy, and Randolph Wall. **Variable and constant regions are separated in the 10-kbase transcription unit coding for immunoglobulin K light chains**. *Proceedings of the National Academy of Sciences of the United States of America*, **75**(12):6044–6048, 1978.

[8] J G Seidman, A Y A Leder, Marshall H Edgell, Fred Polsky, Shirley M Tilghman, David C Tiemeier, and Philip Leder. **Multiple related immunoglobulin variable-region genes identified by cloning and sequence analysis**. *Proceedings of the National Academy of Sciences of the United States of America*, **75**(8):3881–3885, 1978.

[9] H Sakano, K Huppi, G Heinrich, and S Tonegawa. **Sequences at the somatic recombination sites of immunoglobulin light-chain genes**. *Nature*, **280**(July):288–294, 1979.

[10] Edward E Max, J G Seidman, and Philip Leder. **Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene**. *Proceedings of the National Academy of Sciences of the United States of America*, **76**(7):3450–3454, 1979.

[11] F Alt, E Oltz, F Young, J Gorman, G Taccioli, and J Chen. **VDJ recombination**. *Immunology Today*, **13**(8):306–314, 1992.

[12] D G Schatz, M a Oettinger, and D Baltimore. **The V(D)J recombination activating gene, RAG-1.** *Cell*, **59**(6):1035–48, December 1989.

[13] Marjorie A Oettinger, David G Schatz, Carolyn Gorka, and David Baltimore. **RAG-1 and RAG-2, Adjacent Genes That Synergistically Activate V(D)J Recombination**. *Science*, **248**(4962):1517–1523, 1990.

[14] Susumu Tonegawa. **Somatic generation of antibody diversity**. *Nature*, **302**(April):575–581, 1983.

[15] Y. Chien, M. Iwashima, K.B. Kaplan, J.F. Elliott, and M.M. Davis. **A new T-cell receptor gene located within the alpha locus and expressed early in T-cell differentiation**. *Nature*, **327**, 1987.

[16] R Baer, T Boehm, H Yssel, H Spits, and T H Rabbitts. **Complex rearrangements within the human Jdelta-Cdelta / Jalpha-Calpha locus and aberrant recombination between Jalpha segments**. *The EMBO journal*, **7**(6):1661–1668, 1988.

[17] Shingo Hata, Martha Clabby, Peter Devlin, Hergen Spits, Jan E De Vries, and Michael S. Krangel. **Diversity and organization of human T cell receptor delta variable gene segments.** *The Journal of experimental medicine*, **169**(January):41–57, 1989.

[18] Yoshihiro Takihara, Eric Champagne, Ermanno Ciccone, Lorenzo Moretta, Mark Minden, and Tak W Mak. **Organization and orientation of a human T cell receptor delta chain V gene segment that suggests an inversion mechanism is utilized in its rearrangement**. *European journal of immunology*, **19**:571–574, 1989.

[19] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.** *Nucleic acids research*, **33**(Database issue):D256–61, January 2005.

[20] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. **Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes.** *Human molecular genetics*, **23**(22):5866–5878, 2014.

[21] Donna Karolchik, Galt P. Barber, Jonathan Casper, Hiram Clawson, Melissa S. Cline, Mark Diekhans, Timothy R. Dreszer, Pauline a. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel a. Harte, Steve Heitner, Angie S. Hinrichs, Katrina Learned, Brian T. Lee, Chin H. Li, Brian J. Raney, Brooke Rhead, Kate R. Rosenbloom, Cricket a. Sloan, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. **The UCSC Genome Browser database: 2014 update**. *Nucleic Acids Research*, **42**(November 2013):764–770, 2014.

[22] Martin Gellert. **V(D)J recombination: RAG proteins, repair factors, and regulation.** *Annual review of biochemistry*, **71**(D):101–32, January 2002.

[23] Yunmei Ma, Ulrich Pannicke, Klaus Schwarz, and Michael R Lieber. **Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination**. *Cell*, **108**(D):781–794, 2002.

[24] Frederick W Alt and David Baltimore. **Joining of immunoglobulin heavy chain gene segments : Implications from a chromosome with evidence of three D-JH fusions**. *Proceedings of the National Academy of Sciences of the United States of America*, **79**(July):4118–4122, 1982.

[25] Stephen V. Desiderio, George D. Yancopoulos, Michael Paskind, Elise Thomas, Michael A. Boss, Nathaniel Landau, Frederick W. Alt, and David Baltimore. **Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells**. *Nature*, **311**:752–755, 1984.

[26] Toshihisa Komori, Ami Okada, Valerie Stewart, and Frederick W Alt. **Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes**. *Science*, **261**(5125):1171–1175, 1993.

[27] Susan Gilfillan, Andrée Dierich, Marianne Lemeur, Christophe Benoist, Andree Dierich, and Diane Mathis. **Mice lacking TdT: mature animals with an immature lymphocyte repertoire**. *Science*, **261**(5125):1175–1178, 1993.

[28] Oleg Y Borbulevych, Kurt H Piepenbrink, and Brian M Baker. **Conformational melding permits a conserved binding geometry in TCR recognition of foreign and self molecular mimics.** *Journal of immunology (Baltimore, Md. : 1950)*, **186**(5):2950–8, March 2011.

[29] R M Zinkernagel and P C Doherty. **Restriction of in vitro T cell-mediated cytoxicity in lymphocytic chroiomeningitis within a syngeneic or semiallogeneic system**. *Nature*, **248**(April):701–702, 1974.

[30] Emil R Unanue. **Antigen-presenting function of the macrophage**. *Annual review of immunology*, **2**:395–428, 1984.

[31] PJ Bjorkman, MA Saper, B Samraoui, WS Bennett, JL Strominger, and DC Wiley. **Structure of the human class I histocompatibility antigen, HLA-A 2**. *Nature*, **329**(October):506–512, 1987.

[32] Paul M Allen, Gary R Matsueda, Edgar Haber, and Emil R Unanue. **Specificity of the T cell receptor: two different determinants are generated by the same peptide and the I-Ak molecule.** *The Journal of Immunology*, **135**(1):368–373, 1985.

[33] A R M Townsend, J Rothbard, F M Gotch, G Bahadur, D Wraith, and A J McMichael. **The Epitopes of Influenza Nucleoprotein Recognized by Cytotoxic T Lymphocytes Can Be Defined with Short Synthetic Peptides**. *Cell*, **44**:959–968, 1986.

[34] PJ Bjorkman, MA Saper, B Samraoui, WS Bennett, JL Strominger, and DC Wiley. **The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens**. *Nature*, **329**(October):512–518, 1987.

[35] Alain Townsend and Helen Bodmer. **Antigen recognition by class I-restricted T lymphocytes**. *Annual review of immunology*, **7**:601–624, 1989.

[36] Rose Zamoyska. **CD4 and CD8: modulators of T-cell receptor recognition of antigen and of immune responses?** *Current opinion in immunology*, **10**:82–87, 1998.

[37] Markus G Rudolph, Robyn L Stanfield, and Ian a Wilson. **How TCRs bind MHCs, peptides, and coreceptors.** *Annual review of immunology*, **24**:419–66, January 2006.

[38] Jacques FAP Miller. **Events that led to the discovery of T-cell development and function a personal recollection**. *Tissue antigens*, **1959**(March):509–517, 2004.

[39] Maria Ciofani and Juan Carlos Zúñiga Pflücker. **Determining $\gamma\delta$ versus αßT cell development.** *Nature reviews. Immunology*, **10**(9):657–63, September 2010.

[40] Ronald N Germain. **T-cell development and the CD4-CD8 lineage decision.** *Nature reviews. Immunology*, **2**(5):309–22, May 2002.

[41] SF Schlossman. **Antigen recognition: the specificity of T cells involved in the cellular immune response**. *Immunological Reviews*, **10**:97–111, 1972.

[42] Monna Crone, Claus Koch, and Morten Simonsen. **The elusive T cell receptor**. *Immunological Reviews*, **10**:36–57, 1972.

[43] James P Allison, Bradley W Mcintyre, and David Bloch. **Tumor-specific Antigen Of Murine T-lymphoma Defined Antibody With Monoclonal Antibody**. *Journal of Immunology*, **129**(5):2293–2300, 1982.

[44] B Y Stefan C Meuer, Kathleen A Fitzgerald, Rebecca E Hussey, James C Hodgdon, Stuart F Schlossman, and Ellis L Reinherz. **Clonotypic Structures Involved In Antigen-Specific Human T Cell Function**. *Journal of experimental medicine*, **157**(February):705–719, 1983.

[45] K Haskins, John Kappler, and P Marrack. **The major histocompatibility complex-restricted antigen receptor on T cells**. *Journal of Experimental Medicine*, **157**(4):1149–1169, 1983.

[46] B W McIntyre and J P Allison. **The mouse T cell receptor: structural heterogeneity of molecules of normal T cells defined by xenoantiserum.** *Cell*, **34**(3):739–46, October 1983.

[47] J Kappler, R Kubo, K Haskins, J White, and P Marrack. **The mouse T cell receptor: comparison of MHC-restricted receptors on two T cell hybridomas.** *Cell*, **34**(3):727–37, October 1983.

[48] Stephen M. Hedrick, David I. Cohen, Ellen A. Nielsen, and Mark M. Davis. **Isolation of cDNA clones encoding T cell-specific membrane-associated proteins**. *Nature*, **308**:149–153, 1984.

[49] Yusuke Yanagi, Yasunobe Yoshikai, Kathleen Leggett, Stephen P Clark, Ingrid Aleksander, and Tak W Mak. **A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains**. *Nature*, **308**:145–149, 1984.

[50] Stephen M Hedrick, Ellen A Nielsen, Joshua Kavaler, David I Cohen, and Mark M Davis. **Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins**. *Nature*, **308**:153–158, 1984.

[51] G Siu, S P Clark, Y Yoshikai, M Malissen, Y Yanagi, E Strauss, T W Mak, and L Hood. **The human T cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene.** *Cell*, **37**(2):393–401, June 1984.

[52] H Saito, DM Kranz, Y Takagaki, and AC Hayday. **Complete primary structure of a heterodimeric T-cell receptor deduced from cDNA sequences**. *Nature*, **309**:757, 1984.

[53] Y. Chien, D.M. Becker, T. Lindsten, M. Okamura, D.I. Cohen, and M.M. Davis. **A third type of murine T-cell receptor gene**. *Nature*, **312**:31–35, 1984.

[54] Nancy Jones, Jeffrey Leiden, Deno Dialynas, John Fraser, Martha Clabby, Jack L Strominger, David Andrews, William Lane, and James Woody. **Partial Primary Structure of the Alpha and Beta Chains of Human Tumor T-Cell Receptors**. *Science*, **227**(4684):311–314, 1985.

[55] Stefan C Meuer, Oreste Acuto, Rebecca E Hussey, James C Hodgdon, Kathleen A Fitzgerald, Stuart F Schlossman, and Ellis L Reinherz. **Evidence for the T3-associated 90K heterodimer as the T-cell antigen receptor**. *Nature*, **303**:808–810, 1983.

[56] J M Faint, D Pilling, A N Akbar, G D Kitas, P A Bacon, and M Salmon. **Quantitative flow cytometry for the analysis of T cell receptor Vbeta chain expression**. *Journal of immunological methods*, **225**:53–60, 1999.

[57] Stanca M Ciupe, Blythe H Devlin, Mary Louise Markert, and Thomas B Kepler. **Quantification of total T-cell receptor diversity by flow cytometry and spectratyping.** *BMC immunology*, **14**(1):35, August 2013.

[58] Sean C Bendall, Garry P Nolan, Mario Roederer, and Pratip K Chattopadhyay. **A deep profiler's guide to cytometry**. *Trends in immunology*, **33**(7):323–332, 2012.

[59] Isabell Bretschneider, Michael J Clemente, Christian Meisel, Manuel Guerreiro, Mathias Streitz, Werner Hopfenmüller, Jaroslav P Maciejewski, Marcin W Wlodarski, and Hans-Dieter Volk. **Discrimination of T-cell subsets and T-cell receptor repertoire distribution.** *Immunologic research*, **58**(1):20–7, January 2014.

[60] Anita Diu, François Romagné, Catherine Genevée, Corinne Rocher, Jean-Michel Bruneau, Alain David, Françoise Praz, and Thierry Hercend. **Fine specificity of monoclonal antibodies directed at human T cell receptor variable regions : comparison with oligonucleotide-driven amplification for evaluation of Vbeta expression**. *European journal of immunology*, **23**:1422–1429, 1993.

[61] Randall K Saiki, Stephen Scharf, Fred Faloona, Kary B Mullis, and Glenn T Horn. **Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia**. *Science*, **230**(4732):1350–1354, 1985.

[62] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Glenn T Horn, Kary B Mullis, and Henry A Erlich. **Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase**. *Science*, **239**(4839):487–491, 1988.

[63] Y W Choi, B Kotzin, L Herron, J Callahan, P Marrack, and J Kappler. **Interaction of Staphylococcus aureus toxin "superantigens" with human T cells.** *European journal of immunology*, **86**(November):8941–8945, 1989.

[64] William M C Rosenberg, Paul A H Moss, and John I Bell. **Variation in human T cell receptor Vβ and Jβ repertoire: analysis using anchor polymerase chain reaction**. *European journal of immunology*, **22**(2):541–549, 1992.

[65] Ai-Hong Li, Erik Forestier, Richard Rosenquist, and Göran Roos. **Minimal residual disease quantification in childhood acute lymphoblastic leukemia by real-time polymerase chain reaction using the SYBR green dye**. *Experimental hematology*, **30**:1170–1177, 2002.

[66] Madeleine Cochet, Christophe Pannetier, Armelle Regnault, Sylvie Darche, Claude Leclerc, and Phillippe Kourilsky. **Molecular detection and in vivo analysis of the specific T cell response to a protein antigen**. *European journal of immunology*, **22**:2639–2647, 1992.

[67] J. Gorski, Maryam Yassai, Xiaolei Zhu, B. Kissela, B. Kissella, Carolyn Keever, and N. Flomenberg. **Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping.** *Journal of Immunology*, **152**(10):5109, 1994.

[68] Jessica Gorski, Teresa Piatek, Maryam Yassai, and Krystyna Maslanka. **Improvements in repertoire analysis by CDR3 size spectratyping**. *Annals of the New York Academy of Sciences*, **756**:99–102, 1995.

[69] Mala K Maini, Giulia Casorati, Paolo Dellabona, Andreas Wack, and Peter C L Beverley. **T-cell clonality in immune responses**. *Immunology Today*, **5699**(6):262–266, 1999.

[70] Kazuhiko Yamamoto, Hiroko Sakoda, Toshihiro Nakajima, Tomohiro Kato, Mitsuo Okubo, Makoto Dohi, Yutaka Mizushima, Koji Ito, and Kusuki Nishioka. **Accumulation of multiple T cell clonotypes in the synovial lesions of patients with rheumatoid arthritis revealed by a novel clonality analysis**. *International immunology*, **4**(11):1219–1223, 1992.

[71] Máire F Quigley, Jorge R Almeida, David a Price, and Daniel C Douek. **Unbiased molecular analysis of T cell receptor expression using template-switch anchored rt-PCR.** *Current protocols in immunology*, **Chapter 10**:10.33, August 2011.

[72] Pradyot Dash, Jennifer L Mcclaren, Thomas H Oguin Iii, William Rothwell, Brandon Todd, Melissa Y Morris, Jared Becksfort, Cory Reynolds, Scott A Brown, Peter C Doherty, and Paul G Thomas. **Paired analysis of TCRα and TCRβ chains at the single-cell level in mice**. *The Journal of Clinical Investigation*, **121**(1):288–295, 2011.

[73] Song-Min Kim, Latika Bhonsle, Petra Besgen, Jens Nickel, Anna Backes, Kathrin Held, Sigrid Vollmer, Klaus Dornmair, and Joerg C Prinz. **Analysis of the paired TCR α- and β-chains of single human T cells.** *PloS one*, **7**(5):e37338, January 2012.

[74] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni. **Rep-Seq: uncovering the immunological repertoire through next-generation sequencing.** *Immunology*, **135**(3):183–91, March 2012.

[75] J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert A Holt. **Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing.** *Genome research*, **19**(10):1817–24, October 2009.

[76] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. **Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells.** *Blood*, **114**(19):4099–107, November 2009.

[77] René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. **Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes.** *Genome research*, **21**(5):790–7, May 2011.

[78] T. Petteri Arstila, Armanda Casrouge, Véronique Baron, Jos Even, Jean Kanellopoulos, and Philippe Kourilsky. **A Direct Estimate of the Human T Cell Receptor Diversity**. *Science*, **286**(5441):958–961, October 1999.

[79] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard a Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. **Diversity and clonal selection in the human T-cell repertoire.** *Proceedings of the National Academy of Sciences of the United States of America*, August 2014.

