Sequencing, analyzing, and modeling small samples from large T cell repertoires

Bram Gerritsen

Cover A metaphor for the decomposition of T cell repertoires into individual TCR sequences. Concept: Bram Gerritsen. Design: Robbert Gerritsen
Print Gildeprint
ISBN 978-90-393-6930-2

Sequencing, analyzing, and modeling small samples from large T cell repertoires

Sequencen, analyseren, en modelleren van kleine samples uit grote T cel repertoires

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 22 januari 2018 des middags te 4.15 uur

door

Bram Gerritsen

geboren op 16 juli 1982 te Nijmegen

Promotor:	Prof. dr. R.J. De Boer
Copromotoren:	Dr. A. Pandit
	Dr. A.C. Andeweg

The studies described in this thesis were financially supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch government (FES0908).

Assessment committee:

Prof. dr. Benjamin Chain
Prof. dr. Antoine van Kampen
Prof. dr. Grant Lythe
Prof. dr. Paul Thomas
Prof. dr. Aleksandra Walczak

Contents

Chapter 1

General Introduction

1.1 Introduction

In this thesis we study T cell receptor (TCR) repertoires. Every T cell expresses one or two TCR variants on its surface to detect alterations in the body, for example infections or cancer. T cells kill infected cells, strengthen immune responses, and provide immune memory. Using their TCR, T cells recognize antigens in the form of peptides presented by the Major Histocompatibility Complex (MHC) on the surface of other cells, peptide-MHC (pMHC). The pMHCs that are recognized by T cells in an immune response are called "epitopes". T cells that have never encountered their epitope (cognate antigen) are called naïve T cells. Upon encountering their epitope, a naïve T cell expands and differentiates into effector and memory T cells. The effector T cells can kill infected host cells or provide help to other immune cells. Memory T cells last longer and enable a stronger and faster immune response upon secondary encounter of the same epitope (e.g., a later infection by the same or similar pathogen). The collection of T cells and different TCRs, is called a repertoire. T cell repertoires reflect present and past host exposure to pathogens, and are extraordinarily diverse. Assessment of this diversity can give insight into the capacity of the adaptive immune system to mount a specific response. For example, a pathogen may not be killed by the immune system if there are no T cells in the repertoire with a TCR capable of recognizing the pathogen. The immune system can also kill healthy cells (and cause autoimmunity) if naïve T cells respond to self-antigens (i.e., peptides derived from the body's own proteins). Therefore, the immune system has to balance its repertoire, making sure it is diverse enough to recognize any pathogen, but at the same time it should not react to self-antigens.

The human genome encodes about 20,000 genes, far fewer than the millions of different TCRs generated by the adaptive immune system. The TCR is a heterodimer, consisting of $\alpha\beta$, or $\gamma\delta$ chains. T cells expressing the $\alpha\beta$ TCR are most abundant in human, and are the primary focus in this thesis. The TCR is generated through a stochastic process of gene rearrangement, where Variable (V), Diversity (D), and Joining (J), gene segments are joined for the β chain, and V and J gene segments are joined together for the α chain. Nucleotides are randomly inserted and deleted at the junctions where the gene segments are joined into a region called the Complementarity Determining Region 3 (CDR3). The CDR3, which starts near the 3' end of the V gene segment, and extends partly into the J gene segment, is the primary region of the TCR that contacts antigen (Freeman *et al.*, 2009). The recombination process can (theoretically) generate more different TCRs (approximately 10^{61} (Mora and Walczak, 2016)) than there are stars in the visible universe. However, the actual number of different TCRs employed by the

human adaptive immune system is still an open question. In summary, the repertoire is randomly generated and is dynamic because it changes from day to day. By drawing small blood samples (see below) we try to gain a deeper understanding of the complexity and dynamics of TCR repertoires.

1.2 Repertoire Sequencing (Rep-Seq)

The first draft sequence of the human genome was published in 2001 by the Human Genome Project (Lander *et al.*, 2001). It was a huge challenge spanning several years to sequence the about 3 billion basepairs of the human genome. Since then, sequencing technology has progressed to the point that a single sequencer, such as the HiSeq X, can sequence about 1800 genomes a year to 30× coverage (Goodwin *et al.*, 2016). New scientific questions are being addressed using high-throughput sequencing (HTS), with characterization of TCR repertoires being among the most complex and difficult, because repertoires are highly diverse and samples typically contain only a tiny fraction ($1{:}10^6$) of the full repertoire. Moreover, it is difficult to distinguish rare TCR sequences from erroneous variants due to sequencing and PCR errors. In this thesis, only the sequencers from Illumina were used. Illumina sequencing is well established and the most commonly used HTS technique (Heather *et al.*, 2017; Rosati *et al.*, 2017). Below we give a brief overview of how TCR repertoires are sequenced using Illumina sequencing.

1.2.1 Overview of a high throughput sequencing experiment

HTS experiments of T cell repertoires typically consist of three steps. First, genetic material (gDNA or RNA) is isolated from T lymphocytes, second, TCR sequences are amplified using PCR, and third, the resulting amplicons are sequenced. Commonly, only one of the chains of the ($\alpha\beta$ or $\gamma\delta$) TCR heterodimer is sequenced, as pairing information is lost due to pooling of the genetic material. Some methods, such as "pairSEQ" (Howie *et al.*, 2015), attempt to overcome this limitation. Typically, only the β chain is sequenced because most ($97 - 99\%$ (Brady *et al.*, 2010)) T cells express one β chain due to allelic exclusion, and because the β chain also rearranges the D gene, generating more repertoire diversity than the α chain (Woodsworth *et al.*, 2013).

Depending on the study, either gDNA or RNA may be the best starting material, and we here describe some of the considerations. gDNA is more stable than RNA, contains

information about failed TCR rearrangements (Murugan *et al.*, 2012; Rosati *et al.*, 2017), and does not vary in quantity from T cell to T cell like the number of mRNA molecules (Benichou *et al.*, 2012; Mamedov *et al.*, 2013)). However, gDNA requires multiplex primer sets to amplify the 5' V gene end (and the 3' J gene end, or the intron), limiting the finding of novel V genes, and preventing the sequencing of the CDR1 and CDR2 regions (as the V gene primers are generally near the CDR3). Additionally, use of multiple primers can potentially introduce PCR amplification bias for certain V (and J) genes (Benichou *et al.*, 2012; Mamedov *et al.*, 2013; Rosati *et al.*, 2017). RNA allows for more efficient PCR as TCR mRNA comprises a larger proportion of the starting material than TCR gDNA, limiting the loss of rare TCR sequences (Mamedov *et al.*, 2013). More importantly, RNA allows the use of 5'RACE (see below), enabling the quantification of the number of mRNA molecules by application of unique molecular identifiers (UMIs; see below) (Rosati *et al.*, 2017).

The next step is the preparation of a library of TCR cDNA sequences. During library preparation, TCR sequences are amplified by PCR, adapters are added to the sequences, and enough material (amplicons) is produced for the sequencer. The library preparation method used in the studies in this thesis is rapid amplification of 5' complementary DNA ends (5'RACE) (Mamedov *et al.*, 2013; Matz *et al.*, 1999), briefly described here. Starting with mRNA, a primer targeting the TCR constant region, in combination with reverse transcriptase, is used to synthesize the initial cDNA molecules. The reverse transcriptase produces a homopolymer overhang of about 2-5 bases (usually dCTP) at the 3' end of the initial cDNA. A template-switch oligonucleotide that hybridizes with the non-templated homopolymer overhang allows the reverse transcriptase to switch templates and synthesize the complementary DNA strand. Importantly, the constant region and template-switch oligos are independent of specific TCR rearrangements, enabling the amplification of any TCR sequence. One or more additional PCR steps are performed to enrich the sample for TCR specific sequences and to add adapters (Mamedov *et al.*, 2013; Rosati *et al.*, 2017).

Finally, the amplicons (referred to as templates from now) are put on the Illumina sequencer. The templates hybridize with their adapter sequences to oligos that are attached to a solid surface. Through a PCR step, called bridge amplification (Mardis, 2008; Shendure and Ji, 2008), clusters of identical template sequences are generated, followed by sequencing-by-synthesis (SBS). During a SBS cycle, polymerase enzymes incorporate maximally one additional base through the use of modified nucleotides, which contain a fluorophore and are reversibly blocked by a 3'-O-azidomethyl group. The solid surface ("flow cell lane") is imaged, after which the chemical block and fluorophores are

removed, and the next cycle begins (Goodwin *et al.*, 2016; Mardis, 2008; Shendure and Ji, 2008). Illumina software analyzes the images, identifying clusters (of identical template molecules), and identifying which base was incorporated in each cluster based on the fluerescence signal. The stack of images produces for each cluster a string of bases, a "sequence read".

The benefit of high throughput sequencing is that millions of reads can be generated, the downside is that reads (typically 100-300bp) are generally not not long enough to span the full TCR V(D)J region (about 500 bp long). Fortunately, not the whole TCR sequence is required, because a complete TCR chain can be inferred from the CDR3 region nucleotide sequence (about 50 bp), in combination with V and J gene identification (Thomas *et al.*, 2013). Hence, studies use various methods such as size fractionation, PCR primers near the CDR3, and paired-end sequencing to optimally capture the CDR3 region of the TCR. The bioinformatic analysis of the reads needs to take the methodology into account, because identification of V and J genes may become ambiguous due to short read lengths.

1.2.2 Low level sequence processing and analysis

Typically, there are three steps involved to go from raw sequencing data to a set of TCR sequences. First, there is general processing such as demultiplexing and merging of paired-end reads, second, retrieval of TCR sequences from the sequence reads, third, error-correction of the TCR sequences. In recent years, dedicated tools have been developed for Rep-Seq, which perform one or more of these steps (Alamyar *et al.*, 2012b; Gerritsen *et al.*, 2016; Giraud *et al.*, 2014; Giudicelli *et al.*, 2004; Hung *et al.*, 2016; Kuchenbecker *et al.*, 2015; Shugay *et al.*, 2014; Thomas *et al.*, 2013; Vander Heiden *et al.*, 2014; Yang *et al.*, 2015; Ye *et al.*, 2013; Yu *et al.*, 2015; Zhang *et al.*, 2015) (for details see Heather *et al.* (2017)). To retrieve TCR sequences from the sequence data, typically, germline V and J genes are aligned to the reads, and the CDR3 regions are identified (Alamyar *et al.*, 2012b; Gerritsen *et al.*, 2016; Giudicelli *et al.*, 2004; Hung *et al.*, 2016; Kuchenbecker *et al.*, 2015; Thomas *et al.*, 2013; Yang *et al.*, 2015; Ye *et al.*, 2013; Yu *et al.*, 2015; Zhang *et al.*, 2015). Unfortunately, experimental artifacts, such as biased amplification and PCR- and sequencing errors, distort both the number of and abundance of identified TCR sequences, hampering the quantitative analysis and comparison of TCR repertoires (Benichou *et al.*, 2012; Bolotin *et al.*, 2012; Woodsworth *et al.*, 2013). Below, we describe a few of the approaches to deal with these artifacts.

The sequencer outputs raw sequence reads with for every base an associated quality score, which reflects the fidelity of the base call. Base calling errors (i.e. errors introduced during the SBS step) can be removed by discarding sequences containing low quality scores (Nguyen *et al.*, 2011; Warren *et al.*, 2011). Since erroneous TCR sequences tend to be rare in a sample, another method is to remove all rare TCR sequences (Warren *et al.*, 2011). Unfortunately, filtering based on abundance and base quality can lead to the loss of a large proportion of the sequence data (up to 50% or more depending on the platform (Bolotin *et al.*, 2012)). Additionally, PCR errors occur independent of base quality scores and those occurring in the early cycles are amplified along with the correct sequences, potentially leading to abundant erroneous TCR sequences (Heather *et al.*, 2017).

An alternative approach to error correction, is to cluster TCR sequences, merging less abundant TCRs to more abundant TCRs that have similar sequences. This idea is based on a common signature of PCR- and sequencing errors, where abundant TCRs tend to be surrounded by erroneous variants. Some tools have arbirtrary thresholds for sequence similarity (for example a Hamming distance ≤ 2 (Robins *et al.*, 2009)) and ratio between two TCR sequences before merging the less abundant to the more abundant TCR sequence (Bolotin *et al.*, 2012; Kuchenbecker *et al.*, 2015). Depending on the choice of threshold, either too many or too few sequences may be merged during clustering, potentially leading to a loss of genuine TCR sequences. Additionally, it may not be clear what the optimal thresholds are for any given dataset. To retain as many genuine TCR sequences while removing the erroneous ones, we developed a pipeline that automatically chooses its error correction thresholds based on the sequencing data, and varies them for TCR sequences within the data based on a simple statistical model (Gerritsen *et al.*, 2016).

A recent development in Rep-Seq, is the application of unique molecular identifiers (UMIs) (Kivioja *et al.*, 2011), which are short stretches of (about 12) random nucleotides that are incorporated before the first PCR. This enables the quantification of the number of (mRNA) molecules initially present in the sample by counting the number of UMIs associated with a particular TCR, as opposed to counting the number of sequence reads. In addition, sequence and PCR errors can be corrected by collapsing groups of sequence reads associated with the same UMI (Shugay *et al.*, 2014). Several tools incorporate UMI-based error correction, such as MIGEC (Shugay *et al.*, 2014) and Decombinator (Thomas *et al.*, 2013), and UMI error correction can be combined with a clustering approach (for example by combining MIGEC and MiXCR (Bolotin *et al.*, 2015)). Although UMIs are extremely useful for error correction, simply discarding UMIs which have only

few sequence reads (which hampers error correction), can quickly lead to a considerable loss of genuine TCR sequences (Rosati *et al.*, 2017). UMIs themselves can also contain errors, and recently a tool has been published, called UMI-tools (Smith *et al.*, 2017), which specifically attempts to error correct the UMIs themselves.

1.3 Thesis overview

T cells play a crucial role in enabling the immune system to recognize and combat pathogens, from helping B cells to become more specific to a pathogen, to directly killing infected host cells. The T cell receptor (TCR) is the central player that enables a T cell to specifically respond to an epitope, for example when naïve T cells expand and differentiate into effector T cells. The availability of many T cells expressing different TCRs (i.e. a T cell repertoire) enables the adaptive immune system to recognize and deal with novel pathogens, and to quickly dispatch pathogens it has seen before (i.e., a memory response). T cell repertoires (within the memory, naïve, and regulatory T-cell pools) hold a wealth of information about the function and health of the adaptive immune system. In this thesis, we apply bioinformatic methods, develop computational methods, and devise mathematical models to study various aspects of T cell repertoires, from repertoire dynamics to the maintainance of a naïve repertoire.

Characterizing T cell repertoires is challenging, because repertoires are much larger (i.e. more diverse) than the samples that are sequenced. Additionally, TCRs may differ from each other by as little as a single nucleotide, making it difficult to distinguish erroneous sequences from genuine TCRs. We developed a computational pipeline, called Recover TCR (RTCR), specialized in the complete and accurate retrieval of highly diverse TCR repertoires from high throughput sequencing (HTS) data (Chapter 2). RTCR uses a statistical model to correct PCR- and sequencing errors, and it estimates appropriate parameters for the error correction from the data (i.e., a "data-driven" approach). Next, we apply this pipeline to longitudinal HTS data from blood samples of a healthy volunteer, investigating repertoire dynamics when there is no apparent infection or other trigger for an immune response (Chapter 3). We find that the frequencies of large memory TCRB clonotypes fluctuate over time, presumably due to mounted immune responses. Next, to understand how a diverse naïve repertoire is maintained, we develop mathematical models describing the clone-size distribution of the naïve repertoire. Surprisingly, we find that naïve clone-sizes are to a large extent determined by VDJ recombination probabilities (Chapter 4). Finally, using HTS data processed with the RTCR pipeline, we describe

the expressed TRB V, D, J, and C genes in the Ferret, allowing for more detailed future research into the adaptive T cell response of this animal model (Chapter 5).

Chapter 2

RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data

BRAM GERRITSEN, ARIDAMAN PANDIT, ARNO C. ANDEWEG, AND ROB J. DE BOER (2016)

Bioinformatics, 32(20):3098-3106

Abstract

Motivation: High Throughput Sequencing (HTS) has enabled researchers to probe the human T cell receptor (TCR) repertoire, which consists of many rare sequences. Distinguishing between true but rare TCR sequences and variants generated by PCR and sequencing errors remains a formidable challenge. The conventional approach to handle errors is to remove low quality reads, and/or rare TCR sequences. Such filtering discards a large number of true and often rare TCR sequences. However, accurate identification and quantification of rare TCR sequences is essential for repertoire diversity estimation.

Results: We devised a pipeline, called Recover TCR (RTCR), that accurately recovers TCR sequences, including rare TCR sequences, from HTS data (including barcoded data) even at low coverage. RTCR employs a data-driven statistical model to rectify PCR and sequencing errors in an adaptive manner. Using simulations we demonstrate that RTCR can easily adapt to the error profiles of different types of sequencers and exhibits consistently high recall and high precision even at low coverages where other pipelines perform poorly. Using published real data we show that RTCR accurately resolves sequencing errors and outperforms all other pipelines.

Availability: The RTCR pipeline is implemented in Python (v2.7) and C and is freely available at http://uubram.github.io/RTCR/ along with documentation and examples of typical usage.

Contact: b.gerritsen@uu.nl

2.1 Introduction

T cells are crucial to the adaptive immune system, enabling it to recognize almost any pathogen that infects the host while remaining tolerant to many self antigens. The recognition of antigens by T cells is mediated by the T cell receptor (TCR). Through random genetic recombination, the immune system can potentially equip every T cell with a different TCR, allowing it to bind different antigens than other T cells. The different T cells together form a T cell repertoire, which due to its pivotal role in the immune response, is studied extensively in areas such as infectious diseases, cancer, autoimmunity and ageing (Bolotin *et al.*, 2012; Suessmuth *et al.*, 2015; Woodsworth *et al.*, 2013).

Classical TCRs are heterodimers, consisting of $\alpha\beta$ protein chains. The genes encoding the α and β chains are generated via somatic stochastic DNA rearrangements, in which germline variable (V), diversity (D), and joining (J) gene segments recombine (Bassing *et al.*, 2002). Random deletions and non-templated nucleotide insertions occur at the V(D)J junctions, which together with the random selection of gene segments is responsible for generating a full repertoire of TCRs. Theoretically, about 5×10^{11} different TCRβ chains are possible (Robins *et al.*, 2010), which together with the TCRα chains can result in more than 10^{15} distinct TCRs (Davis and Bjorkman, 1988). Because humans have about 10^{12} T cells (Arstila *et al.*, 1999), every individual harbours at most only a small but diverse (Arstila *et al.*, 1999; Qi *et al.*, 2014; Robins *et al.*, 2010; Warren *et al.*, 2011) fraction of this potential repertoire.

High throughput sequencing (HTS) (Holt and Jones, 2008; Shendure and Ji, 2008) is often used to probe the TCR repertoire (Freeman *et al.*, 2009; Klarenbeek *et al.*, 2010; Ndifon *et al.*, 2012; Robins *et al.*, 2009, 2010; Wang *et al.*, 2010; Warren *et al.*, 2011). Analysing the repertoire is challenging because of its high diversity and many rare clonotypes, pushing the boundaries of sequencing technologies. Some of the important confounding factors in quantification and identification of the repertoire are: short read length (e.g. 150bp reads whereas the VDJ region of the TCR is about 500bp), unequal PCR amplification, sequencing errors, and sampling biases (Baum *et al.*, 2012; Calis and Rosenberg, 2014; Nguyen *et al.*, 2011; Robins *et al.*, 2009; Warren *et al.*, 2011). Despite the improvements in sequencing technologies, quantification of true TCR repertoire diversity remains elusive, because the repertoire is heavily undersampled and sequencing errors artificially skew the repertoire.

To analyse HTS data from TCR repertoires, multiple pipelines have been developed. Some of these are: IMGT (Alamyar *et al.*, 2012a), which provides a web interface and

detailed annotations, iSSAKE (Warren *et al.*, 2009), which assembles immune receptors from very short reads, IRmap (Wang *et al.*, 2010), designed for 454 sequencing data, Decombinator (Thomas *et al.*, 2013), designed for fast annotation, MiTCR (Bolotin *et al.*, 2013), MiXCR (Bolotin *et al.*, 2015) and IMSEQ (Kuchenbecker *et al.*, 2015) focusing on error correction, Presto (Vander Heiden *et al.*, 2014) and MiGEC (Shugay *et al.*, 2014) handling reads with unique molecular identifiers ("barcoded" data), and TCRklass (Yang *et al.*, 2015) annotating all reads including those that lack the CDR3 of the TCR. Most of these pipelines filter out low quality reads and/or remove rare TCR sequences. Since most unique TCR sequences are rare, such filtering can cause a massive loss of true TCR sequences.

We developed a pipeline, called Recover TCR (RTCR), that attempts to accurately recover TCR sequences at varying coverage, including rare TCR sequences while maintaining high precision and high recall. Accurate quantification of TCR repertoires is especially important in clinical settings, where low coverage TCR sequencing can be used for cost effectiveness. There are multiple ways to identify sequence errors in HTS data. Some of these are: i) base quality, i.e. a low quality base is more likely to be false than a high quality base, and ii) similarity, i.e. true TCR sequences tend to be surrounded by similar erroneous variants due to PCR and sequencing errors. In RTCR, these strategies are translated into a simple binomial model (Nguyen *et al.*, 2011) together with several heuristics to rationally eliminate PCR and sequencing errors. RTCR automatically sets its parameters based on the data, relieving the user from setting arbitrary parameters. RTCR supports 'barcoded' HTS data, combining barcode-based error correction with its regular error correction.

To measure the performance of RTCR, we compared it to TCRklass, MiTCR, MiXCR, IMSEQ, and MiGEC, using simulated and real HTS datasets. We demonstrate that RTCR can easily adapt to error profiles of different types of sequencers and exhibits consistently high recall and high precision at even low coverage. We benchmark different pipelines using several synthetic TCR HTS datasets generated via realistic PCR and sequencing simulations. We find that RTCR outperforms all other pipelines on recall and matches the high precision of MiTCR, MiXCR and IMSEQ. Using real data we then show that RTCR can accurately resolve apparent sequencing errors which are incompletely resolved by other pipelines.

2.2 Methods

2.2.1 Recover T cell receptor (RTCR) pipeline

RTCR is a pipeline for identification and data-driven error correction of TCR sequences from HTS data. The pipeline was written in Python v2.7 and C, and provides an easy to use command line interface. Below we will explain the steps the RTCR pipeline takes to analyse an HTS dataset.

Reads obtained from HTS are typically too short to span the whole TCR gene and are error-prone. If a read contains the CDR3 region of a TCR, the corresponding TCR gene can be uniquely identified (provided the read also contains enough of the flanking V and J segment nucleotides to unambiguously determine the correct V and J segment). Every base in a read is assigned a 'Phred' score (Q), which indicates the probability (p) of an erroneous base call by the sequencer: $Q = -10 \log_{10} p$. To infer TCR sequences, RTCR aligns germline V and J segments to the reads using an external aligner. We chose Bowtie 2 (Langmead and Salzberg, 2012) as the default aligner for RTCR, because it is fast, accurate (data not shown), and uses Phred scores to score the alignments. The pipeline can easily be configured to use a different aligner. The D segments are not aligned to the reads because it is difficult to align them unambiguously and excluding the D segments does not change the inference of a TCR sequence. RTCR uses the alignments to identify and extract the CDR3 region from every read, annotating it with the V and J segment identified by the aligner. Sets of identical CDR3 sequences are collapsed as follows: i) a single CDR3 sequence is kept and assigned the number of sequences in the set as its abundance, ii) each position in the CDR3 sequence is assigned the highest Phred score found at that position in the set, and iii) the CDR3 sequence is assigned the VJ segment combination most common in the set, breaking ties using the alignment score assigned by the aligner. An option is provided to the user to prevent RTCR from collapsing CDR3 sequences with identical CDR3 but different V and J segment annotation. We chose to collapse all identical CDR3 sequences by default to avoid generation of false TCR sequences due to ambiguity in segment annotation.

It is known that PCR and sequencing experiments can generate errors in some reads which would inflate the number of distinct TCR sequences (Baum *et al.*, 2012; Bolotin *et al.*, 2012; Nguyen *et al.*, 2011). RTCR uses a simple statistical model to estimate the number of erroneous sequences in the data and the total number of errors these sequences may contain. Let ϵ be the probability of an error for a base in a read. If we assume all

bases are independent and are erroneous with the same probability (ϵ), then a sequence (i.e. a string of bases) can be modeled as a set of Bernoulli trials. The probability of having exactly h errors in a sequence of length l is then given by the conventional binomial:

$$p_h = \binom{l}{h} \epsilon^h (1-\epsilon)^{l-h} \quad \text{, for } h \in \{0, 1, \ldots, l\}. \tag{2.1}$$

Next, consider a set of n sequences, each of length l. There are $n \times p_h$ sequences expected to have exactly h errors, and the maximum number of errors expected to occur in at least one sequence is:

$$H = \max(\{h : np_h \geq 1\}). \tag{2.2}$$

Consider for example n as the number of times a particular TCR sequence of length l has been sequenced, then there are expected to be np_0 correct copies of it in the data. The remaining $n(1-p_0)$ erroneous copies are spread across an unknown number of distinct variants and there is expected to be at least one erroneous copy having H mismatches with the TCR sequence.

The QMerge, IMerge and LMerge algorithms that are explained below, use Eq.(2.1) and Eq.(2.2) together with several heuristics to determine which and how many sequences are likely to be erroneous. The algorithms depend primarily on the per base error rate (ϵ). RTCR estimates this error rate (ϵ) from the number of mismatches observed in the aligned germline sequences with the reads. RTCR calculates two separate error rates, one from the V alignments with the CDR3 region, and one from the J alignments with the CDR3 region. To remain conservative in the number of distinct TCR sequences recovered, the higher of the two error rates is assigned to ϵ.

Both QMerge and IMerge group TCR sequences by length and error correct each length group independently. Due to stochasticity and experimental bias, the true number of mismatches in a length group may be higher or lower than expected given the error rate, ϵ, which was calculated using *all* TCR sequences in the HTS dataset. To prevent underestimation of the true number of mismatches in a length group, RTCR combines the information from the alignments and the base quality (Phred) scores to calculate a length group specific error rate (ϵ_l):

$$\epsilon_l = \frac{m_a + m_u}{n}, \tag{2.3}$$

where l is the length of the TCR sequences, n is the number of bases in the length group, m_a is the number of mismatches found in the aligned regions of the TCR sequences in the length group, and m_u is the number of mismatches expected in the unaligned regions of the TCR sequences, estimated using the base quality scores:

$$m_u = \sum_Q u_Q 10^{\frac{-Q}{10\alpha}}, \tag{2.4}$$

where Q is a Phred score, u_Q, is the number of bases in the unaligned regions of the length group with a Phred score of Q, and α is a normalization factor for the Phred scores. Since Phred scores reflect the probability that a base is false, every Phred score can be recalculated by taking all aligned bases with a particular phred score and use the fraction f that was false, to calculate an effective Phred score $Q_{\text{eff}} = -10 \log_{10} f$. The normalization factor α is calculated from the average ratio of observed Phred scores to the effective Phred scores, $\alpha = \sum_Q Q/Q_{\text{eff}}$. Finally, RTCR takes the maximum of the globally calculated error rate (ϵ) and the group specific error rate (ϵ_l), $\max(\epsilon, \epsilon_l)$, as the error rate for the length group in the QMerge and IMerge algorithms.

The error correction algorithms of RTCR described below (including barcode error correction) use the same approach to merging TCR sequences. If two (parent) TCR sequences are merged, a (child) consensus sequence is formed from the highest abundant base at every position, breaking ties by selecting the higher quality base, using 'N' if ties cannot be broken. The algorithms keep track of the frequencies of the parent bases at every position using a position frequency matrix (PFM). TCR(/consensus) sequences are merged by summing their associated PFMs and generating a consensus sequence from the resulting PFM. Hence, the final error corrected TCR sequence is independent of the order in which its parent TCR sequences were merged. Additionally, the error correction algorithms use the PFMs to keep track of the number of mutations that have been performed, by summing the frequencies of the bases that were not selected for the TCR sequences associated with the PFMs.

Quality merge (QMerge) algorithm

QMerge groups sequences by length and merges sequences within each group based upon their abundance and base quality scores. Let n be the total number of sequences of length l under consideration. To prevent RTCR from merging unrelated sequences, QMerge uses Eq.(2.2) and considers all pairs of sequences of length l differing by at

most H bases. We define a 'merge quality score' as the sum of the minimum quality scores of all mismatching bases between two sequences:

$$m = \sum_{i \in \text{mismatches}} \min(q_i, q_i'), \tag{2.5}$$

where q and q' are vectors, each containing the base quality scores of one of the two sequences in the pair; and *mismatches* contains the indices of the mismatching bases. QMerge uses the merge quality score to order the pairs and merge the lowest quality sequences first. We define a 'quality threshold':

$$Q = 10 \log_{10} nl, \tag{2.6}$$

which is the Phred score equivalent to the probability that one in $n \times l$ bases is false. QMerge calculates the merge quality score (m) for every pair of sequences (within Hamming distance H) and considers pairs for which $m \leq Q$.

QMerge traverses sequence pairs in the following order: increasing merge quality score (m), Hamming distance (HD), and decreasing abundance of the more frequent sequence in the pair. A merge is not performed if it requires mutation of a base with a quality score higher than the median Phred score in the data. The child inherits the VJ annotation of the more abundant parent. Its abundance is the sum of the abundances of both parents. After a successful merge, the parent sequences are removed from the data. If the child matches an existing sequence, the child is merged to it. The merge quality score (m) is (re)calculated for all pairs of sequences involving the child. The algorithm halts when there are no more pairs to process or the expected number of false bases ($n \times l \times \epsilon$) in the set have been corrected.

Iterative merge (IMerge) algorithm

After performing QMerge many errors are expected to remain in the data. IMerge attempts to resolve these errors by using a clustering approach where less abundant sequences are merged to more abundant nearby sequences (Fig. 2.1). Like QMerge, IMerge also groups sequences by length and considers each group separately. Because true TCR sequences (green circles in Fig. 2.1) may differ by a single nucleotide substitution, their neighborhoods of erroneous sequences may overlap. IMerge resolves clusters gradually to prevent true TCR sequences from merging to each other. All TCR sequences of

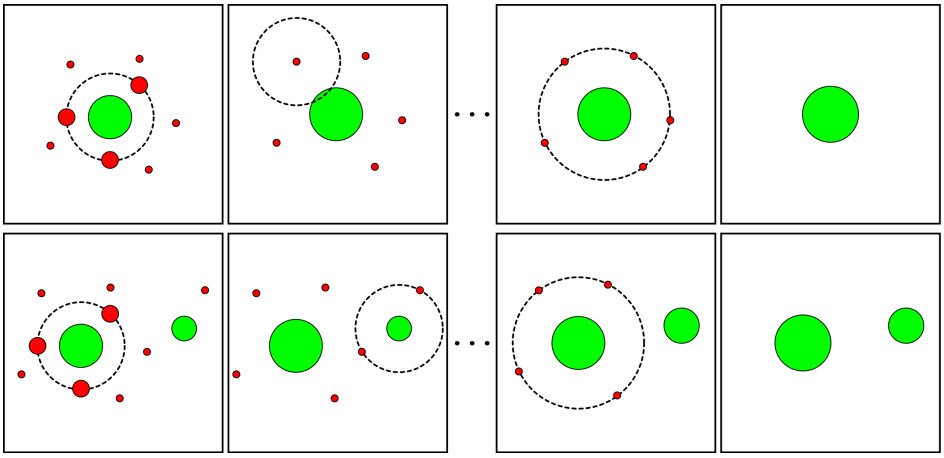


Figure 2.1. Schematic of IMerge algorithm. Top row shows one true TCR sequence (green) surrounded by erroneous variants (red); the diameter of symbols represents the abundance of the TCR sequence. IMerge considers all nearby TCR sequences as potential erroneous variants. Erroneous variants present at Hamming Distance (HD) 1 can get merged to the true TCR sequence depending upon the abundance of the true sequence, abundance of erroneous variant and the error rate determined from the data (top row second column). Once all sequences at HD 1 are considered IMerge iterates over the TCR sequences considering all variants present HD$\leq$2. The bottom row demonstrates a scenario where two true TCR sequences are neighbors but do not get merged due to abundance and HD thresholds defined by Eqs. 2.2 and 2.8.

length l, starting with the most abundant, are allowed to absorb neighboring erroneous sequences (red circles in Fig. 2.1), starting with the rarest, within HD h in their neighborhood. The algorithm begins at $h = 1$ and increases h by one after every iteration over all sequences until there are no more sequences that can be merged.

When IMerge considers a particular TCR sequence of length l, it estimates the true abundance of the TCR sequence. To not underestimate the true abundance of the TCR, IMerge calculates the 99% lower confidence interval using a normal approximation:

$$n = \frac{n_{\text{orig}} + z\sqrt{n_{\text{orig}}(1 - p_0)}}{p_0}, \tag{2.7}$$

where z is the 99% normal quantile, p_0 is the probability of having zero errors in a sequence (defined by Eq.(2.1)), and n_{orig} is the abundance of the TCR sequence in the original data ($n_{\text{orig}} = 0$ for novel TCR sequences generated by QMerge). IMerge considers all neighboring sequences within Hamming distance $H+1$ (where H is defined by Eq.(2.2)) from the TCR sequence. IMerge calculates the expected abundance of a TCR sequence if this TCR sequence would absorb all its erroneous copies within Hamming distance h:

$$N_h = n_{\text{QMerge}} + \frac{n - n_{\text{QMerge}}}{1 - p_0} \sum_{d=1}^{h} p_d, \tag{2.8}$$

where n_{QMerge} is the abundance of a TCR sequence obtained after QMerge; and $1 \leq h \leq H + 1$. IMerge orders all sequences (neighbors) within Hamming distance h by increasing abundance and minimum base quality score, and merges the neighbors to the TCR sequence. A merge is not performed if the resulting abundance of the TCR sequence would exceed N_h, i.e. its expected abundance if it would absorb all its erroneous copies until Hamming distance h (note, $N_H \approx n$). This limiting of the number of neighbors that can be absorbed within Hamming distance h allows true TCR sequences to protect themselves from being absorbed by merging their (erroneous) neighbors (Fig. 2.1). The algorithm halts when no merges can be performed for any TCR sequence.

