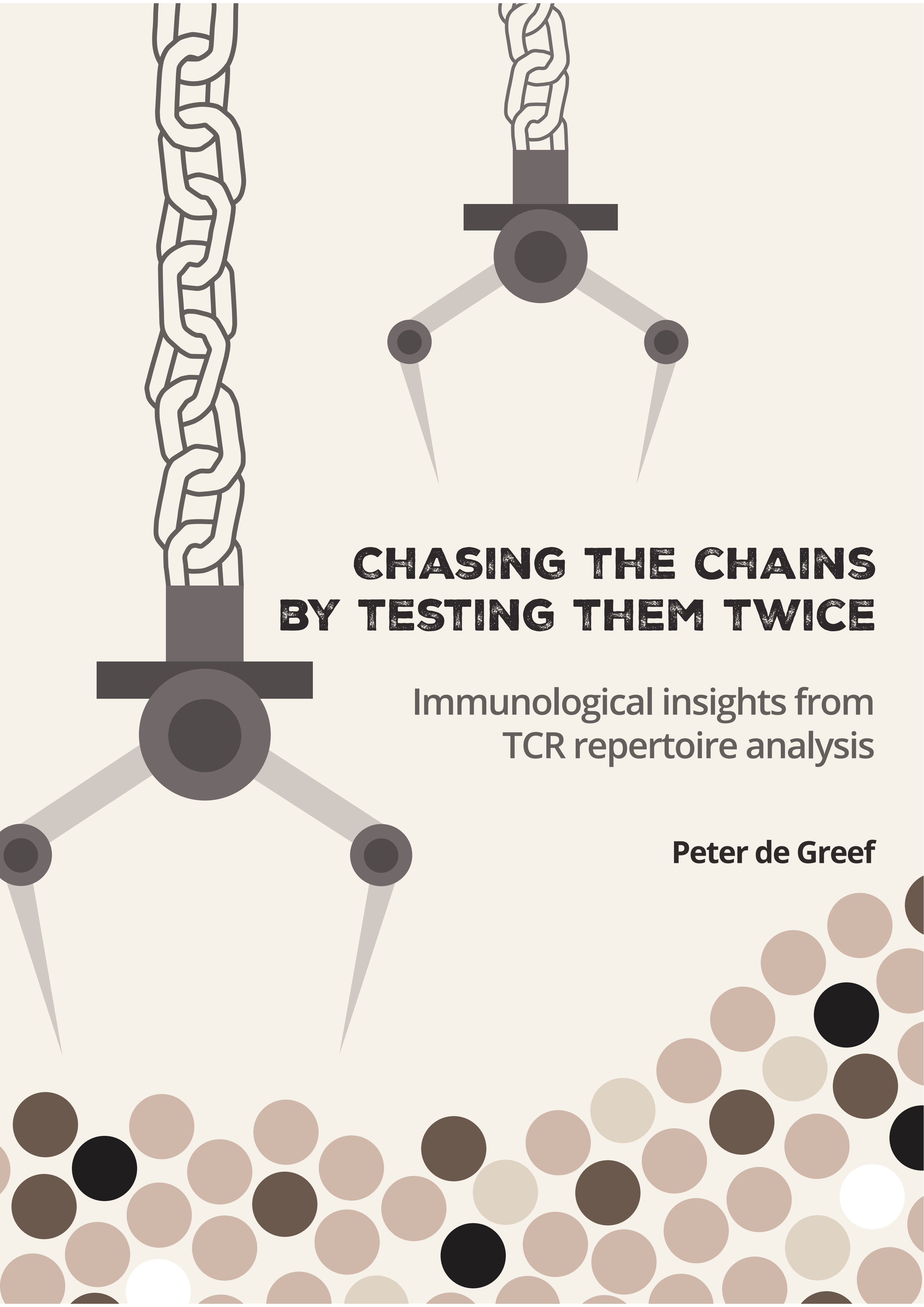


CHASING THE CHAINS BY TESTING THEM TWICE

Immunological insights from TCR repertoire analysis

Peter de Greef

Chasing the chains by testing them twice

Immunological insights from TCR repertoire analysis

Peter de Greef

Peter de Greef
Chasing the chains by testing them twice
Immunological insights from TCR repertoire analysis
PhD thesis, Utrecht University

Cover design: Janneke Wierda and Peter de Greef
Printing: Ridderprint | www.ridderprint.nl
ISBN: 978-90-393-7544-0

Copyright © Peter de Greef, 2023
All rights reserved. No part of this thesis may be reproduced in any form without written permission of the author.

Chasing the chains by testing them twice

Immunological insights from TCR repertoire analysis

Immunologische inzichten door tweemaal te toetsen

Analyse van het TCR-repertoire

(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. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag **25 april 2023**
des middags te **2.15 uur**

door

Peter Cornelis de Greef

geboren op 10 mei 1994 te Wageningen

Promotor:
Prof. dr. R.J. de Boer

Beoordelingscommissie:
Prof. dr. F. van Wijk
Prof. dr. T.N.M. Schumacher
Prof. dr. A. Walczak
Prof. dr. A. Mayer
Prof. dr. G. Yaari

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) aan het Graduate Programme Quantitative Biology - Computational Life Sciences (022.005.023).