[80] Olga V Britanova, Ekaterina V Putintseva, Mikhail Shugay, Ekaterina M Merzlyak, Maria a Turchaninova, Dmitriy B Staroverov, Dmitriy a Bolotin, Sergey Lukyanov, Ekaterina a Bogdanova, Ilgar Z Mamedov, Yuriy B Lebedev, and Dmitriy M Chudakov. **Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling.** *Journal of immunology (Baltimore, Md. : 1950)*, **192**(6):2689–98, March 2014.

[81] P L Klarenbeek, E B M Remmerswaal, I J M Ten Berge, M E Doorenspleet, B D C van Schaik, R E E Esveldt, S D Koch, A Ten Brinke, a H C van Kampen, F J Bemelman, P P Tak, F Baas, N de Vries, and R a W van Lier. **Deep Sequencing of Antiviral T-Cell Responses to HCMV and EBV in Humans Reveals a Stable Repertoire That Is Maintained for Many Years.** *PLoS pathogens*, **8**(9):e1002889, September 2012.

[82] Anna M Sherwood, Cindy Desmarais, Robert J Livingston, Jessica Andriesen, Maximilian Haussler, Christopher S Carlson, and Harlan Robins. **Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after $\alpha\beta$ and $\gamma\delta$ T cell commitment.** *Science translational medicine*, **3**(90):90ra61, July 2011.

[83] Chunlin Wang, Catherine M Sanders, Qunying Yang, Harry W Schroeder, Elijah Wang, Farbod Babrzadeh, Baback Gharizadeh, Richard M Myers, James R Hudson, Ronald W Davis, and Jian Han. **High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets.** *Proceedings of the National Academy of Sciences of the United States of America*, **107**(4):1518–23, January 2010.

[84] JosephJ.C. Thome, Naomi Yudanin, Yoshiaki Ohmura, Masaru Kubota, Boris Grinshpun, Taheri Sathaliyawala, Tomoaki Kato, Harvey Lerner, Yufeng Shen, and DonnaL. Farber. **Spatial Map of Human T Cell Compartmentalization and Maintenance over Decades of Life**. *Cell*, **159**(4):814–828, November 2014.

[85] Pina F I Krell, Susanne Reuther, Ute Fischer, Thomas Keller, Stephan Weber, Michael Gombert, Friedhelm R Schuster, Corinna Asang, Polina Stepensky, Brigitte Strahm, Roland Meisel, and Jens Stoye. **Next-generation-sequencing-spectratyping reveals public T-cell receptor repertoires in pediatric very severe aplastic anemia and identifies a b chain CDR3 sequence associated with hepatitis-induced pathogenesis**. *Haematologica*, **98**(9), 2013.

[86] Junfeng Wu, Dawei Liu, Wenwei Tu, Wenxia Song, and Xiaodong Zhao. **T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome.** *The Journal of allergy and clinical immunology*, August 2014.

[87] Amy E O'Connell, Stefano Volpi, Kerry Dobbs, Claudia Fiorini, Erdyni Tsitsikov, Helen de Boer, Isil B Barlan, Jenny M Despotovic, Francisco J Espinosa-Rosales, I Celine Hanson, Maria G Kanariou, Roxana Martínez-Beckerat, Alvaro Mayorga-Sirera, Carmen Mejia-Carvajal, Nesrine Radwan, Aaron R Weiss, Sung-Yun Pai, Yu Nee Lee, and Luigi D Notarangelo. **Next generation sequencing reveals skewing of the T and B cell receptor repertoires in patients with wiskott-Aldrich syndrome.** *Frontiers in immunology*, **5**(July):340, January 2014.

[88] Xiaomin Yu, Jorge R Almeida, Sam Darko, Mirjam van der Burg, Suk See DeRavin, Harry Malech, Andrew Gennery, Ivan Chinn, Mary Louise Markert, Daniel C Douek, and Joshua D Milner. **Human syndromes of immunodeficiency and dysregulation are characterized by distinct defects in T-cell receptor repertoire development.** *The Journal of allergy and clinical immunology*, **133**(4):1109–15, April 2014.

[89] D. Wu, A. Sherwood, J. R. Fromm, S. S. Winter, K. P. Dunsmore, M. L. Loh, H. a. Greisman, D. E. Sabath, B. L. Wood, and H. Robins. **High-Throughput Sequencing Detects Minimal Residual Disease in Acute T Lymphoblastic Leukemia**. *Science Translational Medicine*, **4**(134):134ra63–134ra63, May 2012.

[90] Malek Faham, Jianbiao Zheng, Martin Moorhead, Victoria E. H. Carlton, Patricia Stow, Elaine Coustan-Smith, Ching-Hon Pui, and Dario Campana. **Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia**. *Blood*, **120**(26):5173–5181, 2012.

[91] Stephan a Grupp, Michael Kalos, David Barrett, Richard Aplenc, David L Porter, Susan R Rheingold, David T Teachey, Anne Chew, Bernd Hauck, J Fraser Wright, Michael C Milone, Bruce L Levine, and Carl H June. **Chimeric antigen receptor-modified T cells for acute lymphoid leukemia.** *The New England journal of medicine*, **368**(16):1509–18, April 2013.

[92] W.-K. Weng, R. Armstrong, S. Arai, C. Desmarais, R. Hoppe, and Y. H. Kim. **Minimal Residual Disease Monitoring with High-Throughput Sequencing of T Cell Receptors in Cutaneous T Cell Lymphoma**. *Science Translational Medicine*, **5**(214):214ra171, December 2013.

[93] Scott D Boyd, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, Birgitte B Simen, Bozena Hanczaruk, Khoa D Nguyen, Kari C Nadeau, Michael Egholm, David B Miklos, James L Zehnder, and Andrew Z Fire. **Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing**. *Science Translational Medicine*, **1**(12):1–8, 2009.

[94] Aaron C Logan, Hong Gao, Chunlin Wang, Bita Sahaf, Carol D Jones, Eleanor L Marshall, Ismael Buño, Randall Armstrong, Andrew Z Fire, Kenneth I Weinberg, Michael Mindrinos, James L Zehnder, Scott D Boyd, Wenzhong Xiao, Ronald W Davis, and David B Miklos. **High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment.** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(52):21194–9, December 2011.

[95] Aaron C Logan, Nikita Vashi, Malek Faham, Victoria Carlton, Katherine Kong, Ismael Buño, Jianbiao Zheng, Martin Moorhead, Mark Klinger, Bing Zhang, Amna Waqar, James L Zehnder, and David B Miklos. **Immunoglobulin and T cell receptor gene high-throughput sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-transplantation relapse and survival.** *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, **20**(9):1307–13, September 2014.

[96] Jonathan a Schumacher, Eric J Duncavage, Timothy L Mosbruger, Philippe M Szankasi, and Todd W Kelley. **A comparison of deep sequencing of TCRG rearrangements vs traditional capillary electrophoresis for assessment of clonality in T-Cell lymphoproliferative disorders.** *American journal of clinical pathology*, **141**(3):348–59, March 2014.

[97] Ilgar Z Mamedov, Olga V Britanova, Dmitriy a Bolotin, Anna V Chkalina, Dmitriy B Staroverov, Ivan V Zvyagin, Alexey a Kotlobay, Maria a Turchaninova, Denis a Fedorenko, Andrew a Novik, George V Sharonov, Sergey Lukyanov, Dmitriy M Chudakov, and Yuri B Lebedev. **Quantitative tracking of T cell clones after haematopoietic stem cell transplantation.** *EMBO molecular medicine*, **3**(4):201–7, April 2011.

[98] O V Britanova, a G Bochkova, D B Staroverov, D a Fedorenko, D a Bolotin, I Z Mamedov, M a Turchaninova, E V Putintseva, a a Kotlobay, S Lukyanov, a a Novik, Y B Lebedev, and D M Chudakov. **First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy.** *Bone marrow transplantation*, **47**(11):1479–81, November 2012.

[99] Jeroen W J van Heijst, Izaskun Ceberio, Lauren B Lipuma, Dane W Samilo, Gloria D Wasilewski, Anne Marie R Gonzales, Jimmy L Nieves, Marcel R M van den Brink, Miguel A Perales, and Eric G Pamer. **Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation.** *Nature medicine*, **19**(3):372–373, February 2013.

[100] Everett H Meyer, Andro R Hsu, Joanna Lilietal, L Andrea, Mareike Florek, James L Zehnder, Sam Strober, Philip Lavori, David B Miklos, David S Johnson, and Robert S Negrin. **A distinct evolution of the T-cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease**. *Blood*, **121**(24):4955–4963, 2013.

[101] Marie-Paule Lefranc, Véronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jérôme Lane, Laetitia Regnier, François Ehrenmann, Gérard Lefranc, and Patrice Duroux. **IMGT, the international ImMunoGeneTics information system.** *Nucleic acids research*, **37**(Database issue):D1006–12, January 2009.

[102] Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. **IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.** *Developmental and comparative immunology*, **27**(1):55–77, January 2003.

[103] Marie-Paule Lefranc. **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR).** *Cold Spring Harbor protocols*, **2011**(6):627–32, June 2011.

[104] Marie-Paule Lefranc, Véronique Giudicelli, Quentin Kaas, Elodie Duprat, Joumana Jabado-Michaloud, Dominique Scaviner, Chantal Ginestoux, Oliver Clément, Denys Chaume, and Gérard Lefranc. **IMGT, the international ImMunoGeneTics information system.** *Nucleic acids research*, **33**(Database issue):D593–7, January 2005.

[105] Elodie Foulquier, Chantal Ginestoux, and Marie-Paule Lefranc. **IMGT Scientific chart: color menu**. http://www.imgt.org/IMGTScientificChart/RepresentationRules/colormenu.php. Accessed: 2015-01-11.

[106] John T. Lis. **Fractionation of DNA Fragments by Polyethylene Glycol Induced Precipitation**. *Methods in Enzymology*, **65**(1974):347–353, 1980.

[107] Illumina. **Preparing DNA Libraries for Sequencing on the MiSeq**, 2013.

[108] Simon Andrews. **FastQC: A quality control tool for high throughput sequence data**. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 2014-06-16.

[109] Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics*, **25**(11):1422–3, June 2009.

[110] Gregory J Hannon. **FASTX-Toolkit: FASTQ/A short-reads pre-processing tools**. http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed: 2015-01-10.

[111] Niclas Thomas, James Heather, Wilfred Ndifon, John Shawe-Taylor, and Benjamin Chain. **Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine.** *Bioinformatics*, **29**(5):542–50, March 2013.

[112] **Decombinator source code**. https://github.com/uclinfectionimmunity/Decombinator/, 2015. Accessed: 2015-01-09.

[113] Marie-Paule Lefranc. **Immunoglobulin and T Cell Receptor Genes: IMGT and the Birth and Rise of Immunoinformatics**. *Frontiers in Immunology*, **5**(February):1–22, 2014.

[114] Alexander Dobin, Carrie a Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics*, **29**(1):15–21, January 2013.

[115] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*, **25**(16):2078–9, August 2009.

[116] James T Robinson, Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. **Integrative genomics viewer**. *Nature biotechnology*, **29**(1):24–26, 2011.

[117] Tanja Magoč and Steven L Salzberg. **FLASH: fast length adjustment of short reads to improve genome assemblies.** *Bioinformatics (Oxford, England)*, **27**(21):2957–63, November 2011.

[118] Jonathan Friedman. **Jensen-Shannon Divergence function**. http://stats.stackexchange.com/posts/94618/revisions. Accessed: 2015-01-11.

[119] James M. Heather, Katharine Best, Niclas Thomas, and Benjamin Chain. **TCR analysis scripts**. http://dx.doi.org/10.6084/m9.figshare.1190861, 2014.

[120] Adrien Six, Maria Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M. Walczak, and Pierre Boudinot. **The Past, Present, and Future of Immune Repertoire Biology The Rise of Next-Generation Repertoire Analysis**. *Frontiers in Immunology*, **4**(November):1–16, 2013.

[121] Lidia Ceriani and Paolo Verme. **The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini**. *The Journal of Economic Inequality*, **10**(3):421–443, June 2011.

[122] C. E. Shannon. **A Mathematical Theory of Communication**. *Bell System Technical Journal*, **XXVII**(3):379, 1948.

[123] Marie-Paule Lefranc. **Nomenclature of the Human T Cell Receptor Genes**. *Current protocols in immunology*, **S40**:1–23, 2000.

[124] Marie-Paule Lefranc. **IMGT Locus on focus**. *Experimental and clinical immunogenetics*, **5535**(15):1–7, 1998.

[125] Santosh K Srivastava and Harlan S Robins. **Palindromic nucleotide analysis in human T cell receptor rearrangements.** *PloS one*, **7**(12):e52250, January 2012.

[126] Dmitry a Bolotin, Ilgar Z Mamedov, Olga V Britanova, Ivan V Zvyagin, Dmitriy Shagin, Svetlana V Ustyugova, Maria a Turchaninova, Sergey Lukyanov, Yury B Lebedev, and Dmitriy M Chudakov. **Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms.** *European journal of immunology*, **42**(11):3073–83, November 2012.

[127] Peipei Liu, Di Liu, Xi Yang, Jing Gao, Yan Chen, Xue Xiao, Fei Liu, Jing Zou, Jun Wu, Juncai Ma, Fangqing Zhao, Xuyu Zhou, George F Gao, and Baoli Zhu. **Characterization of human $\alpha\beta$TCR repertoire and discovery of D-D fusion in TCRβ chains.** *Protein & cell*, May 2014.

[128] Janet M Murray, Terri Messier, Jami Rivers, J Patrick O'Neill, Vernon E Walker, Pamela M Vacek, and Barry a Finette. **VDJ recombinase-mediated TCR β locus gene usage and coding joint processing in peripheral T cells during perinatal and pediatric development.** *Journal of immunology (Baltimore, Md. : 1950)*, **189**(5):2356–64, September 2012.

[129] Vanessa Venturi, Máire F Quigley, Hui Yee Greenaway, Pauline C Ng, Zachary S Ende, Tina McIntosh, Tedi E Asher, Jorge R Almeida, Samuel Levy, David a Price, Miles P Davenport, and Daniel C Douek. **A Mechanism for TCR Sharing between T Cell Subsets and Individuals Revealed by Pyrosequencing.** *Journal of immunology*, March 2011.

[130] Hanjie Li, Congting Ye, Guoli Ji, Xiaohui Wu, Zhe Xiang, Yuanyue Li, Yonghao Cao, Xiaolong Liu, Daniel C Douek, David A Price, and Jiahuai Han. **Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes.** *Journal of immunology*, **189**(5):2404–13, September 2012.

[131] Asaf Madi, Eric Shifrut, Shlomit Reich-zeliger, Hilah Gal, Katharine Best, Wilfred Ndifon, Benjamin Chain, Irun R Cohen, and Nir Friedman. **T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity**. *Genome research*, **24**:1603–1612, 2014.

[132] S Porcelli, C E Yockey, M B Brenner, and S P Balk. **Analysis of T cell antigen receptor (TCR) expression by human peripheral blood CD4-8- alpha/-beta T cells demonstrates preferential use of several V beta genes and an invariant TCR alpha chain.** *The Journal of experimental medicine*, **178**(1):1–16, July 1993.

[133] O Lantz and a Bendelac. **An invariant T cell receptor alpha chain is used by a unique subset of major histocompatibility complex class I-specific CD4+ and CD4-8- T cells in mice and humans.** *The Journal of experimental medicine*, **180**(3):1097–106, September 1994.

[134] Albert Bendelac, Olivier Lantz, Mary E Quimby, Jonathan W Yewdell, Jack R. Bennink, Randy R Brutkiewicz, and Jack R Bennink. **CD1 Recognition by Mouse NK1 + T Lymphocytes**. *Science*, **268**(5212):863–865, 1995.

[135] Florance Tilloy, Emmaneul Treiner, Se-Ho Park, Corinne Garcia, Francois Lemonnier, Henri de la Salle, Albert Bendelac, Marc Bonneville, and Olivier Lantz. **An Invariant T Cell Receptor Alpha Chain Defines a Novel TAP-independent Major Histocompatility Complex Class Ib-restricted alpha/beta T Cell Subpopulation in Mammals**. *Journal of Experimental Medicine*, **189**(12):1907–1921, 1999.

[136] Emmanuel Treiner, Livine Duban, Seiamak Bahram, Mirjana Radosavljevic, Valerie Wanner, Florence Tilloy, Pierre Affaticati, Susan Gilfillan, and Olivier Lantz. **Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1**. *Nature*, **422**(March):1–7, 2003.

[137] Lars Kjer-Nielsen, Onisha Patel, Alexandra J. Corbett, Jerome Le Nours, Bronwyn Meehan, Ligong Liu, Mugdha Bhati, Zhenjun Chen, Lyudmila Kostenko, Rangsima Reantragoon, Nicholas A. Williamson, Anthony W. Purcell, Nadine L. Dudek, Malcolm J. McConville, Richard A. J. OHair, George N. Khairallah, Dale I. Godfrey, David P. Fairlie, Jamie Rossjohn, and James McCluskey. **MR1 presents microbial vitamin B metabolites to MAIT cells**. *Nature*, **491**:717, 2012.