Levenshtein merge (LMerge) algorithm

After the QMerge and IMerge algorithms, indels are expected to remain in the data (sometimes in combination with mismatches). RTCR estimates the expected number of deletions (insertions), n_d (n_i), from the number of deletions (insertions) found in the alignments of germline V and J sequences with the CDR3 region. The LMerge algorithm

is similar to IMerge, with an important difference being that it calculates the Levenshtein instead of Hamming distance between TCR sequences. Similar to the IMerge stopping criteria, the LMerge algorithm also does not introduce more than n_d deletions or n_i insertions.

RTCR performs a post-processing step where CDR3 sequences with unresolved bases ('N') are merged to the nearest CDR3 sequence that differs with it only on the unresolved positions. Finally, CDR3 sequences containing a base quality score below five are discarded. This culling of low quality sequences is performed only after error correction.

2.2.2 Simulation of TCR HTS

We used the probabilistic model of Murugan *et al.* (2012) to simulate the biological process of rearranging V, D, and J segments, generating a synthetic repertoire of 10^4 TCR β chain (TCRB) sequences. Because TCRB sequences are generally too long to be spanned by reads, HTS protocols such as 5′ RACE (Warren *et al.*, 2011) are used to amplify the part of the TCRB sequence containing the CDR3 region. To get sequence lengths similar to the 5′ RACE protocol we included only the first 61 basepairs of the constant region in the synthetic TCR sequences. To mimick the heavy-tailed distribution of TCRB sequences in humans (Mora *et al.*, 2010; Venturi *et al.*, 2011), we expanded the synthetic repertoire to 10^5 sequences according to an empirical TCRB distribution. This distribution was derived from lane SRR060714 of Warren *et al.* (2011) using MiTCR (Bolotin *et al.*, 2013).

HTS protocols involve PCR amplification. PCR can distort sequence abundances (Best *et al.*, 2014) because not all sequences are doubled in a PCR cycle and polymerases can have a sequence bias. Additionally, false but abundant TCR sequences can be formed if mutations occur in early PCR cycles. We simulated a simplified PCR process to introduce imperfect amplification and to generate TCR variants. In every cycle of *in silico* PCR the number of TCR sequences was doubled using sampling with replacement so that some sequences were missed in every doubling. Every TCR sequence was doubled approximately 18 times, with a substitution error rate of 5×10^{-5} (Cline *et al.*, 1996), resulting in $10^5 \times 2^{18}$ synthetic amplicons with about 2% of the amplicons containing one or more PCR errors.

We used the Illumina simulator of ART (Huang *et al.*, 2012) version 2.3.7 to generate paired-end reads from subsets of the synthetic amplicons. The size of the subset de-

termined the fold coverage (i.e. number of reads per 'cell'). For example, a subset of 10^6 amplicons represents 10x coverage, because the synthetic repertoire consisted of 10^5 TCRB sequences (and 10^4 clonotypes), and ART generates at least one paired-end read for every amplicon in the subset. We simulated two recent sequencers, HiSeq 2500 and MiSeq, using the default error profiles provided by ART. For HiSeq (MiSeq) the settings were: read length 150 (250), mean fragment length 200 (500), standard deviation 15 (0). Finally, we merged read pairs as follows: read pairs with less than 18bp paired end overlap were dropped; for overlapping regions a consensus sequence was created by selecting the higher quality base when bases agreed, or if one had $\geq$Q30 and the other $<$Q20, in all other cases an 'N' with Q0 was recorded.

2.2.3 Analysis of TCR HTS

Analyses were performed using TCRklass 0.6.0, MiTCR 1.0.3, MiXCR 1.6, IMSEQ 1.0.1, MiGEC 1.2.3, and RTCR 0.3.0, using the default settings. As 'default settings' for IMSEQ we turned on its clustering based error correction and merging of identical CDR3 sequences with ambiguous segment identification, i.e. "-ma -qc -sc". Before evaluating the performance of each pipeline, non-functional TCR sequences (i.e. those that are out-of-frame or contain a stop-codon) were removed. For the analyses of non-barcoded HTS data (real and simulated), all pipelines, with the exception of MiXCR which uses its own reference, were run with the germline reference sequences of MiTCR. For the analysis of barcoded HTS data, RTCR was run with the V(D)J reference sequences of MiGEC. To compare the error correction of MiGEC and RTCR, the latter was run on the sequences resulting from the Checkout utility of MiGEC so that both had the same starting point. MiGEC was run with an 'overseq' threshold of 5 (i.e. discarding UMI groups with fewer than 5 reads).

Both TCRklass and IMSEQ pipelines report identical CDR3 sequences with different VJ combinations by default which can inflate the false positive rate. To make the reporting equivalent among the pipelines, we collapsed these sequences and summed their counts. Collapsing these sequences had only a minor positive effect on the precision of TCRklass and IMSEQ and no effect on the recall.

2.3 Results and Discussion

HTS produces millions of reads, each potentially containing one or more errors, and retrieving TCR sequences from the reads without performing any error correction results in many false TCR sequences (Baum *et al.*, 2012; Bolotin *et al.*, 2012). Error correction of TCR sequences, especially the CDR3 region of the TCR, is a complex problem because true TCR sequences may differ from each other by as little as a single nucleotide. We developed the RTCR pipeline to accurately retrieve TCR sequences from HTS sequencing data. To test the performance of RTCR we compare it to four other recent pipelines TCRklass (Yang *et al.*, 2015), MiTCR (Bolotin *et al.*, 2013), MiXCR (Bolotin *et al.*, 2015), IMSEQ (Kuchenbecker *et al.*, 2015), and MiGEC (Shugay *et al.*, 2014).

2.3.1 *In silico* TCR HTS data

To determine the accuracy of the TCR pipelines we generated *in silico* sequencing reads (see section 2.2.2) from a simulated TCRB repertoire of 10^5 cells with 10^4 distinct sequences (simulations were performed in triplicate). Since real sequencing experiments differ in quality and coverage (number of reads per cell), we used error profiles of two recent sequencers, HiSeq 2500 and MiSeq, and varied the coverage over a wide range from 1x to 100x.

We compare the pipelines on their recall, i.e. on the fraction of true CDR3s recovered from the HTS dataset, and on their precision, i.e. the fraction of recovered CDR3s that are correct. For all pipelines the recall tends to increase with higher coverage, but for TCRklass this comes at the cost of very low precision (Fig. 2.2). We therefore omit TCRklass from further comparison. MiTCR has very high precision in all datasets (Fig. 2.2 and 2.3), but its recall is relatively low, especially in lower coverage datasets. Similar to MiTCR, IMSEQ has poor recall in the lower coverage (1x, 10x) datasets (Fig. 2.2 and 2.3). Although the recall of IMSEQ is better than that of MiTCR in the MiSEQ datasets, the precision of IMSEQ is lower, especially in the 1x HiSeq 2500 datasets. MiXCR has better recall than IMSEQ in the HiSeq 2500 datasets, but the situation is reversed in the MiSeq datasets. Only RTCR is able to maintain both high precision and high recall in both HiSeq 2500 and MiSeq datasets, showing over 90% precision and recall on average (Table 2.1).

We compared the CDR3 sequences reported by the pipelines to those of the 'true' simu-

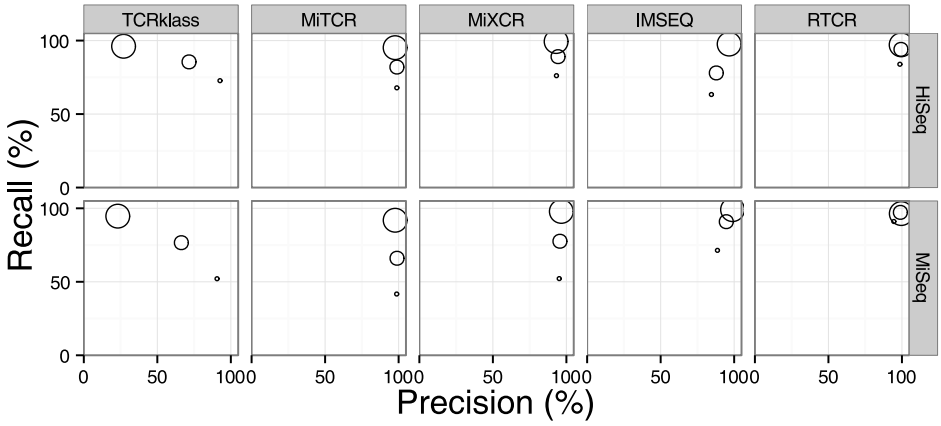


Figure 2.2. Precision and recall of CDR3 sequences retrieved from the same datasets as shown in Figure 2.3. Every data point is an average of three independent datasets. Circles represent the coverage, in order or increasing size: 1x, 10x, 100x. Coverage tends to increase recall, but decrease precision. Accurate analyses result in circles in the upper right corner. On both HiSeq and MiSeq data RTCR is already accurate at the lowest coverage.

lated TCRB repertoire (Fig. 2.3). The horizontal line in each panel depicts the number of CDR3 sequences of the 'true' repertoire that was represented by one or more reads that spanned the CDR3 region. If a bar falls below this line, the pipeline underpredicted the number of CDR3 sequences in the HTS dataset; conversely, if a bar is higher than the black line, the number of CDR3 sequences was overpredicted by the pipeline. To visualize the quality of the reported list of CDR3 sequences, we colored the bars reflecting the fraction of the reported sequences that perfectly matched a CDR3 in the 'true' repertoire (green), that had one mismatch with the most similar true CDR3 sequence (yellow), two mismatches (orange), or more than two mismatches (red). All pipelines, except TCRklass, tend to underreport the true number of sequences. The number of clones reported by RTCR is closest to the true diversity.

Since MiTCR had consistently high precision in all datasets, we attempted to increase the recall of MiTCR by changing several of its parameters. First, we tested the "save my diversity" parameter, but this resulted in a loss of precision with hardly any increase in recall (data not shown). As MiTCR ascribes high quality TCR sequences as core sequences (ignoring the low quality sequences), using a "quality" parameter to differentiate between them, we also attempted to increase the recall by lowering this parameter to 5. Although this lead to an increase in recall (data not shown), it markedly reduced the

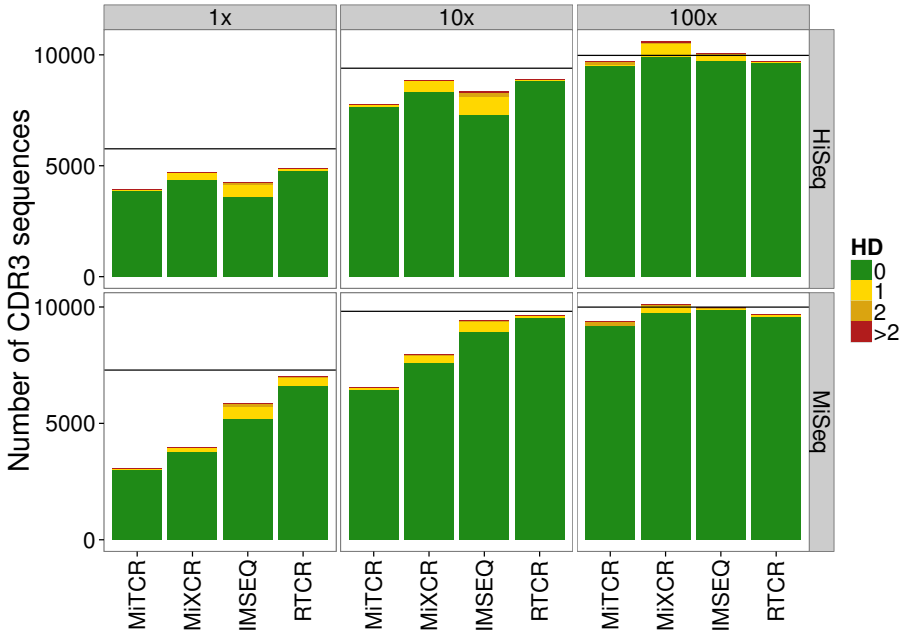


Figure 2.3. Accuracy of CDR3 sequences retrieved by several pipelines from simulated HTS datasets. ART was used to simulate 150bp HiSeq 2500 (top-row) and 250bp MiSeq (bottom row) paired-end reads. The fold coverage of every dataset, was 1x, 10x, or 100x. Due to sampling only a subset of the *in silico* CDR3 sequences was spanned by one or more reads (horizontal black lines) and could potentially be retrieved. The merged paired-end reads of HiSeq 2500 and MiSeq had an average Phred quality of 37 and 35 respectively, and a substitution error rate of about 1%. We ran the analysis on several independent datasets and here show one representative example. HD, Hamming Distance; number of mismatches with the most similar true CDR3 sequence of the same length.

Table 2.1. Average precision and recall of the pipelines on all HiSeq 2500 and MiSeq datasets (1x, 10x, and 100x coverages combined).

Pipeline	HiSeq 2500		MiSeq	
	Precision(%)	Recall(%)	Precision(%)	Recall(%)
TCRklass	63.7	84.8	60.0	74.5
MiTCR	98.4	81.7	98.4	66.5
MiXCR	93.5	88.2	95.8	75.9
IMSEQ	89.7	79.6	94.1	87.1
RTCR	98.4	92.0	97.2	94.6

precision in the MiSeq data. This large difference between MiTCR and MiTCR_Q5 suggests the high precision of MiTCR results from discarding low quality sequences prior to its error correction.

With increasing coverage in the HiSeq 2500 datasets, the recall of RTCR increased from 84% to 97% while precision remained approximately 98% (Fig. 2.3 and 2.2). In the MiSeq datasets, the recall of RTCR increased from 91% to 96% with increasing coverage, and precision ranged from 95% to 99%. Comparing the HiSeq 2500 and MiSeq results of RTCR, the recall varies more in the HiSeq datasets. Closer inspection showed that the recall dropped largely due to a failure to identify the CDR3 sequences in the reads. The recall of RTCR before error correction was about 85% for the 1x coverage HiSeq 2500 datasets. So the lower recall in the HiSeq 2500 1x coverage datasets was not due to overzealous error correction. Different settings for Bowtie 2, or a different aligner, might increase the recall of RTCR.

In summary, the precision and recall of RTCR is more stable across different coverages and sequencers than that of TCRklass, MiTCR, MiXCR, and IMSEQ. Importantly, RTCR had the highest recall in the low coverage datasets (1x and 10x), which markedly reduces sequencing costs, allowing more libraries to be sequenced.

Typically, researchers apply abundance and quality filters to their raw reads. We think these filters should not be used in combination with the advanced data-driven error correction of RTCR. To test the effect of such filters, we ran RTCR on one of the HiSeq 2500 10x simulated datasets, applying either an abundance or quality filter (Fig. 2.4). The abundance filter, which was applied after RTCR analysis, led to a large decrease in recall without a corresponding gain in precision. The quality filter, applied to raw reads (Fig. 2.4, right panel), strongly decreased recall, whereas the precision was either unaffected or decreased somewhat, because RTCR benefits from the additional information provided by low quality reads. Together these results demonstrate that quality and abundance filters can be detrimental to the precision and recall of RTCR.

To test how well the pipelines recover CDR3 abundances, we compared the abundances of the reported CDR3 sequences to their true abundances in the reads (Fig. 2.5). Most pipelines accurately predicted the abundance of identified TCR sequences, but all pipelines missed some low frequency ($\leq 10^{-4}$) TCR sequences in the 10x HiSeq 2500 datasets. MiTCR also missed abundant ($\geq 10^{-3}$) TCR sequences (black circles in Fig. 2.5), suggesting it is too ambitious in its error correction.

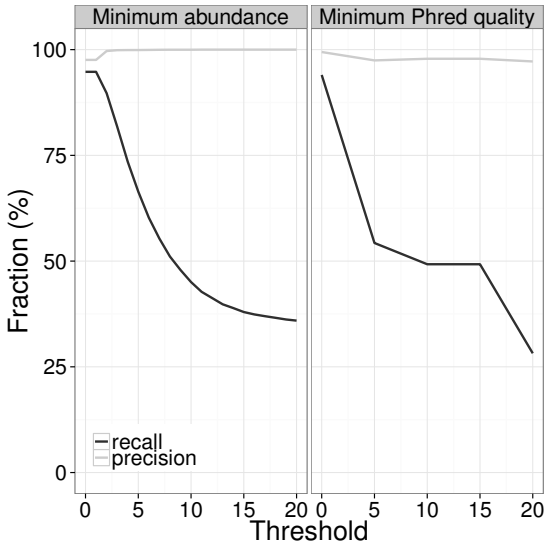


Figure 2.4. Applying naive filters can negatively affect the accuracy of RTCR. Left panel, an abundance threshold was applied to clones reported by RTCR, discarding any clones with a count lower than the threshold. Right panel, to emulate discarding low quality reads, a quality filter was applied to raw sequence data before RTCR analysis, discarding all reads containing one or more bases in the CDR3 region with a Phred score below the threshold. One of the simulated HiSeq 2500 10x datasets was used for both panels.

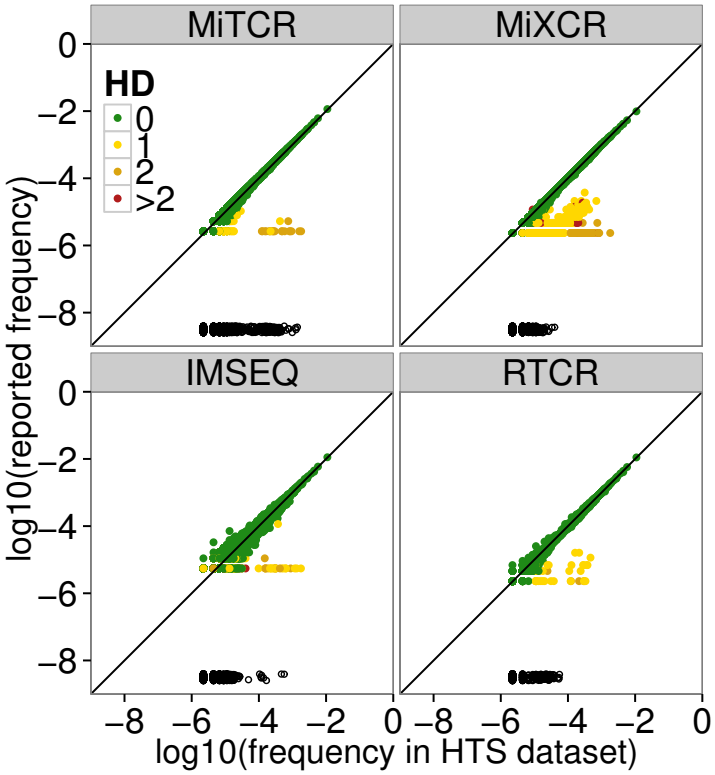


Figure 2.5. Quantitative TCR profiling of a HiSeq 2500 dataset, 10x coverage. CDR3 sequences (dots) retrieved by a pipeline are colored according to HD from the nearest true CDR3 sequence. Black open circles located at the bottom of each panel indicate missed CDR3 sequences. The reported frequency is the relative count assigned to a CDR3 sequence by a pipeline. The frequency of a CDR3 sequence in a HTS dataset is the ratio of reads spanning the particular CDR3 and the total number of reads spanning any CDR3.

2.3.2 Analysis of a published TCR HTS dataset

Having tested TCRklass, MiTCR, MiXCR, IMSEQ, and RTCR on simulated HTS datasets, where correctness of reported CDR3 sequences can be measured directly, we next compared the pipelines using a published TCR HTS dataset (Warren *et al.*, 2011). Warren *et al.* (2011) obtained two blood samples of 20 mL each, one week apart, from a healthy adult male and sequenced these using an Illumina GAIIx Analyzer. Unfortunately, only the quality filtered reads, i.e. those having a CDR3 containing only bases with a quality score of at least Q30, were published. If the filtering removed many true TCR sequences, then this limits the benefit of the error correction of RTCR. To handle any remaining errors in the high fidelity reads, Warren et al. applied an abundance based filter, called D96, removing low-abundance sequence variants comprising a total of 4% of the reads. We analysed the published data of blood draws one and two with MiTCR, MiXCR, IMSEQ, and RTCR, and compared the number of CDR3 sequences reported. We removed all out-of-frame CDR3 sequences and those containing a stop-codon (Warren *et al.*, 2011).

Despite the many errors that may have been removed by the quality filtering, it is likely that different pipelines may not correct all errors. To test this hypothesis we visualized the sequence space around 3 representative abundant sequences in lane SRR060714 (Fig. 2.6). The sequence space around the chosen sequences had progressively less abundant sequences at higher HD, suggesting that the surrounding sequences might be erroneous variants (all 3 columns). The sequence (CSVPGQGGYEQYF) chosen for the first column broke this pattern with a medium abundant sequence (CSVPGQGVYEQYF) of about 600 reads, at HD 3. It is likely a correct sequence (similar to Fig. 2.1) because it was both abundant and assigned a different V gene (V29-1 instead of V20-1). All tested pipelines reported this sequence (Fig. 2.6, for every pipeline in the first column the rightmost circle from the center), but TCRklass and MiTCR also reported many (much) less abundant sequences. This example results suggest MiXCR, IMSEQ, and RTCR are better at correcting PCR errors than MiTCR and TCRklass. However, MiXCR reported fewer clones than the uncorrected ("Raw") data and had a smaller overlap between the blood draws (Table 2.2), suggesting that this pipeline might not be correcting as many PCR errors as IMSEQ and RTCR.

All pipelines reported more sequences than D96 (Table 2.2) and a smaller overlap (as measured by the Jaccard index) between the blood draws. Interestingly, IMSEQ reported considerably fewer CDR3 sequences, and had a lower overlap between the blood draws than RTCR, suggesting that IMSEQ removed true CDR3 sequences.

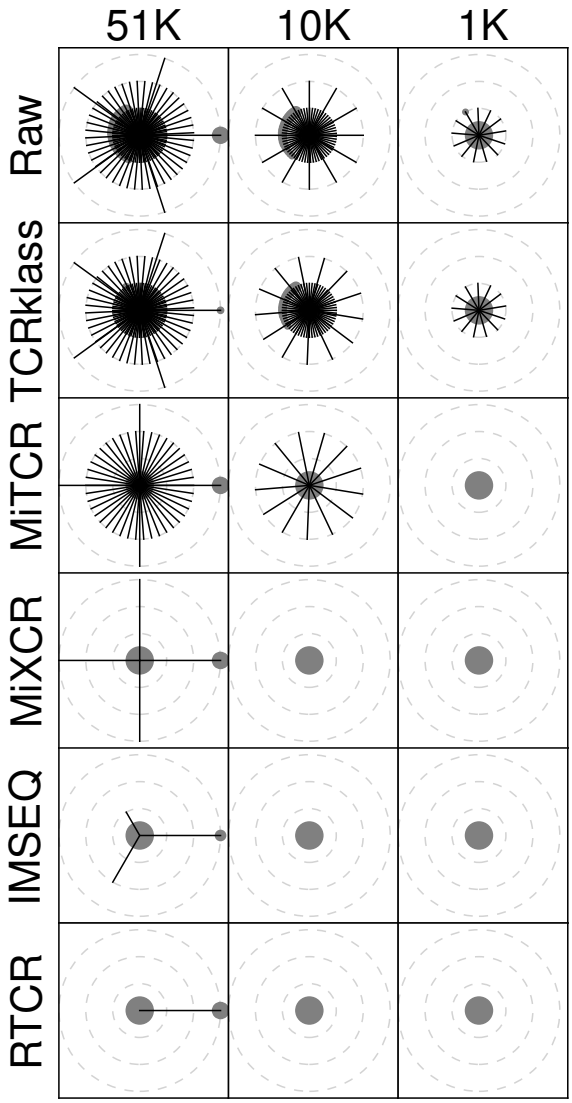


Figure 2.6. The sequence space of 3 abundant sequences from lane SRR060714 of Warren et al. (Warren *et al.*, 2011). The number above each column indicates the abundance of the chosen sequence in the 'raw' data, i.e. the CDR3 sequences identified by RTCR before error correction. Sequences (grey circles) within 3 HD of a central clonotype (columns) are connected (black lines). Circle area is the log-abundance ratio between a sequence and the total abundance of all sequences within the panel.

Table 2.2. CDR3 statistics of several analyses of Warren et al.'s Male 1 dataset. The D96 counts and overlap are from Warren et al. (Warren *et al.*, 2011). Raw: the CDR3 sequences reported by RTCR before error correction. Sequences with stop-codon and out-of-frame sequences have been removed. BD, blood draw.

Pipeline	CDR3 BD 1	CDR3 BD 2	Overlap	Jaccard index
Raw	4,635,984	1,271,640	159,900	0.028
TCRklass	4,404,901	1,192,925	150,829	0.028
MiTCR	1,202,106	490,500	52,561	0.032
MiXCR	1,458,062	687,732	55,142	0.026
IMSEQ	879,442	363,206	38,355	0.032
RTCR	955,694	451,488	47,653	0.035
D96	494,796	352,139	45,150	0.056

2.3.3 Analysis of a published barcoded TCR HTS data

A recent advance in HTS is the addition of unique molecular identifiers ('barcodes') to every template molecule (Kinde *et al.*, 2011; Kivioja *et al.*, 2011). With barcoded HTS data, many PCR and sequencing errors can be corrected by grouping reads with the same barcode together for consensus assembly (Shugay *et al.*, 2014). This also enables direct quantification of the number of template molecules in the input (Kivioja *et al.*, 2011) (i.e. the abundance), reducing the effect of PCR amplification bias on estimation of the true abundances of TCR sequences. RTCR supports the analysis of barcoded HTS data. First, RTCR collapses groups of sequences with the same barcode using consensus assembly. Next, RTCR runs the remainder of the pipeline as it would with non-barcoded data, combining barcode error correction with data-driven quality and frequency-based error correction. As RTCR has additional error correction on top of consensus assembly, it considers even small 'barcode groups' containing a single sequence, which are typically discarded by other pipelines.

To evaluate the performance of RTCR, we used a high quality and extremely deeply sequenced barcoded HTS dataset ("Experiment 1" from (Egorov *et al.*, 2015)), and compared the results to that of MiGEC, a pipeline designed to analyse barcoded data (Egorov *et al.*, 2015; Shugay *et al.*, 2014). Egorov et al. obtained blood from a 50-y-old male donor and divided it into eight replicas of about 4000 PBMCs each. We used the barcoded TCRβ sequences (Illumina MiSeq 2x 150bp paired-end reads) to compare both pipelines. MiGEC reported 236 clones of which most, 229, were also reported by RTCR. RTCR reported many more clones, 2717 in total across the eight replicas. This large difference is not unexpected, because RTCR recovers clones from barcode groups supported by only one sequence. The fact that RTCR recovers more clones that are reproducibly

found in more than one library (Fig. 2.7), suggests that RTCR markedly outperforms MiGEC on recall because reproducible clones are more likely to be real. Importantly, reproducible clones are not guaranteed to have a correct sequence, as PCR errors are highly reproducible (Shugay *et al.*, 2014), and their presence in multiple samples can be due to cross-sample contamination (Mamedov *et al.*, 2013). Additionally, RTCR reported many more non-reproducible clones (2543) than MiGEC (151). However, given that there should be about 2000 T cells present in each replica, and that about half of these are expected to be naive singletons, a diversity of several thousand clones across eight replicas is a very realistic result. In addition, the median Levenshtein distance from the non-reproducible clones to their closest neighbor was 6 (not shown), suggesting these clones are truly different. Together, these results suggest RTCR has a much higher recall than MiGEC. Unfortunately, we cannot quantify the precision and recall, because these measures cannot be reliably estimated in real data.

2.3.4 Performance

On an 8 core Intel Xeon 3.2Ghz 32GB RAM, RTCR takes about 136 minutes (of which Bowtie 2 takes about 63 minutes) to analyse a HiSeq 2500 dataset consisting of 10 million 150bp paired-end reads.

2.4 Conclusion

TCRs exhibit enormous diversity due to somatic recombination. The advent of HTS has enabled us to sequence large number of TCR sequences from an individual. However, HTS is marred by errors and given the TCR diversity it becomes difficult to distinguish between true TCR sequence and erroneous variants. We here present RTCR, a pipeline designed to accurately recover TCR sequences from error-prone HTS data. RTCR performs error correction using a statistical model and estimates the model parameters from the data, relieving the user from setting arbitrary parameters. Using simulations and experimental data, we demonstrate that RTCR can identify, and correct PCR and sequencing errors exhibiting consistently high precision and recall. The high accuracy of RTCR makes it well suited for estimation of repertoire diversity and for disease profiling. Especially in the lower coverage (1x and 10x) simulated datasets, RTCR outperformed all other pipelines. This means that RTCR has the potential to make the analysis of repertoire sequencing data more cost effective.

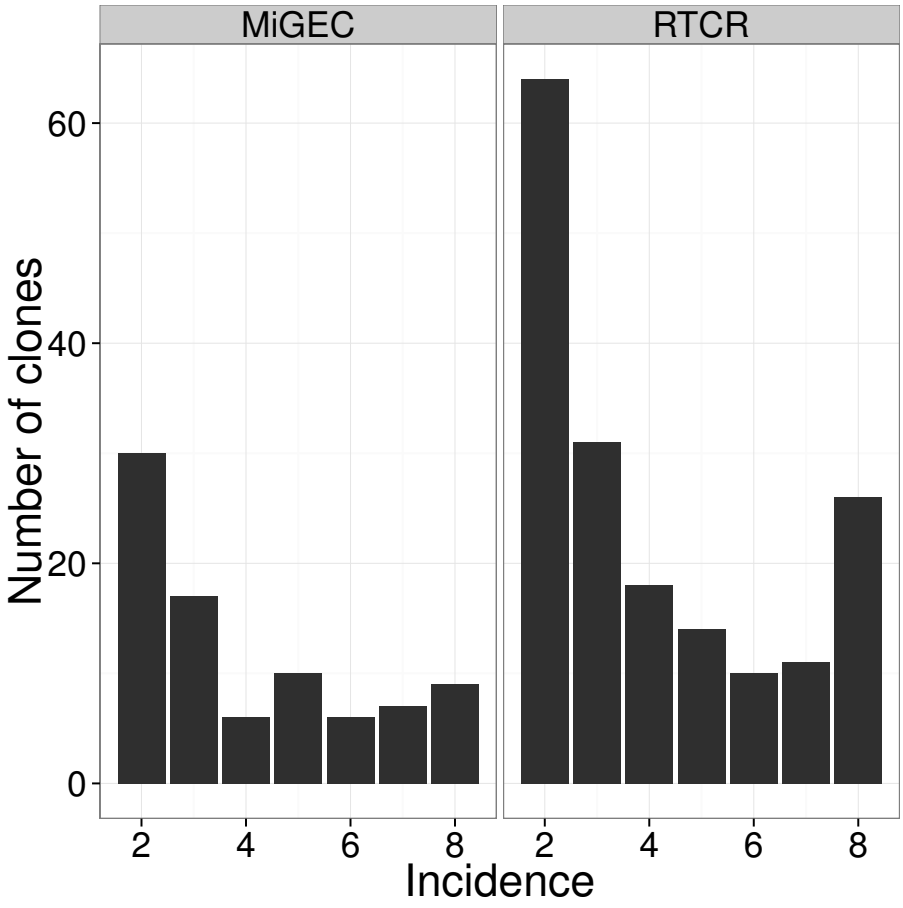


Figure 2.7. Capturing reproducible clones. Left panel, all distinct CDR3 sequences reported by MiGEC in each of the eight replicas of (Egorov *et al.*, 2015) were pooled and their frequencies of occurrence tallied, showing only those clones that occurred in more than one replica. Right panel, the same for RTCR.

Funding: BG and AA were supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch government (FES0908). RdB and AP were supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 317040 (QuanTI).

Chapter 3

The abundance of large memory clonotypes evolves over time in a healthy TCRB repertoire

BRAM GERRITSEN[a], ARIDAMAN PANDIT[a], FATIHA ZAARAOUI-BOUTAHAR[b], MIRJAM C.G.N. VAN DEN HOUT[c], WILFRED F.J. VAN IJCKEN[c], ARNO C. ANDEWEG[b], AND ROB J. DE BOER[a] (2017)

In preparation

[a] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, the Netherlands.*

[b] *Department of Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands.*

[c] *Center for Biomics, Erasmus Medical Center, Rotterdam, the Netherlands.*

Abstract

The T cell receptor (TCR) repertoire is extremely diverse and dynamic, with clonotypes expanding and contracting by self-renewal, clonal expansion, and cell death. We performed high throughput sequencing (HTS) analysis of TCRB mRNA in unsorted blood samples from a healthy human adult, taken minutes, days, and years apart, to determine repertoire dynamics. Clonal abundances of the non-singleton (mostly memory) clonotypes were more similar among blood samples that were taken within a week from each other than those with years between them, and was accompanied by a significant shift in J-gene usage among distinct TCRB sequences. Remarkably, a single clonotype that was undetected in the early samples, became one of the largest clonotypes in the samples that were taken years later. By selecting TCRB sequences with similar increasing dynamics, we found two clusters of "novel expanded" clonotypes. One cluster contained variants of the hugely expanded clonotype, and most of the variants used TRBJ2-1. The other cluster was mainly defined by the use of TRBV7-2 and TRBJ1-5, and a non-germline encoded Glutamine in the CDR3. These results demonstrate that clonal abundances of non-singleton (memory) clonotypes evolve over time in a healthy repertoire, and suggest we found (highly) similar clonotypes expanding in response to the same epitope.