[138] Marco Lepore, Artem Kalinichenko, Alessia Colone, Bhairav Paleja, Amit Singhal, Andreas Tschumi, Bernett Lee, Michael Poidinger, Francesca Zolezzi, Luca Quagliata, Peter Sander, Evan Newell, Antonio Bertoletti, Luigi Terracciano, Gennaro De Libero, and Lucia Mori. **Parallel T-cell cloning and deep sequencing of human MAIT cells reveal stable oligoclonal TCRβ repertoire.** *Nature communications*, **5**(May):3866, January 2014.

[139] Anne G Kasmar, Ildiko van Rhijn, Tan-Yun Cheng, Marie Turner, Chetan Seshadri, Andre Schiefner, Ravi C Kalathur, John W Annand, Annemieke de Jong, John Shires, Luis Leon, Michael Brenner, Ian a Wilson, John D Altman, and D Branch Moody. **CD1b tetramers bind $\alpha\beta$ T cell receptors to identify a mycobacterial glycolipid-reactive T cell repertoire in humans.** *The Journal of experimental medicine*, **208**(9):1741–7, August 2011.

[140] Ildiko Van Rhijn, Anne Kasmar, Annemieke de Jong, Stephanie Gras, Mugdha Bhati, Marieke E Doorenspleet, Niek de Vries, Dale I Godfrey, John D Altman, Wilco de Jager, Jamie Rossjohn, and D Branch Moody. **A conserved human T cell population targets mycobacterial antigens presented by CD1b**. *Nature Immunology*, (June):1–10, June 2013.

[141] B. van Schaik, P. Klarenbeek, M. Doorenspleet, a. van Kampen, D. B. Moody, N. de Vries, and I. Van Rhijn. **Discovery of Invariant T Cells by Next-Generation Sequencing of the Human TCR -Chain Repertoire**. *The Journal of Immunology*, October 2014.

[142] Calman Prussin and Barbara Foster. **TCR V alpha 24 and V beta 11 coexpression defines a human NK1 T cell analog containing a unique Th0 subpopulation.** *The Journal of Immunology*, **159**:5862–5870, 1997.

[143] Cormac Cosgrove, James E. Ussher, Andri Rauch, Kathleen Gartner, Ayako Kurioka, Michael H. Huhn, Krista Adelmann, Yu-Hoi Kang, Joannah R. Fergusson, Peter Simmonds, Philip Goulder, Ted H. Hansen, Julie Fox, Huldrych F. Gunthard, Nina Khanna, Fiona Powrie, Alan Steel, Brian Gazzard, Rodney E. Phillips, John Frater, Holm Uhlig, and Paul Klenerman. **Early and nonreversible decrease of CD161++/MAIT cells in HIV infection**. *Blood*, **121**(6):951–961, 2013.

[144] Marie Malissen, Jeannine Trucy, Evelyne Jouvin-Marche, Pierre-Andre Cazenave, Roland Scollay, and Bernard Malissen. **Regulation of TCR alpha and beta gene allelic exclusion during T-cell development**. *Immunology Today*, **13**(8):315–322, 1992.

[145] M. S. Carter, J. Doskow, P. Morris, S. Li, R. P. Nhim, S. Sandstedt, and M. F. Wilkinson. **A Regulatory Mechanism That Detects Premature Nonsense Codons in T-cell Receptor Transcripts in Vivo Is Reversed by Protein Synthesis Inhibitors in Vitro**. *Journal of Biological Chemistry*, **270**(48):28995–29003, December 1995.

[146] Brenna L Brady, Natalie C Steinel, and Craig H Bassing. **Antigen receptor allelic exclusion: an update and reappraisal.** *Journal of immunology (Baltimore, Md. : 1950)*, **185**(7):3801–8, October 2010.

[147] Salvatore Spicuglia, Aleksandra Pekowska, Joaquin Zacarias-Cabeza, and Pierre Ferrier. **Epigenetic control of Tcrb gene rearrangement.** *Seminars in immunology*, **22**(6):330–6, December 2010.

[148] Shulin Li and Miles F Wilkinson. **Nonsense surveillance in lymphocytes?** *Immunity*, **8**:135–141, 1998.

[149] Joachim Weischenfeldt, Inge Damgaard, David Bryder, Kim Theilgaard-Mönch, Lina A Thoren, Finn Cilius Nielsen, Sten Eirik W Jacobsen, Claus Nerlov, and Bo Torben Porse. **NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements**. *Genes & development*, **22**:1381–1396, 2008.

[150] Nicholas R J Gascoigne and S Munir Alam. **Allelic exclusion of the T cell receptor alpha-chain : developmental regulation of a post-translational event**. *Seminars in immunology*, **11**:337–347, 1999.

[151] Jayanthi P Gudikote and Miles F Wilkinson. **Tcell receptor sequences that elicit strong downregulation of premature termination codonbearing transcripts**. *The EMBO journal*, **21**:125–134, 2002.

[152] Ben F Koop, Lee Rowen, Kai Wang, Chia Lam Kuo, Donald Seto, Johannes A Lenstra, Stephen Howard, Wei Shan, Purnima Deshpande, and Leroy Hood. **The Human T-Cell Receptor TCRAC/TCRDC (Calpha/Cdelta) Region: Organization, Sequence, and Evolution of 97.6 kb of DNA**. *Genomics*, **19**:478–493, 1994.

[153] T Demoulins. **A biased Valpha24+ T-cell repertoire leads to circulating NKT-cell defects in a multiple sclerosis patient at the onset of his disease**. *Immunology Letters*, **90**(2-3):223–228, December 2003.

[154] J. J. Miles, N. a. Borg, R. M. Brennan, F. E. Tynan, L. Kjer-Nielsen, S. L. Silins, M. J. Bell, J. M. Burrows, J. McCluskey, J. Rossjohn, and S. R. Burrows. **TCR Genes Direct MHC Restriction in the Potent Human T Cell Response to a Class I-Bound Viral Epitope**. *The Journal of Immunology*, **177**(10):6804–6814, November 2006.

[155] Monica Boita, Giuseppe Guida, Paola Circosta, Angela Rita, Stefania Stella, Enrico Heffler, Iuliana Badiu, Davide Martorana, Sara Mariani, Giovanni Rolla, and Alessandro Cignetti. **The molecular and functional characterization of clonally expanded CD8 + TCR BV T cells in eosinophilic granulomatosis with polyangiitis (EGPA)**. *Clinical Immunology*, **152**(1-2):152–163, 2014.

[156] KD Bourcier, DG Lim, and YH Ding. **Conserved CDR3 Regions in T-Cell Receptor (TCR) CD8+ T Cells That Recognize the Tax11-19/HLA-A*0201 Complex in a Subject Infected with Human T-Cell Leukemia Virus Type 1: Relationship of T-Cell Fine Specificity and Major Histocompatibility Complex/Peptide**. *Journal of virology*, **75**(20):9836–9843, 2001.

[157] Johan Verhagen, Raphaël Genolet, Graham J Britton, Brian J Stevenson, Catherine a Sabatos-Peyton, Julian Dyson, Immanuel F Luescher, and David C Wraith. **CTLA-4 controls the thymic development of both conventional and regulatory T cells through modulation of the TCR repertoire.** *Proceedings of the National Academy of Sciences of the United States of America*, December 2012.

[158] Sobhan Roy, Dalam Ly, Nan-Sheng Li, John D Altman, Joseph a Piccirilli, D Branch Moody, and Erin J Adams. **Molecular basis of mycobacterial lipid antigen presentation by CD1c and its recognition by $\alpha\beta$ T cells.** *Proceedings of the National Academy of Sciences of the United States of America*, October 2014.

[159] C. Miossec, F.Faure, L. Ferradini, S. Roman-Roman, S. Jitsukawa, S. Ferrini, A. Moretta, F. Triebel, and T. Hercend. **Further analysis of the T cell receptor gamma/delta+ peripheral lymphocyte subset: The Vdelta1 Gene Segment is Expressed with either Calpha or Cdelta**. *Journal of Experimental Medicine*, **171**(April), 1990.

[160] C Miossec, a Caignard, L Ferradini, S Roman-Roman, F Faure, H Michalaki, F Triebel, and T Hercend. **Molecular characterization of human T cell receptor alpha chains including a V delta 1-encoded variable segment.** *European journal of immunology*, **21**(4):1061–4, April 1991.

[161] I Bank, M Book, L Cohen, a Kneller, E Rosental, M Pras, I B Bassat, and a Ben-Nun. **Expansion of a unique subpopulation of cytotoxic T cells that express a C alpha V delta 1 T-cell receptor gene in a patient with severe persistent neutropenia.** *Blood*, **80**(12):3157–63, December 1992.

[162] C Castelli, a Mazzocchi, S Salvi, a Anichini, and M Sensi. **Use of the V delta 1 variable region in the functional T-cell receptor alpha chain of a WT31+ cytotoxic T lymphocyte clone which specifically recognizes HLA-A2 molecule.** *Scandinavian journal of immunology*, **35**(4):487–94, April 1992.

[163] Takamasa Ueno, Hiroko Tomiyama, Mamoru Fujiwara, Shinichi Oka, and Masafumi Takiguchi. **HLA class I-restricted recognition of an HIV-derived epitope peptide by a human T cell receptor alpha chain having a Vdelta1 variable segment.** *European journal of immunology*, **33**(10):2910–6, October 2003.

[164] I Bank, L Cohen, a Kneller, N K De Rosbo, M Book, and a Ben-Nun. **Aberrant T-cell receptor signalling of interferon-gamma- and tumour necrosis factor-alpha-producing cytotoxic CD8+ Vdelta1/Vbeta16 T cells in a patient with chronic neutropenia.** *Scandinavian journal of immunology*, **58**(1):89–98, July 2003.

[165] Xiaoming Sun, Masumichi Saito, Yoshinori Sato, Takayuki Chikata, Takuya Naruto, Tatsuhiko Ozawa, Eiji Kobayashi, Hiroyuki Kishi, Atsushi Muraguchi, and Masafumi Takiguchi. **Unbiased Analysis of TCRα/β Chains at the Single-Cell Level in Human CD8(+) T-Cell Subsets.** *PloS one*, **7**(7):e40386, January 2012.

[166] Xiaoming Sun, Mamoru Fujiwara, Yi Shi, Nozomi Kuse, Hiroyuki Gatanaga, Victor Appay, George F Gao, Shinichi Oka, and Masafumi Takiguchi. **Superimposed epitopes restricted by the same HLA molecule drive distinct HIV-specific CD8+ T cell repertoires.** *Journal of immunology (Baltimore, Md. : 1950)*, **193**(1):77–84, July 2014.

[167] Géraldine Folch and Marie-Paule Lefranc. **Assignment of rearranged cDNAs and gDNAs to germline genes: Human TRDV**, 2001. Accessed: 2014-07-05.

[168] Géraldine Folch, Chantal Ginestoux, and Marie-Paule Lefranc. **Gene table: human (Homo sapiens) TRDV**. `http://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=human&group=TRDV`, 2011. Accessed: 2014-07-05.

[169] S Yokota, T E Hansen-Hagge, and C R Bartram. **T-cell receptor delta gene recombination in common acute lymphoblastic leukemia: preferential usage of V delta 2 and frequent involvement of the J alpha cluster.** *Blood*, **77**(1):141–8, January 1991.

[170] T E Hansen-Hagge, C Mahotka, R E Panzer-Grümayer, H J Reuter, K Schwarz, J J van Dongen, and C R Bartram. **Alternative splicing restricts translation of rearrangements at the T-cell receptor delta/alpha locus.** *Blood*, **85**(7):1888–96, April 1995.

[171] Ann M Carroll, Jill K Slack, and Xiaochun Mu. **V(D)J Recombination Generates a High Frequency of Nonstandrd TCR Ddelta-Associated Rearrangements in Thymocytes**. *The Journal of Immunology*, **150**(6):2222–2230, 1993.

[172] Ferenc Livak. **In vitro and in vivo studies on the generation of the primary T-cell receptor repertoire**. *Immunological reviews*, **200**:23–35, 2004.

[173] Robert E Tillman, Andrea L Wooley, Maureen M Hughes, and Barry P Sleckman. **Regulation of T-cell receptor beta-chain gene assembly by recombination signals : the beyond 12 / 23 restriction**. *Immunological reviews*, **200**:36–43, 2004.

[174] P A H Moss, R J Moots, W M C Rosenberg, S J Rowland-Jones, H C Bodmer, A J McMichael, and J I Bell. **Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide.** *Proceedings of the National Academy of Sciences USA*, **88**(October):8987–8990, 1991.

[175] Ricardo Cibotti, Jean-pierre Cabaniols, Christophe Pannetier, Christiane Delarbre, Isabelle Vergnon, Jean M Kanellopoulos, and Philippe Kourilsky. **Public and private V beta T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice.** *The Journal of experimental medicine*, **180**(September):862–872, 1994.

[176] Victor P Argaet, Christopher W Schmidt, Scott R Burrows, Sharon L Silins, Mike G Kurilla, Denise L Doolan, Andreas Suhrbier, Denis J Moss, Elliot Kieff, Tom B Sculley, and Ihor S Misko. **Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus.** *The Journal of experimental medicine*, **180**(December):2335–2340, 1994.

[177] Stephen J Turner, Peter C Doherty, James McCluskey, and Jamie Rossjohn. **Structural determinants of T-cell receptor bias in immunity.** *Nature Reviews Immunology*, **6**(12):883–94, December 2006.

[178] Vanessa Venturi, David A Price, Daniel C Douek, and Miles P Davenport. **The molecular basis for public T-cell responses?** *Nature Reviews Immunology*, **8**:231–238, 2008.

[179] John J Miles, Daniel C Douek, and David a Price. **Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination.** *Immunology and cell biology*, **89**(3):375–87, March 2011.

[180] Hanjie Li, Congting Ye, Guoli Ji, and Jiahuai Han. **Determinants of public T cell responses.** *Cell research*, **22**(1):33–42, January 2012.

[181] Vanessa Venturi, Katherine Kedzierska, David a Price, Peter C Doherty, Daniel C Douek, Stephen J Turner, and Miles P Davenport. **Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination.** *Proceedings of the National Academy of Sciences of the United States of America*, **103**(49):18691–6, December 2006.

[182] Máire F Quigley, Hui Yee, Vanessa Venturi, Ross Lindsay, Kylie M Quinn, and Robert A Seder. **Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire**. *Proceedings of the National Academy of Sciences USA*, **107**(45):19414–19419, 2010.

[183] a. Murugan, T. Mora, a. M. Walczak, and C. G. Callan. **Statistical inference of the generation probability of T-cell receptors from sequence repertoires**. *Proceedings of the National Academy of Sciences*, September 2012.

[184] Yuval Elhanati, Anand Murugan, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. **Quantifying selection in immune receptor repertoires.** *Proceedings of the National Academy of Sciences of the United States of America*, **111**(27):9875–9880, June 2014.

[185] Dale I Godfrey and Mitchell Kronenberg. **Going both ways : immune regulation via CD1d-dependent NKT cells**. *The Journal of Clinical Investigation*, **114**(10):1379–1388, 2004.

[186] The MHC Sequencing Consortium, B. Aguado, S. Bahram, S. Beck, R.D.Campbell, S. A. Forbes, D. Geraghty, T. Guillaudeux, L. Hood, R. Horton, H. Inoko, M. Janer, C. Jasoni, A. Madan, S. Milne, M. Neville, A. Oka, S. Qin, G. Ribas-Despuig, J. Rogers, L. Rowen, T. Shiina, T. Spies, G. Tamiya, H. Tashiro, J. Trowsdale, Q. Vu, L. Williams, and M. Yamazaki. **Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium.** *Nature*, **401**(6756):921–3, October 1999.

[187] Roger Horton, Laurens Wilming, Vikki Rand, Ruth C Lovering, Elspeth a Bruford, Varsha K Khodiyar, Michael J Lush, Sue Povey, C Conover Talbot, Mathew W Wright, Hester M Wain, John Trowsdale, Andreas Ziegler, and Stephan Beck. **Gene map of the extended human MHC.** *Nature reviews. Genetics*, **5**(12):889–99, December 2004.

[188] Mirjana Radosavljevic and Seiamak Bahram. **In vivo immunogenetics: from MIC to RAET1 loci.** *Immunogenetics*, **55**(1):1–9, April 2003.

[189] Donna G Albertson, Rita Fishpoll, Paul Sherington, Elisabeth Nacheva, and Cesar Milstein. **Sensitive and high resolution in situ hybridization to human chromosomes using biotin labelled probes: assignment of the human thymocyte CD1 antigen genes to**. *The EMBO journal*, **7**(9):2801–2805, 1988.