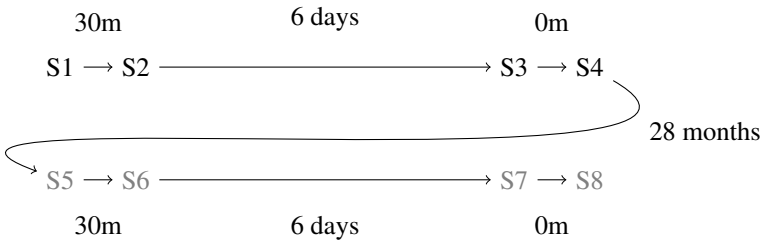


Figure 3.1. Time between samples, m = minutes. Black (upper line), samples belonging to "set 1". Gray (lower line), samples belonging to "set 2".

3.1 Introduction

HTS is applied to study the diversity and shape of the TCR repertoire, which can give information about health and disease state of an individual (Woodsworth *et al.*, 2013). Longitudinal analysis of TCR repertoires has been shown to be useful in understanding several aspects of disease progression and therapy efficacy, like minimum residual disease monitoring (Woodsworth *et al.*, 2013), or estimation of repertoire diversity after hematopoietic stem cell transplantation (HSCT) (Muraro *et al.*, 2014). However, little is known about repertoire dynamics in a healthy individual. To understand the temporal stability of the repertoire, we investigated the repertoire dynamics of a healthy volunteer. We obtained unsorted blood samples at different timepoints and compared the clonal abundances between the samples. We observed that samples taken less than a week apart were more similar than those taken two years apart, and that this resulted from changes in the abundances of the large (non-singleton) clonotypes in the samples.

3.2 Results

Two sets of blood samples were obtained from a healthy 53yo human male (Figure 3.1). Each set consisted of four blood samples that were spaced apart in time by 30 minutes, 6 days, and 0 minutes, respectively. The sets themselves were taken 28 months apart. PBMCs were isolated from each blood sample and total RNA was extracted. The RNA molecules were reverse transcribed to cDNA molecules, to which 13 nucleotide long unique molecular identifiers (UMIs) were attached, followed by PCR amplification and 250 bp paired-end sequencing on an Illumina HiSeq 2500 platform. The resulting reads were merged using PEAR (Zhang *et al.*, 2014) and merged read pairs were annotated and

error-corrected using the RTCR pipeline (Gerritsen *et al.*, 2016).

3.2.1 Diversity of unsorted samples is in-between previous estimates for naïve and memory repertoires

To reduce technical variation, the RNA samples were processed at the same time (see Section 3.4) and multiplexed in a single sequence run. After sequence processing, we observed considerable variation between samples, in the number of CDR3 nucleotide sequences observed, and the number of UMIs sampled (Figure 3.2). Using iNEXT (Hsieh *et al.*, 2016), we computed rarefaction and extrapolation curves (Chao *et al.*, 2014; Colwell *et al.*, 2012) to show the expected number of CDR3 sequences ("species") observed for a given number of UMIs sampled ("individuals"). In ecology, such curves are used to rank the assemblies from which the samples were taken by their diversity (Chao and Jost, 2012). In our case, all the samples are from the same repertoire ("assembly"), and hence we expect these curves to exhibit similar diversity, especially for samples taken immediately after each other, such as samples 7 and 8. Strikingly, the latter has about 6×10^4 distinct TCRB sequences when sampling 2.7×10^5 UMIs, whereas sample 7, when downsampled 1000 times to the same number of UMIs, has on average 34% more distinct TCRB sequences (mean $= 8 \times 10^4$, sd $= 132$) at the same sampling effort. This may suggest that in repertoire sequencing (repseq) experiments, additional sources of variation distort the diversity estimates, and thus diversity estimates based on a single (patient) sample should be interpreted with caution.

Estimation of repertoire richness is challenging, as a large fraction of the CDR3 sequences are expected to be missed by small (and therefore incomplete) blood samples. The number of missed sequences is especially large for heavy-tailed (clonotype) distributions, such as power laws (Mora and Walczak, 2016). We observe in our samples a power-law (i.e. a straight line on a log-log frequency-rank graph; Supplemental figure S3.3). This suggests that the true clonotype distribution is heavy-tailed, if not power-law distributed. Hence, estimations are expected to be much lower than the true diversity of the repertoires from which blood (or tissue) samples were taken.

To get a better estimate of repertoire diversity than is provided by a single sample, we used the presence and absence of clonotypes in all eight samples to estimate the full repertoire diversity using the nonparametric Chao2 estimator (Chao, 1987). The Chao2 estimate is 7×10^6 TCRB sequences for the full repertoire, which is in line with the naïve (10^8) and memory (10^6) estimates of Qi *et al.* (2014), because our unsorted blood

samples are expected to contain both memory and naïve T cells.

3.2.2 Abundances of large (memory) clonotypes change over time

Since memory clonotypes tend to be (much) larger than naïve clonotypes, we consider singleton clonotypes (i.e. clonotypes that have a maximum count of 1 in each sample) part of the naïve repertoire, and the other (non-singleton) clonotypes part of the memory repertoire. We clustered the samples based on the clonal abundances of the memory and naïve clonotypes separately (Figure 3.3A,B). The memory samples that were taken less than a week apart, clustered together, demonstrating a shift in clonal abundances over the long ($>$ 2 year) timespan between sample 4 and sample 5. The naïve samples formed no clusters that were identifiable by time between samples, probably due to undersampling of the naïve repertoire. A dedicated experiment which sorts large numbers ($> 10^6$) of naïve T cells might be able to detect fluctuations in clonal abundances in the naïve repertoire. To test if the clustering of the memory clonotypes was due to novel memory clonotypes being picked up, we also clustered the memory clonotypes that were present in all the samples (Figure 3.3C) and we observed the same clustering. These results demonstrate that the clonal abundance of existing memory clonotypes fluctuates considerably (to such an extent that it is detectable by clustering).

The overlap (Jaccard index) between the samples varied between 3.2% and 5.2%, with the largest overlap being between sample 3 and sample 7, which were taken over 2 years apart. Hence, irrespective of the time between samples, most clonotypes were novel (i.e. not detected in the older sample). Since the abundance (and presence) of memory clonotypes varied over time, we searched for the most abundant novel clonotypes (not present at one time point) in our dataset. For samples separated by two years, the maximum abundance of the novel clonotypes was approximately 1% (Figure 3.4). Remarkably, a single clonotype, "TRBV7-9 CASSLTYGSYNEQFF TRBJ2-1", increased in frequency from less than 0.03% in the first four samples (set 1), to about 1% in samples 5 to 8 (set 2), becoming one of the largest clonotypes in the repertoire (Supplemental figure S3.1). Interestingly, the usage (independent of clonal abundance) of TRBJ2-1 also significantly increased over this period (Supplemental figure S3.2B), suggesting the presence of immune responses involving multiple clonotypes.

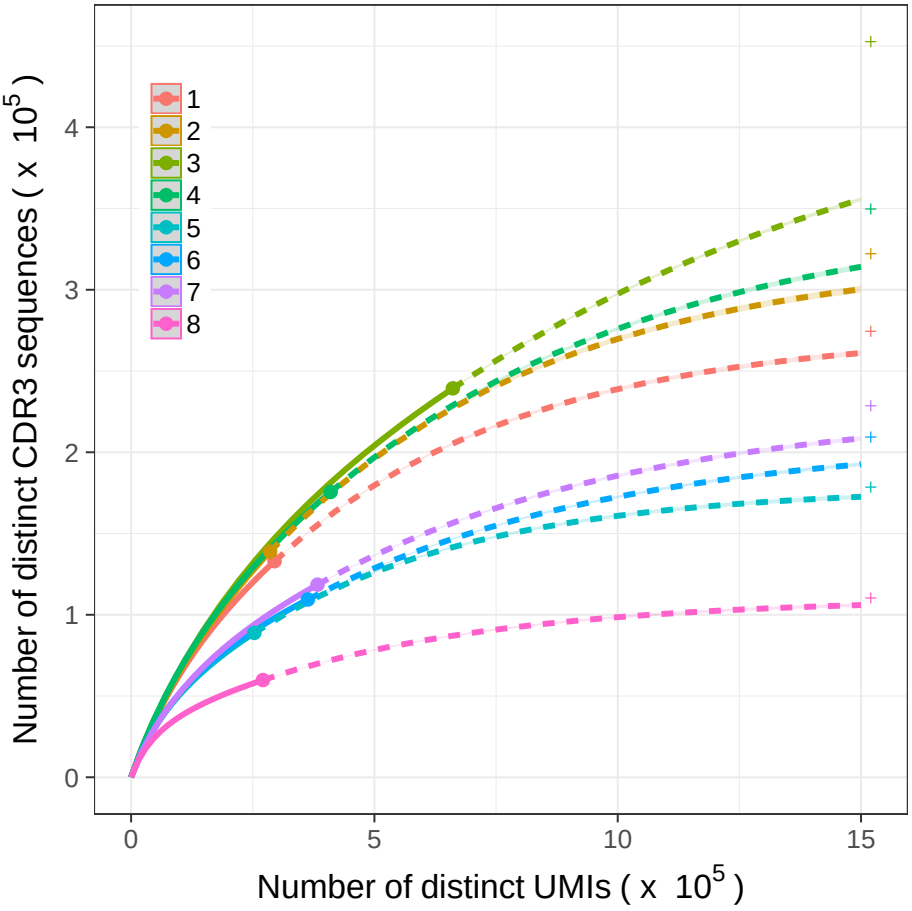


Figure 3.2. Rarefaction (continuous lines) and extrapolation (dashed lines) of the number of distinct CDR3 nucleotide sequences in relation to the number of UMIs sampled (i.e. "sampling effort"). Thick dots denote the observed number of UMIs and distinct CDR3 nt sequences in the 8 samples. Cross shapes ("+") indicate per sample the lower bound estimate (Chao1 (Chao, 1984)) of the number of distinct CDR3 nt sequences in the full repertoire. The combined estimate, based on Chao2 (Chao, 1987), was much higher (7×10^6). 95% Confidence intervals, indicated by shaded areas, were calculated from 200 bootstrap replications. Note, the confidence intervals are very thin and (mostly) do not exceed the width of the dashes.

3.2.3 Characterization of a potential immune response in a healthy volunteer

To discover clonotypes that are potentially involved in the same immune response as the previously identified "expander" (Figure 3.4, Supplemental figure S3.1), we selected other clonotypes that also increase markedly, using the following critera: a four-fold increase in frequency from the first set (samples 1-4) to the second set (samples 5-8) of samples, and a cumulative abundance ≥ 100 TCRB sequences in the second set. This resulted in a set of 79 additional clonotypes (80 including the expander) that, although the difference is small, have a significantly smaller Levenshtein distance to the expander (7.63) than randomly drawn CDR3 amino acid sequences (7.96; $p \leq 0.042$, based on 100k repetitions of sampling 79 CDR3 sequences from the set of all distinct CDR3 sequences observed in the eight samples). This may indicate that there are TCRB clonotypes in this set that are responding to the same epitope as the expander, since TCRs responding to the same epitope tend to share sequence similarities (Dash *et al.*, 2017; Glanville *et al.*, 2017).

To refine our clonotype selection, we clustered the 80 clonotypes based on their k-mer (amino acid triplet) similarity. We selected the two largest clusters of clonotypes sharing 4 or more amino acid triplets (Figure 3.5), and manually aligned those clonotypes that were within a Levenshtein distance of 5 or less from each other (Figure 3.6). The first cluster (Figure 3.6A) contains the largest expander and is mainly defined by the use of TRBJ2-1 (the TRBJ gene that was significantly increased in usage from set 1 to set 2). The second cluster (Figure 3.6B) is defined by TRBV7-2 and TRBJ1-5. Interestingly, clonotypes in this cluster also tend to have a non-germline encoded Glutamine (Q) at position 6 in the CDR3. For both clusters, the combination of sequence similarity and similar temporal pattern suggests a shared specificity for an epitope.

3.3 Discussion

We characterized changes in the TCRB repertoire of a healthy volunteer across a timespan of about two years. Using the nonparametric Chao2 estimator, we estimated a repertoire diversity of about 7×10^6 TCRB sequences, which is in line with previous estimates of the TCRB repertoire by Qi *et al.* (2014). Interestingly, diversity estimated from single blood samples (using the Chao1 estimator) varied strongly (Figure 3.2; $m = 2.66 \times 10^5$, $s = 1.08 \times 10^5$), even when downsampling to a similar number of

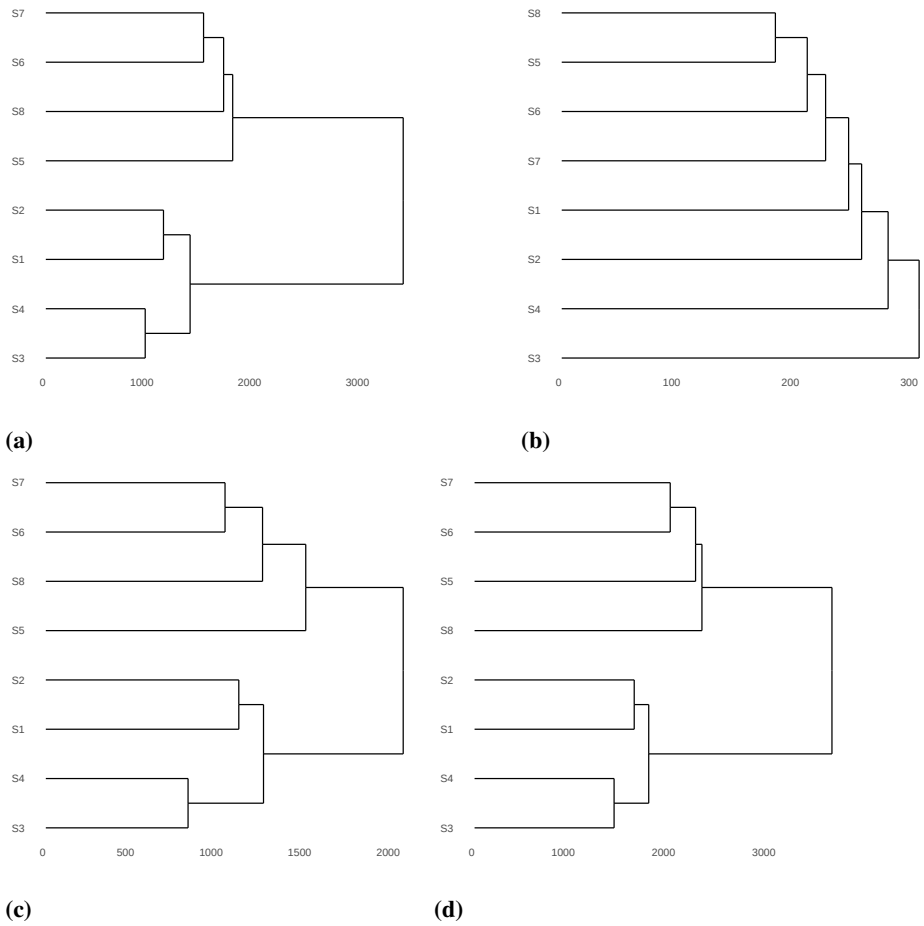
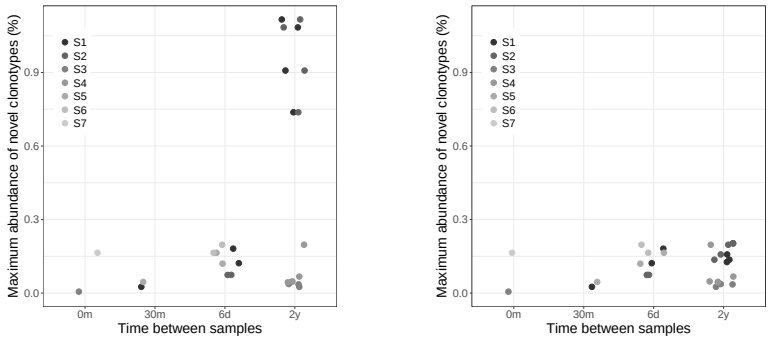


Figure 3.3. Similarity of clonal abundances between samples. (all panels) Average linkage clustering of the average Euclidean distance between samples, based on 1000 times resampling every sample to the smallest sample size (253160 UMIs). (a) Top 1000 most abundant TCRB clonotypes ("memory"). (b) Singletons ("naïve"; approximately 270k clonotypes on average). (c) Memory clonotypes that are present in all samples (1.5k clonotypes on average). (d) All clonotypes (approximately 600k clonotypes on average). (b, d) The average number of clonotypes exceeds the number of UMIs in a single sample, because a clonotype observed in one sample is not necessarily present in other samples.



(a) **(b)**

Figure 3.4. Abundance of novel clonotypes. For every combination of two samples, the maximum abundance among novel clonotypes was recorded (i.e. those clonotypes that are present in the newer sample and were not observed in the older sample; dots) as a function of the time between the samples. The grayscale of a dot indicates the older sample in each comparison. Note, sample 8 is not in the legend because it cannot be the older sample in a comparison with any of the other samples. (a) In comparisons between samples 5-8 and the samples 1-2 there is a large novel clonotype (approximately 1%), which turned out to be a single clonotype ("TRBV7-9 CASSLTYGSYNEQFF TRBJ2-1") that apparently expanded during the study period. (b) Without the expanded clonotype, the maximum abundance of novel clonotypes of the 2y comparisons are similar to the 6d comparisons.

UMIs. This result is indicative of additional sources of variation beyond multinomial sampling of an assembly. For example, the stochasticity of PCR amplification can affect the number of TCRB and which TCRB sequences end up being sequenced (Best *et al.*, 2015). Thus, diversity estimates from a single sample should be interpreted with caution.

Our two most remarkable results are that clonal abundances changed when the time between samples was over two years, and that a single clonotype that was tiny in the early samples, expanded to become one of the largest clonotypes in the repertoire. Using the temporal pattern of this large expanded clonotype, we could identify two clusters, one of which had similar sequences to the expanded clonotype. This suggests we detected an immune response in a healthy individual. Interestingly, despite apparent absence of an infection or another immune response trigger, these (potential) responses resulted in very abundant clonotypes. Additional experiments are required to study the specificity of the clonotypes that clustered together.

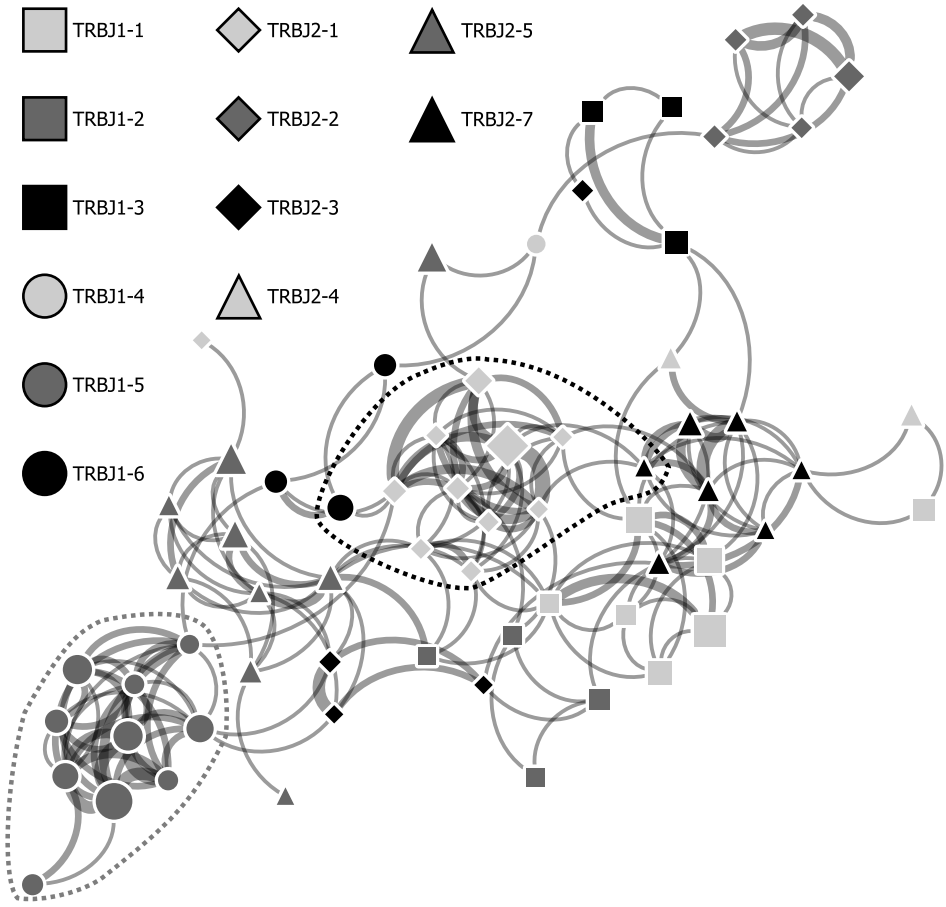


Figure 3.5. Network display of similarity between selected clonotypes (nodes; see Section 3.2.3 for selection criteria). Edge thickness indicates the number of Amino Acid triplets are shared between connected clonotypes. Only clonotypes sharing at least four triplets with another are shown. Two clusters (indicated by black and gray dashed lines) were defined by selecting the first neighbors of the two largest clonotypes in the network. Node diameter reflects log10 of clone-size.

	TRBV	CDR3	TRBJ	S1	S2	S3	S4	S5	S6	S7	S8
(a)	7-9	CASS.LTYG.SYNEQFF	2-1	0	0	170	37	2826	2680	3478	2937
	18	CASSPLSQGNSYNEQFF	2-1	17	15	28	5	54	50	102	70
	9	CASS.VL.D.GYNEQFF	2-1	0	0	0	1	37	73	39	38
	11-2	CASS.LGTS.GSHEQFF	2-1	0	0	3	0	38	41	60	37
	5-4	CASS.LP.S.VGNEQFF	2-1	0	0	3	1	43	48	58	17
	18	CASS.PQ.G.GGNEQFF	2-1	0	0	3	0	32	45	28	35
	7-9	CASS.LRGG.SYNEQFF	2-1	3	21	0	7	18	38	37	9
	28	CASS.LSIQ.GSYEQYF	2-7	0	0	0	0	19	43	34	12
(b)	7-2	CASSLQGNQPQHF	1-5	3	0	76	26	865	1532	1495	976
	7-2	CASSLQENQPQHF	1-5	0	0	16	1	165	323	405	236
	7-2	CASSTQENQPQHF	1-5	0	0	14	4	227	296	407	189
	7-2	CASSLALGQPQHF	1-5	1	0	7	1	122	225	258	148
	7-2	CASSLQGDQPQHF	1-5	0	2	11	11	140	206	244	119
	18	CASS.DTNQPQHF	1-5	14	14	27	20	40	69	102	57
	7-2	CASGLQGDQPQHF	1-5	0	0	2	1	47	33	98	48

Figure 3.6. Examples of clonotypes that are similar in both sequence and dynamics. The clonotypes are from the (a) black cluster, and (b) gray cluster, in Figure 3.5. Only those clonotypes are shown that were at Levenshtein distance ≤ 5 from each other.

3.4 Materials and methods

3.4.1 Donor samples

Blood samples were obtained according to a protocol approved by the Medical Ethics Committee of the Erasmus Medical Center Rotterdam, The Netherlands, from a healthy donor, 53 years of age. Peripheral blood mononuclear cell (PBMCs) were isolated from blood using a standard Ficoll gradient separation protocol.

3.4.2 RNA isolation and TCR amplification

Total RNA was isolated and purified from PBMC using the RNeasy Mini Kit (Qiagen, Hilden, Germany): 250 μl of ethanol was added to the upper aqueous phase of the processed TRIzol samples and directly transferred to the RNeasy spin columns for purification. RNA concentrations and OD 260:280 nm ratios were measured with the NanoDrop® ND-1000 UV-VIS spectrophotometer (NanoDrop Technologies, Wilmington, USA). TCR amplification was performed according to a protocol described by Mamedov *et al.* (2013). Briefly, RNA obtained from the equivalent of 10e6 unsorted PBMCs was reverse transcribed by RACE using a single primer directed to the constant

region, thus avoiding primer bias. Next, two-stage seminested and barcoded PCR amplification was performed on the total yield of cDNA including a size selection / agarose gel purification step after the first PCR of 20 cycles (Mamedov *et al.*, 2013).

3.4.3 Library prep and sequencing

The second-PCR generated amplicons were subjected to high-throughput sequencing according to the instructions of the manufacturers using the Ovation Low Complexity Sequencing System kit from NuGEN (San Carlos, CA, USA) and the Illumina HiSeq2500 platform (PE 250).

3.5 Supplemental information

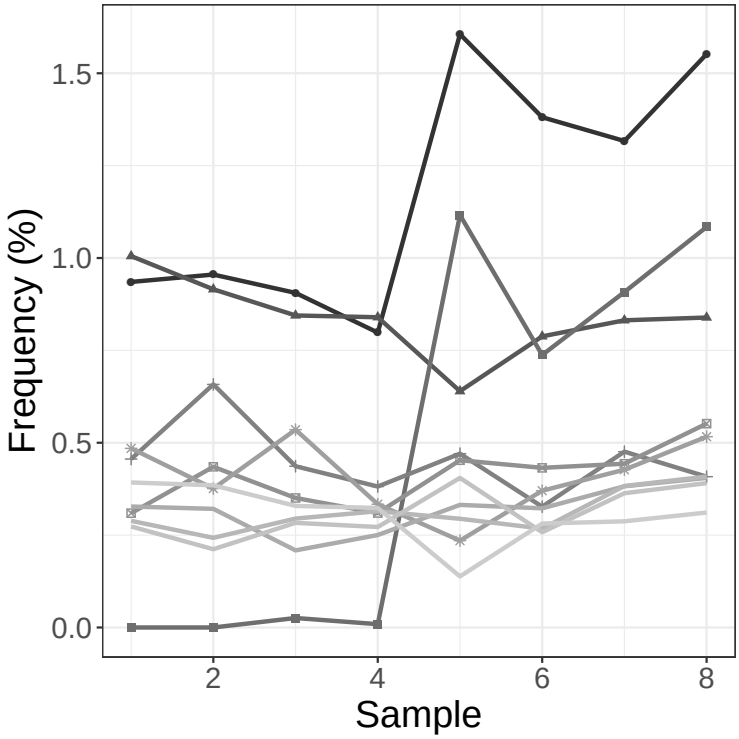
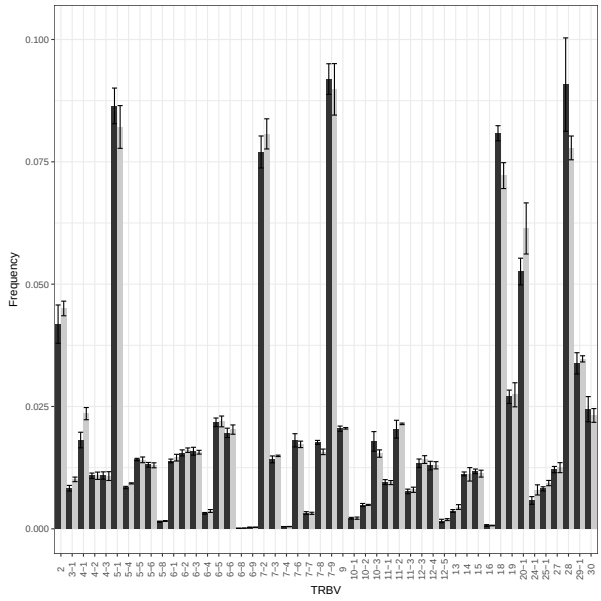
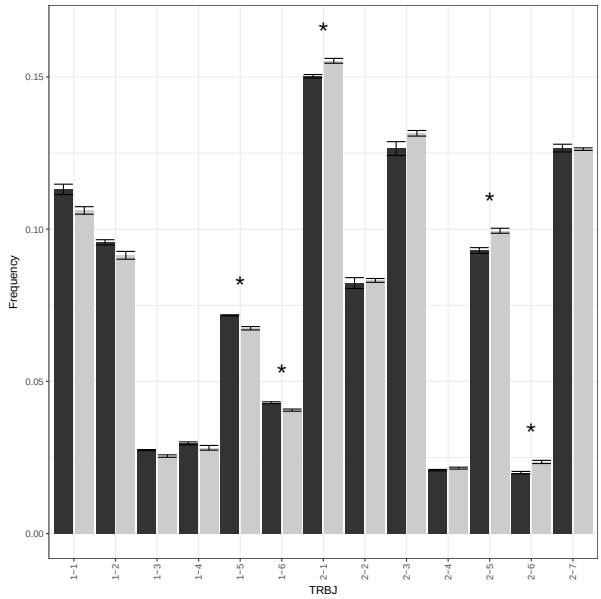


Figure S3.1. Clonal dynamics of the top 10 on average most abundant clonotypes. One clonotype (filled squares; "TRBV7-9 CASSLTYGSYNEQFF TRBJ2-1") increased in abundance from near 0 in samples 1-4 to about 1% in samples 5-8. Another clonotype (filled circles;"TRBV7-9 CASSYGSRSHEQYF TRBJ2-7"), also increased in abundance from the first to second set of samples, but at a more modest 1.6 fold change.



(a)



(b)

Figure S3.2. Average V (a) and J usage (b) among distinct TCRB sequences (i.e. irrespective of clone-size) in samples 1-4 (black bars) and samples 5-8 (grey bars). Error bars indicate standard error of the mean. Significant changes ($p \leq 0.05$) are indicated with "*", based on t-test with Holm multiple testing correction.

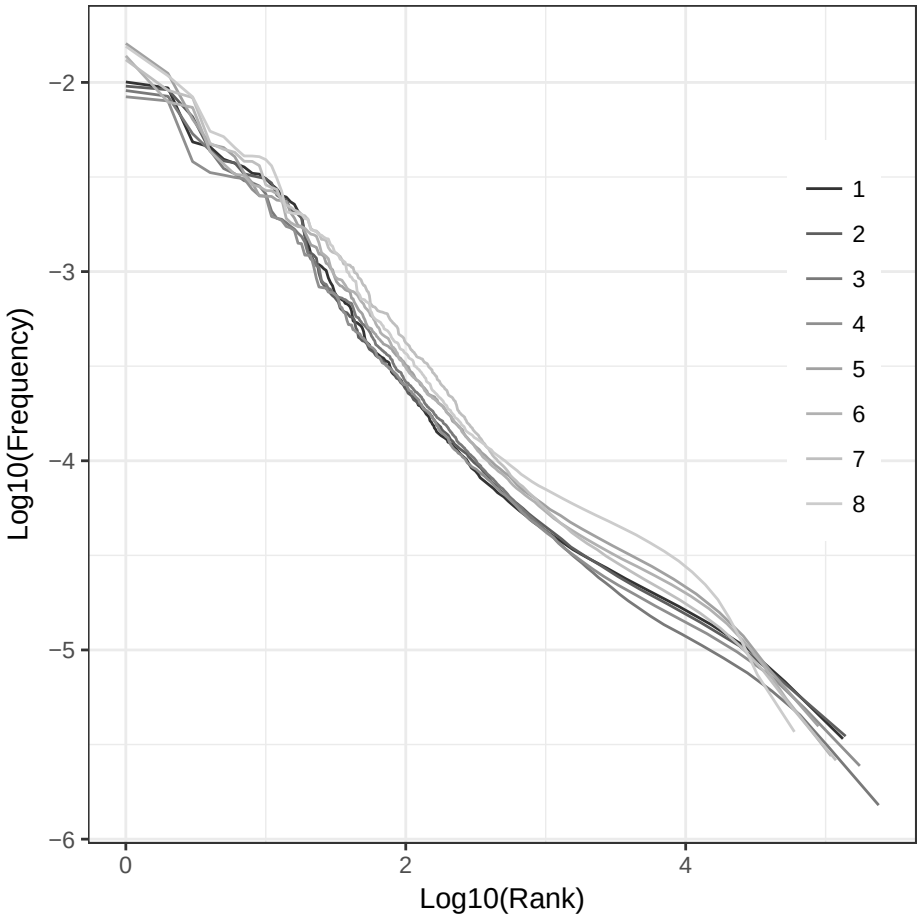


Figure S3.3. Clonesize distribution of every sample.

Chapter 4

VDJ recombination plays a major role in shaping the naïve clone-size distribution

BRAM GERRITSEN[1,a], THERES OAKES[1,b], PETER DE GREEF[1,a], JAMES M. HEATHER[b], RUTGER HERMSEN[a], BENJAMIN CHAIN[b], AND ROB J. DE BOER[a] (2017)

In preparation

[1] These authors contributed equally

[a] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, the Netherlands.*

[b] *Division of Infection and Immunity, University College London, London, United Kingdom.*

Abstract

The clone-size distribution of the human naïve TCR repertoire is unknown, and measuring it is not feasible because of the vast repertoire diversity. We developed simple mathematical models describing the naïve clone-size distribution, and tested these models using Next Generation Sequencing (NGS) of TCR mRNA in blood samples from healthy volunteers. Interpretation of the NGS data is challenging: first, sampling markedly distorts clone-size distributions as even large clonotypes become small and, second, small clonotypes can become large when the number of mRNA molecules varies between naïve T cells in the sample. Our modeling suggested a dedicated experiment splitting blood samples before mRNA extraction. Since a small fraction, but a large number, of distinct TCR sequences is found in several subsamples, we establish that many naïve T cell clonotypes are truly large. We find that TCR sequences (σ) of large naïve clonotypes tend to have a higher probability ($\mathcal{P}(\sigma)$) to be generated by VDJ recombination than small naïve clonotypes. We therefore extend our model by assigning each clonotype its own $\mathcal{P}(\sigma)$, estimated using the model of the Mora and Walczak groups (Marcou *et al.*, 2017; Murugan *et al.*, 2012). We confirm that many TCR sequences should indeed be observed repeatedly, and that there should be significantly more large TCRA than large TCRB clonotypes in small samples. These results demonstrate an unexpectedly large role for the probabilities involved in VDJ recombination in shaping the clone-size distribution of naïve clonotypes, casting doubt on the role that cognate signals play in determining clone-sizes of naïve T cell repertoires.

4.1 Introduction

The human adaptive immune system employs a vast number ($> 10^{11}$ (Clark *et al.*, 1999)) of T lymphocytes (T cells) to detect and dispose of pathogens. Most T cells express a single T cell receptor (TCR) variant with which it recognizes antigen in the form of a short peptide presented to the T cell by the Major Histocompatibility Complex (pMHC) (Davis and Bjorkman, 1988). The TCR has to be specific to distinguish between self- and non-self-pMHC, but due to the large number of possible foreign antigens ($> 20^9$) the TCR should also be sufficiently crossreactive (i.e. recognize multiple different antigens) (Mason, 1998; Sewell, 2012). The actual diversity of the TCR repertoire is unknown, but with improved sequencing techniques, estimates have risen by orders of magnitude from 10^6 (Arstila *et al.*, 1999), 10^7 (Robins *et al.*, 2009), to over 10^8 (Qi *et al.*, 2014).

Generation of $\alpha\beta$-T cell diversity happens in the thymus, where thymocytes randomly rearrange gene segments to generate a TCR (Nikolich-Žugich *et al.*, 2004). This heterodimer is generated by random recombination of Variable (V), Diversity (D), and Joining (J) segments, and V and J segments for TCRB and TCRA sequences, respectively (Davis and Bjorkman, 1988). Most variability arises due to random nucleotide insertions and deletions where the segments are joined (Murugan *et al.*, 2012). This VDJ recombination process leads to a potential number of TCRs (recent estimates range from $> 10^{20}$ (Zarnitsyna *et al.*, 2013) to 10^{61} (Mora and Walczak, 2016)) that vastly outnumbers the realized number of TCRs ($< 10^{12}$). After VDJ recombination, T cells undergo positive and negative selection, which selects those T cells that do not have too high or too low affinity for self-pMHC (McDonald *et al.*, 2015). About 3-5% of thymocytes survive selection and enter the periphery as "naïve" T cells (i.e. T cells that have not yet encountered any foreign cognate antigen) (Merkenschlager *et al.*, 1997).

Since lack of repertoire diversity may lead pathogens to go undetected (Muraro *et al.*, 2014; Nikolich-Žugich *et al.*, 2004; Yager *et al.*, 2008), an important question is how repertoire diversity is maintained throughout life. During aging, the thymic output of new T cells declines because of thymic involution, but this is (largely) compensated by peripheral division (den Braber *et al.*, 2012). The naïve T cell pool slowly declines during aging (Wertheimer *et al.*, 2014), interestingly in CMV$^-$ individuals the CD4 T cell pool is relatively constant while the CD8 T cell pool declines irrespective of CMV status (Wertheimer *et al.*, 2014). The repertoire contracts about 2 to 5 fold in old age (Qi *et al.*, 2014), which may not necessarily result in "holes" in the repertoire. Where in mice the repertoire is maintained primarily by thymic output, in humans peripheral

division is already the main source of T cells in early adulthood (Bains *et al.*, 2009; den Braber *et al.*, 2012). In the periphery T cells compete for cytokines, such as IL-7, and for interaction with self-pMHC (Jenkins *et al.*, 2010; Takada and Jameson, 2009). This competition may reduce repertoire diversity as some clonotypes (i.e. T cells expressing the same TCR) grow larger and outcompete other clonotypes (De Boer and Perelson, 1994). The life and death of T cells (T cell dynamics) leads to differences in the sizes of clonotypes which can be shown by a "clone-size distribution", summarizing how many clonotypes there are at each particular clone-size. Hence, clone-size distributions inform us about T cell dynamics and how diverse repertoires are maintained.

Previous studies have mainly focused on fitness differences between T cells to explain why some naïve clonotypes are larger than others (De Boer and Perelson, 1994, 1995, 1997; Desponds *et al.*, 2016, 2017; Dowling and Hodgkin, 2009; Hapuarachchi *et al.*, 2013; Johnson *et al.*, 2012; Lythe *et al.*, 2016; Stirk *et al.*, 2008, 2010). However, it has been reported that TCR sequences (σ) that are more likely to be generated by VDJ recombination, $\mathcal{P}(\sigma)$, are more likely to be selected in the thymus (Elhanati *et al.*, 2014), and that "public" TCR sequences (i.e. those that occur in multiple individuals) tend to have a higher $\mathcal{P}(\sigma)$ (Elhanati *et al.*, 2014; Hou *et al.*, 2016; Murugan *et al.*, 2012; Ndifon *et al.*, 2012). The latter studies suggest that $\mathcal{P}(\sigma)$ may play a role in determining the clone-size of naïve clonotypes. Here, we investigated the role of $\mathcal{P}(\sigma)$ on the naïve clone-size distribution using NGS of blood samples of healthy human adults and modeling. Thanks to the model of the Mora and Walczak groups (Marcou *et al.*, 2017; Murugan *et al.*, 2012) we were able to determine the $\mathcal{P}(\sigma)$ of the TCR sequences in our NGS data from human adults. Remarkably, despite peripheral division being the main source of T cells, the generation probabilities ($\mathcal{P}(\sigma)$) form a sufficient explanation for the observed differences in naïve clone-sizes in (small) blood samples.

4.2 Results

Quantification of naïve CD4/CD8 α/β TCR clone-size distributions

We obtained T lymphocytes from two healthy human volunteers (30 to 40 years old). T lymphocytes were isolated from blood, and sorted by FACS into naïve ($CD27^{+}CD45RA^{high}$), central memory (CM), effector memory (EM) and RA-positive effector memory (EMRA) $CD4^{+}$ and $CD8^{+}$ T cells (see Figure [supplemental FACS plot]). The TCRα and TCRβ mRNA molecules were reverse transcribed to cDNA molecules,

to which 12 nucleotide long unique molecular identifiers (UMIs) were attached (Kivioja *et al.*, 2011), followed by PCR amplification and sequencing on an Illumina MiSeq platform. For each T cell subset approximately 1 to 3 million sequence reads were obtained for both donors. Sequence reads were processed using the Decombinator pipeline (Thomas *et al.*, 2013), which annotates the sequences and performs error correction by collapsing TCR sequences with identical UMIs. We define "clone-size" as the number of UMIs associated with a particular TCRA or TCRB sequence. Using the UMIs for error correction leads to high fidelity TCR sequences, and using the number of UMIs per TCR sequence (rather than the number of sequence reads) as clone-size, reduces potential bias introduced by unequal PCR amplification (Best *et al.*, 2015; Kivioja *et al.*, 2011; Shugay *et al.*, 2014). Since PCR- and sequencing-errors can artificially inflate the clone-size, we performed an additional UMI-count correction step (for details see Section 4.4.1). Briefly, UMIs associated with the same TCR sequence, but differing by one or a few nucleotide substitutions were collapsed (depending on a threshold defined by the clone-size, i.e. the number of UMIs supporting the clonotype). Finally, to enrich the naïve TCR datasets for true naïve TCR sequences, we removed TCR sequences that also occurred in the non-naïve (i.e. CM, EM, EMRA) subsets from the $CD27^{+}CD45RA^{high}$ sequence data. Note, we nevertheless expect antigen-experienced clonotypes to be present in this highly purified $CD27^{+}CD45RA^{high}$ subset, since memory cells expressing a naïve phenotype have been described in several studies (Gattinoni *et al.*, 2011; Lugli *et al.*, 2013a,b; Marraco *et al.*, 2015).

4.2.1 A neutral model to describe the naïve T cell pool

We developed a simple model in which individual T cells have the same chance of survival, that is irrespective of their TCR, which makes this a "neutral" model (Hubbell, 2001; Sloan *et al.*, 2006). This "Markov Chain" model is a stochastic birth-death process having only 3 parameters: pool size, N, thymic release size, k, and thymic production probability, θ. Studying human adults, the pool size, N, is assumed to be constant, which implies that exit from the pool by cell death or activation is perfectly balanced by thymic production (θ) and peripheral division ($1-\theta$). Thus, every time a naïve T cell leaves the pool, there is a probability $\frac{\theta}{k}$ that the thymus produces a unique clonotype of exactly k cells, and a probability $1-\theta$ that a T cell in the periphery divides. Competition between naïve T cells for survival signals has no effect on the shape of the naïve clone-size distribution, and a clonotype produced by the thymus is never produced again (See section 4.4.2 for the equations), in this model.

For TCRB ($k = 100$) we use a larger thymic release size than for TCRA ($k = 10$), because thymocytes rearrange the β chain first, followed by 6-8 divisions, before re-arranging the α chain, resulting in an expected average of 62-102 T cells sharing the same TCRB after thymic selection (Gonçalves *et al.*, 2017). For $CD4^+$ and $CD8^+$ naïve T cells we use pool sizes of $N = 7.5 \times 10^{10}$ and $N = 2.5 \times 10^{10}$, respectively.

After many births and deaths of naïve T cells, the clone-size distribution will approach a "steady state", where the fraction of T cells produced by the thymus is given by the thymic production probability (θ). Because of thymic involution, this fraction declines throughout life from an estimated 20% at age 25 to approximately 2% at age 75 (den Braber *et al.*, 2012). Due to this variation and uncertainty in the exact values of the thymic production, we compute the steady state clone-size distribution for low ($\theta = 2\%$), medium ($\theta = 10\%$), and high ($\theta = 50\%$) thymic production in a healthy human adult (Figure 4.1A).

The model predicts a highly diverse repertoire of naïve clonotypes, i.e. on the order of 10^{10} TCRA and 10^9 TCRB clonotypes at medium thymic output. There are many more large TCRB clonotypes than TCRA clonotypes, because the thymic release size (k) is higher for TCRB than for TCRA. As expected, a high thymic production leads to the highest diversity, but even with low thymic output, the repertoire still contains over 10^8 TCRB clonotypes, which concurs with conservative experimental estimates (Qi *et al.*, 2014).

4.2.2 Sampling distorts the clone-size distribution

Since it is not feasible to sequence the full TCR repertoire, we use the neutral model to calculate the expected clone-size distribution within a typical small sample, and compare this prediction to distributions obtained from blood samples. Sampling a small fraction of the repertoire markedly distorts the clone-sizes, because even large clonotypes are expected to be small in the sample, becoming indistinguishable from the many small clonotypes that by chance end up in the sample. As a result, thymic production, which has a clear effect on the full naïve clone-size distribution (Figure 4.1A), has a negligible effect on the distribution in a small (fraction $s \approx 10^{-6}$) sample (Figure 4.1B). To show how strong the sampling effect can be, we employ the solution of the model to reveal that $\hat{F}_i \approx F_i(\frac{s}{\theta})^i$, where $\hat{F}_i$ and F_i are the number of clonotypes containing i cells in the sample and repertoire, respectively, θ is the probability that a new cell originates from the thymus, and s is the fraction of the repertoire that was sampled (For more details see

Section 4.4.4). We conclude that sampling makes it difficult to distinguish different full clone-size distributions as they tend to converge to very similar sample distributions.

4.2.3 More large naïve clonotypes observed than predicted

In all subsets (i.e. CD4/CD8 and TCRα/β) the neutral model predicts that there should be no clonotypes larger than 2, whereas a high number of clonotypes (on average 357; 0.75%) with a UMI count higher than 2 were observed in each subset in both volunteers (Figure 4.1B). The UMI count may not be a true reflection of the number of T cells with a particular TCR sequence, because the number of mRNA molecules may vary among naïve T cells (Calis and Rosenberg, 2014). This means that some T cells are expected to be represented by multiple UMIs, and others by one or none. To take TCR-mRNA variation into account, we fitted a gamma distribution to the data using the predicted sample clone-size distribution with a medium thymic output ($\theta = 10\%$) as a basis (Figure 4.1C;Supplemental figure S4.1). The resulting fit captures the overall shape of the naïve clone-size distribution in a blood sample, however the number of singletons is underestimated by 2 to 3 fold, which is more than the variation seen in other blood samples from the same volunteer (see below in Figure S4.2). These results suggest that additional factors play a role in shaping the naïve clone-size distribution.

4.2.4 VDJ recombination generates some sequences more often than others

Using the VDJ recombination model of Marcou *et al.* (2017), we calculated the generation probabilities ($\mathcal{P}(\sigma)$) of all TCR sequences in our datasets. The naïve clone-size is positively correlated with the probability its TCR sequence is produced by VDJ recombination (Figure 4.2). On average the TCR sequences of the largest naïve clonotypes have a 75-fold higher generation probability than the smallest naïve clonotypes, a trend that is not observed in the non-naïve subsets (Figure 4.2). Thus, differential production of TCR sequences is an obvious factor missing from our model.

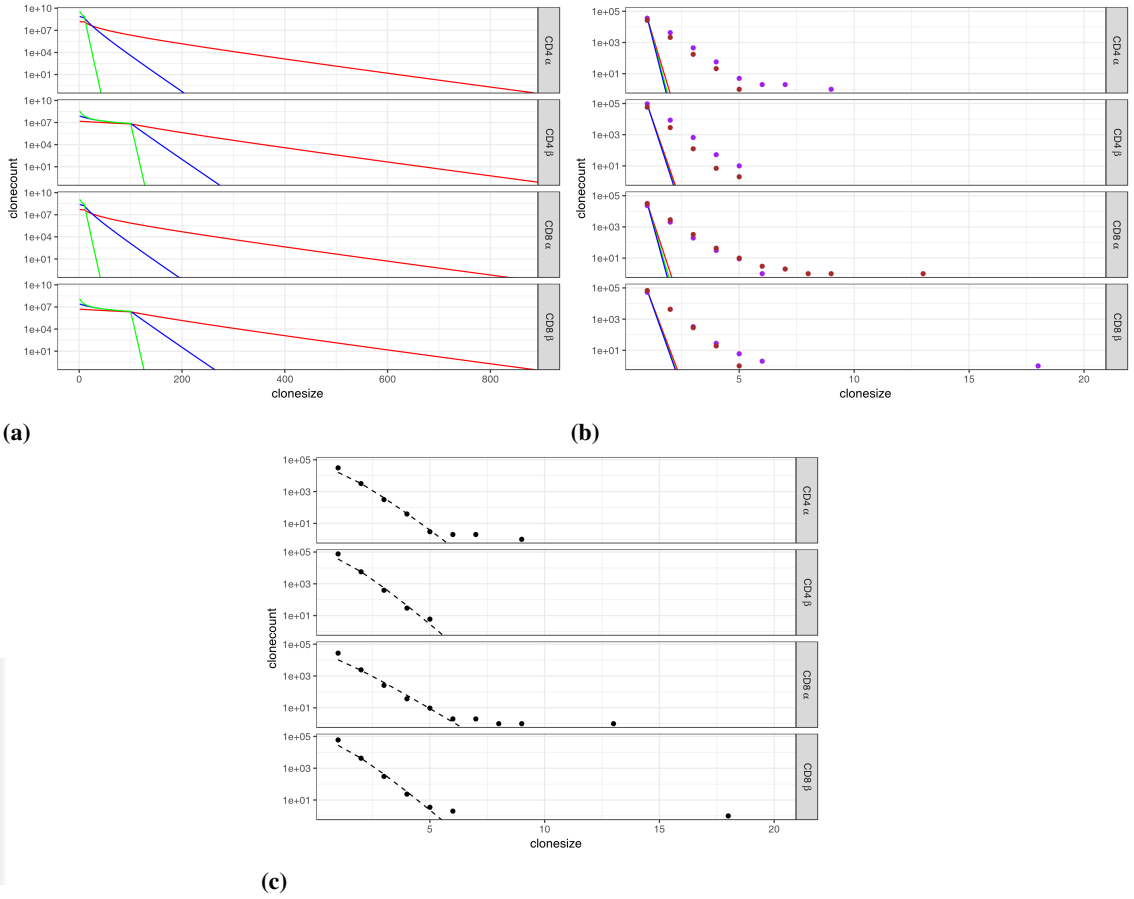


Figure 4.1. A model assuming neutral dynamics predicts even smaller naïve T cell clonotypes than we observe in small human samples. (a) clone-size distributions of full body naïve T cell repertoires as predicted by neutral dynamics with a low ($\theta = 2\%$, red), medium ($\theta = 10\%$, blue), or high ($\theta = 50\%$, green) thymic output. Even with low thymic output naïve T cell clonotypes are predicted to become no larger than about 800 cells. (b) Comparison between naïve clone-size distributions in small samples taken from two volunteers (purple and brown dots), and similar samples from the neutral model. The sample size for the model was the average number of UMI counts of the two volunteers for every subset. (c) The neutral model fits clonotypes larger than 1 well when we fit a gamma distribution to the average clone-size distribution of the volunteers to account for possible variation in the number of TCR mRNA molecules in naïve T cells.

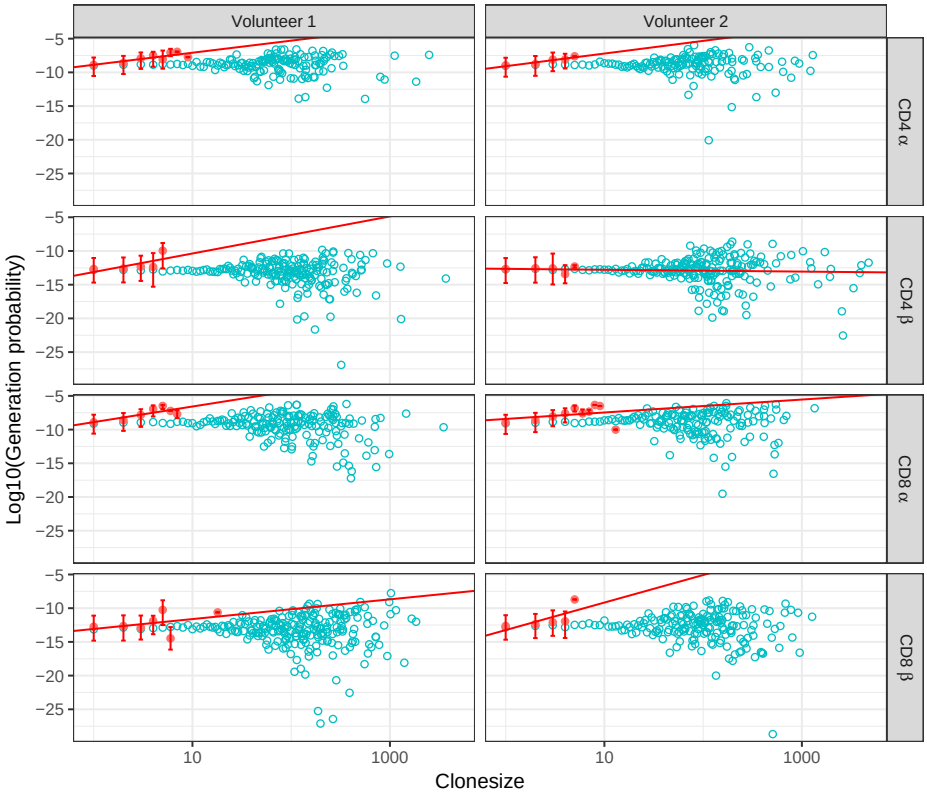


Figure 4.2. Large naïve T cell clonotypes tend to have a higher generation probability than small naïve T cell clonotypes. For every TCR sequence, σ, the generation probability, $P(\sigma)$, was calculated using IGoR (Marcou *et al.*, 2017). The median $P(\sigma)$ was recorded for all observed clone-sizes of naïve clonotypes (red; including 25% and 75% quartiles) and non-naïve clonotypes (green).

4.2.5 Neutral model with differential TCR generation probabilities predicts existence of large naïve clonotypes

To take the association between the size of a naïve clonotype and its generation probability, $\mathcal{P}(\sigma)$, into account, we parameterise an extension of our neutral model using the generation probabilities of Marcou *et al.* (2017) (see Section 4.4.3). This extension just means every time the virtual thymus produces a clonotype, there is a non-zero chance that this clonotype already exists in the naïve T cell pool, and since survival chances remain the same for all T cells in the periphery, this remains a neutral model (Compared to the NCM of Hubbell, TCR generation probabilities correspond to species immigration probabilities).

The steady state clone-size distribution of the above described "differential production" model is depicted in Figure 4.3A, for low ($\theta = 2\%$), medium ($\theta = 10\%$), and high ($\theta = 50\%$) thymic production. Naïve TCRA and TCRB clonotypes can become larger than 10^5 and 10^3 cells (Figure 4.3A), respectively, whereas clone-sizes were never expected to be larger than 800 cells in the "equal production" model (Figure 4.1A). Since the generation probability of TCRA sequences is on average over 200-fold higher than that of TCRB sequences (Supplemental figure S4.3), many more large TCRA clonotypes than large TCRB clonotypes are predicted (which reverses the prediction of the equal production model). Clonotypes smaller than the thymic release size (k) tend to be affected by thymic production (θ) in a manner similar to the equal production model (Figures 4.1A, 4.3A). This is expected, because small clonotypes tend to be produced only once.

Despite the much larger clonotypes in the full clone-size distribution, the differential production model predicts that in a small ($s \approx 10^{-6}$) sample naïve clonotypes should be no larger than 3 cells, whereas, on average, 40 (0.094%) clonotypes larger than that are observed in every subset in the blood samples (Figure 4.3B). Since this discrepancy might again be due to variation in the number of TCR mRNA molecules in T cells, we fitted a gamma distribution to the data using $\theta = 10\%$ thymic production as a basis. The fit of the model (Figure 4.3C) is very similar to that of the old model (Figure 4.1C), suggesting that a single blood sample is insufficient to distinguish these models. This prompted us to perform a new experiment.

4.2.6 Split blood sample experiment confirms that large naïve clonotypes exist and that there are more large naïve TCRA clonotypes than large naïve TCRB clonotypes

To remove the uncertainty stemming from possible mRNA variation among naïve T cells, and test directly whether large clonotypes in a sample are also abundant in the full repertoire, we performed a dedicated experiment, splitting a blood sample into three subsamples before mRNA extraction, and counted in how many subsamples each TCR sequence occurs (i.e. the "incidence"). Since each subsample constitutes a small fraction of the full repertoire (approximately 10^5 cells out of 10^{11} naïve T cells), clonotypes observed in more than one subsample are expected to be large in the full repertoire (i.e., $> 10^5$ and $> 7 \times 10^5$ cells for incidence 2 and 3, respectively). On average 3004 (1.6%) clonotypes in the TCRA, and 953 (0.24%) clonotypes in the TCRB subsets, were observed in more than one subsample (Figure 4.4A), establishing the existence of many large naïve clonotypes in the full repertoire.

Without generation probabilities ($\mathcal{P}(\sigma)$) the neutral model predicts few (196 in all subsets together) clonotypes to occur in more than one subsample (Figure 4.4A, blue bars), whereas with generation probabilities many (12776) such clonotypes are predicted to be present (Figure 4.4A, green bars), demonstrating that clonotypes occur in multiple subsamples because of their $\mathcal{P}(\sigma)$, rather than by a stochastic birth-death process. This result demonstrates that probabilities involved in VDJ recombination, $\mathcal{P}(\sigma)$, provide a sufficient explanation for the presence of large naïve T cell clonotypes in the full repertoire.

As predicted by the differential production model, naïve TCRA clonotypes with a higher incidence have a right-shifted $\mathcal{P}(\sigma)$ distribution compared to TCRA clonotypes with a lower incidence (Figure 4.4B). This means large TCRA clonotypes tend to have a higher $\mathcal{P}(\sigma)$ than small TCRA clonotypes in the full repertoire. A remarkable observation is that for the naïve TCRB clonotypes, incidence 2 clonotypes have a right-shifted $\mathcal{P}(\sigma)$ distribution compared to the incidence 2 clonotypes, but that the incidence 3 clonotypes have a *left-shifted* distribution compared to the incidence 2 clonotypes (Figure 4.4B), breaking the trend that large clonotypes tend to have a higher $\mathcal{P}(\sigma)$ than small clonotypes in the full repertoire. Unexpectedly, this observation is perfectly compatible with the differential production model as this model predicts almost no TCRB clonotypes to occur in all three subsamples (Figure 4.4A, green bars). This means the model predicts that the about 400 incidence 3 TCRB clonotypes that we observe should be large for another

reason, e.g., these clonotypes could be memory cells expressing naïve markers (here CD27^{+}CD45RA^{high}, (Gattinoni *et al.*, 2011; Lugli *et al.*, 2013a,b; Marraco *et al.*, 2015)) which are not expected to have a high $\mathcal{P}(\sigma)$. Indeed, the observed incidence 3 TCRB clonotypes tend to have a lower $\mathcal{P}(\sigma)$ than even the incidence 2 TCRB clonotypes.

In summary, the sequence data confirms the predictions of the differential production model that naïve clonotypes are large, and that there are more large naïve TCRA clonotypes than large naïve TCRB clonotypes. Together, these results demonstrate a major role for generation probabilities involved in VDJ recombination in shaping the clone-size distribution of naïve T cells.

4.3 Discussion

To determine the factors that shape the clone-size distribution of the naïve T cell pool, we developed simple mathematical models, and tested these models using NGS of TCR mRNA in single blood samples from healthy volunteers. A simple neutral birth-death process predicts small blood samples to contain mostly singleton naïve clonotypes, and is unable to explain the small fraction, but large number, of large clonotypes, that we observe. Interestingly, the abundant naïve clonotypes in a sample tended to have a higher generation probability, $\mathcal{P}(\sigma)$, than less abundant naïve clonotypes. To remove uncertainty stemming from variation in the number of mRNA molecules in the naïve T cells in a sample, we performed a dedicated experiment, splitting a blood sample into multiple subsamples, before mRNA extraction. Clonotypes observed in multiple subsamples can safely be considered large, because these subsamples constitute only a very small fraction of the full repertoire. We extended the neutral model by assigning each clonotype its own $\mathcal{P}(\sigma)$, which confirmed that many TCR sequences should be observed repeatedly, and that there should be significantly more large TCRA than large TCRB clonotypes in small samples. Remarkably, the generation probabilities form a sufficient explanation for the observed naïve clone-sizes, and the number of clonotypes appearing repeatedly in the subsamples. The most unexpected correct prediction of the model is that very few TCRB sequences should be present in all three subsamples, which is confirmed by the finding that the TCRB sequences that we do observe three times tend to have normal generation probabilities, and are therefore expected to be antigen-experienced cells expressing CD27 and CD45RA (Gattinoni *et al.*, 2011; Lugli *et al.*, 2013a,b; Marraco *et al.*, 2015). These results demonstrate an unexpectedly large role for the probabilities involved in VDJ recombination in shaping the clone-size distribution of naïve T cells, casting doubt

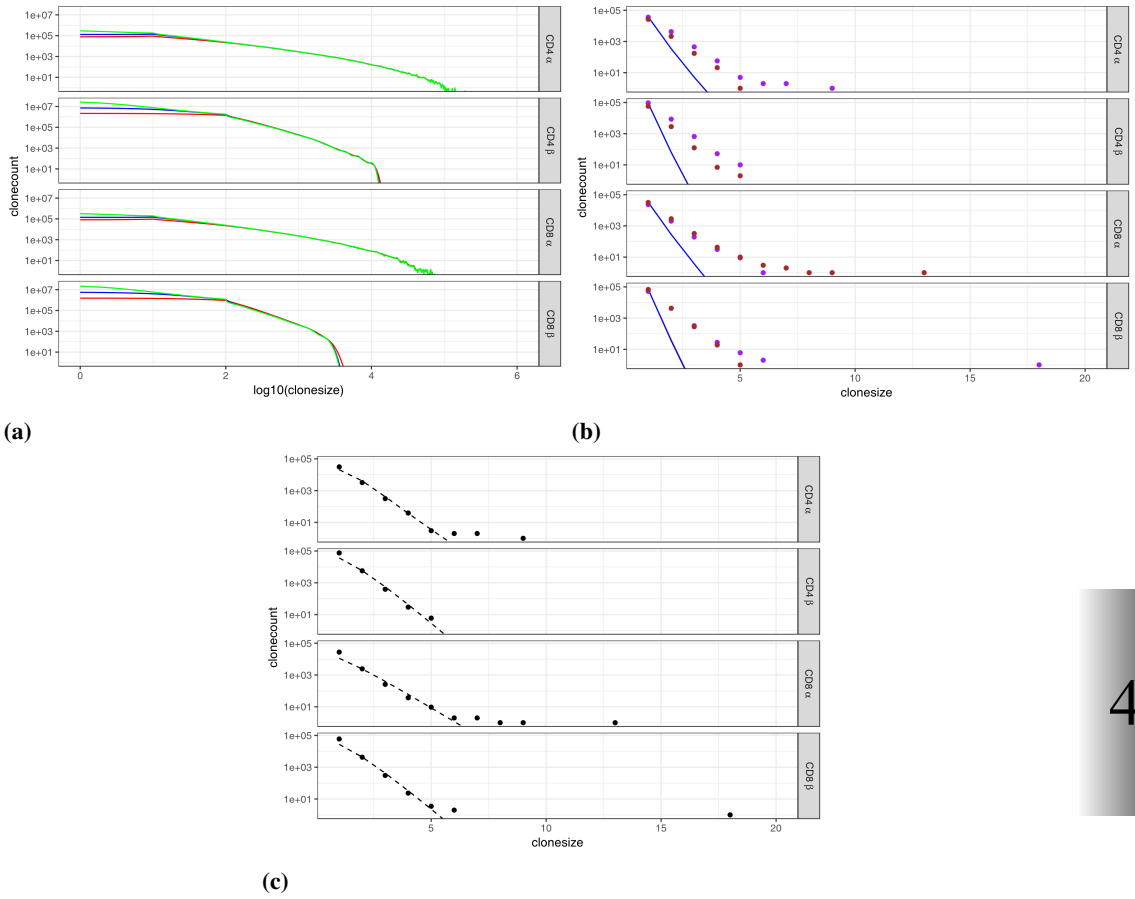
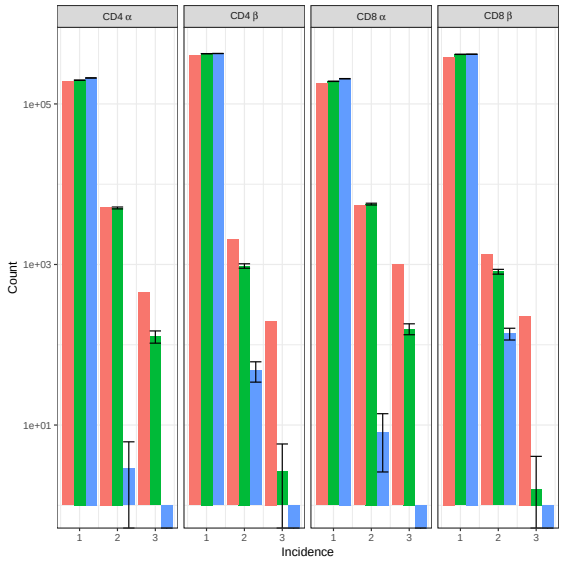
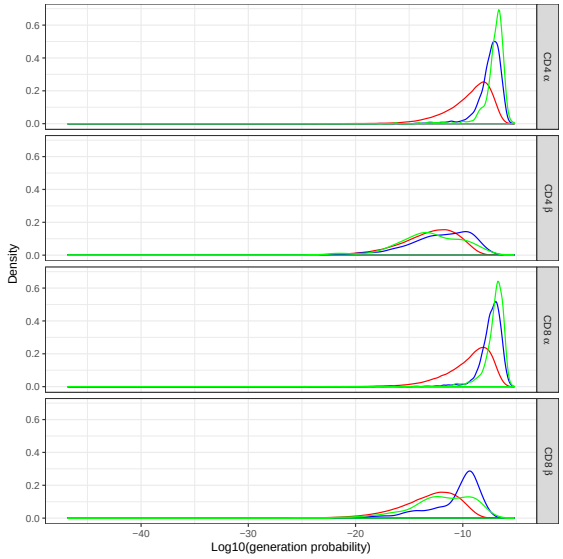


Figure 4.3. Neutral model with differential TCR generation probabilities predicts very large naïve T cell clonotypes. Following the format of Figure 4.1, (a) Clone-size distributions based on the generation probabilities of both volunteers for each subset (e.g. CD4 α) respectively. (b) The inclusion of generation probabilities predicts larger naïve clonotypes for the α chain than for the β chain, a pattern that is reflected by the experimentally observed clone-sizes. (c) Including a gamma distribution to model mRNA content variation enables the model to predict more of the larger clonotypes in the sample.



(a)



(b)

Figure 4.4. The number of clonotypes occurring once, twice, or three times in a split blood sample. (a) Comparison between observed clonal incidence (red) and expected clonal incidence according to the neutral model with (green) and without generation probabilities (blue). (b) Distribution of generation probabilities of clonotypes that occur in one (red), two (blue), and three (green) blood samples.

on the effect of (TCR-dependent) fitness differences between naïve T cells in the periphery on the naïve clone-size distribution. Finally, we find that the total abundance of the clonotypes in the subsampling experiment correlates positively with their incidence (Supplemental figure S4.4), confirming that the abundances estimated by the number of UMIs per clonotype provide a reasonable estimate of clone-size, and again that some naïve clonotypes are truly large in the full repertoire.

Previously, differences in TCRs, resulting in higher renewal rates and/or longer expected life spans of the most competitive clonotypes, have been used to explain variation in naïve clone-sizes (De Boer and Perelson, 1994, 1995, 1997; Desponds *et al.*, 2016, 2017; Hapuarachchi *et al.*, 2013; Lythe *et al.*, 2016; Stirk *et al.*, 2008, 2010). On the contrary, our neutral differential production model assumes that TCR-dependent fitness differences do not strongly determine the size of naïve clonotypes. This is the first model that uses generation probabilities, $\mathcal{P}(\sigma)$, to explain observed naïve clone-size distributions, not requiring cognate interactions. Moreover, if cognate interactions were to play a large role they would confound the effects of $\mathcal{P}(\sigma)$, and we would not expect to observe the positive correlation between $\mathcal{P}(\sigma)$ and clone-size (Figures 4.2,4.4B).