[190] Keiichiro Hashimoto, Momoki Hirai, and Yoshikazu Kurosawa. **A gene outside the human MHC related to classical HLA class I genes**. *Science*, **269**(5224):693–695, 1995.

[191] a N Akbar, L Terry, A Timms, P C Beverley, and G Janossy. **Loss of CD45R and gain of UCHL1 reactivity is a feature of primed T cells.** *Journal of immunology (Baltimore, Md. : 1950)*, **140**(7):2171–8, April 1988.

[192] M Merkenschlager and P C Beverley. **Evidence for differential expression of CD45 isoforms by precursors for memory-dependent and independent cytotoxic responses: human CD8 memory CTLp selectively express CD45RO (UCHL1).** *International immunology*, **1**(4):450–9, January 1989.

[193] Federica Sallusto, Jens Geginat, and Antonio Lanzavecchia. **Central memory and effector memory T cell subsets: function, generation, and maintenance.** *Annual review of immunology*, **22**:745–63, January 2004.

[194] M K Maini, L R Wedderburn, F C Hall, a Wack, G Casorati, and P C Beverley. **A comparison of two techniques for the molecular tracking of specific T-cell responses; CD4+ human T-cell clones persist in a stable hierarchy but at a lower frequency than clones in the CD8+ population.** *Immunology*, **94**(4):529–35, August 1998.

[195] K. E. Foulds, L. a. Zenewicz, D. J. Shedlock, J. Jiang, a. E. Troy, and H. Shen. **Cutting Edge: CD4 and CD8 T Cells Are Intrinsically Different in Their Proliferative Responses**. *The Journal of Immunology*, **168**(4):1528–1532, February 2002.

[196] Hung Sia Teh, Pawel Kisielow, Bernadette Scott, Hiroyuki Kishi, Yasushi Uematsu, Horst Bluthmann, and Harald von Boehmer. **Thymic major histocompatibility complex antigens and the $\alpha\beta$ T-cell receptor determine the CD4/CD8 phenotype of T cells**. *Nature*, **335**(September):229–233, 1988.

[197] Susan Chan, Margarida Correia-Neves, Christophe Benoist, and Diane Mathis. **CD4/CD8 lineage commitment: matching fate with competence**. *Immunological reviews*, **165**:195–207, 1998.

[198] Paul L Klarenbeek, Paul P Tak, Barbera D C van Schaik, Aeilko H Zwinderman, Marja E Jakobs, Zhuoli Zhang, Antoine H C van Kampen, René a W van Lier, Frank Baas, and Niek de Vries. **Human T-cell memory consists mainly of unexpanded clones.** *Immunology letters*, **133**(1):42–8, September 2010.

[199] Mitchell S Cairo and John E Wagner. **Placental and/or umbilical cord blood: an alternative source of hematopoietic stem cells for transplantation**. *Blood*, **90**(12):4665–4678, 1997.

[200] L Garderet, N Dulphy, C Douay, N Chalumeau, V Schaeffer, M T Zilber, a Lim, J Even, N Mooney, C Gelin, E Gluckman, D Charron, and a Toubert. **The umbilical cord blood alphabeta T-cell repertoire: characteristics of a polyclonal and naive but completely formed repertoire.** *Blood*, **91**(1):340–6, January 1998.

[201] Erez Rechavi, Atar Lev, Yu Nee Lee, Amos J Simon, Yoav Yinon, Schlomo Lipitz, Ninette Amariglio, Boaz Weisz, Luigi D Notarangelo, and Raz Somech. **Timely and spatially regulated maturation of B and T cell repertoire during human fetal development**. *Science translational medicine*, **7**(276):1–12, 2015.

[202] M. A. Hall, J L Reid, and J S Lanchbury. **The distribution of human TCR junctional region lengths shifts with age in both CD4 and CD8 T cells.** *International immunology*, **10**(10):1407–19, October 1998.

[203] Molly Bogue, Susan Gilfillan, Christophe Benoist, and Diane Mathis. **Regulation of N-region diversity in antigen receptors through thymocyte differentiation and thymus ontogeny**. *Proceedings of the National Academy of Sciences USA*, **89**(November):11011–11015, 1992.

[204] a Bonati, P Zanelli, S Ferrari, a Plebani, B Starcich, M Savi, and T M Neri. **T-cell receptor beta-chain gene rearrangement and expression during human thymic ontogenesis.** *Blood*, **79**(6):1472–83, March 1992.

[205] Brendan Marshall, Ruth Schulz, Min Zhou, and Andrew Mellor. **Alternative Splicing and Hypermutation of a Nonproductively Rearranged TCR α -Chain in a T Cell Hybridoma**. *The Journal of Immunology*, **162**:871–877, 1999.

[206] Nicholas J McGlincy and Christopher W J Smith. **Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense?** *Trends in biochemical sciences*, **33**(8):385–93, August 2008.

[207] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. **The real cost of sequencing: higher than you think!** *Genome biology*, **12**(8):125, January 2011.

[208] Anne Eugster, Annett Lindner, Anne-Kristin Heninger, Carmen Wilhelm, Sevina Dietz, Mara Catani, Anette-G Ziegler, and Ezio Bonifacio. **Measuring T cell receptor and T cell gene expression diversity in antigen-responsive human CD4(+) T cells.** *Journal of immunological methods*, **400-401**:13–22, November 2013.

[209] Janko Nikolich-Zugich, Mark K Slifka, and Ilhem Messaoudi. **The many important facets of T-cell repertoire diversity.** *Nature reviews. Immunology*, **4**(2):123–32, February 2004.

[210] Jeffrey M. Leiden. **Transcriptional regulation of T cell receptor genes**. *Annual review of immunology*, **11**(1):539–570, 1993.

[211] Lee Rowen, Ben F Koop, and Leroy Hood. **The complete 685-kilobase DNA sequence of the human β T cell receptor locus**. *Science*, **272**(5269):1755–1762, 1996.

[212] T Lindsten, C H June, and C B Thompson. **Transcription of T cell antigen receptor genes is induced by protein kinase C activation.** *Journal of Immunology*, **141**(5):1769–74, September 1988.

[213] Florence Paillard, Ghislaine Sterkers, Georges Bismuth, Elisabeth Gomard, and Catherine Vaquero. **Lymphokine mRNA and T cell multireceptor mRNA of the Ig super gene family are reciprocally modulated during human T cell activation**. *European journal of immunology*, **18**:1643–1646, 1988.

[214] G. P. Lennon, J. E. Sillibourne, E. Furrie, M. J. Doherty, and R. a. Kay. **Antigen Triggering Selectively Increases TCRBV Gene Transcription**. *The Journal of Immunology*, **165**(4):2020–2027, August 2000.

[215] Alexandra Gallard, Gilles Foucras, Christiane Coureau, and Jean-Charles Guéry. **Tracking T cell clonotypes in complex T lymphocyte populations by real-time quantitative PCR using fluorogenic complementarity-determining region-3-specific probes.** *Journal of immunological methods*, **270**(2):269–80, December 2002.

[216] Yasuhiro Minami, Lawrence E Samelson, and Richard D Klausner. **Internalization and Cycling of the T Cell Antigen Receptor**. *The Journal of biological chemistry*, **262**(27):13342–13347, 1987.

[217] Victoria L Crotzer, Allan S Mabardy, Arthur Weiss, and Frances M Brodsky. **T cell receptor engagement leads to phosphorylation of clathrin heavy chain during receptor internalization.** *The Journal of experimental medicine*, **199**(7):981–91, April 2004.

[218] MS Gottlieb, HM Schanker, PT Fan, A Saxon, and JD Weisman. **Pneumocystis pneumoniaLos Angeles**. *Morbidity and Mortality Weekly Report*, **30**(21):250–252, 1981.

[219] R M Du Bois and M A Branthwaite. **Primary Pneumocystis carinii and cytomegalovirus infections**. *The Lancet*, **12**:1339, 1981.

[220] Vincent Quagliarello. **The Acquired Immunodeficiency Syndrome: Current Status**. *The Yale Journal of Biology and Medicine*, **55**:443–452, 1982.

[221] Yuan Chang, Ethel Cesarman, Melissa S Pessin, Frank Lee, Janice Culpepper, M Knowles, Patrick S Moore, and Daniel M Knowles. **Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma**. *Science*, **266**(5192):1865–1869, 1994.

[222] PS Moore and Y Chang. **Detection of herpesvirus-like DNA sequences in Kaposi's sarcoma in patients with and those without HIV infection**. *New England Journal of Medicine*, **332**(18):1181–1185, 1995.

[223] CDC. **Epidemiologic Notes and Reports Update on Kaposi's Sarcoma and Opportunistic Infections in Previously Healthy Persons**. *Morbidity and Mortality Weekly Report*, **31**(19):1–2, 1982.

[224] CDC. **Current Trends Update on Acquired Immune Deficiency Syndrome (AIDS)**. *Morbidity and Mortality Weekly Report*, **31**(37), 1982.

[225] F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. **Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS)**. *Science*, **220**(4599):868–871, 1983.

[226] Mikulas Popovic, M . G . Sarngadharan, Elizabeth Read, and Robert C . Gallo. **Detection , Isolation , and Continuous Production of Cytopathic Retroviruses (HTLV-III) from Patients with AIDS and Pre-AIDS**. *Science*, **224**(4648):497–500, 1984.

[227] Jay A . Levy, Anthony D . Hoffman, Susan M . Kramer, Jill A . Landis, Joni M . Shimabukuro, and Lyndon S. Oshiro. **Isolation of Lymphocytopathic Retroviruses from San Francisco Patients with AIDS**. *Science*, **225**(4664):840–842, 1984.

[228] L Ratner, R.C. Gallo, and F. Wong-Staal. **HTLV-III, LAV, ARV are variants of same AIDS virus**. *Nature*, **313**:636–637, 1985.

[229] Jean L . Marx. **A Virus by Any Other Name**. *Science*, **227**(4693):1449–1451, 1985.

[230] John Coffin, Ashley Haase, Jay A Levy, Luc Montagnier, Steven Oroszlan, Natalie Teich, Howard Temin, Kumao Toyoshima, Harold Varmus, Peter Vogt, and Robin Weiss. **What to call the AIDS virus?** *Nature*, **321**:10, 1986.

[231] Pierre Sonigo, Marc Alizon, Katherine Staskus, David Klatzmann, Stewart Cole, Olivier Danos, Ernest Retzel, Pierre Tiollais, Ashley Haase, and Simon Wain-Hobs. **Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus**. *Cell*, **42**(August):369–382, 1985.

[232] Andrew R Moss and Peter Bacchetti. **Natural history of HIV infection**. *AIDS*, **3**:55–61, 1989.

[233] Saleta Sierra, Bernd Kupfer, and Rolf Kaiser. **Basics of the virology of HIV-1 and its replication.** *Journal of clinical virology*, **34**(4):233–44, December 2005.

[234] Janet Embretson, Mary Zupancic, Jorge L Ribas, Allen Burke, Paul Racz, Klara Tenner-Racz, and Ashley T Haase. **Massive covert infection of helper T lymphocytes and macrophages by HIV during the incubation period of AIDS**. *Nature*, **362**(359-362), 1993.

[235] Giuseppe Pantaleo, Cecilia Graziosi, James F Demarest, Luca Butini, Maria Montroni, Cecil H Fox, Jan M Orenstein, Donald P Kotler, and Anthony S Fauci. **HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease**. *Nature*, **362**(March):355–358, 1993.

[236] Angus G Dalgleish, Peter CL Beverley, Paul R Clapham, Dorothy H Crawford, Melvyn F Greaves, and Robin A Weiss. **The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus**. *Nature*, **312**(December):763–767, 1984.

[237] Paul Jay Maddon, Angus G Dalgleish, J Steven Mcdougal, Paul R Clapham, Robin A Weiss, and Richard Axel. **The T4 Gene Encodes the AIDS Virus Receptor and Is Expressed in the Immune System and the Brain**. *Cell*, **47**:333–348, 1986.

[238] J. B. Alimonti. **Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS**. *Journal of General Virology*, **84**(7):1649–1661, July 2003.

[239] Gilad Doitsh, Marielle Cavrois, Kara G Lassen, Orlando Zepeda, Zhiyuan Yang, Mario L Santiago, Andrew M Hebbeler, and Warner C Greene. **Abortive HIV infection mediates CD4 T cell depletion and inflammation in human lymphoid tissue.** *Cell*, **143**(5):789–801, November 2010.

[240] Gilad Doitsh, Nicole L K Galloway, Xin Geng, Zhiyuan Yang, Kathryn M Monroe, Orlando Zepeda, Peter W Hunt, Hiroyu Hatano, Stefanie Sowinski, Isa Muñoz Arias, and Warner C Greene. **Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection.** *Nature*, **505**:509–514, December 2013.

[241] Steven G Deeks. **HIV infection, inflammation, immunosenescence, and aging.** *Annual review of medicine*, **62**:141–55, January 2011.

[242] Jamie D K Wilson, Graham S Ogg, Rachel L Allen, Philip J R Goulder, Anthony Kelleher, Andy K Sewell, Christopher A O O'Callaghan, Sarah L Rowland-Jones, Margaret F C Callan, and Andrew J McMichael. **Oligoclonal Expansions of CD8+ T Cells in Chronic HIV Infection Are Antigen Specific**. *Journal of experimental medicine*, **188**(4):785–790, July 1998.

[243] Vijay K Kuchroo, Ana C Anderson, and Constantinos Petrovas. **Coinhibitory receptors and CD8 T cell exhaustion in chronic infections.** *Current opinion in HIV and AIDS*, **9**(5):439–45, September 2014.

[244] Persephone Borrow, Hanna Lewicki, Beatrice H Hahn, George M Shaw, and Michael B A Oldstone. **Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection.** *Journal of virology*, **68**(9):6103–6110, 1994.

[245] Michel R. Klein, Card A. van Baalen, Agnes M. Holwerda, Susana R. Kerkhof Garde, Richard J. Bende, Ireneus P. M. Keet, Jan-Karel M. Eeftinck-Schattenkerk, Albert D. M. E. Osterhaus, Hanneke Schuitemaker, and Frank Miedema. **Gag-specific cytotoxic T lymphocyte responses during the clinical course of HIV-1 infection: a longitudinal analysis of rapid progressors and long-term asymptomatics.** *The Journal of experimental medicine*, **181**(April):1365–1372, 1995.

[246] J. Cao, J. McNevin, U. Malhotra, and M. J. McElrath. **Evolution of CD8+ T Cell Immunity and Viral Escape Following Acute HIV-1 Infection**. *The Journal of Immunology*, **171**(7):3837–3846, September 2003.

[247] Hendrik Streeck and Douglas F Nixon. **T cell immunity in acute HIV-1 infection.** *The Journal of infectious diseases*, **202 Suppl**:S302–8, October 2010.

[248] Tao Dong, Guillaume Stewart-Jones, Nan Chen, Philippa Easterbrook, Xiaoning Xu, Laura Papagno, Victor Appay, Michael Weekes, Chris Conlon, Celsa Spina, Susan Little, Gavin Screaton, Anton van der Merwe, Douglas D Richman, Andrew J McMichael, E Yvonne Jones, and Sarah L Rowland-Jones. **HIV-specific cytotoxic T cells from long-term survivors select a unique T cell receptor.** *The Journal of experimental medicine*, **200**(12):1547–57, December 2004.

[249] Jorge R Almeida, David A. Price, Laura Papagno, Zaïna Aït Arkoub, Delphine Sauce, Ethan Bornstein, Tedi E Asher, Assia Samri, Aurélie Schnuriger, Ioannis Theodorou, Dominique Costagliola, Christine Rouzioux, Henri Agut, Anne-Geneviève Marcelin, Daniel Douek, Brigitte Autran, and Victor Appay. **Superior control of HIV-1 replication by CD8+ T cells is reflected by their avidity, polyfunctionality, and clonal turnover.** *The Journal of experimental medicine*, **204**(10):2473–2485, October 2007.

[250] Asier Sáez-Cirión, Christine Lacabaratz, Olivier Lambotte, Pierre Versmisse, Alejandra Urrutia, Faroudy Boufassa, Françoise Barré-Sinoussi, Jean-François Delfraissy, Martine Sinet, Gianfranco Pancino, and Alain Venet. **HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype.** *Proceedings of the National Academy of Sciences of the United States of America*, **104**(16):6776–81, April 2007.

[251] Mary Carrington and Stephen J O'Brien. **The influence of HLA genotype on AIDS.** *Annual review of medicine*, **54**:535–51, January 2003.

[252] Hiroaki Mitsuya, Kent J Weinhold, Phillip A Furman, Marty H St. Clair, Sandra Nusinoff Lehrman, Robert C Gallo, Dani Bolognesi, David W Barry, and Samuel Broder. **3'-Azido-3'-deoxythymidine (BW A509U): An antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III / lymphadenopathy-associated virus in vitro**. *Proceedings of the National Academy of Sciences USA*, **82**(October):7096–7100, 1985.

[253] Robert Yarchoan, Kent J Weinhold, H K I M Lyerly, Edward Gelmann, Robert M Blum, Gene M Shearer, Hiroaki Mitsuya, Jerry M Collins, Charles E Myers, Raymond W Klecker, Phillip D Markham, David T Durack, S Nusinoff Lehrman, David W Barry, Margaret A Fischl, Robert C Gallo, Dani P Bolognesi, and Samuel Broder. **Administration of 3'-azido-3'-deoxythymidine, an inhibitor of HTLV-III/LAV replication, to patients with AIDS or AIDS-related complex**. *The Lancet*, **327**(8481):575–580, 1986.

[254] Brenden A Larder and Sharon D Kemp. **Multiple Mutations in HIV-1 Reverse Transcriptase Confer High-Level Resistance to Zidovudine (AZT)**. *Science*, **246**(4934):1155–1158, 1989.

[255] Robert Yarchoan, Hiroaki Mitsuya, and Samuel Broder. **Strategies for the Combination Therapy of HIV**. *Journal of acquired immune deficiency syndromes*, **3**(Suppl 2):S99–S103, 1990.

[256] R D Moore and R E Chaisson. **Natural history of HIV infection in the era of combination antiretroviral therapy.** *AIDS*, **13**(14):1933–42, October 1999.

[257] Ian Williams, Duncan Churchill, Jane Anderson, Marta Boffito, Mark Bower, Gus Cairns, Kate Cwynarksi, Simon Edwards, Sarah Fidler, Martin Fisher, Andrew Freedmna, Anna Maria Geretti, Yvonne Gilleece, Rob Horne, Margaret Johnson, Saye Khoo, Clifford Leen, Neal Marshall, Mark Nelson, Chloe Orkin, Nicholas Paton, Frank Post, Anton Pozniak, Caroline Sabin, Roy Trevelion, Andrew Ustianowski, John Walsh, Laura Waters, Edmund Wilkins, Alan Winston, and Mike Youle. **British HIV Association guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2012**. *HIV Medicine*, **15**(Suppl 1):1–85, 2014.

[258] Antiretroviral Therapy Cohort Collaboration. **Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies**. *Lancet*, **372**(9635):293–299, 2008.

[259] World Health Organization. **Global Update on the health sector response to HIV, 2014**. Technical report, 2014.

[260] L Imberti, A Sottini, A Bettinardi, M Puoti, and D Primi. **Selective depletion in HIV infection of T cells that bear specific T cell receptor V beta sequences.** *Science (New York, N.Y.)*, **254**(5033):860–2, November 1991.

[261] Jeffrey Laurence, Andrew S. Hodstev, and David N. Posnett. **Superantigen implicated in dependence of HIV-1 replication in T cells on TCR V beta expression**. *Nature*, **358**:255–259, 1992.

[262] P Bahadoran, F Rieux-Laucat, F Le Deist, S Blanche, A Fischer, and J P de Villartay. **Lack of selective V beta deletion in peripheral CD4+ T cells of human immunodeficiency virus-infected infants.** *European journal of immunology*, **23**(8):2041–4, August 1993.

[263] Naget Rebai, Giuseppe Pantaleo, James F Demarest, Cristina Ciurli, Hugo Soudeyns, Joseph W Adelsberger, Mauro Vaccarezza, Robert E Walker, Rafick P Sekaly, and Anthony S Fauci. **Analysis of the T-cell receptor beta-chain variable-region repertoire in monozygotic twins discordant for human immunodeficieny virus: Evidence for perturbations of specific Vbeta segments in CD4+ T cells of the virus-positive twins**. *Proceedings of the National Academy of Sciences*, **91**(February):1529–1533, 1994.

[264] M. Connors, J.A. Kovacs, S. Krevat, J.C. Gea-Banacloche, M.C. Sneller, M. Flanigan, J.A. Metcalf, R.E. Walker, J. Falloon, and M. Baseler. **HIV infection induces changes in CD4+ T-cell phenotype and depletions within the CD4+ T-cell repertoire that are not immediately restored by antiviral or immune-based therapies**. *Nature medicine*, **3**(5):533–540, 1997.

[265] J C Gea-Banacloche, L Martino, J M Mican, C W Hallahan, M Baseler, R Stevens, L Lambert, M Polis, H C Lane, and M Connors. **Longitudinal changes in CD4+ T cell antigen receptor diversity and naive/memory cell phenotype during 9 to 26 months of antiretroviral therapy of HIV-infected patients.** *AIDS research and human retroviruses*, **16**(17):1877–86, November 2000.

[266] Monica Kharbanda, Thomas W Mccloskey, Rajendra Pahwa, Mei Sun, and Savita Pahwa. **Alterations in T-Cell Receptor Vbeta Repertoire of CD4 and CD8 T Lymphocytes in Human Immunodeficiency Virus-Infected Children**. *Clinical and diagnostic laboratory immunology*, **10**(1):53–58, 2003.

[267] S Kostense, F M Raaphorst, D W Notermans, J Joling, B Hooibrink, N G Pakker, S A Danner, J M Teale, and F Miedema. **Diversity of the T-cell receptor BV repertoire in HIV-1-infected patients reflects the biphasic CD4+ T-cell repopulation kinetics during highly active antiretroviral therapy.** *AIDS*, **12**(18):F235–40, December 1998.

[268] Joseph A Conrad, Ramesh K Ramalingam, Coley B Duncan, Rita M Smith, Jie Wei, Louise Barnett, Brenna C Simons, Shelly L Lorey, and Spyros A Kalams. **Antiretroviral therapy reduces the magnitude and T cell receptor repertoire diversity of HIV-specific T cell responses without changing T cell clonotype dominance.** *Journal of virology*, **86**(8):4213–21, April 2012.

[269] Paul D Baum, Jennifer J Young, Diane Schmidt, Qianjun Zhang, Rebecca Hoh, Michael Busch, Jeffrey Martin, Steven Deeks, and Joseph M McCune. **Blood T cell receptor diversity decreases during the course of HIV infection but the potential for a diverse repertoire persists.** *Blood*, **119**:3469–3477, February 2012.

[270] Sergio Serrano-Villar, Talia Sainz, Sulggi A Lee, Peter W Hunt, Elizabeth Sinclair, Barbara L Shacklett, April L Ferre, Timothy L Hayes, Ma Somsouk, Priscilla Y Hsue, Mark L Van Natta, Curtis L Meinert, Michael M Lederman, Hiroyu Hatano, Vivek Jain, Yong Huang, Frederick M Hecht, Jeffrey N Martin, Joseph M McCune, Santiago Moreno, and Steven G Deeks. **HIV-infected individuals with low CD4/CD8 ratio despite effective antiretroviral therapy exhibit altered T cell subsets, heightened CD8+ T cell activation, and increased risk of non-AIDS morbidity and mortality.** *PLoS pathogens*, **10**(5):e1004078, May 2014.

[271] W. Ndifon, H. Gal, E. Shifrut, R. Aharoni, N. Yissachar, N. Waysbort, S. Reich-Zeliger, R. Arnon, and N. Friedman. **Chromatin conformation governs T-cell receptor J gene segment usage**. *Proceedings of the National Academy of Sciences*, **109**(39):15865–15870, September 2012.

[272] Stephen C. Threlkeld, Peggy A. Wentworth, Spyros A. Kalams, Barbara M. Wilkes, Debbie J. Ruhl, Elissa Keogh, John Sidney, Scott Southwood, Bruce D. Walker, and Alessandro Sette. **Degenerate and promiscuous recognition by CTL of peptides presented by the MHC class I A3-like superfamily**. *Journal of immunology*, **159**(4):1648–1657, 1997.

[273] T. Ueno, H. Tomiyama, and M. Takiguchi. **Single T Cell Receptor-Mediated Recognition of an Identical HIV-Derived Peptide Presented by Multiple HLA Class I Molecules**. *The Journal of Immunology*, **169**(9):4961–4969, November 2002.

[274] Geraldine M A Gillespie, Guillaume Stewart-Jones, Jaya Rengasamy, Tara Beattie, Job J Bwayo, Francis A Plummer, Rupert Kaul, Andrew J McMichael, Philippa Easterbrook, Tao Dong, E Yvonne Jones, and Sarah L Rowland-Jones. **Strong TCR conservation and altered T cell cross-reactivity characterize a B*57-restricted immune response in HIV-1 infection.** *Journal of Immunology*, **177**(6):3893–3902, September 2006.

[275] Vanessa Venturi, Hui Yee Chin, Tedi E Asher, Kristin Ladell, Phillip Scheinberg, Ethan Bornstein, David van Bockel, Anthony D Kelleher, Daniel C Douek, David A Price, and Miles P Davenport. **TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV.** *Journal of immunology (Baltimore, Md. : 1950)*, **181**(11):7853–7862, December 2008.

[276] Christoph T Berger, Nicole Frahm, David a Price, Beatriz Mothe, Musie Ghebremichael, Kari L Hartman, Leah M Henry, Jason M Brenchley, Laura E Ruff, Vanessa Venturi, Florencia Pereyra, John Sidney, Alessandro Sette, Daniel C Douek, Bruce D Walker, Daniel E Kaufmann, and Christian Brander. **High-functional-avidity cytotoxic T lymphocyte responses to HLA-B-restricted Gag-derived epitopes associated with relative HIV control.** *Journal of virology*, **85**(18):9334–45, September 2011.

[277] Daniel Mendoza, Cassandra Royce, Laura E Ruff, David R Ambrozak, Máire F Quigley, Thurston Dang, Vanessa Venturi, David A Price, Daniel C Douek, Stephen A Migueles, and Mark Connors. **HLA B*5701-positive long-term nonprogressors/elite controllers are not distinguished from progressors by the clonal composition of HIV-specific CD8+ T cells.** *Journal of virology*, **86**(7):4014–8, April 2012.

[278] Chihiro Motozono, Nozomi Kuse, Xiaoming Sun, Pierre J Rizkallah, Anna Fuller, Shinichi Oka, David K Cole, Andrew K Sewell, and Masafumi Takiguchi. **Molecular Basis of a Dominant T Cell Response to an HIV Reverse Transcriptase 8-mer Epitope Presented by the Protective Allele HLA-B*51:01.** *Journal of immunology (Baltimore, Md. : 1950)*, **192**(7):3428–34, April 2014.

[279] N. Khan, N. Shariff, M. Cobbold, R. Bruton, J. A. Ainsworth, A. J. Sinclair, L. Nayak, and P. A. H. Moss. **Cytomegalovirus Seropositivity Drives the CD8 T Cell Repertoire Toward Greater Clonality in Healthy Elderly Individuals**. *The Journal of Immunology*, **169**(4):1984–1992, August 2002.

[280] Qin Ouyang, Wolfgang M Wagner, Steffen Walter, Claudia A Müller, Anders Wikby, Geraldine Aubert, Tatjana Klatt, Stefan Stevanovic, Tony Dodi, and Graham Pawelec. **An age-related increase in the number of**

CD8+ T cells carrying receptors for an immunodominant EpsteinBarr virus (EBV) epitope is counteracted by a decreased frequency of their antigen-specific responsiveness. *Mechanisms of Ageing and Development*, **124**(4):477–485, April 2003.

[281] Rangsima Reantragoon, Alexandra J Corbett, Isaac G Sakala, Nicholas a Gherardin, John B Furness, Zhenjun Chen, Sidonia B G Eckle, Adam P Uldrich, Richard W Birkinshaw, Onisha Patel, Lyudmila Kostenko, Bronwyn Meehan, Katherine Kedzierska, Ligong Liu, David P Fairlie, Ted H Hansen, Dale I Godfrey, Jamie Rossjohn, James McCluskey, and Lars Kjer-Nielsen. **Antigen-loaded MR1 tetramers define T cell receptor heterogeneity in mucosal-associated invariant T cells.** *The Journal of experimental medicine*, **210**(11):2305–20, October 2013.

[282] Hui Yee Greenaway, Benedict Ng, David a Price, Daniel C Douek, Miles P Davenport, and Vanessa Venturi. **NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination.** *Immunobiology*, **218**(2):213–24, February 2013.

[283] Demin Li and Xiao-Ning Xu. **NKT cells in HIV-1 infection.** *Cell research*, **18**(8):817–22, August 2008.

[284] H. J. J. van der Vliet, B. M. E. von Blomberg, M. D. Hazenberg, N. Nishi, S. a. Otto, B. H. van Benthem, M. Prins, F. a. Claessen, a. J. M. van den Eertwegh, G. Giaccone, F. Miedema, R. J. Scheper, and H. M. Pinedo. **Selective Decrease in Circulating V 24+V 11+ NKT Cells During HIV Type 1 Infection**. *The Journal of Immunology*, **168**(3):1490–1495, February 2002.

[285] Caroline S Fernandez, Anthony D Kelleher, Robert Finlayson, Dale I Godfrey, and Stephen J Kent. **NKT cell depletion in humans during early HIV infection.** *Immunology and cell biology*, (January):1–13, April 2014.

[286] Evan W Newell and Mark M Davis. **Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells.** *Nature biotechnology*, **32**(2):149–57, February 2014.

[287] Brinda Emu, Walter J Moretto, Rebecca Hoh, Melissa Krone, Jeffrey N Martin, Douglas F Nixon, Steven G Deeks, and Joseph M McCune. **Composition and Function of T Cell Subpopulations Are Slow to Change Despite Effective Antiretroviral Treatment of HIV Disease.** *PloS one*, **9**(1):e85613, January 2014.

[288] Joseph P Casazza, Michael R Betts, Louis J Picker, and Richard A Koup. **Decay Kinetics of Human Immunodeficiency Virus-Specific CD8 + T Cells in Peripheral Blood after Initiation of Highly Active Antiretroviral Therapy**. *Journal of virology*, **75**(14):6508–6516, 2001.

[289] Jane Deayton, Amanda Mocroft, Pauline Wilson, Vincent C Emery, Margaret A Johnson, and Paul D Griffiths. **Loss of cytomegalovirus (CMV) viraemia following highly active antiretroviral therapy in the absence of specific anti-CMV therapy**. *AIDS*, **13**:1203–1206, 1999.

[290] S Worrell, J Deayton, P Hayes, VC Emery, F Gotch, B Gazzard, and E-L Larsson-Sciard. **Molecular correlates in AIDS patients following antiretroviral therapy: diversified Tcell receptor repertoires and in vivo control of cytomegalovirus replication**. *HIV Medicine*, **2**:11–19, 2001.

[291] T. Yokosuka. **Predominant Role of T Cell Receptor (TCR)-alpha Chain in Forming Preimmune TCR Repertoire Revealed by Clonal TCR Reconstitution System**. *Journal of Experimental Medicine*, **195**(8):991–1001, April 2002.

[292] Xiaoling Liang, Luise U Weigand, Ingrid G Schuster, Elfriede Eppinger, Judith C van der Griendt, Andrea Schub, Matthias Leisegang, Daniel Sommermeyer, Florian Anderl, Yanyan Han, Joachim Ellwart, Andreas Moosmann, Dirk H Busch, Wolfgang Uckert, Christian Peschel, and Angela M Krackhardt. **A single TCR alpha-chain with dominant peptide recognition in the allorestricted HER2/neu-specific T cell repertoire.** *Journal of immunology (Baltimore, Md. : 1950)*, **184**(3):1617–29, February 2010.

[293] Michael R. Betts, David R. Ambrozak, Daniel C. Douek, Sebastian Bonhoeffer, Jason M. Brenchley, Joseph P. Casazza, Richard A. Koup, and Louis J. Picker. **Analysis of total human immunodeficiency virus (HIV)-specific CD4+ and CD8+ T-cell responses: relationship to viral load in untreated HIV infection**. *Journal of virology*, **75**(24):11983–11991, 2001.

[294] Laura Papagno, Celsa a Spina, Arnaud Marchant, Mariolina Salio, Nathalie Rufer, Susan Little, Tao Dong, Gillian Chesney, Anele Waters, Philippa Easterbrook, P Rod Dunbar, Dawn Shepherd, Vincenzo Cerundolo, Vincent Emery, Paul Griffiths, Christopher Conlon, Andrew J McMichael, Douglas D Richman, Sarah L Rowland-Jones, and Victor Appay. **Immune activation and CD8+ T-cell differentiation towards senescence in HIV-1 infection.** *PLoS biology*, **2**(2):E20, February 2004.

[295] Sonia Bastidas, Frederik Graw, Miranda Z Smith, Herbert Kuster, Huldrych F Günthard, and Annette Oxenius. **CD8+ T cells are activated in an antigen-independent manner in HIV-infected individuals.** *Journal of immunology (Baltimore, Md. : 1950)*, **192**(4):1732–44, February 2014.