We show large naïve clonotypes tend to have a higher $\mathcal{P}(\sigma)$, implying such clonotypes should reappear in other individuals, i.e. that they should be "public". Public clonotypes tend to be made more often by the thymus through convergent recombination (Ndifon *et al.*, 2012; Venturi *et al.*, 2006, 2008, 2011) (i.e. multiple nucleotide sequences coding for the same amino acid sequence) or are more likely rearranged by VDJ recombination (i.e. higher $\mathcal{P}(\sigma)$) (Elhanati *et al.*, 2014; Hou *et al.*, 2016; Murugan *et al.*, 2012; Ndifon *et al.*, 2012). Since these clonotypes (with high $\mathcal{P}(\sigma)$) tend to have few insertions and deletions, they tend to be close to germline. Interestingly, during infancy the repertoire is initially populated by TCR sequences that are also close to germline, essentially containing specificities that might be considered innate (Sethna *et al.*, 2017). Given the ubiquity of clonotypes that are close to germline, there might be evolutionary pressure on the germline sequences, possibly through selection for specificities against particular pathogens. Together, these results emphasize the need to determine $\mathcal{P}(\sigma)$ of the TCR sequences that are shared between individuals, and raise questions on the potential benefits of having a small fraction, but a large number, of large clonotypes in the naïve repertoire.

4.4 Supplemental information

4.4.1 Sequence analysis

We used the Decombinator pipeline (Thomas *et al.*, 2013) to demultiplex, annotate, and error-correct the raw sequencing reads. Our reads contain 12 bp long Unique Molecular Identifiers (UMIs) that can be used to identify which TCR sequences are derived from the same cDNA molecule. To error correct TCR sequences, Decombinator collapses those TCR sequences that are similar and are associated with the same UMI. The pipeline also error corrects UMIs, collapsing those UMIs that are associated with the same TCR sequence and differ from each other by 2 or fewer sequence edits (i.e. the default barcode threshold). This error correction assumes it is unlikely for any clonotype, irrespective of clone-size, to contain two UMIs that are nearly identical, concluding the UMIs are different because of PCR or sequencing errors. To improve upon the reliability of the reported clone-sizes, we set the barcode threshold to 0 and developed a new UMI error correction algorithm.

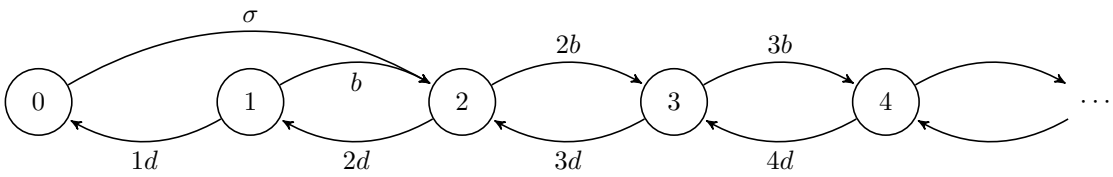
Consider a clonotype supported by i different UMIs, i.e. clone-size i. The Hamming distance, H, between two random UMIs (assuming uniform base frequencies) can be represented by a binomial random variable, $H \sim \mathrm{B}(n, p)$, where $n = 12$ and $p = \frac{3}{4}$. There are $\binom{i}{2}$ distinct comparisons between the i UMIs, and assuming that every comparison is independent, the distribution of observed Hamming distances is $n_i(h) = \binom{i}{2}\mathcal{P}(H = h)$. To determine whether two UMIs are unexpectedly similar, we define a threshold distance that depends on the clone-size (i):

$$D_\alpha = \max(\{d : \sum_{h=1}^{d} n_i(h) \leq \alpha\}) \,. \tag{4.1}$$

Our algorithm error corrects UMIs for the clonotype follows: From $d = 1$ to $d = D_\alpha$, for all UMI pairs with $H \leq d$, add read count of the less abundant UMI to the more abundant UMI and remove the former. Finally, we applied this algorithm to every clonotype in our sequence data using $\alpha = 5\%$.

4.4.2 Neutral model with equal production probabilities

To model naïve T cell dynamics we developed a model that is similar to the Neutral Community Model (NCM) of Hubbell (Hubbell, 2001). Naïve T cells, viewed through an ecological lense, are individuals, and all naïve T cells sharing the same TCR are part of the same species (here, clonotype). Neutrality, as defined by Hubbell, means that all species have the same per capita probability of birth and death. In the equal production model we assume that all clonotypes have an equal probability of being produced by the thymus. Once produced, the same clonotype cannot be produced again, i.e. the thymus produces each clonotype only once. The model can be represented by a Markov chain, as is shown in the following example:



The thymus, state $i = 0$, produces all clonotypes at the same clone-size, k naïve T cells, and at a rate $\sigma = \frac{\theta}{k}$, releasing the produced clonotypes into the naïve T cell pool of N cells ($N = 7.5 \times 10^{10}$ and $N = 2.5 \times 10^{10}$ for CD4$^+$ and CD8$^+$ T cells, respectively). In the above example $k = 2$, i.e. the clone-size at which clonotypes enter the pool. The transition rates of clonotypes only depend on their size (i.e. neutrality assumption), hence, the per capita birth rate, $b = \frac{1-\theta}{N}$, and per capita death rate, $d = \frac{1}{N}$, are the same for all clonotypes. At steady state, the rate, σ, at which clonotypes enter the naïve pool, should equal the rate, dF_1, at which clonotypes leave the pool, hence $F_1 = \frac{\sigma}{d} = \frac{\theta}{k}N$, where F_i is the number of clonotypes of clone-size i. The rate at which clonotypes of size i divide and die depends on the total number of T cells all F_i clonotypes contain: $N_i = iF_i$. For clone-sizes up to size k, transitions from $i-1$ to i, at steady state, balance the rate at which clonotypes of size i die, dN_i, with the rate at which clonotypes enter the pool, σ, and the rate at which clonotypes of size $i-1$ divide, bN_{i-1}, i.e. $dN_i = \sigma + bN_{i-1}$. The analytical solution of this recurrence relation is:

$$iF_i = \frac{N - N(1-\theta)^i}{k}, \quad \text{for } 2 \leq i \leq k\,. \tag{4.2}$$

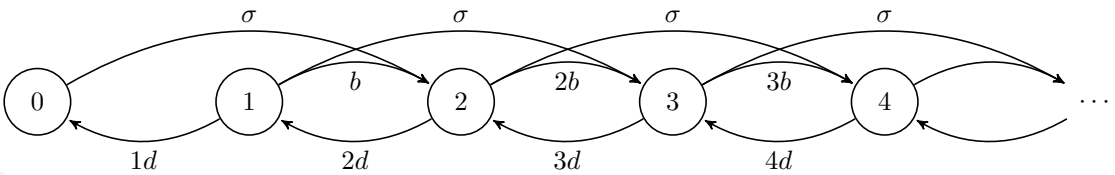
For clonotypes larger than k, only birth and death of clonotypes need to balance between states $i-1$ and i (as there is no flux from the thymus): $dN_i = bN_{i-1}$. This recurrence

relation has the following analytical solution:

$$iF_i = kF_k(1-\theta)^{i-k}, \quad \text{for } k \leq i \leq N \,. \tag{4.3}$$

4.4.3 Neutral model with differential production probabilities

In the differential production model, we consider the probability that VDJ recombination generates a particular clonotype, $\mathcal{P}(\sigma)$. The thymic selection process is not explicitly modeled, meaning that the model assumes that the thymic production probability (i.e. after thymic selection) positively correlates with $\mathcal{P}(\sigma)$ (shown by Elhanati *et al.* (2014)). For a given clonotype, with generation probability $\mathcal{P}(\sigma)$, the probability it will be observed at a particular clone-size can be represented by a Markov chain, see example below:



The probability that the thymus produces the clonotype is $\sigma = \mathcal{P}(\sigma)\frac{\theta}{k}$. At any clone-size the thymus can produce this clonotype, causing the clonotype to increase size by k cells ($k = 2$ in the above example). The per capita birth rate, $b = \frac{1-\theta}{N}$, and per capita death rate, $d = \frac{1}{N}$, are equivalent to those of the neutral production model, and are again the same for all clonotypes. The rate at which the clonotype loses a T cell depends on the per capita death rate, d, the number of naïve T cells in the clonotype, i, and the probability, S_i, that the clonotype is at state i, i.e. $\mathcal{P}(i \to i-1) = diS_i$. Note, the probability the clonotype is at any clone-size is $\sum_{i=0}^{N} S_i = 1$. At steady state, the transition rates between clone-size $i-1$ and i balances birth, death, and thymic production:

$$diS_i = b(i-1)S_{i-1} + \sigma \sum_{j=\max(i-k,0)}^{i-1} S_j, \quad \text{for } 1 \leq i \leq N \,. \tag{4.4}$$

To calculate the clone-size probability distribution for a single clonotype, we set $S_0 = 1$, compute the above recurrence relation until a predetermined maximum clone-size, c, and normalize the resulting distribution to 1. We typically used $c = \max(10000, 4\mathcal{P}(\sigma)N)$

for all distributions using the differential production model.

We simulated two clonotypes with different $\mathcal{P}(\sigma)$ and compared their empirical clone-size probability distribution to the corresponding distribution calculated using Equation 4.4 (Supplemental figure S4.5). The simulation and the recurrence relation (Equation 4.4) are in close agreement.

To calculate the full clone-size distribution for a particular subset (e.g. CD4-α naïve T cells), we sum the clone-size probability distribution for every $\mathcal{P}(\sigma)$ observed in the subset, weighting the distributions by the number of times their corresponding $\mathcal{P}(\sigma)$ was observed in the subset. The summed clone-size distribution is normalized to contain N cells, by multiplying the distribution by $\frac{N}{\sum_{i=0}^{N} iF_i}$, where F_i is the number of clonotypes at clone-size i.

4.4.4 Sampling from clone-size distributions

Consider a naïve T cell pool of N T cells, distributed according to some clone-size distribution, F, where F_i is the number of clonotypes of clone-size i. From this pool a fraction, s, is sampled (sample size is sN). Assuming the naïve pool is large and well-mixed, the number of T cells, x, sampled from a particular clonotype with clone-size j, can be approximately represented by a binomial random variable, $X_j \sim \mathrm{B}\,(n = j, p = s)$. The expected sample clone-size distribution, $\hat{F}$, is then given by the following equation:

$$\hat{F}_i = \sum_{j=i}^{N} F_j \mathcal{P}(X_j = i)\ . \tag{4.5}$$

We used the above equation to compute the sample distributions for both the equal and differential production models (e.g. Figures 4.1B,4.3B).

For the neutral equal production model, with $k = 1$, we found an analytical solution to Equation 4.5:

$$\hat{F}_i = F_i \left(\frac{s}{s + (1-s)\theta} \right)^i \tag{4.6}$$

Since s is typically very small, this equation can be simplified to $\hat{F}_i \approx F_i (\frac{s}{\theta})^i$ (as $s \ll \theta$),

which clearly shows sampling can strongly distort clone-size distributions.

4.4.5 Accounting for mRNA variation among naïve T cells

The number of distinct UMIs sequenced in a blood sample, n, is an upper bound on the number of naïve T cells of which an mRNA molecule was tagged with a UMI and sequenced. The actual number of naïve T cells, represented by the n observed UMIs, is likely lower. We use a discretized gamma distribution, $f(x) = \mathcal{P}(x \leq X \leq x + 1)$, where $X \sim \Gamma(\alpha, \beta)$, to describe the probability of a naïve T cell in the sample to both get a UMI and to be sequenced. We fitted the discretized gamma distribution to the log-transformed observed sample distribution, using the expected sample distribution of naïve T cells from the modeled full distribution as a basis (Figures 4.1C,4.3C,S4.1).

For both the equal and differential production models the number of singletons was underestimated. This lack of fit might be a result of a combination of the fitting procedure used, the gamma distribution, or the models. We decided to perform a dedicated experiment, splitting a blood sample into multiple subsamples, to remove the confounding effect of mRNA variation.

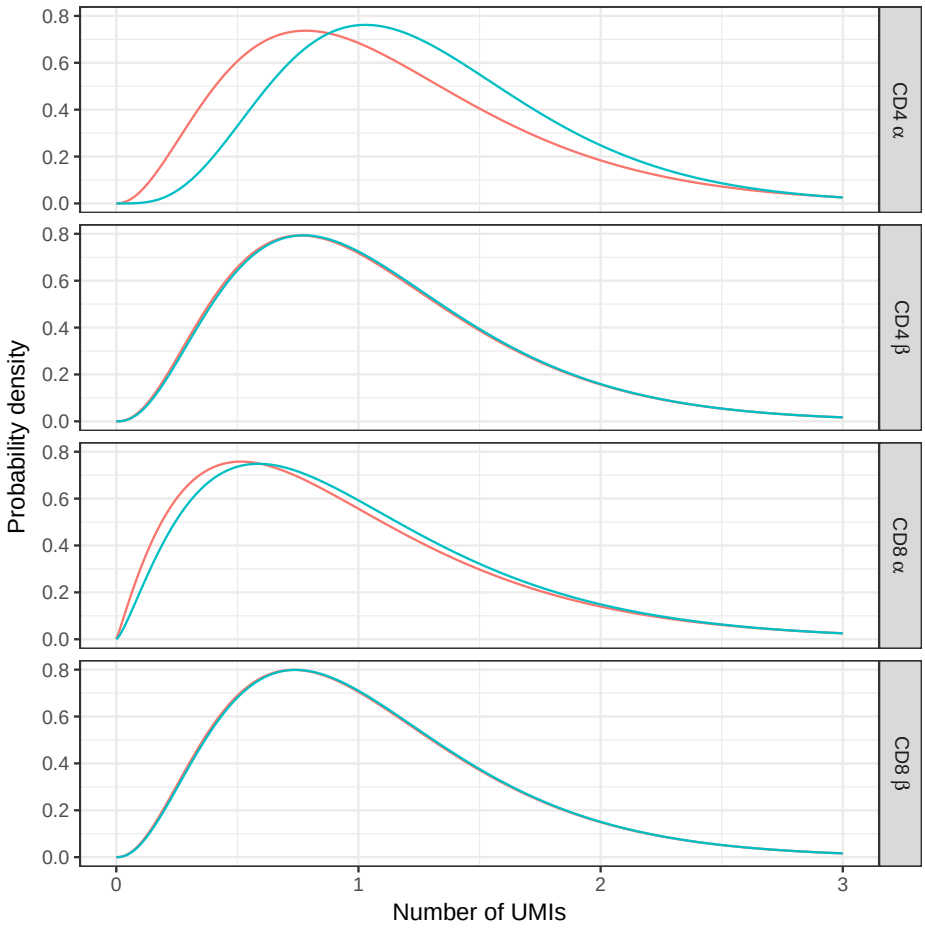


Figure S4.1. Gamma distributions describing the probability a virtual T cell receives a UMI and is sequenced. Neutral model with (green) and without (red) differential production probabilities.

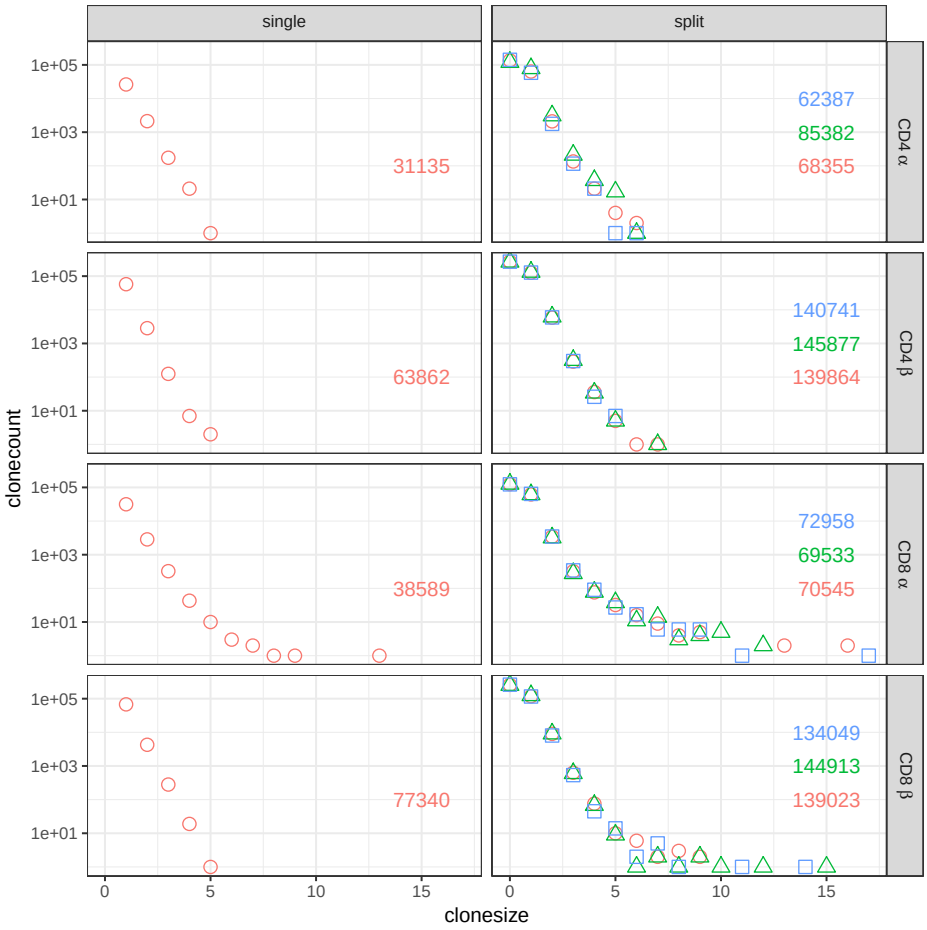


Figure S4.2. Naïve clone-size distributions of blood samples from volunteer 2. (right side of each panel) Total number of UMIs. (Left) Single blood sample experiment. (Right) Split blood sample experiment.

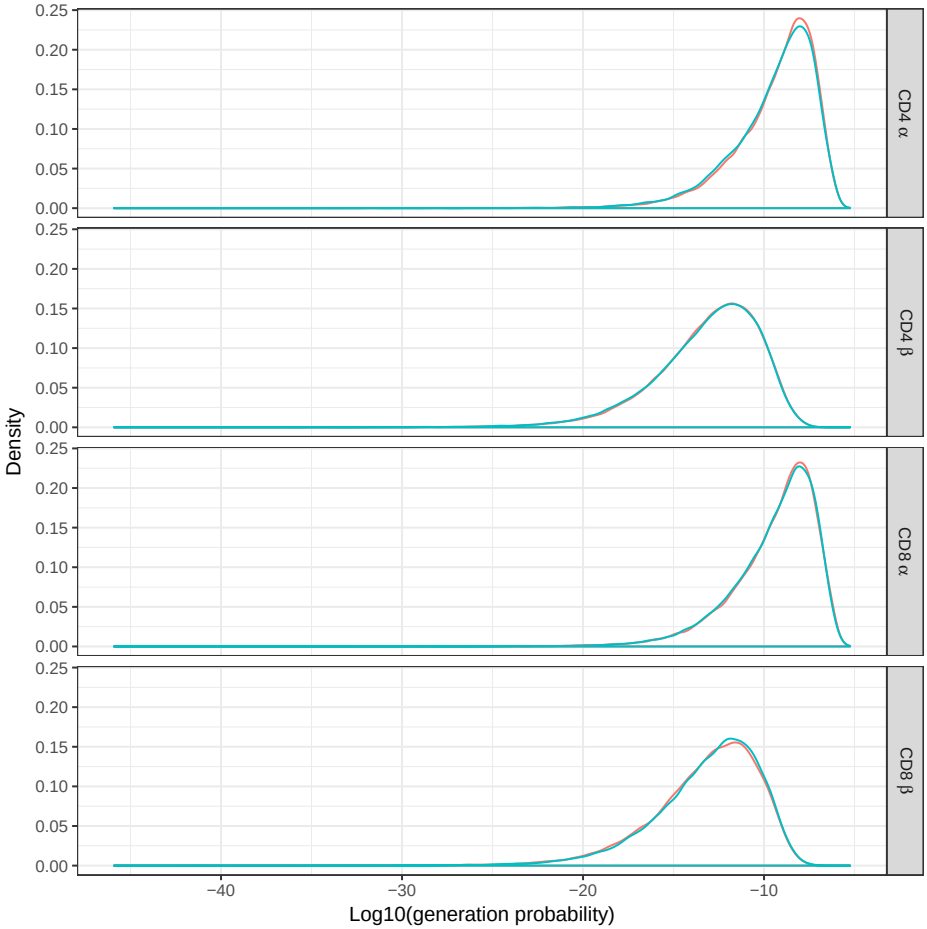


Figure S4.3. Comparison between the distributions of generation probabilities ($\mathcal{P}(\sigma)$) of volunteer 1 (red) and 2 (green).

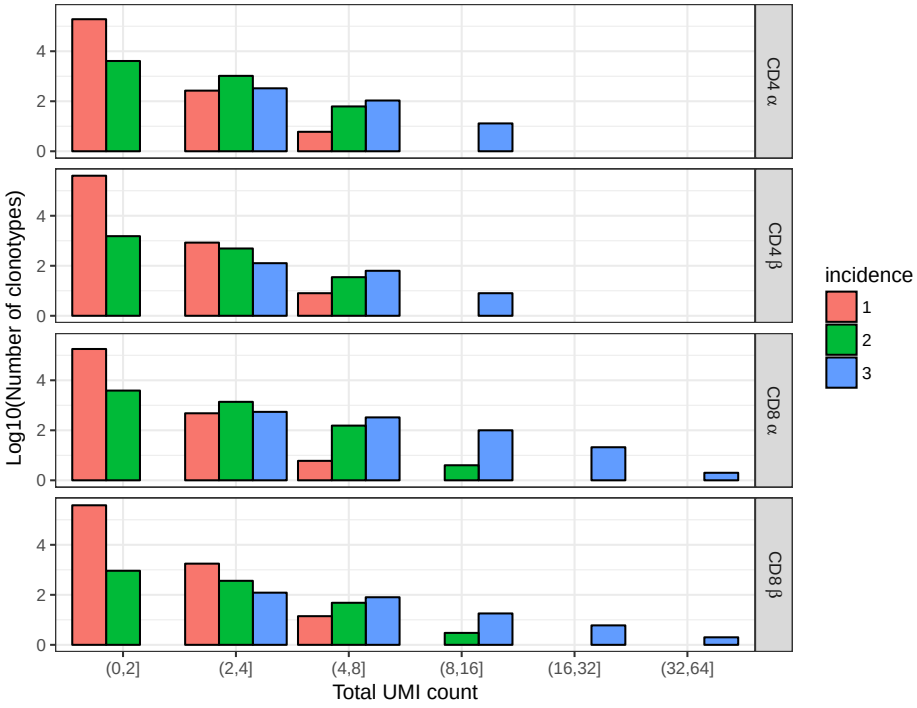


Figure S4.4. Comparison between abundance and incidence of naïve clonotypes of volunteer 2. For every clonotype, the total number of UMIs across the three subsamples was summed ("Total UMI count"). All clonotypes were binned according to their Total UMI count and for each bin the relative fraction of clonotypes occuring in one, two, or all three subsamples (red, green and blue, respectively) is shown.

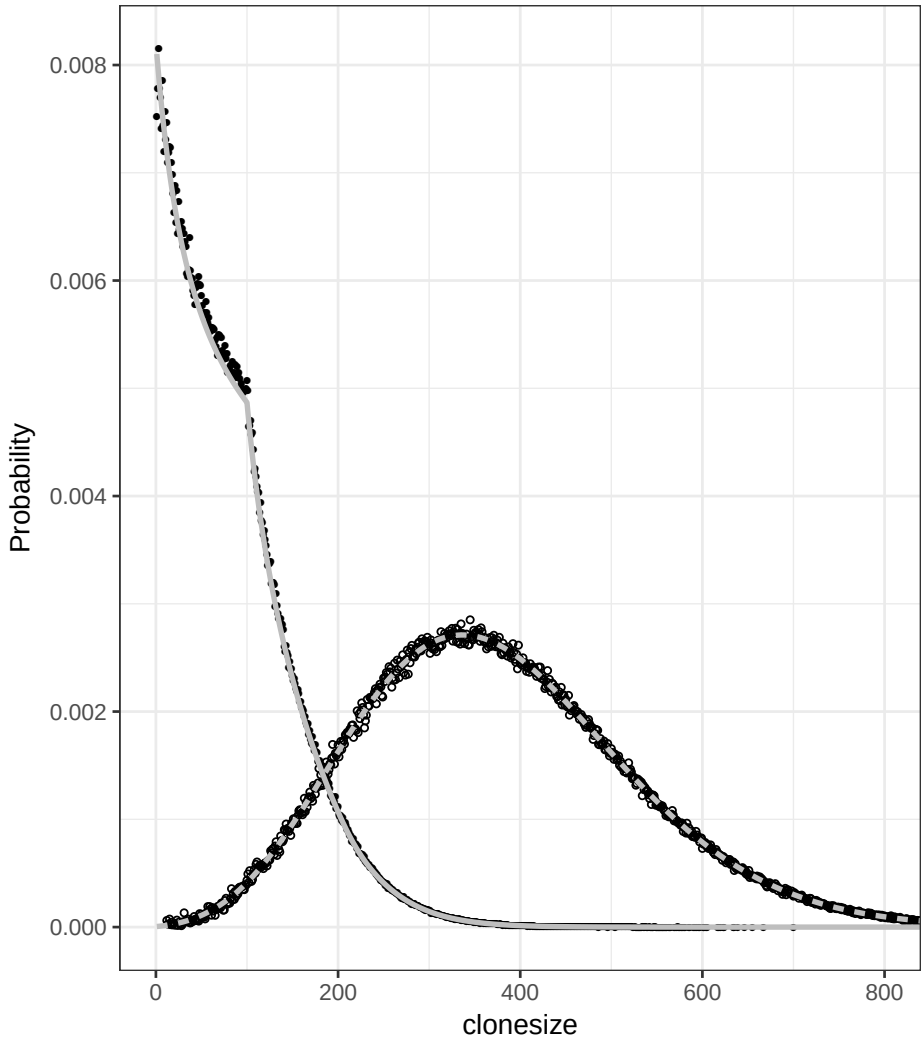


Figure S4.5. Comparison between Gillespie simulation of the differential production model and its recurrence relation (Equation 4.4). Simulations were started at a clone-size of $\mathcal{P}(\sigma)N$ and run for 10^6 steps, recording the final clone-size. For each clonotype, $\mathcal{P}(\sigma) = 10^{-9}$ (filled circles, continuous line) and $\mathcal{P}(\sigma) = 5 * 10^{-9}$ (open circles, dashed line), 10^7 simulations were run, showing the fraction of simulations resulting in a particular clone-size (dots) and the computed recurrence relation (lines).

Chapter 5

Identification of expressed V, D, J, and C genes in the TRB locus of the ferret

BRAM GERRITSEN[a], ARIDAMAN PANDIT[a], FATIHA ZAARAOUI-BOUTAHAR[b], MIRJAM C.G.N. VAN DEN HOUT[c], WILFRED F.J. VAN IJCKEN[c], ROB J. DE BOER[a], AND ARNO C. ANDEWEG[b] (2017)

In preparation

[a] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, the Netherlands.*

[b] *Department of Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands.*

[c] *Center for Biomics, Erasmus Medical Center, Rotterdam, the Netherlands.*

Abstract

The domestic ferret, *Mustela putorius furo*, is an important mammalian animal model to study human respiratory infection. However, insufficient genomic annotation hampers detailed studies of Ferret T cell responses. We analyzed the published T cell Receptor Beta (TRB) locus and performed high throughput sequencing (HTS) of peripheral blood of four healthy adult ferrets to identify expressed V, D, J, and C genes. The HTS data is used as a guide to manually curate the expressed V, D, J, and C genes. The ferret locus appeared to be most similar to that of the dog. Like other mammalian TRB loci, the ferret TRB locus contains a library of variable genes located upstream of two D-J-C gene clusters, followed by a (in the ferret non-functional) V gene with an inverted transcriptional orientation.

5.1 Introduction

The ferret is an important mammalian animal model to study human respiratory infections. The ferret model is well suited to study the pathogenicity and transmissibility of e.g. Coronaviruses (SARS), Pneumoviridae (RSV) and Orthomyxoviruses that include human and avian influenza viruses (Enkirch and von Messling, 2015; Oh and Hurt, 2016). Ferrets are an attractive mammalian model for these infections since ferrets and humans share similar lung physiology and (viral) receptor distribution (Belser *et al.*, 2011; van Riel *et al.*, 2006). A significant drawback of the ferret model is a lack of ferret specific reagents for detailed studies of the host immune response to these pathogens. However, the use of the ferret model has increased over the years, and its usage, with the recent publication of the ferret (draft) genome (Peng *et al.*, 2014), is likely to increase even further. Currently, little is known about the T cell receptor (TCR) repertoire of the ferret, limiting analyses of the immune responses of the ferret to influenza virus and other pathogens.

TCRs are important in mediating recognition of peptide antigens presented to T lymphocytes via the peptide-MHC complex (Davis and Bjorkman, 1988). Conventional TCRs are $\alpha\beta$ or $\gamma\delta$ heterodimers, that are formed by somatic rearrangement of Variable (V), Diversity (D), and Joining (J) gene segments for the β and δ chains, and V and J gene segments for the α and γ chains (Davis and Bjorkman, 1988). Although the ratio between $\alpha\beta$ and $\gamma\delta$ T cell subsets is not known for the ferret, the $\alpha\beta$ T cells are much more common than $\gamma\delta$ T cells in both human and dog (Mineccia *et al.*, 2012). The β chain (at least in humans) tends to make more contact with the peptide antigen than the α chain (Glanville *et al.*, 2017), making the TRB locus an interesting first candidate to annotate in detail.

In this study, we annotate the expressed V, D, J, and C genes in the ferret TRB locus by combining genomic information from the locus with HTS of the ferret TRB repertoire. We find that the TRB locus of the ferret has a similar structure to that of other mammalian TRB loci, such as mouse and human (Glusman *et al.*, 2001), bovine (Connelley *et al.*, 2009), dog (Mineccia *et al.*, 2012), and rabbit (Antonacci *et al.*, 2014): a library of V genes, followed by two (or three in bovine) D-J-C clusters. Each cluster consists of one D gene, six or seven (six in ferret) J genes, and a single C gene. The D-J-C clusters are followed by a V gene with an inverted transcriptional orientiation. We also include a phylogenetic analysis, showing that the ferret V and J genes are indeed most closely related to those of the dog. The ferret locus is small like that of the dog, about 300Kb,

and has a (largely) conserved synteny with the dog locus. Our annotation of the Ferret TRB locus can help studies of T cell responses to, for example, influenza infection in the Ferret.

5.2 Materials and methods

5.2.1 Animals

Four surplus cryopreserved blood samples were obtained from an influenza vaccination-challenge study (Bodewes *et al.*, 2010). The control blood samples originated from 6 to 12 months old healthy outbred female ferrets (*Mustela putorius furo*).

5.2.2 Sequencing and sequence analysis

Peripheral blood mononuclear cells (PBMCs) were isolated from the blood of four healthy thirteen-month-old female ferrets using a standard Ficoll gradient separation protocol. Subsequently, total RNA was isolated and purified using the RNeasy Mini Kit (Qiagen, Hilden, Germany): 250 μl of ethanol was added to the upper aqueous phase of the processed TRIzol samples and directly transferred to the RNeasy spin columns for purification. RNA concentrations and OD 260:280 nm ratios were measured with the NanoDrop® ND-1000 UV-VIS spectrophotometer (NanoDrop Technologies, Wilmington, USA). TCR amplification was performed according to a protocol described by Mamedov et al. (Mamedov *et al.*, 2013). Briefly, RNA obtained from unsorted PBMCs was reverse transcribed by RACE using a single primer directed to the constant region. Twelve nucleotide long unique molecular identifiers (UMIs) were incorporated during cDNA synthesis (Kivioja *et al.*, 2011). Subsequently, two-stage seminested and barcoded PCR amplification was performed including a size selection / agarose gel purification step after the first PCR (Mamedov *et al.*, 2013). The resulting TCR amplicons were were subjected to high-throughput sequencing according to the instructions of the manufacturers using the Ovation Low Complexity Sequencing System kit from NuGEN (San Carlos, CA, USA) and the Illumina MiSeq platform (PE 300). All sequence reads having the same UMI were collapsed into consensus sequences using the RTCR pipeline (Gerritsen *et al.*, 2016). After describing the V and J sequences in the ferret TRB locus, the RTCR pipeline was used to annotate the sequences, and perform additional error

correction.

5.2.3 Nomenclature

Potential TRB V, D, J, or C genes were assigned a temporary unique identifier, "tmpXY" (e.g. tmpV10), where X denotes the gene type and Y is a unique number within the gene type. Confirmed TRB genes were given a name according to IMGT nomenclature established for other species (most notably the dog). Ferret TRBV genes were assigned subgroup numbers according to the TRBV genes they were most similar to in the IMGT GENE DB (Giudicelli *et al.*, 2005), which contains the following species: *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Canis lupus familiaris*, and *Oncorhynchus mykiss*. In all cases the dog TRBV genes were most similar to the corresponding ferret TRBV genes, except for TRBV3 (tmpV23) and TRBV27 (tmpV3), where Macaca and human were more similar to the ferret than the dog, respectively. Some V genes were assigned the same subgroup (e.g. "TRBV12") based on similarity to another species. Only tmpV15 and tmpV17 had their subgroup reassigned (it is now based on their position in the locus), because they had less than 75% similarity to another ferret TRB gene in the same subgroup, tmpV21 (72%) and tmpV4 (62.2%), respectively. The TRB J, D, and C genes were assigned subgroups based on their position in the ferret TRB locus.

5.2.4 Functionality

The functionality of the V, D, J, and C genes was determined according to the following criteria: a) identification of the L-PART1 upstream of V-EXON for V genes, b) determination of proper RS sequences (V-RS, 5' D-RS, 3' D-RS, 5' J-RS), c) determination of proper acceptor and donor splicing sites, d) absence of stop codons and frameshifts in the coding regions of the genes.