[296] Jason M Brenchley, David A Price, Timothy W Schacker, Tedi E Asher, Guido Silvestri, Srinivas Rao, Zachary Kazzaz, Ethan Bornstein, Olivier Lambotte, Daniel Altmann, Bruce R Blazar, Benigno Rodriguez, Leia Teixeira-Johnson, Alan Landay, Jeffrey N Martin, Frederick M Hecht, Louis J Picker, Michael M Lederman, Steven G Deeks, and Daniel C Douek. **Microbial translocation is a cause of systemic immune activation in chronic HIV infection.** *Nature medicine*, **12**(12):1365–71, December 2006.

[297] Saurabh Mehandru, Michael A Poles, Klara Tenner-Racz, Patrick Jean-Pierre, Victoria Manuelli, Peter Lopez, Anita Shet, Andrea Low, Hiroshi Mohri, Daniel Boden, Paul Racz, and Martin Markowitz. **Lack of Mucosal Immune Reconstitution during Prolonged Treatment of Acute and Early HIV-1 Infection**. *PLoS Medicine*, **3**(12):2335–2348, 2006.

[298] F Maldarelli, X Wu, L Su, F R Simonetti, W Shao, S Hill, J Spindler, A L Ferris, J W Mellors, M F Kearney, J M Coffin, and S H Hughes. **HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells.** *Science*, **345**:179–83, 2014.

[299] LillianB. Cohn, IsraelT. Silva, ThiagoY. Oliveira, RafaelA. Rosales, EricaH. Parrish, GeraldH. Learn, BeatriceH. Hahn, JulieL. Czartoski, M.Juliana McElrath, Clara Lehmann, Florian Klein, Marina Caskey, BruceD. Walker, JanetD. Siliciano, RobertF. Siliciano, Mila Jankovic, and MichelC. Nussenzweig. **HIV-1 Integration Landscape during Latent and Active Infection**. *Cell*, **160**(3):420–432, 2015.

[300] Matthias Egger, Bernard Hirschel, Patrick Francioli, Philippe Sudre, Marc Wirz, Markus Flepp, Martin Rickenbach, Raffaele Malinverni, Pietro Vernazza, and Manuel Battegay. **Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study**. *BMJ*, **315**:1194–1199, 1997.

[301] Matthias Egger, Margaret May, Geneviève Chêne, Andrew N Phillips, Bruno Ledergerber, François Dabis, Dominique Costagliola, Antonella D Arminio Monforte, Frank De Wolf, Peter Reiss, Jens D Lundgren, Amy C Justice, Schlomo Staszewski, Catherine Leport, Robert S Hogg, Caroline A Sabin, M John Gill, Bernd Salzberger, and Jonathan A C Sterne. **Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy : a collaborative analysis of prospective studies**. *The Lancet*, **360**:119–129, 2002.

[302] Joseph C Mudd and Michael M Lederman. **CD8 T cell persistence in treated HIV infection.** *Current opinion in HIV and AIDS*, **9**(5):500–5, September 2014.

[303] Victor Leung, Jennifer Gillis, Janet Raboud, Curtis Cooper, Robert S Hogg, Mona R Loutfy, Nima Machouf, Julio S G Montaner, Sean B Rourke, Chris Tsoukas, and Marina B Klein. **Predictors of CD4:CD8 ratio normalization and its effect on health outcomes in the era of combination antiretroviral therapy.** *PloS one*, **8**(10):e77665, January 2013.

[304] Edwin Leeansyah, Anupama Ganesh, F Quigley, Anders So, Jan Andersson, Peter W Hunt, Ma Somsouk, Steven G Deeks, Jeffrey N Martin, Markus Moll, Barbara L Shacklett, and Johan K Sandberg. **Activation, exhaustion, and persistent decline of the antimicrobial MR1-restricted MAIT-cell population in chronic HIV-1 infection**. *Blood*, **121**(7):1124–1135, 2013.

[305] Johanna Maria Eberhard, Philip Hartjen, Silke Kummer, Reinhold E Schmidt, Maximilian Bockhorn, Clara Lehmann, Ashwin Balagopal, Joachim Hauber, Jan van Lunzen, and Julian Schulze Zur Wiesch. **CD161+ MAIT Cells Are Severely Reduced in Peripheral Blood and Lymph Nodes of HIV-Infected Individuals Independently of Disease Progression.** *PloS one*, **9**(11):e111323, January 2014.

[306] O-Jin Lee, Young-Nan Cho, Seung-Jung Kee, Moon-Ju Kim, Hye-Mi Jin, Sung-Ji Lee, Ki-Jeong Park, Tae-Jong Kim, Shin-Seok Lee, Yong-Soo Kwon, Nacksung Kim, Myung-Geun Shin, Jong-Hee Shin, Soon-Pal Suh, Dong-Wook Ryang, and Yong-Wook Park. **Circulating mucosal-associated invariant T cell levels and their cytokine levels in healthy adults.** *Experimental gerontology*, **49**:47–54, January 2014.

[307] Alison Motsinger, David W Haas, Aleksandar K Stanic, Luc Van Kaer, Sebastian Joyce, and Derya Unutmaz. **CD1d-restricted human natural killer T cells are highly susceptible to human immunodeficiency virus 1 infection.** *The Journal of experimental medicine*, **195**(7):869–79, April 2002.

[308] Nadine G Pakker, Daan W Notermans, Rob J De Boer, Marijke T L Roos, Frank De Wolf, Andrew Hill, John M Leonard, Sven A Danner, Frank Miedema, and Peter T A Schellekens. **Biphasic kinetics of peripheral blood T cells after triple combination therapy in HIV-1 infection: a composite of redistribution and proliferation**. *Nature medicine*, **4**(2):208–214, 1998.

[309] Taisheng Li, Ning Wu, Yi Dai, Zhifeng Qiu, Yang Han, Jing Xie, Ting Zhu, and Yanling Li. **Reduced thymic output is a major mechanism of immune reconstitution failure in HIV-infected patients after long-term antiretroviral therapy.** *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, **53**(9):944–51, November 2011.

[310] Ineke den Braber, Tendai Mugwagwa, Nienke Vrisekoop, Liset Westera, Ramona Mögling, Anne Bregje de Boer, Neeltje Willems, Elise H R Schrijver, Gerrit Spierenburg, Koos Gaiser, Erik Mul, Sigrid a Otto, An F C Ruiter, Mariette T Ackermans, Frank Miedema, José a M Borghans, Rob J de Boer, and Kiki Tesselaar. **Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans.** *Immunity*, **36**(2):288–97, February 2012.

[311] J.D. Watson and F.H.C. Crick. **Molecular structure of nucleic acids**. *Nature*, **171**(4356):709–756, 1953.

[312] Doris T. Zallen. **Despite Franklin's work, Wilkins earned his Nobel**. *Nature*, **425**(September):15, 2003.

[313] F. Sanger, G.G. Brownlee, and B.G. Barrell. **A two-dimensional fractionation procedure for radioactive nucleotides**. *Journal of Molecular Biology*, **13**(2):373–IN4, September 1965.

[314] GG Brownlee and F Sanger. **Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli**. *Journal of molecular biology*, **23**:337–353, 1967.

[315] JM Adams, PGN Jeppesen, F Sanger, and BG Barrell. **Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA**. *Nature*, **228**(September):1009–1014, 1969.

[316] W Min-Jou, G Haegeman, M Ysebaert, and W Fiers. **Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein**. *Nature*, **237**(May):82–88, 1972.

[317] W Fiers, R Contreras, F Duerinck, G Haegeman, D Iserentant, J Merregaert, W Min-Jou, F Molemans, A Raeymaekers, A Van den Berghe, G Volckaert, and M Ysebaert. **Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene**. *Nature*, **260**:500–507, 1976.

[318] Ray Wu. **Nucleotide sequence analysis of DNA**. *Nature*, **51**:501–521, 1972.

[319] R Padmanabhan and Ray Wu. **Nucleotide Sequence Analysis of DNA**. *Biochemical and biophysical research communications*, **48**(5):1295–1302, 1972.

[320] F Sanger, J E Donelson, A R Coulson, H Kossel, and D Fischer. **Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA**. *Proceedings of the National Academy of Sciences USA*, **70**(4):1209–1213, 1973.

[321] R Padmanabhan, Ernest Jay, and R Wu. **Chemical Synthesis of a Primer and Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4**. *Proceedings of the National Academy of Sciences USA*, **71**(6):2510–2514, 1974.

[322] F Sanger and AR Coulson. **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *Journal of molecular biology*, **94**:441–448, 1975.

[323] Allan M Maxam and Walter Gilbert. **A new method for sequencing DNA**. *Proceedings of the National Academy of Sciences USA*, **74**(2):560–564, 1977.

[324] A. R. Coulson Fred Sanger, S Nicklen. **DNA sequencing with chain-terminating**. *Proceedings of the National Academy of Sciences*, **74**(12):5463–5467, 1977.

[325] F. Sanger, G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, J.C. Fiddes, C.A. Hutchison, P.M. Slocombe, and M. Smith. **Nucleotide sequence of bacteriophage phi X174 DNA**. *Nature*, **265**:687–695, 1977.

[326] ZG Chidgeavadze and R. Sh. Beabealashvilli. **2', 3'-Dideoxy-3'aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases.** *Nucleic acids research*, **12**(3):1671–1686, 1984.

[327] Lloyd M Smith, Steven Fung, Michael W Hunkapiller, Tim J Hunkapiller, and Leroy E Hood. **The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus synthesis of fluorescent DNA primers for use in DNA sequence analysis**. *Nucleic acids research*, **13**(7):2399–2412, 1985.

[328] Wilhelm Ansorge, Brian S Sproat, Josef Stegemann, and Christian Schwager. **A non-radioactive automated method for DNA sequence determination**. *Journal of Biochemical and Biophysical Methods*, **13**:315–323, 1986.

[329] Wilhelm Ansorge, Brian Sproat, Josef Stegemann, Christian Schwager, and Martin Zenke. **Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis**. *Nucleic acids research*, **15**(11):4593–4602, 1987.

[330] James M Prober, George L Trainor, Rudy J Dam, Frank W Hobbs, Charles W Robertson, Robert J Zagursky, Anthony J Cocuzza, Mark A Jensen, and Kirk Baumeister. **DNA Sequencing with Rapid for System Fluorescent Chain-Terminating Dideoxynuclcotides**. *Science*, **238**(4825):336–341, 1987.

[331] Hideki Kambara, Tetsuo Nishikawa, Yoshiko Katayama, and Tomoaki Yamaguchi. **Optimization of parameters in a DNA sequenator using fluorescence detection**. *Nature biotechnology*, **6**:816–820, 1988.

[332] Harold Swerdlow and Raymond Gesteland. **Capillary gel electrophoresis for rapid, high resolution DNA sequencing**. *Nucleic acids research*, **18**(6):1415–1419, 1990.

[333] JA Luckey and Howard Drossman. **High speed DNA sequencing by capillary electrophoresis**. *Nucleic acids research*, **18**(15):4417–4421, 1990.

[334] T Hunkapiller, RJ Kaiser, BF Koop, and L Hood. **Large-scale and automated DNA sequence determination**. *Science*, **254**(5028):59–67, 1991.

[335] R Staden. **A strategy of DNA sequencing employing computer programs**. *Nucleic acids research*, **6**(7):2601–2610, 1979.

[336] Stephen Anderson. **Shotgun DNA sequencing using cloned DNase I-generated fragments**. *Nucleic Acids Research*, **9**(13):3015–3027, 1981.

[337] David A Jackson, Robert H Symonst, and Paul Berg. **Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli**. *Proceedings of the National Academy of Sciencess of the USA*, **69**(10):2904–2909, 1972.

[338] Stanley N Cohen, Annie C Y Chang, Herbert W Boyert, and Robert B Hellingt. **Construction of Biologically Functional Bacterial Plasmids In Vitro**. *Proceedings of the National Academy of Sciencess of the USA*, **70**(11):3240–3244, 1973.

[339] H Klenow and I Henningsen. **Selective Elimination of the Exonuclease Activity of the Deoxyribonucleic Acid Polymerase from Escherichia coli B by Limited Proteolysis**. *Proceedings of the National Academy of Sciencess of the USA*, **65**(2):168–175, 1970.

[340] Cheng-Yao Chen. **DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present.** *Frontiers in microbiology*, **5**(June):305, January 2014.

[341] Lloyd M Smith, Jane Z Sanders, Robert J Kaiser, Peter Hughes, Chriss Dodd, Charles R Connel, Cheryl Heiner, Stephen B H Kent, and Leroy E Hood. **Fluorescence detection in automated DNA sequence analysis**. *Nature*, **321**(June):674–679, 1986.

[342] Wilhelm J Ansorge. **Next-generation DNA sequencing techniques.** *New biotechnology*, **25**(4):195–203, April 2009.

[343] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Sandra W. Clifton, Patrick J. Minx, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Marvin Frazier, Christopher Elkin, Edward Uberbacher, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Todd Taylor, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Jean Weissenbach, Roland Heilig, William Saurin, Patrick Wincker, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Andreas Rump, Catherine Robert, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, JoAnn Dubois, Douglas R. Smith, Jun Gu, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Shizen Qin, Hong Mei Lee, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Leroy Hood, Lee Rowen, Anup Madan, The Institute for Systems Biology:, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Gabriele Nordsiek, Helmut Blöcker, Klaus Hornischer, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church,

Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, and Aristides Patrinos. **Initial sequencing and analysis of the human genome**. *Nature*, **409**(February):860–921, 2001.

[344] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C a Evans, R a Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, a G Clark, J Nadeau, V a McKusick, N Zinder, a J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, a E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K a Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, a K Naik, V a Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. **The sequence of the human genome.** *Science*, **291**(5507):1304–51, February 2001.

[345] På l Nyrén and Arne Lundin. **Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis**. *Analytical biochemistry*, **509**:504–509, 1985.

[346] Edward David Hyman. **A New Method for Sequencing DNA**. *Analytical biochemistry*, **174**:423–436, 1988.

[347] På l Nyrén. **Enzymatic method for continuous monitoring of DNA polymerase activity**. *Analytical biochemistry*, **238**:235–238, 1987.

[348] Mostafa Ronaghi, Samer Karamohamed, Bertil Pettersson, Mathias Uhlen, and På l Nyrén. **Real-Time DNA Sequencing Using Detection of Pyrophosphate Release**. *Analytical biochemistry*, **242**:84–89, 1996.

[349] Mostafa Ronaghi, Mathias Uhlen, and På l Nyrén. **A Sequencing Method Based on Real-Time Pyrophosphate**. *Science*, **281**(5375):363–365, July 1998.

[350] Marcel Margulies, Michael Egholm, WE Altman, and Said Attiya. **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature*, **437**(7057):376–380, 2005.

[351] Dan S Tawfik and Andrew D Griffiths. **Man-made cell-like compartments for molecular evolution**. *Nature biotechnology*, **16**:652–656, 1998.

[352] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul a Kravitz, Dana a Busam, Karen Y Beeson, Tina C McIntosh, Karin a Remington, Josep F Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E Frazier, Stephen W Scherer, Robert L Strausberg, and J Craig Venter. **The diploid genome sequence of an individual human.** *PLoS biology*, **5**(10):e254, September 2007.

[353] David a Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing-zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard a Gibbs, and Jonathan M Rothberg. **The complete genome of an individual by massively parallel DNA sequencing.** *Nature*, **452**(7189):872–6, April 2008.

[354] Karl V Voelkerding, Shale a Dames, and Jacob D Durtschi. **Next-generation sequencing: from basic research to diagnostics.** *Clinical chemistry*, **55**(4):641–58, April 2009.

[355] Jay Shendure and Hanlee Ji. **Next-generation DNA sequencing.** *Nature biotechnology*, **26**(10):1135–45, October 2008.

[356] Milan Fedurco, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.** *Nucleic acids research*, **34**(3):e22, January 2006.

[357] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant,

Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, ScottM. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, AndrewA. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. ONeill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, ComeRaczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M.Seoane, Richard J. Shaw, Mitch K. Shiver, StevenW.Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie VandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, JingwenWang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, and Rogers{, Jane and Mullikin, James C. and Hurles, Matthew E. and McCooke{, Nick J. and West, John S. and Oaks, Frank L. and Lundberg, Peter L. and Klenerman, David and Durbin, Richard and Smith, Anthony J. **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature*, **456**(6):53–9, November 2008.

[358] Gerardo Turcatti, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi. **A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.** *Nucleic acids research*, **36**(4):e25, March 2008.

[359] Shankar Balasubramanian. **Sequencing nucleic acids: from chemistry to medicine**. *Chemical Communications*, **47**(26):7281–7286, 2011.

[360] Michael a Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC genomics*, **13**(1):341, January 2012.

[361] Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, and Konstantinos T Konstantinidis. **Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample.** *PloS one*, **7**(2):e30087, January 2012.

[362] Kevin Judd Mckernan, Heather E Peckham, Gina L Costa, F Mclaughlin, Yutao Fu, Eric F Tsung, Christopher R Clouser, Cisyla Duncan, Jeffrey K Ichikawa, Clarence C Lee, Zheng Zhang, Swati S Ranade, T Dimalanta, Fiona C Hyland, Tanya D Sokolsky, Lei Zhang, Andrew Sheridan, Haoning Fu, Cynthia L Hendrickson, Bin Li, Lev Kotler, Jeremy R Stuart, A Malek, Jonathan M Manning, Alena A Antipova, Damon S Perez, P Moore, Kathleen C Hayashibara, Michael R Lyons, Robert E Beaudoin, E Coleman, Michael W Laptewicz, Adam E Sannicandro, Michael D Rhodes, Rajesh K Gottimukkala, Shan Yang, Vineet Bafna, Ali Bashir, Andrew Macbride, Can Alkan, Jeffrey M Kidd, Evan E Eichler, Martin G Reese, M Francisco, De La Vega, and Alan P Blanchard. **Sequence and structural variation in a human genome uncovered by short-read , massively parallel ligation sequencing using two-base encoding**. *Genome Research*, **19**:1527–1541, 2009.

[363] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science*, **309**(5741):1728–32, September 2005.

[364] H P J Buermans and J T den Dunnen. **Next generation sequencing technology: Advances and applications.** *Biochimica et biophysica acta*, **1842**(10):1932–1941, July 2014.

[365] Travis C Glenn. **Field guide to next-generation DNA sequencers.** *Molecular ecology resources*, **11**(5):759–69, September 2011.

[366] Radoje Drmanac, Andrew B Sparks, Matthew J Callow, Aaron L Halpern, Norman L Burns, Bahram G Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P Pant, Jonathan Baccash, Adam P Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C Ebert, Coleen R Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E McBride, Matt Morenzoni, Robert E Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock a Peters, Joe Peterson, Charit L Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W Shannon, Conrad G Sheppy, Michel Sun, Joseph V Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R Oliphant, William C Banyai, Bruce Martin, Dennis G Ballinger, George M Church, and Clifford a Reid. **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.** *Science (New York, N.Y.)*, **327**(5961):78–81, January 2010.

[367] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, Jeremy Hoon, Jan F Simons, David Marran, Jason W Myers, John F Davidson, Annika Branting, John R

Nobile, Bernard P Puc, David Light, Travis a Clark, Martin Huber, Jeffrey T Branciforte, Isaac B Stoner, Simon E Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline a Fidanza, Eugeni Namsaraev, Kevin J McKernan, Alan Williams, G Thomas Roth, and James Bustillo. **An integrated semiconductor device enabling non-optical genome sequencing.** *Nature*, **475**(7356):348–52, July 2011.

[368] Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. **Performance comparison of benchtop high-throughput sequencing platforms.** *Nature biotechnology*, **30**(5):434–9, May 2012.

[369] LD Stein. **The case for cloud computing in genome informatics**. *Genome Biology*, **11**:207, 2010.

[370] William J Greenleaf and Arend Sidow. **The future of sequencing : convergence of intelligent design and market Darwinism**. *Genome biology*, **15**:303, 2014.

[371] Illumina. **Illuminas Board Unanimously Rejects Roches Unsolicited Tender Offer as Inadequate**. http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1658034, 2012. Accessed: 2014-10-19.

[372] Eric E Schadt, Steve Turner, and Andrew Kasarskis. **A window into third-generation sequencing.** *Human molecular genetics*, **19**(R2):R227–40, October 2010.

[373] Thomas P Niedringhaus, Denitsa Milanova, Matthew B Kerby, Michael P Snyder, and Annelise E Barron. **Landscape of next-generation sequencing technologies**. *Analytical chemistry*, **83**(12):4327–4341, 2011.

[374] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. **Sequencing technologies and genome sequencing.** *Journal of applied genetics*, **52**(4):413–35, November 2011.

[375] Ivo Glynne Gut. **New sequencing technologies.** *Clinical & translational oncology*, **15**(11):879–81, November 2013.

[376] Ido Braslavsky, Benedict Hebert, Emil Kartalov, and Stephen R Quake. **Sequence information can be obtained from single DNA molecules**. *Proceedings of the National Academy of Sciencess of the USA*, **100**(7):3960–3964, 2003.

[377] Timothy D Harris, Phillip R Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, Jennifer Colonell, James Dimeo, J William Efcavitch, Eldar Giladi, Jaime Gill, John Healy, Mirna Jarosz, Dan Lapen, Keith Moulton, Stephen R Quake, Kathleen Steinmann, Edward Thayer, Anastasia Tyurina, Rebecca Ward, Howard Weiss, and Zheng Xie. **Single-molecule DNA sequencing of a viral genome.** *Science (New York, N.Y.)*, **320**(5872):106–9, April 2008.

[378] Jayson Bowers, Judith Mitchell, Eric Beer, Philip R Buzby, Marie Causey, J William Efcavitch, Mirna Jarosz, Edyta Krzymanska-Olejnik, Li Kung, Doron Lipson, Geoffrey M Lowman, Subramanian Marappan, Peter McInerney, Adam Platt, Atanu Roy, Suhaib M Siddiqi, Kathleen Steinmann, and John F Thompson. **Virtual terminator nucleotides for next-generation DNA sequencing.** *Nature methods*, **6**(8):593–5, August 2009.

[379] GenomeWeb. **Helicos BioSciences Files for Chapter 11 Bankruptcy Protection**, 2012. Accessed: 2014-10-20.

[380] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. **Ten years of next-generation sequencing technology**. *Trends in Genetics*, **30**(9), August 2014.

[381] M J Levene, J Korlach, S W Turner, M Foquet, H G Craighead, and W W Webb. **Zero-mode waveguides for single-molecule analysis at high concentrations.** *Science*, **299**(5607):682–6, January 2003.

[382] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, and Geoff Otto. **Real-time DNA sequencing from single polymerase molecules**. *Science*, **323**(January):133–138, 2009.

[383] Benjamin a Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson a Clark, Jonas Korlach, and Stephen W Turner. **Direct detection of DNA methylation during single-molecule, real-time sequencing.** *Nature methods*, **7**(6):461–5, June 2010.

[384] Farzin Haque, Jinghong Li, Hai-Chen Wu, Xing-Jie Liang, and Peizuan Guo. **Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA**. *Nano today*, **8**(1):56–74, 2013.

[385] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. **Characterization of individual polynucleotide molecules using a membrane channel**. *Proceedings of the National Academy of Sciencess of the USA*, **93**(November):13770–13773, 1996.

[386] Jiali Li, Derek Stein, Ciaran Mcmullan, and Daniel Branton. **Ion-beam sculpting at nanometre length scales**. *Nature*, **412**(July):2–5, 2001.

[387] C Dekker. **Solid-state nanopores**. *Nature nanotechnology*, **2**:209–215, 2007.

[388] James Clarke, HC Wu, Lakmal Jayasinghe, and Alpesh Patel. **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nature nanotechnology*, **4**(April):265–270, 2009.

[389] Michael Eisenstein. **Oxford Nanopore announcement sets sequencing sector abuzz.** *Nature biotechnology*, **30**(4):295–6, April 2012.

[390] N. J. Loman and a. R. Quinlan. **Poretools: a toolkit for analyzing nanopore sequence data**. *Bioinformatics*, pages 1–3, August 2014.

[391] Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean, Robert Riehn, Gautam V Soni, Vincent Tabardcossa, and Meni Wanunu. **The potential and challenges of nanopore sequencing**. *Nature biotechnology*, **26**(10):1146–1153, 2008.

[392] F Paillard, G Sterkers, and C Vaquero. **Transcriptional and post-transcriptional regulation of TcR, CD4 and CD8 gene expression during activation of normal human T lymphocytes.** *The EMBO journal*, **9**(6):1867–72, June 1990.

[393] Florence Paillard, Ghyslaine Sterkers, and Catherine Vaquero. **Correlation between up-regulation of lymphokine mRNA dn down-regulation of TcR, CD4, CD8 and lack mRNA as shwon by the effect of CsA on activated T lymphocytes**. *Biochemical and biophysical research communications*, **186**(2):603–611, 1992.

[394] Michael A Frohman, Michael K Dush, and Gail R Martin. **Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer**. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(December):8998–9002, 1988.

[395] MT Suzuki and SJ Giovannoni. **Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.** *Applied and environmental microbiology*, **62**(2):625–630, 1996.

[396] Martin F Polz and Colleen M Cavanaugh. **Bias in Template-to-Product Ratios in Multitemplate PCR Bias in Template-to-Product Ratios in Multitemplate PCR**. *Applied and environmental microbiology*, **64**(10):3724–3730, 1998.

[397] Lisa Föhse, Janine Suffner, Karsten Suhre, Benjamin Wahl, Cornelia Lindner, Chun-Wei Lee, Susanne Schmitz, Jan D Haas, Stella Lamprecht, Christian Koenecke, André Bleich, Günter J Hämmerling, Bernard Malissen, Sebastian Suerbaum, Reinhold Förster, and Immo Prinz. **High TCR Diversity ensures optimal Function and Homeostasis of FoxP3(+) Regulatory T cells.** *European journal of immunology*, **41**(11):1–37, September 2011.

[398] Harlan S Robins, Cindy Desmarais, Jessica Matthis, Robert Livingston, Jessica Andriesen, Helena Reijonen, Christopher Carlson, Gerold Nepom, Cassian Yee, and Karen Cerosaletti. **Ultra-sensitive detection of rare T cell clones**. *Journal of Immunological Methods*, **375**(1-2):14–19, January 2012.

[399] Christopher S Carlson, Ryan O Emerson, Anna M Sherwood, Cindy Desmarais, Moon-Wook Chung, Joseph M Parsons, Michelle S Steen, Marissa a Lamadrid-Herrmannsfeldt, David W Williamson, Robert J Livingston, David Wu, Brent L Wood, Mark J Rieder, and Harlan Robins. **Using synthetic templates to design an unbiased multiplex PCR assay.** *Nature communications*, **4**:2680, October 2013.

[400] Shuo Li, Marie-Paule Lefranc, John J Miles, Eltaf Alamyar, Véronique Giudicelli, Patrice Duroux, J Douglas Freeman, Vincent D a Corbin, Jean-Pierre Scheerlinck, Michael a Frohman, Paul U Cameron, Magdalena Plebanski, Bruce Loveland, Scott R Burrows, Anthony T Papenfuss, and Eric J Gowans. **IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling.** *Nature communications*, **4**(May):2333, September 2013.

[401] W M Schmidt and M W Mueller. **CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs.** *Nucleic acids research*, **27**(21):e31, November 1999.

[402] Pawel Zajac, Saiful Islam, Hannah Hochgerner, Peter Lönnerberg, and Sten Linnarsson. **Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases.** *PloS one*, **8**(12):e85270, January 2013.

[403] Eli Gilboa, Sudha W Mitra, Stephen Goff, and David Baltimore. **A Detailed Model of Reverse Transcription of Crucial Aspects and Tests**. *Cell*, **18**(September):93–100, 1979.

[404] D G Peters, a B Kassam, H Yonas, E H O'Hare, R E Ferrell, and a M Brufsky. **Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite.** *Nucleic acids research*, **27**(24):e39, December 1999.

[405] Phuong Nguyen, Jing Ma, Deqing Pei, Caroline Obert, Cheng Cheng, and Terrence L Geiger. **Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire.** *BMC genomics*, **12**:106, January 2011.

[406] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. **Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment**. *Genome Research*, **8**(3):175–185, March 1998.

[407] Brent Ewing and Phil Green. **Base-calling of Automated Sequencer Traces Using Phred. II. error probabilities**. *Genome research*, **8**(3):186, 1998.

[408] T H Rabbitts, M P Lefranc, M a Stinson, J E Sims, J Schroder, M Steinmetz, N L Spurr, E Solomon, and P N Goodfellow. **The chromosomal location of T-cell receptor genes and a T cell rearranging gene: possible correlation with specific translocations in human T cell leukaemia.** *The EMBO journal*, **4**(6):1461–5, June 1985.

[409] J. E. Sims, A. Tunnacliffe, W. J. Smith, and T. H. Rabbitts. **Complexity of human T-cell antigen receptor beta-chain constant- and variable-region genes**. *Nature*, **312**:515–545, 1984.

[410] U Schneider, H U Schwenk, and G Bornkamm. **Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma.** *International journal of cancer. Journal international du cancer*, **19**(5):621–6, May 1977.

[411] Robert T Abraham and Arthur Weiss. **Jurkat T cells and development of the T-cell receptor signalling paradigm.** *Nature reviews. Immunology*, **4**(4):301–8, April 2004.

[412] H Abts, M Emmerich, S Miltenyi, a Radbruch, and H Tesch. **CD20 positive human B lymphocytes separated with the magnetic cell sorter (MACS) can be induced to proliferation and antibody secretion in vitro.** *Journal of immunological methods*, **125**(1-2):19–28, December 1989.

[413] S Miltenyi, W Müller, W Weichel, and A Radbruch. **High gradient magnetic cell separation with MACS.** *Cytometry*, **11**(2):231–238, 1990.

[414] Xi Yang, Di Liu, Fei Liu, Jun Wu, Jing Zou, Xue Xiao, Fangqing Zhao, and Baoli Zhu. **HTQC: a fast quality control toolkit for Illumina sequencing data.** *BMC bioinformatics*, **14**(1):33, January 2013.

[415] Yaniv Erlich, Partha P Mitra, W Richard Mccombie, and Gregory J Hannon. **Alta-Cyclic : a self- optimizing base caller for next-generation sequencing**. *Nature Methods*, **5**(8):679–682, 2008.

[416] Christian Ledergerber and Christophe Dessimoz. **Base-calling for next-generation sequencing platforms.** *Briefings in bioinformatics*, **12**(5):489–97, September 2011.

[417] N R Landau, T P St John, I L Weissman, S C Wolf, a E Silverstone, and D Baltimore. **Cloning of terminal transferase cDNA by antibody screening.** *Proceedings of the National Academy of Sciences of the United States of America*, **81**(18):5836–40, September 1984.

[418] Yan Wang, Dennis E Prosen, Li Mei, John C Sullivan, Michael Finney, and Peter B Vander Horn. **A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro.** *Nucleic acids research*, **32**(3):1197–207, January 2004.

[419] Peter McInerney, Paul Adams, and Masood Z Hadi. **Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase.** *Molecular biology international*, **2014**:287430, January 2014.

[420] Aron Fazekas, Royce Steeves, and Steven Newmaster. **Improving sequencing quality from PCR products containing long mononucleotide repeats.** *BioTechniques*, **48**(4):277–85, April 2010.

[421] E Viguera, D Canceill, and S D Ehrlich. **Replication slippage involves DNA polymerase pausing and dissociation.** *The EMBO journal*, **20**(10):2587–95, May 2001.

[422] Roche. **Sequencing Solutions Technical Note**. http://roche-biochem.jp/prima/pdf/TechNote_WhatisFlowPatternB_August2013.pdf, 2013.

[423] Susan M Huse, Julie a Huber, Hilary G Morrison, Mitchell L Sogin, and David Mark Welch. **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome biology*, **8**(7):R143, January 2007.

[424] Illumina. **Illumina Announces MiSeq(TM) Personal Sequencing System**, 2011.

[425] Illumina. **Off-Line Basecaller Software v1.6 User Guide Rev**, 2009.

[426] Felix Krueger, Simon R Andrews, and Cameron S Osborne. **Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling.** *PloS one*, **6**(1):e16607, January 2011.

[427] S P Balk, E C Ebert, R L Blumenthal, F V McDermott, K W Wucherpfennig, S B Landau, and R S Blumberg. **Oligoclonal expansion and CD1 recognition by human intestinal intraepithelial lymphocytes.** *Science (New York, N.Y.)*, **253**(5026):1411–5, September 1991.

[428] Sébastien Wälchli, Geir Å ge Lø set, Shraddha Kumari, Jorunn Nergå rd Johansen, Weiwen Yang, Inger Sandlie, and Johanna Olweus. **A practical approach to T-cell receptor cloning and expression.** *PloS one*, **6**(11):e27930, January 2011.

[429] Genome Web. **Roche Shutting Down 454 Sequencing Business**. http://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business, 2013. Accessed: 2014-09-19.

[430] Brian C. Schaefer. **Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends**. *Analytical biochemistry*, **277**:255–273, 1995.

[431] K Maruyama and S Sugano. **Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides.** *Gene*, **138**(1-2):171–4, January 1994.

[432] J B Edwards, J Delort, and J Mallet. **Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5' ends of mRNAs and for constructing cDNA libraries by in vitro amplification.** *Nucleic acids research*, **19**(19):5227–32, October 1991.

[433] Anthony B Troutt, Michael G McHeyzer-Williams, Bali Pulendran, and G. J. V. Nossal. **Ligation-anchored PCR: a simple amplification technique with single-sided specificity.** *Proceedings of the National Academy of Sciencess of the USA*, **89**(October):9823–9825, 1992.