5.3 Results

5.3.1 Identification of expressed ferret TRB V, D, J, and C genes

Consensus sequences, formed by collapsing all sequence reads sharing the same UMI (here referred to as a "UMI-group"), were aligned against the ferret draft genome (MusPutFur1.0; Genbank GCF_000215625.1). Sequences that were the consensus of UMI-groups containing more than 1 sequence read, tended to target the region between the ferret MOXD2 (Ensembl ID ENSMPUG00000002940.1) and EPHB6 (ENSMPUG00000008478.1) genes (Supplemental figure S5.4). This result is in line with the TRB regions of a diverse set of mammals, i.e., cattle, human, mouse, and dog, which are (also) flanked by MOXD2 and EPHB6 (Antonacci *et al.*, 2014). Consensus sequences from UMI-groups of size 2 and up were aligned against the ferret genome. Next, we searched in the TRB locus for expressed V, D, J, or C genes in the regions with over 50× coverage (see Figure 5.1).

Using the HTS data we were able to identify 23 ferret TRBV genes that had over 50× coverage over part of their region. Reads aligning to the same region were collapsed, and the resulting consensus sequence was used to identify the various features of a V-GENE, such as the location of the V-RS, SPLICE-DONOR, and SPLICE-ACCEPTOR sites. In cases where the consensus was unclear, either due to low coverage and/or poor agreement, we used the the LIGMotif programme (Lane *et al.*, 2010) in combination with manual curation to identify the various gene features. If the consensus disagreed with the genome at a particular position, the consensus base is reported, but only if the consensus was well-supported (Figure 5.2). In total we found 34 ferret TRBV genes, of which we classified 16 as pseudogenes (47%), and one (TRBV22 (tmpV6)) as an ORF because it does not have the conserved Cysteine at position 104. Without this Cysteine, the TCRβ chain cannot make a disulfide bridge with the Cysteine at position 23, possibly leading to an improperly folded TCRβ chain.

Interestingly, in the functional V gene TRBV19 (tmpV8), we discovered a non-functional splice variant (see Figure S5.1 for an example), leading to a frameshift and the loss of the FR1 and CDR1 regions of the V-EXON. This splice variant occurs on average in more than 25% of the sequence reads that align to TRBV19 (tmpV8) in each of the four ferrets. To be able to detect this non-functional variant, sequence reads should extend nearly 200 bp into the V-EXON region. As reads typically do not extend this far into the V gene in repertoire sequencing experiments, this may lead to an overestimation of the

expression level of V genes having such non-functional splice variants. So repertoire sequencing experiments should use reads that are long enough to detect the splice variants, or perform a bias correction based on dedicated experiments quantifying the proportion of non-functional splice variants in the ferret population.

Similarly to the V genes, we manually curated the D, J, and C genes using the HTS data as a guide (Figures 5.3 and 5.4). The general structural organization of the ferret TRB locus (see Figure 5.1) is similar to that of the dog (see Ref. (Mineccia *et al.*, 2012) for the locus of *Canis lupus familiaris*). Like in the dog, the ferret TRB locus spans about 300Kb, which is much less than the 650Kb long locus of humans. The ferret TRB locus contains a region of V genes followed by two D-J-C clusters. Both D-J-C clusters span about 7Kb and consist of one D gene, six J genes, followed by a C gene. About 11Kb downstream of the second D-J-C cluster there is a TRBV gene (TRBV30 (tmpV1)) with an inverted transcriptional orientation. In the ferret, we classified TRBV30 as a pseudogene because of an improper DONOR-SPLICE (due to nucleotide deletion: "TAG-tggt" instead of "TAGgtggt") and a too short V-SPACER (21bp instead 23bp). Instead, the homologous TRBV30 in both human and dog is functional.

The ferret TRBD1 and TRBD2 genes are 12 bp and 15 bp long, respectively, and both genes are productively read in all 3 coding phases. The ferret J genes are between 44-53 bp long and conserve the FGXG motif, required for a functional J gene. The only exception is TRBJ1-4 (tmpJ9), which has an unusual motif ("FASG"; Figure 5.3b), identical to the homologous gene, TRBJ1-4, in the dog. The TRBJ2-5 (tmpJ2) is classified as a pseudogene primarily due to an improper J-HEPTAMER, as the "gtg" at the 3' end, essential for RS recognition (Mineccia *et al.*, 2012), has mutated (Figure 5.3b). Although TRBJ1-3 (tmpJ10) is also a pseudogene (because of a stop codon), it can still lead to functional transcripts by VDJ recombination, as it is used in about 3% of the TCRB clonotypes (Figure 5.6).

Like other mammalian species such as human, mouse, dog, and rabbit (Antonacci *et al.*, 2014), the ferret TRBC genes consist of 4 exons each (Figure 5.4). The TRBC genes of the ferret are identical to each other for the first 2 exons (EX1 and EX2), and differ by only 2 nt in EX3 and by 2 nt in EX4. The FG loop is one amino acid longer than the longest TRBC FG loop described by IMGT (Lefranc *et al.*, 2005). We extended the numbering of the FG loop to accommodate the additional amino acid (Figure 5.4). Both TRBC genes appear to be functional, having proper acceptor and donor splice sites for each exon, not containing any stop codon or frameshifts.

5.3.2 Phylogenetic analysis of the ferret TRBV and TRBJ genes

We aligned the ferret TRBV (V-REGION) and TRBJ (J-REGION) amino acid sequences to human and dog TRBV and TRBJ sequences, and constructed a phylogenetic tree using a maximum likelihood approach (Guindon and Gascuel, 2003) (Figure 5.5). We included only the ferret V and J genes that were functional, or were pseudogenes that could lead to functional transcripts. All ferret TRBV genes cluster with the closest dog TRBV gene that is in the same subgroup, except for the ferret TRBV8 and TRBV23 genes. The latter two had less than 75% identity to a similar gene (TRBV5 and TRBV26, respectively) in the same monophyletic group, and therefore TRBV8 and TRBV23 were named according to their position in the TRB locus. Most ferret TRBJ genes are more closely related to the dog TRBJ genes than to the human TRBJ genes, except for TRBJ1-5 and TRBJ2-6, which are closer to human TRBJ1-5 and TRBJ2-7, respectively.

5.3.3 Analysis of the ferret TRBV and TRBJ usage

After identifying the ferret TRBV and TRBJ genes, we used the RTCR pipeline to annotate and error correct the sequence reads. Overall the TRBV and TRBJ usage is very similar among the four ferrets (Figure 5.6). Interestingly, TRBV8, classified as a pseudogene due to a stop codon at the 3' end of the CDR3, is the most common TRBV gene in the repertoire. Similarly, the pseudogene TRBJ1-3 also produces functional transcripts (in about 3% of all transcripts; Figure 5.6).

5.3.4 Discussion

We combined HTS and genome analysis to describe the (expressed) TR genes in the TRB locus of the ferret. The genomic organization of the ferret TRB locus is very similar to that described in other mammals such as human, mouse, dog, and rabbit (Antonacci *et al.*, 2014; Mineccia *et al.*, 2012): the locus is flanked by MOXD2 and EPHB6 at the 5' and 3' ends, respectively, and consists of a library of V genes followed by two D-J-C clusters, followed by a V gene, which is non-functional in the ferret, with an inverted transcriptional orientation. Thus, the ferret confirms the strong organizational conservation of mammalian TRB loci. The ferret and dog TRB loci are closely related, because the ferret and the dog are in the same mammalian order, the Carnivora. Like in the dog, the ferret TRB locus is relatively small (300Kb) containing between 30 to 40

(35) TRBV genes of which nearly half (17) are pseudogenes. The ferret also expresses TRBV and TRBJ pseudogenes that nonetheless lead to functional transcripts, because their stop codons are deleted during VDJ recombination.

As previously described (Mineccia *et al.*, 2012), the CDR3 length distribution is highly conserved, which in this study is also confirmed as the ferret and the human have nearly identical CDR3 length distributions (Supplemental Figure S5.2). Despite the relatively low number of V genes in the TRB locus of the ferret, the ferret repertoire is highly diverse as there is hardly any overlap in TCRβ chains between the ferrets (Supplemental Figure S5.3). Within ferrets the TCRβ chain overlap between samples is about 50%, probably reflecting the presence of memory clonotypes in the blood that have expanded due to antigen stimulation.

Although the ferret is an important animal model in research on respiratory infections, its adaptive immune responses were until now poorly characterized (Enkirch and von Messling, 2015). Our characterization of expressed TRB genes in the ferret, were required to form a starting point for detailed analyses of the cellular immune responses of ferrets in health and disease.

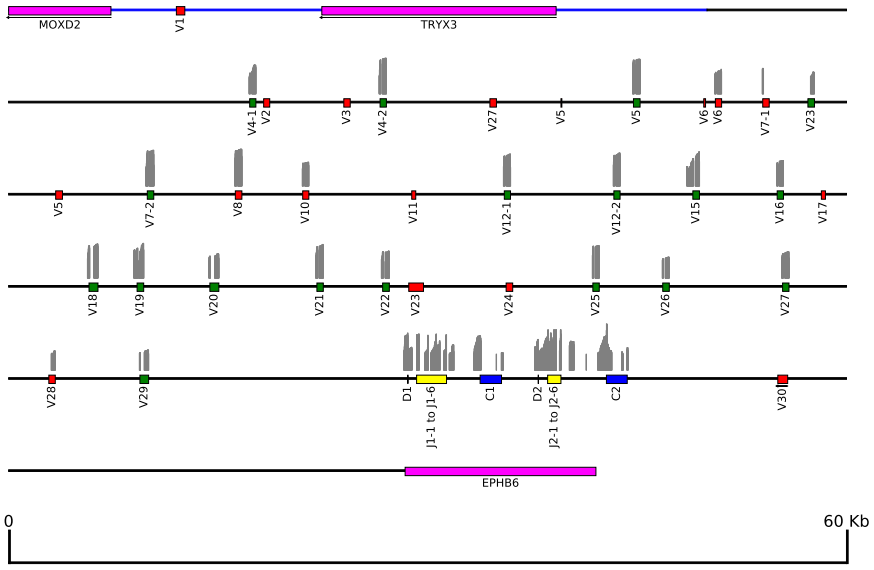


Figure 5.1. Schematic representation of the genomic organization of V, D, J, and C genes in the ferret TRB locus. Scaffolds GL896904.1 and GL897291.1 are shown as blue and black horizontal lines, respectively. Boxes are to scale, and show the following regions for the various genes: V genes (green; red for pseudogenes), from the 5' start of L-PART1 to the 3' end of V-REGION; for D genes (black), the D-region; for J genes (yellow), from 5' start of the J-REGION of TRBJ1-1 until 3' end of the J-REGION of TRBJ1-6 and similarly for TRBJ2; for C genes (blue), from 5' start of EX1 until 3' end of EX4. A few non-TRB genes are included: the canonical start and end genes of the TRB locus, MOXD2 and EPHB6, respectively, and TRYX3. Genomic coverage from the HTS (grey bars; bar width is 10bp) is shown on a log10 scale, excluding bars that have less than 50× coverage on average.

```
                                                              FR1-IMGT                CDR1-IMGT          FR2-IMGT            CDR2-IMGT                     FR3-IMGT                         CDR3-IMGT
                                                               (1-26)                  (27-38)            (39-55)             (56-65)                      (66-104)                         (105-115)
                                                         A                B              BC             C         C'            C'C"          C"          D            E             F          FG
                                                       (1-15)          (16-26)         (27-38)       (39-46)   (47-55)        (56-65)       (66-74)     (75-84)      (85-96)       (97-104)    (105-115)
                                                   --------------> ---------->                     -------> -------->                    ---------> ----------> ------------> ------->
gene name                 L-REGION                 1        10   15 16     23 26  27          38   3941  46 47      55   56       65   66      74 75   80 84 85  89     96 97    104  105 111
                                                   |........|....| |......|..|   |..........|    |.|....| |.......|    |........|    |.......| |....|...| |...|......| |......|   |.....|....
TRBV1 (tmpV41)            MGPPSILWVWVLVATMSTGAT    ASLVEQRPRWVPVAC GQAKTLHCILR  NSQ......YPW    MSWYQQDL QGSLQVLAT    LRS....PGD    EEVVSVP.G ADYRVRRVNS TELRLHVANMTQ G..LYCTC   ST.........  [6.6.2]
TRBV4-1 (tmpV24)          MGCRLLCWVVLCMLWAVSTGN    ..GVTQTPKHLVSGT GRSVTLKCKQH  LGH.......NA    MYWYKQSV QKPPKLMLA    YSY....KDL    LQNETAS.S RFLPDRS.NN SQLDLQIKALEP EDSALYLC   ASSK.......  [5.6.4]
TRBV4-2 (tmpV22)          MGCRLPYWVALCLLGAVSM      DPGVIQTPKHLVTGT TRSVTLKCKQH  LGH.......NA    MYWYKQSV QKPPKLMLA    YNY....KNL    FENATAS.S RFLPDRS.NN SQLDLQIKALEP EDSALYLC   ASSK.......  [5.6.4]
TRBV5 (tmpV21)            MGSRLLCCVALCLLGAGPVA     .SGVTQTPRHLIKTR GQKVTLRCSFV  SGH.......LS    VYWYQQVL GQGPRFLIQ    YYD....RKE    RDKGDMP.E RFSARQF.SN SSSQLDLDLLEL EDSALYLC   ASSE.......  [5.6.4]
TRBV7-2 (tmpV16)          MDPRLLCWVTLCLLGQGHTG     .PGVSQFPRHRVTER GQNVTLGCDPI  SGH.......TV    LYWYRQTT GKGLEFLVY    FQN....TDA    VDKSGMTND RLSVQRS.KG TNSTLQIQRAEQ GDSAVYLC   ASSP.......  [5.6.4]
TRBV8 (tmpV15)   MRRPVLPGSARGTQLLCWVSLCLLGAGPA     GAEVIQTPRHLIQGS GGTAVLTCQPV  SGH.......NS    VRWYQQTL GQGPKFLIE    YFA....QAP    RDKGDMP.E RFSAQQF.SN YSSQLNLSLLEP GDSALYLC   ASS*.......  [5.6.4]
TRBV12-1 (tmpV13)         MGTLALCLLGVLCFLWVEFT     DAGVIQTPRHKVTKM GQEVTLKCQPI  SGH.......EG    LFWYSQTS VQGPKLLIS    FNN....EAP    IDDSGMPKE RFSAEIS.NK SLSTLKIKSTEP GDSASYLC   ASSI.......  [5.6.4]
TRBV12-2 (tmpV12)         MATRLLCCVALCLLGREFI      DAGVIQTPRHNVTKM GQEVTLKCQPI  SGH.......AA    LYWYSQTS VQGPKLLIY    FNN....QAP    IDDSGMPKE RFSAEIS.NK SLSTLKIKPTEP GDSATYLC   ASSV.......  [5.6.4]
TRBV15 (tmpV11)           MAPELLHCVAFCLLGAGLM      GAMVIQSSRYQVTRV GKPVNLSCSQN  MNH.......DS    MYWYQQKL NQAPKLLLY    YYD....KNF    NREAETS.D NFQASRP.NT SFCSLNILSPGL GDSGVYLC   ASSK.......  [5.6.4]
TRBV16 (tmpV10)           MCPVFICSLVLWLLNAGAL      HAEVTQTPGHLVKGK GQKAKMHCVPI  KGH.......SY    VDWYQQIP AKEFKFFIS    FQN....DAI    FDKTGMPEK RFSAACP.QN ASCSLEIKPTEL QDSAVYFC   ASSE.......  [5.6.4]
TRBV18 (tmpV9)            MGTRVLCWVVIWLLGAGLM      NAGVTQNPRHLVIRT GQEATLKCSPE  KGH.......VY    FYWYQQPQ GESLKFMVY    LQN...AHTV    .DDSGMSKE RFSTNIS.EE GQSILKIQPAEL GDSAVYFC   ASSL.......  [5.7.4]
TRBV19 (tmpV8)            MGNQVICCVALCLLRAGTT      DGGITQTPKYLLREE GREVTLKCEQD  FNH.......DY    MYWYRQDP GQGPRLIYY    SPL....KDD    VQRGDIP.E GYFGSRG.KK TIFSLAVTSTRK NHTALYLC   ASSR.......  [5.6.4]
TRBV20 (tmpV30)           MLLLLLLLGPSSGL           GARVSQDPSAVICKS GASVQIECRAV  DLQ......VQT    VFGYRQLP KQGLTLMAT    SNVG...SSA    THEPDFPKA KFPISHP.NL IFFSLTVTSASP ADSSLYLC   GA.........  [6.7.2]
TRBV21 (tmpV7)            MFFRLLCCMALFSWGAGSL      DPNVTQTPGYLVKGK GQKAKMECVPI  KGH.......SY    VFWYHRKL EEEFKFLVY    LQD....KDI    VDKIEGFDN QFSAECP.KN SPCTLEINSTEP GDSALYMC   ASNK.......  [5.6.4]
TRBV22 (tmpV6)            MSSWALCYVALCLLGTGP       THAEIYQMPAFLLTR AGR.AVTLECK  QNL.....RYNA    MYWYWQDP GQSLRLIYY    STV....EKD    VQRGDIT.E GYNVSRE.EK GLFPLTMNLAHT NQTGVYLC   SGSA.......  [7.6.4]
TRBV23 (tmpV17)           MNLNLFCVTLWLLGAGPM       EAEITQTPRYSIIQI GTKKTLECSQD  MNH.......FA    MFWYRQDP GQGLRLIYY    SSD....SPS    TTKGDVA.E GYSVSRK.EQ KFSPLTLESAST NQTSVYLC   ASSE.......  [5.6.4]
TRBV25 (tmpV5)            MAVRLLCYVALYLLEADLV      DAGITQTPRHCVSGT GKKIVLECSQT  MGY.......EN    MYWYRQDP GKALQLIHY    SYG....VNN    TENEELS.S GSTVSRL.RK EVFPLTLESASP SQTSLYLC   ASSE.......  [5.6.4]
TRBV26 (tmpV4)            MNNRILCCVVIYLISVGLKG     .AVVTQFPKHRIVGP GKELTLQCLQD  MNY.......VL    MYWYRQDP GFGLQLIYY    STG....ARN    FEKGDVPQG YRVSRKD..L QSFSLTLESAST NQTSVYLC   TSSE.......  [5.6.4]
TRBV27 (tmpV3)            MAFRLLGCVVLCLLGAGPV      EAGVTQTPRHFITRT GRQLTLYCSQD  MDH.......DV    MYWYRQDP GVGLKLIYF    SRN....VKF    IEVGDVP.D GYSVLRK.EK KDFPLTLKSTGT NQTSLYLC   ASSV.......  [5.6.4]
TRBV29 (tmpV2)            MLPCLLLLLGQGSVFST        ..LLSQEPRRDICQC GTSMTIQCETD  TQV.......SL    MYWYRQLP GQSLILIAT    ANQG...MEA    TYESGFTKE KFPISRP.TL TFSSLTVNNMSF EDSSFYLC   SAE........  [5.7.3]
```

Figure 5.2. Protein display of the ferret consensus TRBV genes, showing only functional genes and in-frame pseudogenes without a stop codon 5' of the CDR3. Alignment of the V-REGIONs, performed using DomainGapAlign (Ehrenmann and Lefranc, 2011; Ehrenmann *et al.*, 2010), is displayed according to IMGT unique numbering for V-REGION (Lefranc *et al.*, 2003), and the amino acid length of CDR-IMGT is indicated in square brackets. A mutation (dark gray, non-synonymous; light gray, synonymous) between the ferret genome and the transcripts was accepted if the mutation occurred in the transcripts at a frequency of 50% or more in a total of at least 50 transcripts.

(a)

gene name	5′D-NONAMER	5′D-SPACER	5′D-HEPTAMER	D-REGION	3′D-HEPTAMER	3′D-SPACER	3′D-NONAMER
	GGTTTTTGT	**********	CACTGTG		CACAGTG	**********	ACAAAAACC
TRBD1 (tmpD2)	gttttttgt	acaaagatgtaa	cattgtg	GGGACAGGGGGC G T G G G Q G D R G	cacggtg	attcaactctatgggaaagcttt	acaaaaacc
TRBD2 (tmpD1)	cattttgt	atcctgttgtaa	cattgtg	GGAGCTGGGGGGGGG G A G G G E L G G S W G G	cacgatg	actcaggtagaggggtgctttt	acaaaagc

(b)

gene name	J-NONAMER	J-SPACER	J-HEPTAMER	J-REGION	DONOR-SPLICE
	GGTTTTTGT	**********	CACTGTG		
TRBJ1-1 (tmpJ12)	gtttgtttc	cctgccccacat	cactgtg	TGAACACTGAACTTTTCTTTGGAGAAGGCACCAGGCTCACAGTTATAG N T E L F F G E G T R L T V I	gtaaga
TRBJ1-2 (tmpJ11)	catttgaga	gtggccgtatgc	tgatgtg	CTATGATTTTACCTTCGGCCCAGGGACCAAGCTGACGGTTGTCG Y D F T F G P G T K L T V V	gtaagg
TRBJ1-3 (tmpJ10)	ggttctgaa	gtggatctggga	ggctgtg	CTCTTGAGACACCCTCCATTTTGGGGAGGGGACCCGGCTCACTGTTGTAG S * D T L H F G E G T R L T V V	gtaagt
TRBJ1-4 (tmpJ9)	agttttcct	accgggctgcag	tgttgtg	TAACAATGAAAAACTGTATTTTGCAAGTGGAACCAAGCTGTCCGTCCTGG N N E K L Y F A S G T K L S V L	gtaagt
TRBJ1-5 (tmpJ8)	gggtttgtc	acaccctcgctg	tgctgtg	GAGCAACCAGGCCCAGCACTTTGGAGATGGTACTCAACTCTTGGTCCTCG S N Q A Q H F G D G T Q L L V L	gtaagg
TRBJ1-6 (tmpJ7)	ggttttacc	acggctgcctgc	agctgtg	TTCCTATAACTCGCCGCTCTACTTTGGGATGGGCACCAGGCTCACCGTGACAG S Y N S P L Y F G M G T R L T V T	gtatgg
TRBJ2-1 (tmpJ6)	aaattcttg	gcagcccctgac	cactgtg	CTCATACAGCGAGCAGCACTTTGGGCCAGGGACCCGGCTCACAGTACTAG S Y S E Q H F G P G T R L T V L	gtaaga
TRBJ2-2 (tmpJ5)	ggtttgtgc	ctggctccccag	agctgtg	CAAACACAGGACAGCTGTACTTTGGGGCCGGTTCCAAGCTGACGGTTCTGG N T G Q L Y F G A G S K L T V L	gtaagg
TRBJ2-3 (tmpJ4)	gattcttgt	gctggggctccg	ggccgtg	AGTGGAGAAACTCAGTATTTCGGCAGGGGCACCCGGCTGACGGTGCTAG S G E T Q Y F G R G T R L T V L	gtaagc
TRBJ2-4 (tmpJ3)	tgtttttgt	gctgagccaggg	ggctgtg	AGCCAAAATACCCAGTACTTCGGCGCGGGGACCCGGCTGACGGTTCTAG S Q N T Q Y F G A G T R L T V L	gtgagc
TRBJ2-5 (tmpJ2)	tctctcgga	gaggggcccgg	ggccctg	CTTTGCAGCCTCTGCCCTGTCCTTCGGGGCCGGCAGCCGGCTGACCGTCCTGG F A A S A L S F G A G S R L T V L	gtgggt
TRBJ2-6 (tmpJ1)	ggtttgcgc	gcgggtctgggc	ctctgtg	TTCCTACGAGCAGTATTTCGGCTCCGGCACCAGGCTCACGGTCATAG S Y E Q Y F G S G T R L T V I	gtgaga

Figure 5.3. Nucleotide and deduced amino acide sequences of the ferret TRBD (a) and TRBJ (b) genes. TRBJ1-3 and TRBJ2-5 are classified as pseudogenes because of a stop codon and improper J-RS, respectively.

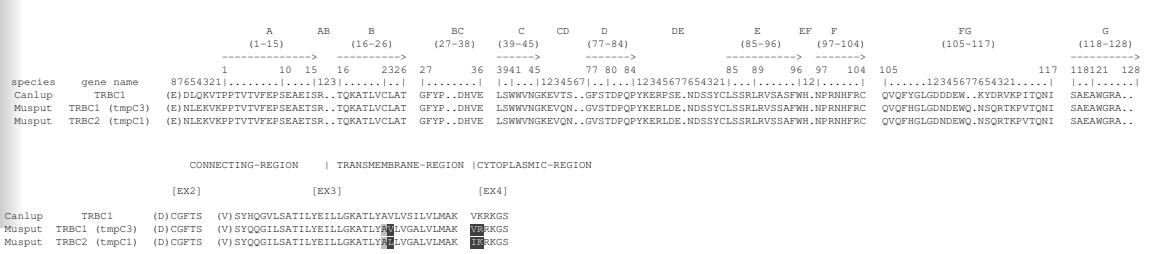
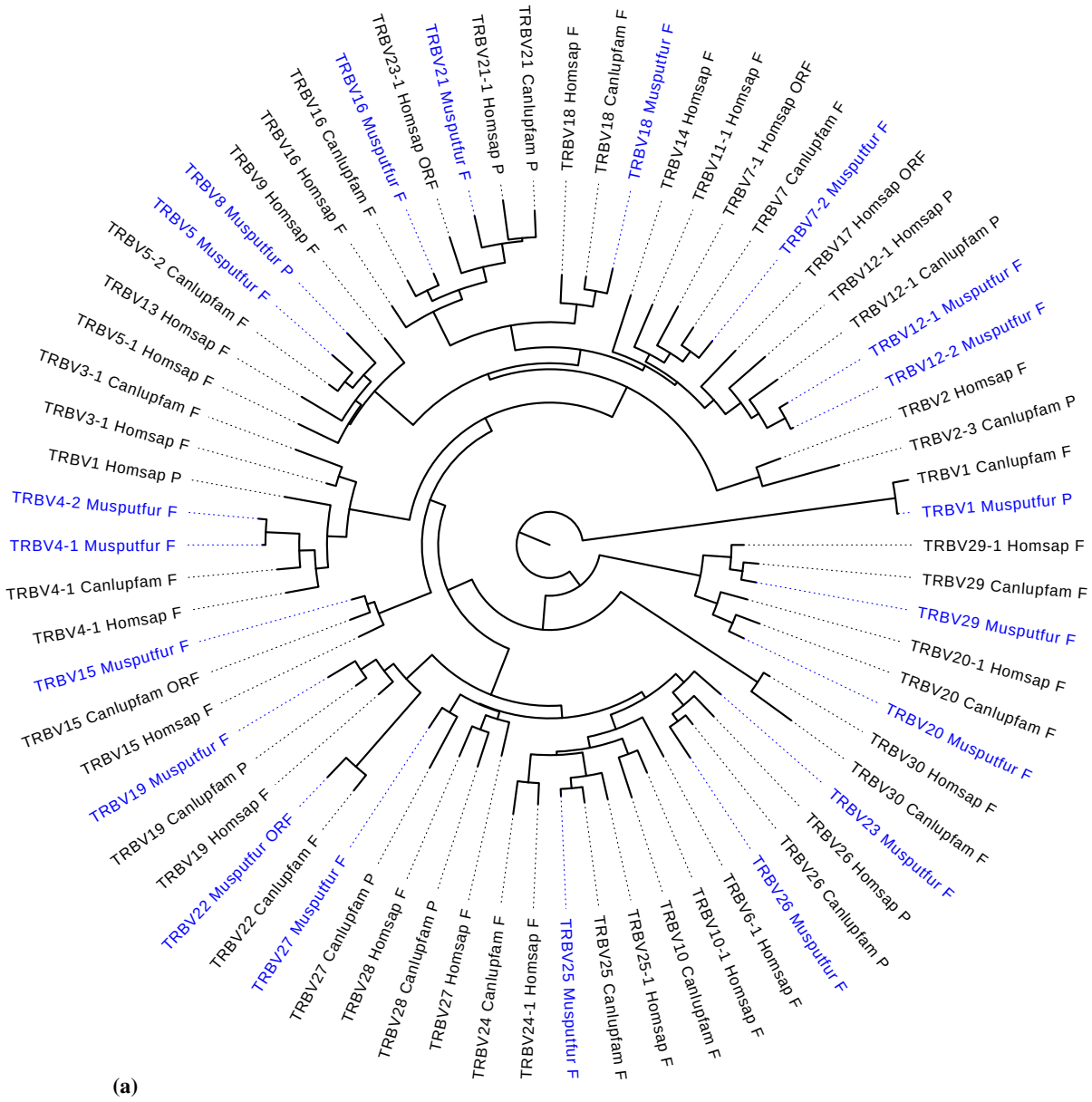
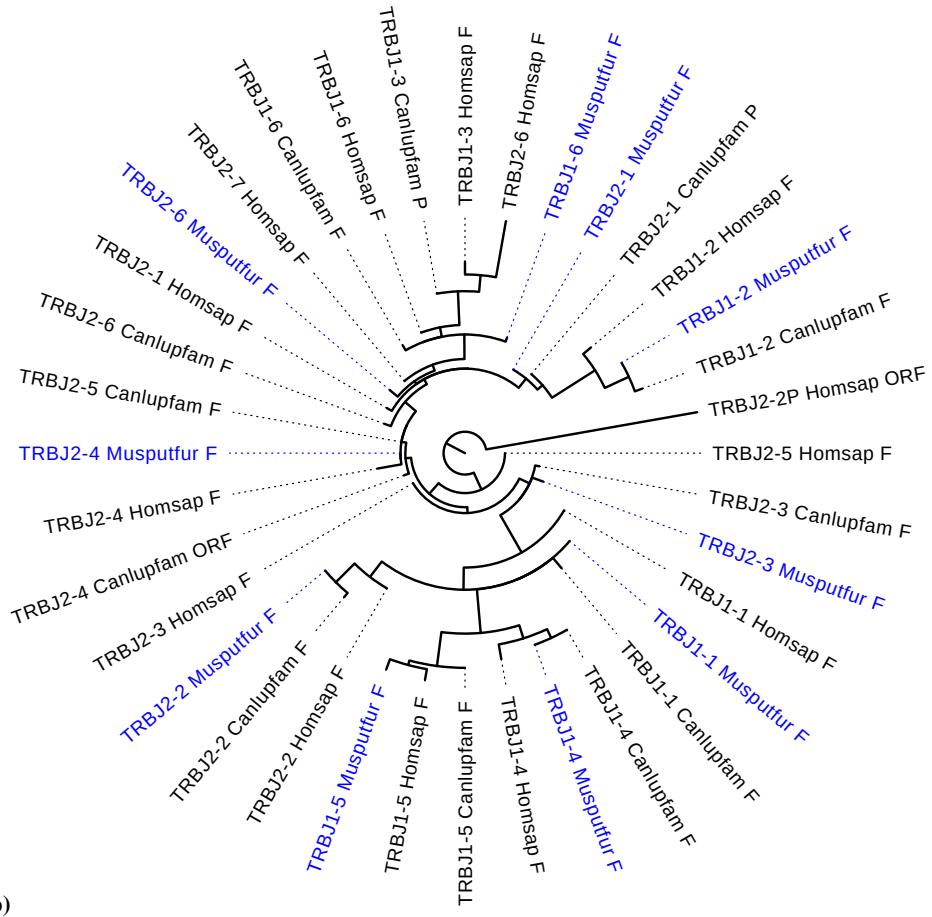


Figure 5.4. Protein display of the ferret TRBC genes, including dog TRBC1, according to the IMGT unique numbering for C-DOMAIN (Lefranc *et al.*, 2005). Nucleotide differences between TRBC1 and TRBC2 exons of the ferret are shown in light gray and dark gray, for synonymous and non-synonymous substitutions, respectively.


TRBV21 Canlupfam P
TRBV18 Homsap F
TRBV18 Canlupfam F
TRBV18 Musputfur F
TRBV14 Homsap F
TRBV11-1 Homsap F
TRBV7-1 Homsap ORF
TRBV7 Canlupfam F
TRBV7-2 Musputfur F
TRBV17 Homsap ORF
TRBV12-1 Homsap P
TRBV12-1 Canlupfam P
TRBV12-1 Musputfur F
TRBV12-2 Musputfur F
TRBV2 Homsap F
TRBV2-3 Canlupfam P
TRBV1 Canlupfam F
TRBV1 Musputfur P
TRBV29-1 Homsap F
TRBV29 Canlupfam F
TRBV29 Musputfur F
TRBV20-1 Homsap F
TRBV20 Canlupfam F
TRBV20 Musputfur F
TRBV30 Homsap F
TRBV30 Canlupfam F
TRBV23 Musputfur F
TRBV26 Homsap P
TRBV26 Canlupfam P
TRBV26 Musputfur F
TRBV6-1 Homsap F
TRBV10-1 Canlupfam F
TRBV10 Canlupfam F
TRBV25-1 Homsap F
TRBV25 Canlupfam F
TRBV25 Musputfur F
TRBV24-1 Homsap F
TRBV24 Canlupfam F
TRBV27 Homsap F
TRBV28 Canlupfam P
TRBV28 Homsap F
TRBV27 Canlupfam P
TRBV27 Musputfur F
TRBV22 Canlupfam F
TRBV22 Musputfur ORF
TRBV19 Homsap F
TRBV19 Canlupfam P
TRBV19 Musputfur F
TRBV15 Homsap F
TRBV15 Canlupfam ORF
TRBV15 Musputfur F
TRBV4-1 Homsap F
TRBV4-1 Canlupfam F
TRBV4-1 Musputfur F
TRBV4-2 Musputfur F
TRBV1 Homsap P
TRBV3-1 Homsap F
TRBV3-1 Canlupfam F
TRBV5-1 Homsap F
TRBV13 Homsap F
TRBV5-2 Canlupfam F
TRBV5 Musputfur F
TRBV8 Musputfur P
TRBV9 Homsap F
TRBV16 Homsap F
TRBV16 Canlupfam F
TRBV16 Musputfur F
TRBV23-1 Homsap ORF
TRBV21 Musputfur F
TRBV21-1 Homsap P

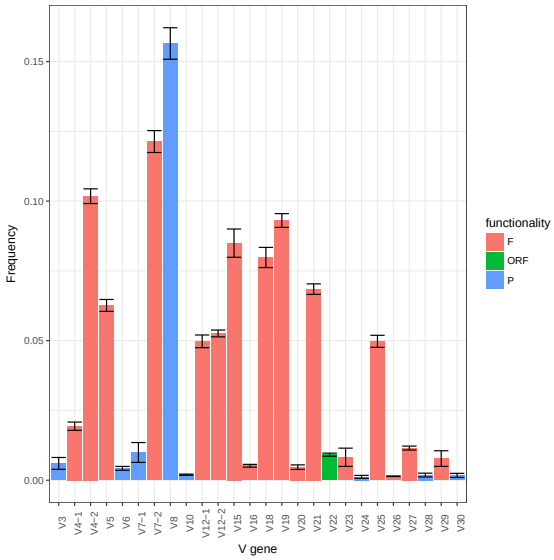
(a)

0.3

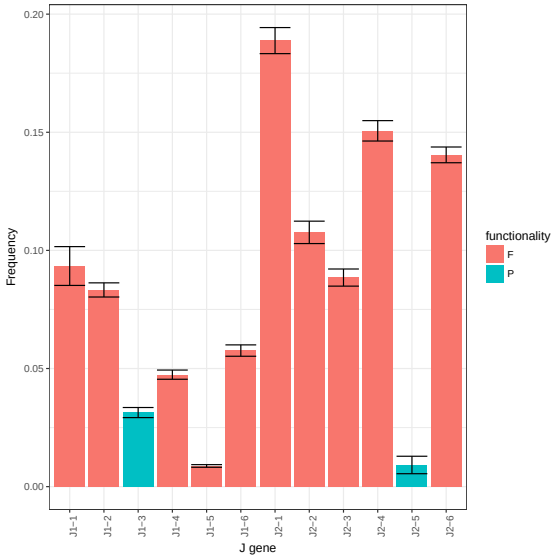


(b)

Figure 5.5. Phylogenetic trees of the dog, human, and ferret TRBV (a) and TRBJ (b) genes. Unrooted phylogenies were inferred from V-REGION (a) and J-REGION (b) amino acid sequences using the PhyML programme (Guindon and Gascuel, 2003). Ferret genes are indicated in blue, human and dog genes in black.