[434] Illumina. **Low-Diversity Sequencing on the Illumina**. http://www.illumina.com/documents/products/technotes/technote_low_diversity_rta.pdf, 2013.

[435] Glenn K Fu, Jing Hu, Pei-Hua Wang, and Stephen P a Fodor. **Counting individual DNA molecules by the stochastic attachment of diverse labels.** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(22):9026–31, May 2011.

[436] James a Casbon, Robert J Osborne, Sydney Brenner, and Conrad P Lichtenstein. **A method for counting PCR template molecules with application to next-generation sequencing.** *Nucleic acids research*, **39**(12):e81, July 2011.

[437] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. **Counting absolute numbers of molecules using unique molecular identifier**. *Nature Methods*, **9**(1):1–5, 2012.

[438] Joshua A. Weinstein, Ning Jiang, Richard A. White, Daniel S. Fisher, and Stephen R. Quake. **High-throughput sequencing of the zebrafish antibody repertoire.** *Science (New York, N.Y.)*, **324**(5928):807–10, May 2009.

[439] AP Bird. **DNA methylation and the frequency of CpG in animal DNA**. *Nucleic Acids Research*, **8**(7):1499–1504, 1980.

[440] Bradley Hall, John M Micheletti, Pooja Satya, Krystal Ogle, Jack Pollard, and Andrew D Ellington. **Design, synthesis, and amplification of DNA pools for in vitro selection.** *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 24**(October):Unit 24.2, October 2009.

[441] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. **Biases in Illumina transcriptome sequencing caused by random hexamer priming.** *Nucleic acids research*, **38**(12):e131, July 2010.

[442] Peter C Fitzgerald, Andrey Shlyakhtenko, Alain A Mir, and Charles Vinson. **Clustering of DNA sequences in human promoters**. *Genome research*, **14**:1562–1574, 2004.

[443] Vladimir B Bajic, Sin Lam Tan, Alan Christoffels, Christian Schönbach, Leonard Lipovich, Liang Yang, Oliver Hofmann, Adele Kruger, Winston Hide, Chikatoshi Kai, Jun Kawai, David a Hume, Piero Carninci, and Yoshihide Hayashizaki. **Mice and men: their promoter properties.** *PLoS genetics*, **2**(4):e54, April 2006.

[444] Simon J Van Heeringen, Waseem Akhtar, Ulrike G Jacobi, Robert C Akkers, Yutaka Suzuki, and Gert Jan C Veenstra. **Nucleotide composition-linked divergence of vertebrate core promoter architecture**. *Genome research*, **21**:410–421, 2011.

[445] Fanglei Zhuang, Ryan T Fuchs, Zhiyi Sun, Yu Zheng, and G Brett Robb. **Structural bias in T4 RNA ligase-mediated 3'-adapter ligation.** *Nucleic acids research*, **40**(7):e54, April 2012.

[446] M P Lefranc, V Giudicelli, C Busin, J Bodmer, W Müller, R Bontrop, M Lemaitre, a Malik, and D Chaume. **IMGT, the International ImMunoGeneTics database.** *Nucleic acids research*, **26**(1):297–303, January 1998.

[447] Altaf Alamyar, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie-paule Lefranc. **IMGT/HIGHV-QUEST: THE IMGT WEB PORTAL FOR IMMUNOGLOBULIN (IG) OR ANTIBODY AND T CELL RECEPTOR (TR) ANALYSIS FROM NGS HIGH THROUGHPUT AND DEEP SEQUENCING**. *Immunome Research*, **8**(1):2, May 2012.

[448] W. James Kent. **BLAT – The BLAST-Like Alignment Tool**. *Genome Research*, **12**(4):656–664, March 2002.

[449] Stephen F Altschup, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. **Basic Local Alignment Search Tool**. *Journal of Molecular Biology*, **215**:403–410, 1990.

[450] M.S. Waterman, T.F. Smith, and W.A. Beyer. **Some Biological Sequence Metrics ***. *Advances in Mathematics*, **20**:367–387, 1976.

[451] Alfred V. Aho and Margaret J. Corasick. **Efficient string matching: an aid to bibliographic search**. *Communications of the ACM*, **18**(6):333–340, June 1975.

[452] Saima Hasib, Mahak Motwani, and Amit Saxena. **Importance of Aho-Corasick String Matching Algorithm in Real World Applications**. *International Journal of Computer Science and Information Technologies*, **4**(3):467–469, 2013.

[453] Véronique Giudicelli and Marie-Paule Lefranc. **Ontology for immunogenetics: the IMGT-ONTOLOGY**. *Bioinformatics*, **15**(12):1047–1054, 1999.

[454] Dmitriy a Bolotin, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, Maria a Turchaninova, Ivan V Zvyagin, Olga V Britanova, and Dmitriy M Chudakov. **MiTCR: software for T-cell receptor sequencing data analysis.** *Nature methods*, (8), July 2013.

[455] Stephanie Gras, Zhenjun Chen, John J Miles, Yu Chih Liu, Melissa J Bell, Lucy C Sullivan, Lars Kjer-Nielsen, Rebekah M Brennan, Jacqueline M Burrows, Michelle a Neller, Rajiv Khanna, Anthony W Purcell, Andrew G Brooks, James McCluskey, Jamie Rossjohn, and Scott R Burrows. **Allelic polymorphism in the T cell receptor and its impact on immune responses.** *The Journal of experimental medicine*, **207**(7):1555–67, July 2010.

[456] Rebekah M Brennan, Jacqueline M Burrows, Thomas Elliott, Michelle a Neller, Stephanie Gras, Jamie Rossjohn, John J Miles, and Scott R Burrows. **Missense single nucleotide polymorphisms in the human T cell receptor loci control variable gene usage in the T cell repertoire.** *British journal of haematology*, pages 1–4, March 2014.

[457] Wenjing Pan, Miranda Byrne-Steele, Chunlin Wang, Stanley Lu, Scott Clemmons, Robert J Zahorchak, and Jian Han. **DNA polymerase preference determines PCR priming efficiency.** *BMC biotechnology*, **14**(1):10, January 2014.

[458] Katsuyuki Shiroguchi, Tony Z Jia, Peter a Sims, and X Sunney Xie. **Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes.** *Proceedings of the National Academy of Sciences of the United States of America*, **109**(4):1347–52, January 2012.

[459] Ilgar Z. Mamedov, Olga V. Britanova, Ivan V. Zvyagin, Maria a. Turchaninova, Dmitriy a. Bolotin, Ekaterina V. Putintseva, Yuriy B. Lebedev, and Dmitriy M. Chudakov. **Preparing Unbiased T-Cell Receptor and Antibody cDNA Libraries for the Deep Next Generation Sequencing Profiling**. *Frontiers in Immunology*, **4**(December):1–10, 2013.

[460] Mikhail Shugay, Olga V Britanova, Ekaterina M Merzlyak, Maria a Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy a Bolotin, Dmitry B Staroverov, Ekaterina V Putintseva, Karla Plevova, Carsten Linnemann, Dmitriy Shagin, Sarka Pospisilova, Sergey Lukyanov, Ton N Schumacher, and Dmitriy M Chudakov. **Towards error-free profiling of immune repertoires.** *Nature methods*, (April):1–5, May 2014.

[461] Takahiro Kanagawa. **Bias and artifacts in multitemplate polymerase chain reactions (PCR)**. *Journal of Bioscience and Bioengineering*, **96**(4):317–323, 2003.

[462] Xiaoyun Qiu, Liyou Wu, Heshu Huang, Patrick E Mcdonel, Anthony V Palumbo, James M Tiedje, and Jizhong Zhou. **Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning**. *Applied and environmental microbiology*, **67**(2):880–887, 2001.

[463] E Pienaar, M Theron, M Nelson, and HJ Viljoen. **A quantitative model of error accumulation during PCR amplification**. *Computational biology and chemistry*, **30**(2):102–111, 2006.

[464] **immunoSEQ**. `http://www.immunoseq.com/`. Accessed: 2014-10-16.

[465] **iRepertoire**. `http://www.irepertoire.com`. Accessed: 2014-10-16.

[466] Charles A Janeway. **Approaching the Asymptote ? Evolution and Revolution in Immunology.** *Cold Spring Harbor Symposia on Quantitative Biology*, **45**:1–13, 1989.

[467] Marielle C Gold, Stefania Cerri, Susan Smyk-Pearson, Meghan E Cansler, Todd M Vogt, Jacob Delepine, Ervina Winata, Gwendolyn M Swarbrick, Wei-Jen Chua, Yik Y L Yu, Olivier Lantz, Matthew S Cook, Megan D Null, David B Jacoby, Melanie J Harriff, Deborah a Lewinsohn, Ted H Hansen, and David M Lewinsohn. **Human mucosal associated invariant T cells detect bacterially infected cells.** *PLoS biology*, **8**(6):e1000407, January 2010.

[468] By Mark Exley, Jorge Garcia, Steven P Balk, and Steven Porcelli. **Requirements for CD1d Recognition by Human Invariant Valpha24+ CD4-CD8- T Cells**. *Journal of experimental medicine*, **186**(1):109–120, 1997.

[469] M Han, L Harrison, P Kehn, K Stevenson, J Currier, and M a Robinson. **Invariant or highly conserved TCR alpha are expressed on double-negative (CD3+CD4-CD8-) and CD8+ T cells.** *Journal of immunology (Baltimore, Md. : 1950)*, **163**(1):301–11, July 1999.

[470] Joseph P Sanderson, Kathrin Waldburger-Hauri, Diana Garzón, Gediminas Matulis, Salah Mansour, Nicholas J Pumphrey, Nikolai Lissin, Peter M Villiger, Bent Jakobsen, José D Faraldo-Gómez, and Stephan D Gadola. **Natural variations at position 93 of the invariant Vα24-Jα18 α chain of human iNKT-cell TCRs strongly impact on CD1d binding.** *European journal of immunology*, **42**(1):248–55, January 2012.

[471] Giuseppe Pantaleo, James F. Demarest, Hugo Soudeyns, Cecilia Graziosi, Francois Denis, Joseph W. Adelsberger, Persephone Borrow, Michael S. Saag, George M. Shaw, Rafick P. Sekaly, and Anthony S. Fauci. **Major expansion of CD8+ T cells with a predominant Vbeta usage during the primary immune response to HIV**. *Nature*, **370**:463–467, 1994.

[472] W Kolowos, M Schmitt, M Herrman, E Harrer, P Löw, J R Kalden, and T Harrer. **Biased TCR repertoire in HIV-1-infected patients due to clonal expansion of HIV-1-reverse transcriptase-specific CTL clones.** *Journal of Immunology*, **162**(12):7525–33, June 1999.

[473] Z C Kou, J S Puhr, M Rojas, W T McCormack, M M Goodenow, and J W Sleasman. **T-Cell receptor Vbeta repertoire CDR3 length diversity differs within CD45RA and CD45RO T-cell subsets in healthy and human immunodeficiency virus-infected children.** *Clinical and diagnostic laboratory immunology*, **7**(6):953–9, November 2000.

[474] Sang Kyung Lee, Zhan Xu, Judy Lieberman, and Premlata Shankar. **The functional CD8 T cell response to HIV becomes type-specific in progressive disease**. *The Journal of Clinical Investigation*, **110**(9):1339–1347, 2002.

[475] Mathias Lichterfeld, Katie L Williams, Stanley K Mui, Shivani S Shah, Bianca R Mothe, Alessandro Sette, Arthur Kim, Mary N Johnston, Nicole Burgett, Nicole Frahm, Daniel Cohen, Christian Brander, Eric S Rosenberg, Bruce D Walker, Marcus Altfeld, and Xu G Yu. **T cell receptor cross-recognition of an HIV-1 CD8+ T cell epitope presented by closely related alleles from the HLA-A3 superfamily.** *International immunology*, **18**(7):1179–88, July 2006.

[476] Dirk Meyer-Olson, Kristen W Brady, Melissa T Bartman, Kristin M O'Sullivan, Brenna C Simons, Joseph A Conrad, Coley B Duncan, Shelly Lorey, Atif Siddique, Rika Draenert, Marylyn Addo, Marcus Altfeld, Eric Rosenberg, Todd M Allen, Bruce D Walker, and Spyros A Kalams. **Fluctuations of functionally distinct CD8+ T-cell clonotypes demonstrate flexibility of the HIV-specific TCR repertoire.** *Blood*, **107**(6):2373–83, March 2006.

[477] Mathias Lichterfeld, Xu G Yu, Stanley K Mui, Katie L Williams, Alicja Trocha, Mark A Brockman, Rachel L Allgaier, Michael T Waring, Tomohiko Koibuchi, Mary N Johnston, Daniel Cohen, Todd M Allen, Eric S Rosenberg, Bruce D Walker, and Marcus Altfeld. **Selective depletion of high-avidity human immunodeficiency virus type 1 (HIV-1)-specific CD8+ T cells after early HIV-1 infection.** *Journal of virology*, **81**(8):4199–214, April 2007.

[478] Xu G Yu, Mathias Lichterfeld, Katie L Williams, Javier Martinez-Picado, and Bruce D Walker. **Random T-cell receptor recruitment in human immunodeficiency virus type 1 (HIV-1)-specific CD8+ T cells from genetically identical twins infected with the same HIV-1 strain.** *Journal of virology*, **81**(22):12666–9, November 2007.

[479] Brenna C. Simons, Scott E. VanCompernolle, Rita M. Smith, Jie Wei, Louise Barnett, Shelly L. Lorey, Dirk Meyer-Olson, and Spyros A. Kalams. **Despite Biased TRBV Gene Usage against a Dominant HLA Recognition of Circulating Epitope Variants**. *The Journal of Immunology*, **181**:5137–5146, 2008.

[480] Maria Candela Iglesias, Jorge R Almeida, Solène Fastenackels, David J van Bockel, Masao Hashimoto, Vanessa Venturi, Emma Gostick, Alejandra Urrutia, Linda Wooldridge, Mathew Clement, Stéphanie Gras, Pascal G Wilmann, Brigitte Autran, Arnaud Moris, Jamie Rossjohn, Miles P Davenport, Masafumi Takiguchi, Christian Brander, Daniel C Douek, Anthony D Kelleher, David a Price, and Victor Appay. **Escape from highly effective public CD8+ T-cell clonotypes by HIV.** *Blood*, **118**(8):2138–49, August 2011.

[481] Crystal Y Chen, Dan Huang, Shuyu Yao, Lisa Halliday, Gucheng Zeng, Richard C Wang, and Zheng W Chen. **IL-2 simultaneously expands Foxp3+ T regulatory and T effector cells and confers resistance to severe tuberculosis (TB): implicative Treg-T effector cooperation in immunity to TB.** *Journal of immunology (Baltimore, Md. : 1950)*, **188**(9):4278–88, May 2012.

[482] Joseph A. Conrad, Ramesh K. Ramalingam, Rita M. Smith, Louise Barnett, Shelly L. Lorey, Jie Wei, Brenna C. Simons, Shanmugalakshmi Sadagopal, Dirk Meyer-Olson, and Spyros A. Kalams. **Dominant clonotypes within HIV-specific T cell responses are PD-1hi and CD127low reactivity and display reduced variant cross-reactivity**. *Journal of Immunology*, **186**(12):6871–6885, 2012.

[483] G. B. Stewart-Jones, P. Simpson, P. Anton van der Merwe, P. Easterbrook, A. J. McMichael, S. L. Rowland-Jones, E. Y. Jones, and G. M. Gillespie. **Structural features underlying T-cell receptor sensitivity to concealed MHC class I micropolymorphisms**. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1–10, November 2012.

[484] I. V. Zvyagin, M. V. Pogorelyy, M. E. Ivanova, E. A. Komech, M. Shugay, D. A. Bolotin, A. A. Shelenkov, A. A. Kurnosov, D. B. Staroverov, D. M. Chudakov, Y. B. Lebedev, and I. Z. Mamedov. **Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing**. *Proceedings of the National Academy of Sciences*, **111**(16):5980–5985, April 2014.

[485] James M. Heather. **Known specificity human TCR CDR3 sequences**. http://dx.doi.org/10.6084/m9.figshare.1153817, 2014.

[486] Jian Han, David C Swan, Sharon J Smith, Shanjuan H Lum, Susan E Sefers, Elizabeth R Unger, and Yi-Wei Tang. **Simultaneous Amplification and Identification of 25 Human Papillomavirus Types with Templex Technology Simultaneous Amplification and Identification of 25 Human Papillomavirus Types with Templex Technology**. *Journal of clinical microbiology*, **44**(11):4157–62, November 2006.

[487] Jian Han. **Molecular differential diagnoses of infectious diseases: is the future now?** In Yiwei Tang and Charles W Stratton, editors, *Advanced Technologies in Diagnostic Microbiology*, chapter Molecular. Springer Publishing Company, New York, 2006.

[488] J Han. **Amplicon rescue multiplex polymerase chain reaction for amplification of multiple targets (Patent US7999092)**. http://www.google.com/patents/US7999092, 2011.