(a)



(b)

Figure 5.6. TRBV (a) and TRBJ (b) gene usage in the ferret. Bars indicate average fraction of distinct TCR sequences encoding a particular gene. Error bars indicate SEM across 16 HTS datasets (4 ferrets total, of which 3 ferrets are represented by 5 repertoire sequencing runs each).

5.4 Supplemental information

```
    UMI          L-PART1                          V-EXON (TRBV19 (tmpV8))                          J-REGION (TRBJ2-1 (tmpJ6))  EX1 (TRBC1 or TRBC2)
CTCATTTTTTCA  MGNQVICCVALCLLRA?  DPGQGPRLIYYSPLKDDVQRGDIPEGYFGSRGKKTIFSLAVTSTRKNHTALYLCA  AQGLRRG  EQHFGPGTRLTVL              ENLEKVKPPTVTVFEPSEAE
```

Figure S5.1. Protein sequence of a potential non-functional splice variant of TRBV19 (tmpV8). The displayed sequence is from ferret GR5, and is a consensus of 774 sequence reads sharing the UMI shown on the left. "?" denotes an incomplete codon (here a single guanine), resulting in a frameshift.

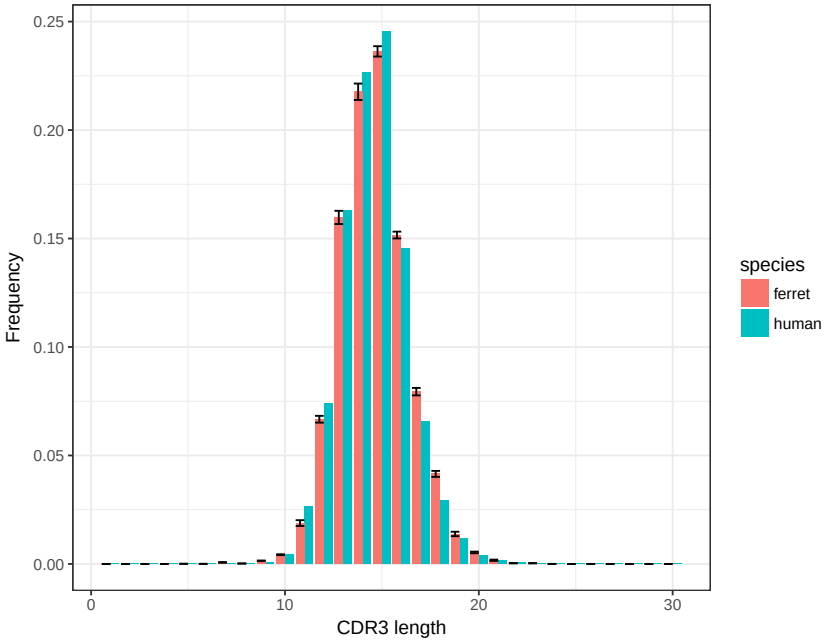


Figure S5.2. Comparison of TRBV CDR3 length distribution of the ferret and the human. (red), average frequency of CDR3 lengths 1 through 30 across all ferret HTS datasets (11 TCR sequences longer than 30AA are not shown), (green), CDR3 length frequencies of a single human adult male (unpublished data; 174753 CDR3 lengths total, 46 TCR sequences longer than 30AA are not shown).

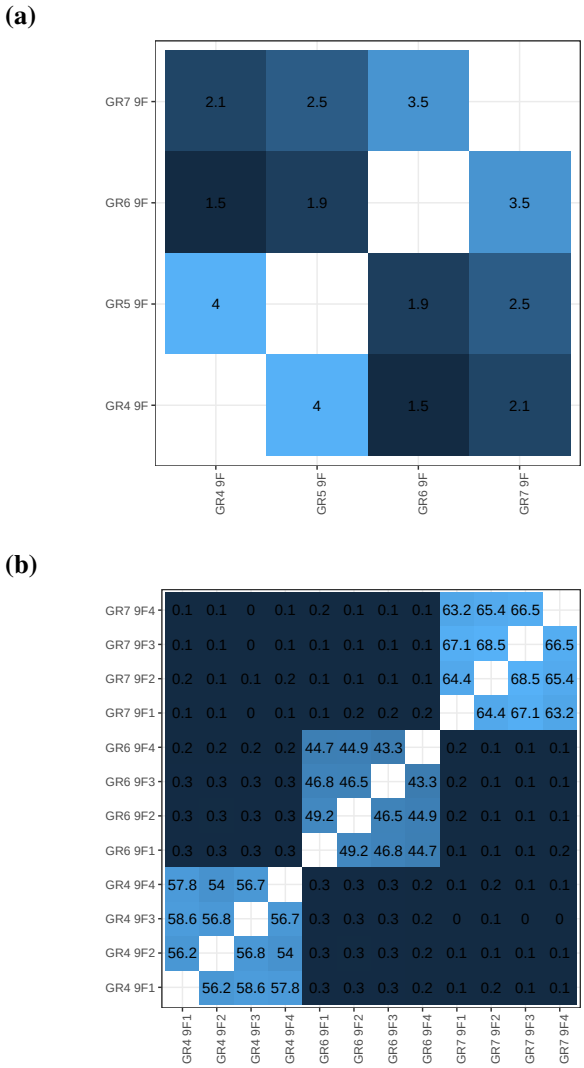


Figure S5.3. Percentage overlap of CDR3 amino acid sequences between the different HTS datasets. The overlap between two samples was calculated using the Jaccard index, i.e. the fraction of total distinct CDR3 sequences that are shared between the two samples. The HTS datasets with the 9F primer (a) contained an order of magnitude more CDR3 sequences than the other datasets (b).

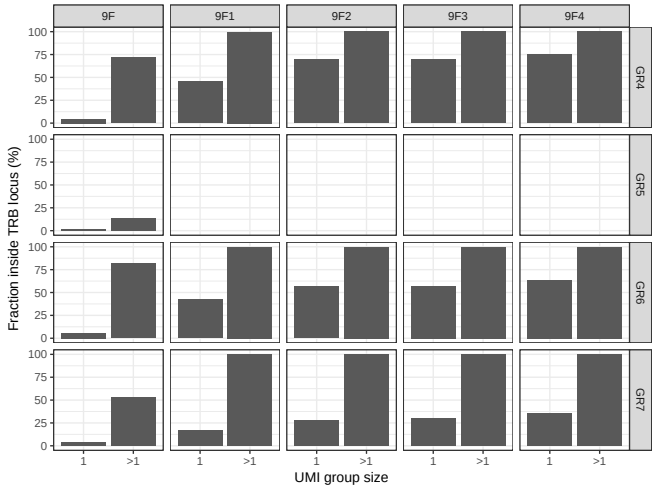


Figure S5.4. UMI corrected sequences from UMI groups containing more than one sequence read tend to align to the ferret TRB locus.

(a)

gene name	Functionality	Positions	Positions (L-PART1)
TRBV1	P	37945-37999 *	37395-37681 *
TRBV3	P	560476-560524	560064-560358
TRBV4-1	F	567216-567264	566820-567114
TRBV4-2	F	557879-557927	557474-557768
TRBV5	F	539733-539781	539306-539598
TRBV6	P	533867-533915	533468-533768
TRBV7-1	P	530475-530523	530058-530355
TRBV7-2	F	514549-514597	514133-514430
TRBV8	P	508215-508293	507823-508116
TRBV10	P	503423-503468	503028-503323
TRBV12-1	F	488979-489030	488586-488883
TRBV12-2	F	481136-481184	480738-481035
TRBV15	F	475480-475528	475055-475349
TRBV16	F	469441-469489	469037-469334
TRBV18	F	458724-458772	458129-458426
TRBV19	F	455280-455328	454854-455148
TRBV20	F	450086-450116	449473-449773
TRBV21	F	442403-442451	441992-442289
TRBV22	ORF	437695-437743	437263-437556
TRBV23	F	527243-527288	526817-527111
TRBV24	P	428852-428900	428428-428720
TRBV25	F	422641-422689	422229-422523
TRBV26	F	417637-417685	417214-417507
TRBV27	F	409053-409101	408643-408937
TRBV28	P	401601-401649	401186-401482
TRBV29	F	395092-395125	394502-394802
TRBV30	P	348736-348796	349151-349444

(b)

gene name	Functionality	Positions
TRBD1	F	375930-375941
TRBD2	F	366587-366601
TRBJ1-1	F	375264-375311
TRBJ1-2	F	375133-375176
TRBJ1-3	P	374480-374529
TRBJ1-4	F	373887-373936
TRBJ1-5	F	373641-373690
TRBJ1-6	F	373168-373220
TRBJ2-1	F	365878-365927
TRBJ2-2	F	365681-365731
TRBJ2-3	F	365442-365490
TRBJ2-4	F	365307-365355
TRBJ2-5	P	365198-365250
TRBJ2-6	F	364974-365020

(c)

gene name	Functionality	Exons	Positions
TRBC1	F	EX1	370373-370762
		EX2	369818-369835
		EX3	369562-369668
		EX4	369229-369246
TRBC2	F	EX1	361328-361717
		EX2	360773-360790
		EX3	360517-360623
		EX4	360227-360244

Figure S5.5. Description of the V (a), D and J (b), and C (c) genes in the ferret TRB locus. Genomic positions are relative to scaffold GL897291.1 or GL896904.1 (denoted with "*").

Chapter 6

Discussion

6.1 Discussion

In this thesis we applied bioinformatic methods, developed computational methods, and devised mathematical models to study various aspects of T cell repertoires. We will here discuss a few of the questions that arose in our studies, and finish with a brief outlook.

6.1.1 Analysis of high-throughput repertoire sequencing data

We developed a pipeline for the complete and accurate retrieval of TCR sequences from HTS data in Chapter 2, and using synthetic datasets we demonstrated that that RTCR outperformed other tools for the retrieval of TCR sequences from HTS data. A valid criticism is that the test data was synthetic and that real data may contain errors (such as chimeric sequences) that are not captured by the synthetic sequences. Unfortunately, there are no gold standard (synthetic or real) datasets for the validation of sequence error correction tools like RTCR (Heather *et al.*, 2017). Recently, Afzal *et al.* (2017) have compared ten state-of-the-art TCR analysis tools and found that RTCR was the most accurate tool across multiple *in silico* datasets having different error rates.

Certain aspects of the pipeline can definitely be improved. The first would be the generation of personalized germline reference datasets for each dataset. RTCR estimates the error rate in the HTS data using alignments with a germline reference of V and J sequences. If the data contains alleles not present in the dataset, then the error rate could be overestimated as every mismatch with the germline is considered an error. A second improvement would be the error correction of UMI sequences, because erroneous UMIs can artificially inflate clone-sizes. This is also one reason our pipeline was not used in Chapter 4, where we developed a separate algorithm for the detection and correction of artificially inflated clone-sizes due to erroneous UMIs. As an aside, note that clustering-based error correction is not perfect and in some cases can artificially inflate clone-sizes by merging the wrong sequences together. Since we were demonstrating the existence of large naïve clonotypes in Chapter 4, we removed this variable by not applying a clustering-based error correction like that of RTCR. Finally, a third improvement would be increasing the speed of annotation and error correction, as with the ever increasing size of HTS datasets, the application of our pipeline can become cumbersome.

6.1.2 Estimation of repertoire richness

The TCR repertoire is incredibly diverse, but the actual number of different TCRs ("richness") is an open question. There is considerable interest in measuring the richness of repertoires in both health and disease. However, sequencing a complete T cell repertoire is not feasible. Therefore, the number of distinct TCRs in the full repertoire has to be estimated from a small sample (typically 10^5 or 10^6 T cells out of approximately $10^{11} - 10^{12}$ (Arstila *et al.*, 1999; Clark *et al.*, 1999) T cells). This is analogous to the "unseen species problem" in ecology, in which the question is that if one has sampled n individuals (T cells), and observed S species (clonotypes), how many new species, U, are expected to be observed if an additional m individuals were sampled. Total repertoire richness is then estimated by $\lim_{m\to\infty} S + \hat{U}$, with $\hat{U}$ an estimator of U.

Compared to estimation of species richness from ecological data, estimation of repertoire richness from HTS data faces particular challenges. First, current HTS techniques require PCR amplification, which distorts the relationship between the number of TCR sequences observed and the number of T cells ("individuals") that were sampled, in a manner that can change from one experiment to the next (Best *et al.*, 2015; Laydon *et al.*, 2015). In combination with deeper sequencing, the amplification can lead to a "false saturation" effect, where additional sequence reads increases the size but not the number of observed clonotypes, which reflects that the sample is exhaustively sequenced, but not the whole repertoire ("assembly") (Laydon *et al.*, 2015; Warren *et al.*, 2011). Second, PCR- and sequencing errors can artificially inflate the number of clonotypes ("species") observed, and further distort clone-sizes (Bolotin *et al.*, 2012; Laydon *et al.*, 2015). Third, nowadays, only one of the two chains of the TCR heterodimer is typically sequenced, obscuring an unknown part of repertoire diversity stemming from the other chain.

Richness estimators tend to be particularly sensitive to the number of rare clonotypes, which are most affected by artificial diversity due to PCR- and sequencing errors, emphasizing the need to correct the observed clone-sizes, either directly (Chiu and Chao, 2016), or by correcting the sequences themselves. An example of the former is a method by Chiu and Chao (2016), which replaces the observed number of singletons, f_1, with an estimate of the number of genuine singletons in the data, based on f_2, f_3, and f_4, before estimating the diversity using a richness estimator such as Chao1 (Chao, 1984). For the error correction of the sequences we do not have a gold standard, but error correction strategies that discard rare clonotypes based on arbitrary thresholds may lead to the loss of genuine rare clonotypes, and as a result, lead to a huge underestimation of repertoire diversity. One essential technique for error correction, is the application of Unique Mo-

lecular Identifiers (UMIs; Chapter 1) (Kivioja *et al.*, 2011), which can markedly reduce PCR amplification bias and enable correction of many PCR- and sequencing errors. Remaining errors in the data can be handled by additional error correction strategies such as those employed by RTCR. Finally, there are alternatives to HTS, such as single cell TCR sequencing (scTCRseq) (Papalexi and Satija, 2017; Redmond *et al.*, 2016), which allows enumeration of the individual T cells and the associated TCR heterodimer. However, the maximum number of cells per experiment is currently no more than 10^4 (with Drop-Seq (Papalexi and Satija, 2017)), which is 2 to 3 orders of magnitude less than is possible in HTS samples.

Even without errors in the HTS data, there are two additional properties of repertoire sequencing that can confound the estimation of species richness. First, the samples are tiny (10^6) in comparison to the full repertoire (10^{12} T cells), and second, the repertoire is heavy-tailed (Mora and Walczak, 2016) (i.e., rare clonotypes are common), meaning that the majority of species are small and are unlikely to be sampled. Mora and Walczak (2016) illustrated, using a power-law distribution, that most estimators (such as Chao1 and Chao2 (Chao, 1987)) are not designed for heavy-tailed distributions, and tend to underestimate the true richness of the repertoire. Interestingly, Orlitsky *et al.* (2016) recently proved that without any assumptions on the shape of the underlying species (clone-size) distribution, no guarantees on estimation performance can be given for any estimator, when extrapolating from a sample of n individuals to more than $m = n\log(n)$ individuals. Extrapolation to the number of T cells in the human body ($n\log(n) \approx 10^{12}$) would require a minimum sample size of about 4% of the T cells. With about 2-3% of T cells in the blood (Di Rosa and Pabst, 2005), this translates to sampling more than the entire blood volume. This example illustrates that HTS samples are so small in comparison to the whole repertoire that assumptions about the form of the clone-size distribution are required for proper richness estimation. Even if tiny samples were theoretically sufficient for extrapolation to the whole repertoire, the rich biology of T cells migrating between blood and tissue, differentiating, expanding, and dying, should probably be taken into account for proper richness estimation of T cell repertoires. Currently, estimates of naïve repertoire richness vary between 10^8 (TCRB only, based on Chao2) (Qi *et al.*, 2014), and 10^{11} clonotypes (based on modeling of clonal dynamics (Lythe *et al.*, 2016), or assuming a power-law distribution (Mora and Walczak, 2016)).

Repertoire richness does not directly reflect the number of different antigens that a repertoire can recognize. An interesting recent development, is the study of the association between a TCR and the antigens it can recognize. Progress into this direction has been made, for example by Thomas *et al.* (2014), who compared repertoires based on frequen-

cies of stretches of 3 or 4 amino acids, and showed repertoires of different mice clustered together based on time after an immunization and immunization status. Additionallly, Dash *et al.* (2017) and Glanville *et al.* (2017) showed TCR sequences recognizing the same epitope tend to have similar CDR3 amino acid sequences.

6.1.3 Why study the TRB locus of the Ferret?

To gain deeper insight into the dynamics of T cell repertoires during infection, we intend to longitudinally follow viral respiratory infections in the Ferret using high-throughput sequencing of its T cell repertoire. We chose the Ferret, because the Ferret is an important animal model for studying human respiratory infections and because Ferrets can be sampled repeatedly. Since there was no germline reference of the TRB locus available for the Ferret, we generated this germline reference ourselves by studying the TRB locus (see Chapter 5). This will enable us to study T cell repertoire dynamics during, for example, an influenza infection in the Ferret, in the near future.

6.1.4 Why are there large naïve clonotypes?

In Chapter 4 we studied the clone-sizes of naïve T (Tn) cells, and observed thousands of naïve clonotypes in more than one (small) blood sample, demonstrating that a small fraction, but a large number, of the naïve clonotypes in the samples are large in the full repertoire. The question we tried to answer was how naïve clonotypes became so large. Since we observed a positive correlation between the naïve clone-size and the probability with which the VDJ recombination process produces the associated TCRB sequence (P_{gen}), we developed a simple mathematical model that describes a random birth death process. In this model some clonotypes are made more often by the thymus using the P_{gen} values determined in previous studies (Marcou *et al.*, 2017; Murugan *et al.*, 2012). In line with our observations, the model predicted the existence of numerous naïve clonotypes large enough to be observed in multiple (small) samples. Thus some naïve T cell clonotypes are large because they are "easy to make". Let us now consider the potential functionality of having large naïve clonotypes in the repertoire.

The prenatal T cell repertoire is initially populated with near-germline T cells as the enzyme responsible for random insertions, Terminal deoxynucleotidyl-transferase (TdT), is suppressed (George and Schroeder, 1992). Recently, Pogorelyy *et al.* (2017) showed that zero-insertion prenatal T cells are long-lived (37 year lifespan). Such long-lived

prenatal T cells might explain the large clonotypes we observed (Figure 4.4A) in excess of what our model predicted. Pogorelyy *et al.* (2017) also observed large clonotypes with low P_{gen} in cord-blood, but were unable to determine the lifespan of these clonotypes. If these large low P_{gen} cord-blood clonotypes are also long-lived, they might explain the presence of low P_{gen} clonotypes among the largest clonotypes we observed (Figure 4.4).

In mice, neonatal near-germline T cells have been shown to be more cross-reactive than their adult (less germline-like) counterparts (Gavin and Bevan, 1995). If the same is true for humans, the prenatal near-germline T cells in early life may constitute a "minimal" repertoire with broad recognition of antigens (Venturi *et al.*, 2013). This suggests that large naïve clonotypes (which tend to be germline-like) have a function of providing early protection. This could also explain their function in adulthood (early in the immune response).

Similar to how some naïve clonotypes have higher frequency within an individual, some naïve clonotypes have higher prevalence in the population. The higher prevalence appears to be governed by the same mechanisms, namely convergent recombination (i.e. the variety of recombination events generating the same TCR nucleotide or amino acid sequence) and higher P_{gen} (recombinatorial bias) (Elhanati *et al.*, 2014; Ndifon *et al.*, 2012; Quigley *et al.*, 2010; Venturi *et al.*, 2011; Warren *et al.*, 2011). This puts large naïve clonotypes on a spectrum of TCR sharing, from the universal TCRs such as invariant Natural Killer T (NKT; Type 1 only) and Mucosal Associated Invariant T (MAIT) TCRα, to the sporadic rarely shared small clonotypes (in the naïve repertoire) (Venturi *et al.*, 2013). One hypothesis could be that large naïve T cell clonotypes, at least those that are prevalent in the population, have specificities for particular antigens, similar to NKT and MAIT cells. Both NKT (Type 1 only) and MAIT cells have a semi-invariant evolutionarily conserved TCRα chain (Kjer-Nielsen *et al.*, 2006; Le Bourhis *et al.*, 2010), of which some α chains are extremely prevalent in the population (Venturi *et al.*, 2013). The high prevalence of some near-germline large naïve clonotypes result from evolutionary pressure on germline sequences shaping the specificities of some clonotypes, e.g. by enabling recognition of common pathogens. Since the precursor frequency of these clonotypes is high, they might be able to respond relatively quickly. One important caveat of this hypothesis is that, in two different individuals challenged with the same pathogen, the T cells of these individuals can be presented with completely different antigens by the MHC, depending on their HLA-types. The HLA is highly polymorphic, for example, the probability of two random individuals having the same HLA-A and HLA-B haplotype is only 0.001% (Calis, 2012). This probability is different for MAIT and NKT cells, as these innate-like subsets recognize (nearly) nonpolymorphic receptors, CD1 and MR1,

respectively, and the antigens presented to them can generally be expected to be the same between two unrelated individuals.

Another possibility is that most large naïve clonotypes are not genuine naïve clonotypes. There are other T cell subsets, such as regulatory T cells (Tregs), and stem-cell memory T cells (Tscm) that express common markers used to identify naïve T cells (such as CD45RA, CCR7, CD62L, and CD27) (Caramalho *et al.*, 2015; Gattinoni *et al.*, 2017; Hsieh *et al.*, 2012). Tscm cells are expanded clonotypes, and recently it was shown by Miyama *et al.* (2017), clonotypes specific for a cytomegalovirus peptide, were highly prevalent among HLA-matched individuals (and also occurred in a non-matched individual) and were abundant in the Tscm subset. This suggests the Tscm subset might be partly responsible for observed overlaps between supposedly naïve repertoires of different individuals, casting doubt on the existence of large genuine naïve clonotypes that are prevalent in the population. Finally, it could be that there is no special function for large naïve clonotypes in the adult repertoire, and their large clone-sizes are the result of convergent recombination and biases in the recombination machinery. Their observed self-reactivity (Madi *et al.*, 2014) might be a reflection of their promiscuity, a remnant of their function in providing broad recognition in early childhood.

6.1.5 Outlook

During our studies, sequencing technology improved at a fast pace (Goodwin *et al.*, 2016). For example, application of UMIs is becoming routine in Rep-Seq, and new technologies are entering the field, such as single cell sequencing (Papalexi and Satija, 2017; Redmond *et al.*, 2016) and single molecule sequencing (such as the minION (Lu *et al.*, 2016)). Together with the ever decreasing cost of sequencing, personalized medicine based on repertoire sequencing is on the horizon.

With improved sequencing techniques and flow cytometry, more repertoires are being sequenced and new T cell subsets are defined. A major open question is the relationship between these subsets. Do these subsets represent terminally differentiated states, or are they transiently induced by environmental cues (or anything in-between)? Additionally, the large number of TCR sequences provides new opportunities. For example, databases are being created that contain TCR sequences and their associated epitopes, HLA-restriction, and disease involvement (Shugay *et al.*, 2017; Tickotsky *et al.*, 2017). This might also help to build classifiers to shed light on another key question in immunology: can we predict the antigens a TCR will bind to from its amino acid sequence?

Bibliography

Afzal, S., Gil-Farina, I., Gabriel, R., Ahmad, S., von Kalle, C., Schmidt, M., and Fronza, R. 2017. Systematic comparative study of computational methods for T-cell receptor sequencing data analysis. *Briefings in Bioinformatics*, page bbx111. (Cited on page 96.)

Alamyar, E., Duroux, P., Lefranc, M.-P., and Giudicelli, V. 2012a. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In *Immunogenetics*, pages 569–604. Springer. (Cited on page 11.)

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., Lefranc, M.-P., *et al.* 2012b. 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 res*, 8(1): 26. (Cited on page 5.)

Antonacci, R., Giannico, F., Ciccarese, S., and Massari, S. 2014. Genomic characteristics of the T cell receptor (TRB) locus in the rabbit (oryctolagus cuniculus) revealed by comparative and phylogenetic analyses. *Immunogenetics*, 66: 255–266. (Cited on pages 77, 80, 81, and 82.)

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. 1999. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, 286(5441): 958–961. (Cited on pages 11, 51, and 97.)

Bains, I., Antia, R., Callard, R., and Yates, A. J. 2009. Quantifying the development of the peripheral naive CD4+ T-cell pool in humans. *Blood*, 113: 5480–5487. (Cited on page 52.)

Bassing, C. H., Swat, W., and Alt, F. W. 2002. The mechanism and regulation of chromosomal V(D)J recombination. *Cell*, 109(2): S45–S55. (Cited on page 11.)

Baum, P. D., Venturi, V., and Price, D. A. 2012. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *European journal of immunology*, 42(11): 2834–2839. (Cited on pages 11, 13, and 21.)

Belser, J. A., Katz, J. M., and Tumpey, T. M. 2011. The ferret as a model organism to study influenza A virus infection. *Disease models & mechanisms*, 4: 575–579. (Cited on page 77.)

Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. 2012. Rep-seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135: 183–191. (Cited on pages 4 and 5.)

Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J., and Chain, B. 2014. Sequence and primer independent stochastic heterogeneity in PCR amplification efficiency revealed

by single molecule barcoding. *bioRxiv*, page 011411. (Cited on page 19.)

Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J., and Chain, B. 2015. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific reports*, 5: 14629. (Cited on pages 41, 53, and 97.)

Bodewes, R., Kreijtz, J. H. C. M., van Amerongen, G., Geelhoed-Mieras, M. M., Verburgh, R. J., Heldens, J. G. M., Bedwell, J., van den Brand, J. M. A., Kuiken, T., van Baalen, C. A., Fouchier, R. A. M., Osterhaus, A. D. M. E., and Rimmelzwaan, G. F. 2010. A single immunization with CoVaccine HT-adjuvanted H5N1 influenza virus vaccine induces protective cellular and humoral immune responses in ferrets. *Journal of virology*, 84: 7943–7952. (Cited on page 78.)

Bolotin, D. A., Mamedov, I. Z., Britanova, O. V., Zvyagin, I. V., Shagin, D., Ustyugova, S. V., Turchaninova, M. A., Lukyanov, S., Lebedev, Y. B., and Chudakov, D. M. 2012. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *European journal of immunology*, 42: 3073–3083. (Cited on pages 5, 6, 11, 13, 21, and 97.)

Bolotin, D. A., Shugay, M., Mamedov, I. Z., Putintseva, E. V., Turchaninova, M. A., Zvyagin, I. V., Britanova, O. V., and Chudakov, D. M. 2013. MiTCR: software for T-cell receptor sequencing data analysis. *Nature methods*, 10(9): 813–814. (Cited on pages 12, 19, and 21.)

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., and Chudakov, D. M. 2015. MiXCR: software for comprehensive adaptive immunity profiling. *Nature methods*, 12: 380–381. (Cited on pages 6, 12, and 21.)

Brady, B. L., Steinel, N. C., and Bassing, C. H. 2010. Antigen receptor allelic exclusion: an update and reappraisal. *The Journal of Immunology*, 185(7): 3801–3808. (Cited on page 3.)

Calis, J. 2012. *Bound to be an Epitope: Determinants of the T-cell response to MHC-I presented peptides*. Ph.D. thesis, University Utrecht. (Cited on page 100.)

Calis, J. J. A. and Rosenberg, B. R. 2014. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in immunology*, 35: 581–590. (Cited on pages 11 and 55.)

Caramalho, Í., Nunes-Cabaço, H., Foxall, R. B., and Sousa, A. E. 2015. Regulatory T-cell development in the human thymus. *Frontiers in immunology*, 6. (Cited on page 101.)

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270. (Cited on pages 38 and 97.)

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791. (Cited on pages 36, 38, and 98.)

Chao, A. and Jost, L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93: 2533–2547. (Cited on page 36.)

Chao, A., Gotelli, N. J., Hsieh, T. C., Sande, E. L., Ma, K. H., Colwell, R. K., and Ellison, A. M. 2014. Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84: 45–67. (Cited on page 36.)

Chiu, C.-H. and Chao, A. 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, 4: e1634. (Cited on page 97.)

Clark, D. R., de Boer, R. J., Wolthers, K. C., and Miedema, F. 1999. T cell dynamics in

HIV-1 infection. *Advances in immunology*, 73: 301–327. (Cited on pages 51 and 97.)

Cline, J., Braman, J. C., and Hogrefe, H. H. 1996. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research*, 24(18): 3546–3551. (Cited on page 19.)

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., and Longino, J. T. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of plant ecology*, 5(1): 3–21. (Cited on page 36.)

Connelley, T., Aerts, J., Law, A., and Morrison, W. I. 2009. Genomic analysis reveals extensive gene duplication within the bovine TRB locus. *BMC genomics*, 10(1): 192. (Cited on page 77.)

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H., Kedzierska, K., *et al.* 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661): 89–+. (Cited on pages 39 and 99.)

Davis, M. M. and Bjorkman, P. J. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181): 395–402. (Cited on pages 11, 51, and 77.)

De Boer, R. J. and Perelson, A. S. 1994. T cell repertoires and competitive exclusion. *Journal of theoretical biology*, 169: 375–390. (Cited on pages 52 and 63.)

De Boer, R. J. and Perelson, A. S. 1995. Towards a general function describing T cell proliferation. *Journal of theoretical biology*, 175: 567–576. (Cited on pages 52 and 63.)

De Boer, R. J. and Perelson, A. S. 1997. Competitive control of the self-renewing T cell repertoire. *International immunology*, 9: 779–790. (Cited on pages 52 and 63.)

den Braber, I., Mugwagwa, T., Vrisekoop, N., Westera, L., Mögling, R., de Boer, A. B., Willems, N., Schrijver, E. H. R., Spierenburg, G., Gaiser, K., Mul, E., Otto, S. A., Ruiter, A. F. C., Ackermans, M. T., Miedema, F., Borghans, J. A. M., de Boer, R. J., and Tesselaar, K. 2012. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36: 288–297. (Cited on pages 51, 52, and 54.)

Desponds, J., Mora, T., and Walczak, A. M. 2016. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 113: 274–279. (Cited on pages 52 and 63.)

Desponds, J., Mayer, A., Mora, T., and Walczak, A. M. 2017. Population dynamics of immune repertoires. *arXiv preprint arXiv:1703.00226*. (Cited on pages 52 and 63.)

Di Rosa, F. and Pabst, R. 2005. The bone marrow: a nest for migratory memory T cells. *Trends in immunology*, 26(7): 360–366. (Cited on page 98.)

Dowling, M. R. and Hodgkin, P. D. 2009. Modelling naive T-cell homeostasis: consequences of heritable cellular lifespan during ageing. *Immunology and cell biology*, 87: 445–456. (Cited on page 52.)

Egorov, E. S., Merzlyak, E. M., Shelenkov, A. A., Britanova, O. V., Sharonov, G. V., Staroverov, D. B., Bolotin, D. A., Davydov, A. N., Barsova, E., Lebedev, Y. B., Shugay, M., and Chudakov, D. M. 2015. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *Journal of immunology (Baltimore, Md. : 1950)*, 194: 6155–6163. (Cited on pages 29 and 31.)

Ehrenmann, F. and Lefranc, M.-P. 2011. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harbor protocols*, 2011: 737–749. (Cited on page 85.)

Ehrenmann, F., Kaas, Q., and Lefranc, M.-P. 2010. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic acids research*, 38: D301–D307. (Cited on page 85.)

Elhanati, Y., Murugan, A., Callan, C. G., Mora, T., and Walczak, A. M. 2014. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 111: 9875–9880. (Cited on pages 52, 63, 66, and 100.)

Enkirch, T. and von Messling, V. 2015. Ferret models of viral pathogenesis. *Virology*, 479-480: 259–270. (Cited on pages 77 and 83.)

Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., and Holt, R. A. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research*, 19(10): 1817–1824. (Cited on pages 2 and 11.)

Gattinoni, L., Lugli, E., Ji, Y., Pos, Z., Paulos, C. M., Quigley, M. F., Almeida, J. R., Gostick, E., Yu, Z., Carpenito, C., Wang, E., Douek, D. C., Price, D. A., June, C. H., Marincola, F. M., Roederer, M., and Restifo, N. P. 2011. A human memory T cell subset with stem cell-like properties. *Nature medicine*, 17: 1290–1297. (Cited on pages 53 and 60.)

Gattinoni, L., Speiser, D. E., Lichterfeld, M., and Bonini, C. 2017. T memory stem cells in health and disease. *Nature medicine*, 23(1): 18–27. (Cited on page 101.)

Gavin, M. A. and Bevan, M. J. 1995. Increased peptide promiscuity provides a rationale for the lack of N regions in the neonatal T cell repertoire. *Immunity*, 3(6): 793–800. (Cited on page 100.)

George, J. F. and Schroeder, H. 1992. Developmental regulation of D beta reading frame and junctional diversity in T cell receptor-beta transcripts from human thymus. *The Journal of Immunology*, 148(4): 1230–1239. (Cited on page 99.)

Gerritsen, B., Pandit, A., Andeweg, A. C., and De Boer, R. J. 2016. RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data. *Bioinformatics*, 32(20): 3098–3106. (Cited on pages 5, 6, 36, and 78.)

Giraud, M., Salson, M., Duez, M., Villenet, C., Quief, S., Caillault, A., Grardel, N., Roumier, C., Preudhomme, C., and Figeac, M. 2014. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC genomics*, 15: 409. (Cited on page 5.)

Giudicelli, V., Chaume, D., and Lefranc, M.-P. 2004. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic acids research*, 32: W435–W440. (Cited on page 5.)

Giudicelli, V., Chaume, D., and Lefranc, M.-P. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic acids research*, 33: D256–D261. (Cited on page 79.)

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas, N., Arlehamn, C. S. L., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., and Davis, M. M. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547: 94–98. (Cited on pages 39, 77, and 99.)

Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J. C., Smit, A. F., Wang, K., Koop, B. F., and Hood, L. 2001. Comparative genomics of the human and mouse T cell receptor loci. *Immunity*, 15(3): 337–349. (Cited on page 77.)

Gonçalves, P., Ferrarini, M., Molina-París, C., Lythe, G., Vasseur, F., Lim, A., Rocha, B., and Azogui, O. 2017. A new mechanism shapes the naïve CD8(+) T cell repertoire:

the selection for full diversity. *Molecular immunology*, 85: 66–80. (Cited on page 54.)

Goodwin, S., McPherson, J. D., and McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6): 333–351. (Cited on pages 3, 5, and 101.)

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52: 696–704. (Cited on pages 82 and 88.)

Hapuarachchi, T., Lewis, J., and Callard, R. E. 2013. A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. *Frontiers in immunology*, 4: 366. (Cited on pages 52 and 63.)

Heather, J. M., Ismail, M., Oakes, T., and Chain, B. 2017. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in bioinformatics*, page bbw138. (Cited on pages 3, 5, 6, and 96.)

Holt, R. A. and Jones, S. J. M. 2008. The new paradigm of flow cell sequencing. *Genome Res.*, 18(6): 839–846. (Cited on page 11.)

Hou, X., Wang, M., Lu, C., Xie, Q., Cui, G., Chen, J., Du, Y., Dai, Y., and Diao, H. 2016. Analysis of the repertoire features of TCR beta chain CDR3 in human by high-throughput sequencing. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology*, 39: 651–667. (Cited on pages 52 and 63.)

Howie, B., Sherwood, A. M., Berkebile, A. D., Berka, J., Emerson, R. O., Williamson, D. W., Kirsch, I., Vignali, M., Rieder, M. J., Carlson, C. S., and Robins, H. S. 2015. High-throughput pairing of T cell receptor α and β sequences. *Science translational medicine*, 7: 301ra131. (Cited on page 3.)

Hsieh, C.-S., Lee, H.-M., and Lio, C.-W. J. 2012. Selection of regulatory T cells in the thymus. *Nature reviews. Immunology*, 12: 157–167. (Cited on page 101.)

Hsieh, T., Ma, K., and Chao, A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (hill numbers). *Methods in Ecology and Evolution*, 7(12): 1451–1456. (Cited on page 36.)

Huang, W., Li, L., Myers, J. R., and Marth, G. T. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4): 593–594. (Cited on page 19.)

Hubbell, S. P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32) (Monographs in Population Biology)*. Princeton University Press. (Cited on pages 53 and 65.)

Hung, S.-J., Chen, Y.-L., Chu, C.-H., Lee, C.-C., Chen, W.-L., Lin, Y.-L., Lin, M.-C., Ho, C.-L., and Liu, T. 2016. TRIg: a robust alignment pipeline for non-regular T-cell receptor and immunoglobulin sequences. *BMC bioinformatics*, 17(1): 433. (Cited on page 5.)

Jenkins, M. K., Chu, H. H., McLachlan, J. B., and Moon, J. J. 2010. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. *Annual review of immunology*, 28: 275–294. (Cited on page 52.)

Johnson, P. L. F., Yates, A. J., Goronzy, J. J., and Antia, R. 2012. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 t-cell diversity with age. *Proceedings of the National Academy of Sciences of the United States of America*, 109: 21432–21437. (Cited on page 52.)

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., and Vogelstein, B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *PNAS*,

108(23): 9530–9535. (Cited on page 29.)

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. 2011. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9: 72–74. (Cited on pages 6, 29, 53, 78, and 98.)

Kjer-Nielsen, L., Borg, N. A., Pellicci, D. G., Beddoe, T., Kostenko, L., Clements, C. S., Williamson, N. A., Smyth, M. J., Besra, G. S., Reid, H. H., *et al.* 2006. A structural basis for selection and cross-species reactivity of the semi-invariant NKT cell receptor in CD1d/glycolipid recognition. *Journal of Experimental Medicine*, 203(3): 661–673. (Cited on page 100.)

Klarenbeek, P. L., Tak, P. P., van Schaik, B. D., Zwinderman, A. H., Jakobs, M. E., Zhang, Z., van Kampen, A. H., van Lier, R. A., Baas, F., and de Vries, N. 2010. Human T-cell memory consists mainly of unexpanded clones. *Immunology letters*, 133(1): 42–48. (Cited on page 11.)

Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A. U., Babel, N., Reinert, K., and Robinson, P. N. 2015. IMSEQ–a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics*, 31(18): 2963–2971. (Cited on pages 5, 6, 12, and 21.)

Lander, E. S. *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature*, 409: 860–921. (Cited on page 3.)

Lane, J., Duroux, P., and Lefranc, M.-P. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC bioinformatics*, 11: 223. (Cited on page 80.)

Langmead, B. and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357–359. (Cited on page 13.)

Laydon, D. J., Bangham, C. R. M., and Asquith, B. 2015. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370. (Cited on page 97.)

Le Bourhis, L., Martin, E., Péguillet, I., Guihot, A., Froux, N., Coré, M., Lévy, E., Dusseaux, M., Meyssonnier, V., Premel, V., *et al.* 2010. Antimicrobial activity of mucosal-associated invariant T cells. *Nature immunology*, 11(8): 701–708. (Cited on page 100.)

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental and comparative immunology*, 27: 55–77. (Cited on page 85.)

Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piédade, I., Rouard, M., Foulquier, E., Thouvenin, V., and Lefranc, G. 2005. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Developmental and comparative immunology*, 29: 185–203. (Cited on pages 81 and 86.)

Lu, H., Giordano, F., and Ning, Z. 2016. Oxford nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 14(5): 265–279. (Cited on page 101.)

Lugli, E., Gattinoni, L., Roberto, A., Mavilio, D., Price, D. A., Restifo, N. P., and Roederer, M. 2013a. Identification, isolation and in vitro expansion of human and nonhuman primate T stem cell memory cells. *Nature protocols*, 8: 33–42. (Cited on

pages 53 and 60.)

Lugli, E., Dominguez, M. H., Gattinoni, L., Chattopadhyay, P. K., Bolton, D. L., Song, K., Klatt, N. R., Brenchley, J. M., Vaccari, M., Gostick, E., Price, D. A., Waldmann, T. A., Restifo, N. P., Franchini, G., and Roederer, M. 2013b. Superior T memory stem cell persistence supports long-lived T cell memory. *The Journal of clinical investigation*, 123: 594–599. (Cited on pages 53 and 60.)

Lythe, G., Callard, R. E., Hoare, R. L., and Molina-París, C. 2016. How many TCR clonotypes does a body maintain? *Journal of theoretical biology*, 389: 214–224. (Cited on pages 52, 63, and 98.)

Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I. R., and Friedman, N. 2014. 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. (Cited on page 101.)

Mamedov, I. Z., Britanova, O. V., Zvyagin, I. V., Turchaninova, M. A., Bolotin, D. A., Putintseva, E. V., Lebedev, Y. B., and Chudakov, D. M. 2013. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Frontiers in immunology*, 4: 456. (Cited on pages 4, 30, 43, 44, and 78.)

Marcou, Q., Mora, T., and Walczak, A. M. 2017. IGoR: a tool for high-throughput immune repertoire analysis. *arXiv preprint arXiv:1705.08246*. (Cited on pages 50, 52, 55, 57, 58, and 99.)

Mardis, E. R. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9: 387–402. (Cited on pages 4 and 5.)

Marraco, S. A. F., Soneson, C., Cagnon, L., Gannon, P. O., Allard, M., Maillard, S. A., Montandon, N., Rufer, N., Waldvogel, S., Delorenzi, M., *et al.* 2015. Long-lasting stem cell–like memory CD8+ T cells with a naïve-like profile upon yellow fever vaccination. *Science translational medicine*, 7(282): 282ra48–282ra48. (Cited on pages 53 and 60.)

Mason, D. 1998. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology today*, 19: 395–404. (Cited on page 51.)

Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L., and Chenchik, A. 1999. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic acids research*, 27: 1558–1560. (Cited on page 4.)

McDonald, B. D., Bunker, J. J., Erickson, S. A., Oh-Hora, M., and Bendelac, A. 2015. Crossreactive $\alpha\beta$ T cell receptors are the predominant targets of thymocyte negative selection. *Immunity*, 43: 859–869. (Cited on page 51.)

Merkenschlager, M., Graf, D., Lovatt, M., Bommhardt, U., Zamoyska, R., and Fisher, A. G. 1997. How many thymocytes audition for selection? *The Journal of experimental medicine*, 186: 1149–1158. (Cited on page 51.)

Mineccia, M., Massari, S., Linguiti, G., Ceci, L., Ciccarese, S., and Antonacci, R. 2012. New insight into the genomic structure of dog T cell receptor beta (TRB) locus inferred from expression analysis. *Developmental and comparative immunology*, 37: 279–293. (Cited on pages 77, 81, 82, and 83.)

Miyama, T., Kawase, T., Kitaura, K., Chishaki, R., Shibata, M., Oshima, K., Hamana, H., Kishi, H., Muraguchi, A., Kuzushima, K., *et al.* 2017. Highly functional T-cell receptor repertoires are abundant in stem memory T cells and highly shared among individuals. *Scientific Reports*, 7. (Cited on page 101.)

Mora, T. and Walczak, A. 2016. Quantifying lymphocyte receptor diversity. *arXiv preprint arXiv:1604.00487*. (Cited on pages 2, 36, 51, and 98.)

Mora, T., Walczak, A. M., Bialek, W., and Callan, C. G. 2010. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12): 5405–5410. (Cited on page 19.)

Muraro, P. A., Robins, H., Malhotra, S., Howell, M., Phippard, D., Desmarais, C., de Paula Alves Sousa, A., Griffith, L. M., Lim, N., Nash, R. A., and Turka, L. A. 2014. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *The Journal of clinical investigation*, 124: 1168–1172. (Cited on pages 35 and 51.)

Murugan, A., Mora, T., Walczak, A. M., and Callan, C. G. 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 109: 16161–16166. (Cited on pages 4, 19, 50, 51, 52, 63, and 99.)

Ndifon, W., Gal, H., Shifrut, E., Aharoni, R., Yissachar, N., Waysbort, N., Reich-Zeliger, S., Arnon, R., and Friedman, N. 2012. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proceedings of the National Academy of Sciences of the United States of America*, 109: 15865–15870. (Cited on pages 11, 52, 63, and 100.)

Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., and Geiger, T. L. 2011. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC genomics*, 12: 106. (Cited on pages 6, 11, 12, and 13.)

Nikolich-Žugich, J., Slifka, M. K., and Messaoudi, I. 2004. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2): 123–132. (Cited on page 51.)

Oh, D. Y. and Hurt, A. C. 2016. Using the ferret as an animal model for investigating influenza antiviral effectiveness. *Frontiers in microbiology*, 7: 80. (Cited on page 77.)

Orlitsky, A., Suresh, A. T., and Wu, Y. 2016. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47): 13283–13288. (Cited on page 98.)

Papalexi, E. and Satija, R. 2017. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature reviews. Immunology*. (Cited on pages 98 and 101.)

Peng, X., Alföldi, J., Gori, K., Eisfeld, A. J., Tyler, S. R., Tisoncik-Go, J., Brawand, D., Law, G. L., Skunca, N., Hatta, M., Gasper, D. J., Kelly, S. M., Chang, J., Thomas, M. J., Johnson, J., Berlin, A. M., Lara, M., Russell, P., Swofford, R., Turner-Maier, J., Young, S., Hourlier, T., Aken, B., Searle, S., Sun, X., Yi, Y., Suresh, M., Tumpey, T. M., Siepel, A., Wisely, S. M., Dessimoz, C., Kawaoka, Y., Birren, B. W., Lindblad-Toh, K., Di Palma, F., Engelhardt, J. F., Palermo, R. E., and Katze, M. G. 2014. The draft genome sequence of the ferret (mustela putorius furo) facilitates study of human respiratory disease. *Nature biotechnology*, 32: 1250–1255. (Cited on page 77.)

Pogorelyy, M. V., Elhanati, Y., Marcou, Q., Sycheva, A. L., Komech, E. A., Nazarov, V. I., Britanova, O. V., Chudakov, D. M., Mamedov, I. Z., Lebedev, Y. B., Mora, T., and Walczak, A. M. 2017. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS computational biology*, 13: e1005572. (Cited on pages 99 and 100.)

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., and Goronzy, J. J. 2014. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36): 13139–13144. (Cited on pages 11, 36, 39, 51, 54, and 98.)

Quigley, M. F., Greenaway, H. Y., Venturi, V., Lindsay, R., Quinn, K. M., Seder, R. A., Douek, D. C., Davenport, M. P., and Price, D. A. 2010. Convergent recombination

shapes the clonotypic landscape of the naive T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 107: 19414–19419. (Cited on page 100.)

Redmond, D., Poran, A., and Elemento, O. 2016. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome medicine*, 8(1): 80. (Cited on pages 98 and 101.)

Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., and Carlson, C. S. 2009. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19): 4099–4107. (Cited on pages 6, 11, and 51.)

Robins, H. S., Srivastava, S. K., Campregher, P. V., Turtle, C. J., Andriesen, J., Riddell, S. R., Carlson, C. S., and Warren, E. H. 2010. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Science translational medicine*, 2(47): 47ra64–47ra64. (Cited on page 11.)

Rosati, E., Dowds, C. M., Liaskou, E., Henriksen, E. K. K., Karlsen, T. H., and Franke, A. 2017. Overview of methodologies for T-cell receptor repertoire analysis. *BMC biotechnology*, 17(1): 61. (Cited on pages 3, 4, and 7.)

Sethna, Z., Elhanati, Y., Dudgeon, C. S., Callan, C. G., Levine, A. J., Mora, T., and Walczak, A. M. 2017. Insights into immune system development and function from mouse T-cell repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 114: 2253–2258. (Cited on page 63.)

Sewell, A. K. 2012. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9): 669–677. (Cited on page 51.)

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nature biotechnology*, 26: 1135–1145. (Cited on pages 4, 5, and 11.)

Shugay, M., Britanova, O. V., Merzlyak, E. M., Turchaninova, M. A., Mamedov, I. Z., Tuganbaev, T. R., Bolotin, D. A., Staroverov, D. B., Putintseva, E. V., Plevova, K., Linnemann, C., Shagin, D., Pospisilova, S., Lukyanov, S., Schumacher, T. N., and Chudakov, D. M. 2014. Towards error-free profiling of immune repertoires. *Nature methods*, 11: 653–655. (Cited on pages 5, 6, 12, 21, 29, 30, and 53.)

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., *et al.* 2017. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*. (Cited on page 101.)

Sloan, W. T., Lunn, M., Woodcock, S., Head, I. M., Nee, S., and Curtis, T. P. 2006. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental microbiology*, 8: 732–740. (Cited on page 53.)

Smith, T., Heger, A., and Sudbery, I. 2017. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, 27: 491–499. (Cited on page 7.)

Stirk, E. R., Molina-París, C., and van den Berg, H. A. 2008. Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of theoretical biology*, 255: 237–249. (Cited on pages 52 and 63.)

Stirk, E. R., Lythe, G., van den Berg, H. A., and Molina-París, C. 2010. Stochastic competitive exclusion in the maintenance of the naïve T cell repertoire. *Journal of theoretical biology*, 265: 396–410. (Cited on pages 52 and 63.)

Suessmuth, Y., Mukherjee, R., Watkins, B., Koura, D. T., Finstermeier, K., Desmarais, C., Stempora, L., Horan, J. T., Langston, A., Qayed, M., *et al.* 2015. CMV reactiva-

tion drives posttransplant T-cell reconstitution and results in defects in the underlying TCRβ repertoire. *Blood*, 125(25): 3835–3850. (Cited on page 11.)

Takada, K. and Jameson, S. C. 2009. Naive T cell homeostasis: from awareness of space to a sense of place. *Nature reviews. Immunology*, 9: 823–832. (Cited on page 52.)

Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., and Chain, B. 2013. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics (Oxford, England)*, 29: 542–550. (Cited on pages 5, 6, 12, 53, and 64.)

Thomas, N., Best, K., Cinelli, M., Reich-Zeliger, S., Gal, H., Shifrut, E., Madi, A., Friedman, N., Shawe-Taylor, J., and Chain, B. 2014. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 30(22): 3181–3188. (Cited on page 98.)

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. 2017. McPAS-TCR: A manually-curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*. (Cited on page 101.)

van Riel, D., Munster, V. J., de Wit, E., Rimmelzwaan, G. F., Fouchier, R. A. M., Osterhaus, A. D. M. E., and Kuiken, T. 2006. H5N1 virus attachment to lower respiratory tract. *Science (New York, N.Y.)*, 312: 399. (Cited on page 77.)

Vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N., O'Connor, K. C., Hafler, D. A., Vigneault, F., and Kleinstein, S. H. 2014. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13): 1930–1932. (Cited on pages 5 and 12.)

Venturi, V., Kedzierska, K., Price, D. A., Doherty, P. C., Douek, D. C., Turner, S. J., and Davenport, M. P. 2006. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 103: 18691–18696. (Cited on page 63.)

Venturi, V., Price, D. A., Douek, D. C., and Davenport, M. P. 2008. The molecular basis for public T-cell responses? *Nature reviews. Immunology*, 8: 231–238. (Cited on page 63.)

Venturi, V., Quigley, M. F., Greenaway, H. Y., Ng, P. C., Ende, Z. S., McIntosh, T., Asher, T. E., Almeida, J. R., Levy, S., Price, D. A., Davenport, M. P., and Douek, D. C. 2011. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *Journal of immunology (Baltimore, Md. : 1950)*, 186: 4285–4294. (Cited on pages 19, 63, and 100.)

Venturi, V., Rudd, B. D., and Davenport, M. P. 2013. Specificity, promiscuity, and precursor frequency in immunoreceptors. *Current opinion in immunology*, 25: 639–645. (Cited on page 100.)

Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R. M., Hudson, J. R., Davis, R. W., *et al.* 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences*, 107(4): 1518–1523. (Cited on pages 11 and 12.)

Warren, R. L., Nelson, B. H., and Holt, R. A. 2009. Profiling model T-cell metagenomes with short reads. *Bioinformatics*, 25(4): 458–464. (Cited on page 12.)

Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R., and Holt, R. A. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size

of at least 1 million clonotypes. *Genome research*, 21: 790–797. (Cited on pages 6, 11, 19, 27, 28, 29, 97, and 100.)

Wertheimer, A. M., Bennett, M. S., Park, B., Uhrlaub, J. L., Martinez, C., Pulko, V., Currier, N. L., Nikolich-Žugich, D., Kaye, J., and Nikolich-Žugich, J. 2014. Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans. *Journal of immunology (Baltimore, Md. : 1950)*, 192: 2143–2155. (Cited on page 51.)

Woodsworth, D. J., Castellarin, M., and Holt, R. A. 2013. Sequence analysis of T-cell repertoires in health and disease. *Genome medicine*, 5: 98. (Cited on pages 3, 5, 11, and 35.)

Yager, E. J., Ahmed, M., Lanzer, K., Randall, T. D., Woodland, D. L., and Blackman, M. A. 2008. Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *The Journal of experimental medicine*, 205: 711–723. (Cited on page 51.)

Yang, X., Liu, D., Lv, N., Zhao, F., Liu, F., Zou, J., Chen, Y., Xiao, X., Wu, J., Liu, P., *et al.* 2015. TCRklass: a new k-string–based algorithm for human and mouse TCR repertoire characterization. *The Journal of Immunology*, 194(1): 446–454. (Cited on pages 5, 12, and 21.)

Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, 41: W34–W40. (Cited on page 5.)

Yu, Y., Ceredig, R., and Seoighe, C. 2015. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic acids research*, 44(4): e31–e31. (Cited on page 5.)

Zarnitsyna, V. I., Evavold, B. D., Schoettle, L. N., Blattman, J. N., and Antia, R. 2013. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in immunology*, 4: 485. (Cited on page 51.)

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. 2014. PEAR: a fast and accurate illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, 30: 614–620. (Cited on page 35.)

Zhang, W., Du, Y., Su, Z., Wang, C., Zeng, X., Zhang, R., Hong, X., Nie, C., Wu, J., Cao, H., *et al.* 2015. IMonitor: a robust pipeline for TCR and BCR repertoire analysis. *Genetics*, 201(2): 459–472. (Cited on page 5.)

Samenvatting

T cellen vormen een belangrijk onderdeel van het menselijk afweersysteem. Ze helpen het lichaam zichzelf te beschermen tegen binnendringers zoals virussen en bacteriën, en ook maken ze zieke lichaamscellen, zoals kankercellen, dood. T cellen herkennen een zieke lichaamscel door het oppervlak van de zieke cel af te tasten met een soort voelspriet, genaamd de T cel receptor (TCR). Herkenning van een zieke cel via aftasten van het celoppervlak is mogelijk, omdat alle lichaamscellen (met een celkern), hun (eiwit-)inhoud afbreken en als een soort jas aan de buitenkant dragen. Een cel die door een virus geinfecteerd is, presenteert aan zijn celoppervlak stukjes eiwit (peptides) die van het virus afkomstig zijn. Een T cel waarvan de TCR bindt aan de peptides aan de buitenkant van een lichaamscel, en dus als een soort sleutel in een slot past, kan overgaan tot het doden van de lichaamscel. T cellen moeten dus onderscheid maken tussen lichaamseigen (self) en lichaamsvreemde (non-self) peptides om te voorkomen dat ze gezonde lichaamscellen doden. Dit wordt bereikt door vrijwel iedere T cel uit te rusten met een eigen TCR die alleen kan binden aan lichaamsvreemde peptides. Omdat er zeer veel peptides mogelijk zijn ($> 20^9 \approx 10^{12}$), zijn er zeer veel T cellen (10^{12}) nodig, waarvan vrijwel ieder een andere TCR draagt, om alle mogelijke lichaamsvreemde peptides te kunnen herkennen. Gezamenlijk vormen de T cellen een “T cel repertoire”. Door T cel repertoires te analyseren, hopen we meer te weten te komen hoe precies het menselijk lichaam zichzelf beschermt, wat mogelijk kan helpen bij het vinden van oplossingen wanneer het lichaam niet of onvoldoende in staat is zichzelf te beschermen.

De identiteit van de TCR die een T cel draagt kan bepaald worden aan de hand van het TCR gen dat in het genoom van de T cel beschreven staat. Het bepalen van de volgorde van de nucleotides (A, C, T, en G) in een gen, wordt “sequencing” genoemd. Door grote sprongen in de sequencing techniek van de afgelopen 10 jaar, is het nu mogelijk om van miljoenen T cellen tegelijk de TCR te sequencen. Hoewel een sample uit een T cel

repertoire dus uit een miljoen T cellen kan bestaan, spreken we van een klein sample, omdat het sample zelf slechts ongeveer een miljoenste van het totale repertoire bevat. In deze thesis sequencen, en analyseren we dus kleine samples uit grote T cel repertoires.

Een sequencing experiment levert vele sequenties op, maar soms kloppen de sequenties niet doordat er leesfouten worden gemaakt. Omdat verschillende TCR genen sterk op elkaar kunnen lijken, is het lastig om een echt TCR gen te onderscheiden van een variant die is ontstaan door een leesfout. In veel gevallen lost men dit op door sequenties van een lage kwaliteit weg te gooien en sequenties die veel op elkaar lijken samen te voegen. Dit kan leiden tot het verlies van zeer veel (tot wel de helft) echte TCR sequenties. In hoofdstuk 2 beschrijven we een tool, Recover T cell receptor (RTCR), die we hebben ontwikkeld om zoveel mogelijk echte TCR sequenties uit een sequencing experiment te halen. Deze tool maakt gebruik van een simpel statistisch model in combinatie met heuristieken om automatisch te bepalen welke sequenties echt zijn en welke leesfouten bevatten. Recentelijk heeft een onafhankelijke onderzoeksgroep RTCR en gerelateerde tools vergeleken en RTCR bevonden als de meest accurate tool op dit moment.

In hoodstuk 3 passen we onze software (RTCR) toe op echte sequencing data en analyseren we hoe het repertoire van een gezond persoon verandert over een periode van ruim twee jaar. Dit is interessant, omdat het repertoire niet statisch is: nieuwe T cellen worden iedere dag aangemaakt, er gaan T cellen dood, en tijdens een infectie gaan de T cellen die de lichaamsvreemde peptides herkennen zich enorm vermeerderen. Dit is ook een van de redenen dat mensen meestal niet voor de tweede keer ziek worden door dezelfde ziekteverwekker, er staat namelijk al een leger T cellen klaar om de ziekteverwekker aan te pakken. Hoe veranderlijk het T cel repertoire van een gezond persoon is, is niet precies bekend. We observeren in onze samples dat sommige groepen T cellen die dezelfde TCR dragen, zogenaamde “clonotypes”, in de eerste samples niet gezien worden, maar in de samples die twee jaar later zijn afgenomen, tot zelfs 1% van de T cellen in beslag kunnen nemen. Dit laat zien dat in een gezond persoon het repertoire sterk kan veranderen.

In hoofdstuk 4 onderzoeken we hoe het kan dat sommige clonotypes meer T cellen bevatten dan andere en verschillende mensen dezelfde T cellen (met dezelfde TCR) kunnen produceren. Deze observaties zijn opvallend vanwege twee redenen. Ten eerste, zoals sneeuwvlokken altijd anders zijn, lijken bijna alle T cellen een andere TCR te dragen. Dit is ook logisch, gezien de grote hoeveelheid verschillende peptides die herkend moeten kunnen worden. Ten tweede, tijdens de ontwikkeling van een T cel wordt zijn TCR gen op een willekeurige manier aangemaakt via een proces dat VDJ recombinatie heet. Dit proces maakt meer verschillende TCRs mogelijk dan er sterren zijn in het zichtbare uni-

versum. Een bekende hypothese voor de verschillende aantallen T cellen in clonotypes is dat de ene T cel beter overleeft dan de andere. Via homeostatische celdeling kan een clonotype groter worden. Wij hebben echter gekozen om de systematiek die gevonden is in VDJ recombinatie, te vertalen naar een simpel model dat de grootte van clonotypes voorspelt. Dit model laat zien dat op basis van wat we weten van VDJ recombinatie, we al zeer grote clonotypes mogen verwachten. Ook laten we zien dat dit in overeenkomst is met de samples die wij hebben afgenomen van gezonde personen, waarbij grotere clonotypes vaak een TCR hebben die relatief gemakkelijk gemaakt kan worden door VDJ recombinatie.

Ten slotte, in hoofdstuk 5, verleggen we ons onderzoek naar de fret, een veelgebruikt modelorganisme voor infecties. In vervolgonderzoek willen we de fret gebruiken om te zien hoe T cel repertoires veranderen tijdens een infectie. Echter, zonder kennis hoe de fret zijn TCR genen genereert is het niet mogelijk het repertoire te analyseren. In hoofdstuk 5, analyseren we het zogenaamde TRB locus van de Fret en beschrijven we de gen segmenten die de Fret gebruikt om zijn TCR genen te maken. Hiermee leggen we de basis voor toekomstig onderzoek naar T cel repertoires in dit belangrijke modelorganisme.

In deze thesis hebben we onderzoek gedaan aan kleine samples uit grote T cel repertoires. Via sequencing, analyse, en modellering van de kleine samples zijn we meer te weten gekomen over hoe T cel repertoires werken.

Curriculum vitæ

Bram Gerritsen was born on July 16th, 1982 in Nijmegen, the Netherlands. From 2001 until 2007 he studied at Radboud University in Nijmegen, obtaining a Bachelor of Science and Master of Science degree in Molecular Life Sciences and Bioinformatics, respectively. In 2007-2008 he went backpacking in Australia, where he worked as a computer programmer for the government for about 5 months. From August 2008 until February 2012, he worked as a bioinformatician for the Dutch cancer institute (NKI) in Amsterdam. In May 2012, he started his PhD research at the Theoretical Biology & Bioinformatics group at Utrecht University, under the supervision of Prof.dr. Rob J. De Boer, Dr. Aridaman Pandit, and Dr. Arno C. Andeweg. The results of his PhD research are described in this thesis.

List of Publications

Gerritsen, Bram*, A. Pandit, A. C. Andeweg, and R. J. de Boer, "RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data," *Bioinformatics*, vol. 32, pp. 3098–3106, Oct. 2016.
* Corresponding author

Gerritsen, Bram and A. Pandit, "The memory of a killer T cell: models of CD8(+) T cell differentiation," *Immunol. Cell Biol.*, vol. 94, pp. 236–241, Mar. 2016.

H. N. Kløverpris, R. McGregor, J. E. McLaren, K. Ladell, M. Harndahl, A. Stryhn, J. M. Carlson, C. Koofhethile, **Gerritsen, Bram**, C. Keşmir, F. Chen, L. Riddell, G. Luzzi, A. Leslie, B. D. Walker, T. Ndung'u, S. Buus, D. A. Price, and P. J. Goulder, "CD8+ TCR Bias and Immunodominance in HIV-1 Infection," *J. Immunol.*, vol. 194, pp. 5329–5345, June 2015.

P. A. Possik, J. Müller, C. Gerlach, J. C. N. Kenski, X. Huang, A. Shahrabi, O. Krijgsman, J.-Y. Song, M. A. Smit, **Gerritsen, Bram**, C. Lieftink, K. Kemper, M. Michaut, R. L. Beijersbergen, L. Wessels, T. N. Schumacher, and D. S. Peeper, "Parallel in vivo and in vitro melanoma RNAi dropout screens reveal synthetic lethality between hypoxia and DNA damage response inhibition," *Cell Rep*, vol. 9, pp. 1375–1386, Nov. 2014.

M. C. Gold, J. E. McLaren, J. A. Reistetter, S. Smyk-Pearson, K. Ladell, G. M. Swarbrick, Y. Y. L. Yu, T. H. Hansen, O. Lund, M. Nielsen, **Gerritsen, Bram**, C. Kesmir, J. J. Miles, D. A. Lewinsohn, D. A. Price, and D. M. Lewinsohn, "MR1-restricted MAIT cells display ligand discrimination and pathogen selectivity through distinct T cell receptor usage," *J. Exp. Med.*, vol. 211, pp. 1601–1610, July 2014.

Acknowledgements

About 5 to 6 years ago, I met with Leila and spoke with her about doing a PhD. She highly recommended the Theoretical Biology and Bioinformatics (TBB) group in Utrecht. She mentioned a professor called Rob de Boer, according to her one of the best supervisors one could wish for. Right now, looking back to that moment, I can say that I whole-heartedly agree. Rob, thank you very much for the years of guidance, I learned a lot!

What Leila could not know, was that I would have Aridaman Pandit as co-supervisor. If she did, there is no way she would not have mentioned him. I was lucky enough both my supervisors were great. Aridaman, what is interesting, is how you changed over the years. You learned a lot yourself, becoming an even better scientist and supervisor. Essentially, you became more like Rob, but with your own personality and flair. I really enjoyed our time together, both in the lab and outside. In short, thank you!

During my PhD we collaborated with Arno Andeweg, a virologist from the Erasmus University. Arno, over the years you turned from a collaborator into a co-promotor. I'm grateful for the many interesting meetings we had, and for the useful feedback you gave on the thesis.

I have also collaborated with Can Keşmir on several projects. Can, I enjoyed working together with you a lot, thank you.

While working on my projects, I have had many useful discussions. Hilje, Chris, Laurens, Gijs, Ewald, thank you all for your mathematical, statistical, and computational insights. I learned a lot from you. And I'm grateful to the master students that I had the privilege to supervise: Emre, Eduardo, and Peter. Bram (van Dijk), thank you for helping with the finishing touches on the cover. Jan kees, without your tireless management of the computers, little computation would be possible.

At the TBB there are many bright and friendly people, and it was a privilege to get to know them.

Paulien Hogeweg, thank you for founding the TBB group, and also for the wonderful courses that you teach.

Ewald and Guillem, thank you both for being my friends and paranymphs.

And finally, thanks to my family,
Robbert, bedankt voor het maken van een hele mooie cover. Pa en ma, heel erg bedankt dat jullie er altijd voor mij zijn. Door jullie is het me gelukt de thesis op tijd af te krijgen.