Contents

1 General introduction 7

2 An extremely broad distribution of naive T-cell clone sizes 15

3 TCRβ rearrangements without a D segment 51

4 Comparing TCR diversity between sampled repertoires 71

5 Using TCR sequencing to follow the vaccination response 81

6 The TCR directs naive T cells to a preferred lymph node 103

7 General discussion 125

References 135

Samenvatting 145

Curriculum Vitae 148

List of Publications 149

Acknowledgements 150

1

General introduction

Adaptive immunity is an important layer of defence against infection

Across the tree of life, many strategies exist to protect organisms against infection with pathogens like harmful bacteria and viruses. These defence strategies together form the immune system, which consists of several lines of defence. A first obstacle is formed by mechanical barriers like the skin and chemical barriers such as the gastric acid that causes the low pH in the stomach. Pathogens passing these barriers get often cleared by the second layer of protection: the innate immune system that functions through the action of many specialised cells, including macrophages, neutrophils, and natural killer cells. The third line of defence, the adaptive immune system, is specific to vertebrates, like humans and mice. This complex system of billions of specialised B cells and T cells has the task to distinguish between self and foreign proteins. When a protein, or a short protein fragment called a peptide, binds to a specific B cell or T cell, it is called an antigen. An important task of the adaptive immune system is to 'remember' antigens after the first exposure. This memory allows the host to be prepared against recurrent infection with the same or a similar pathogen. In this thesis we focus on the repertoire of T cells in humans and mice.

The T cells, that play a major role in adaptive immunity, have a T-cell receptor (TCR) with which they can bind to a specific set of antigens. The antigens that T cells can bind to are peptides, which are presented in the context of the Major Histocompatibility Complex (MHC) proteins on the cell surface. Once a naive T cell binds such a peptide/MHC complex on an antigen-presenting cell with sufficient affinity, it can be activated. This activation induces clonal expansion, which means that the cell starts dividing, by which the number of T cells with that TCR quickly increases. These cells mainly employ effector functions, which vary across T-cell subsets. The cytotoxic T cells, that express the CD8 protein on their cell surface, function by killing infected cells, which limits further spreading of the pathogen. The main task of helper T cells, that express the CD4 protein on the cell surface, is to stimulate the action of other immune cells, such as B cells. While most expanded T cells die after pathogen clearance, a subpopulation of specific cells remains. These cells adopt a memory T-cell phenotype and enable a quicker response upon repeated exposure to the same antigen. Thus, the memory of the adaptive immune system relies on activation and expansion of cells with a receptor that makes them specific for certain antigens.

In order to prevent large infection by a virus or bacteria, the adaptive immune system should respond to most harmful foreign antigens. At the same time, a response against proteins from the host's body can be detrimental, as this may lead to autoimmunity. Thus, the T cells in the body need to be very specific, each binding only a small set of the total diversity of possible antigens (Borghans *et al.*, 1999). In addition, as there are many different foreign antigens, there should be a wide diversity of cells, each with such a specific receptor. In the case of T cells, their TCR is very specific and is generated by a process called V(D)J recombination (Hayday *et al.*, 1985). This process happens in the thymus and generates the α and β chains, or the γ and δ chains, that form the TCR of $\alpha\beta$ and $\gamma\delta$ T cells, respectively. Here, we focus on the TCRs of the $\alpha\beta$ T cells, comprising the vast majority of the T-cell pool in humans (Roden *et al.*, 2008) and mice (Castillo-González *et al.*, 2021).

Generation and selection of the TCR repertoire in the thymus

The α chain of the αβ TCR is composed of several parts, specifically the Variable (V) and Joining (J) gene segments and the constant (C) region. The TRA gene locus contains an array of many V and J segments. During recombination, the activity of the recombination activating gene proteins RAG1 and RAG2 fuses one such V segment to a J segment. Diversity not only arises due to the many combinations of V and J segments, but also by deletions that can occur at both sides of the junction where the segments are joined, and the non-templated (N) nucleotides that can be inserted at the junction. The enzyme involved in the insertion of N nucleotides is terminal deoxynucleotidyl transferase (TdT), which is downregulated during early development in humans (Benedict *et al.*, 2000) and mice (Gregoire *et al.*, 1979). The steps of the VJ recombination process are stochastic and yield a rearranged TRA gene that is made up of the V gene segment, potentially a stretch of random N nucleotides, the J gene segment, and the C region, which together code for the TCR α chain. The β chain of the αβ TCR is generated in a similar process, but typically involves VDJ recombination, as the TRB gene locus contains an additional Diversity (D) gene segment that is recombined between the V and J segments. So, in the β chain there are two junctions at which nucleotides can be deleted and/or N nucleotides can be inserted. Thus, the TCR specificity of αβ T cells follows from a stochastic recombination process that happens in the thymus before they enter the periphery.

Since there are many combinations of V, D, and J segments, many ways in which the random deletions and insertions can affect the junctions, and many combinations of α and β chains, the potential αβ-TCR diversity is extremely high. Theoretical estimates of this number are in the order of 10^{61} (Mora and Walczak, 2018), which vastly outnumbers the 10^{12} T cells in a human body (Jenkins *et al.*, 2010; Trepel, 1974). Hence, only a small subset of the potential TCR diversity can be produced during the lifetime of an individual. It should be noted that the deletions and N-nucleotides at the junctions often lead to a frameshift, after which the sequence encoded by the V and J segments are no longer in frame, or lead to premature stop codons. The resulting TCR would not be functional in such cases. Cells in which this happens may still generate a functional TCR, as they can undergo V(D)J-recombination on both chromosomes containing the TRA and TRB gene loci. Allelic exclusion mechanisms exist to prevent the recombination of two functional TCR β chains in a single cell (Khor and Sleckman, 2002). These mechanisms are less complete for the TCR α chain, although additional mechanisms result in phenotypic allelic exclusion (Alam and Gascoigne, 1998; Gascoigne and Alam, 1999; Niederberger *et al.*, 2003), such that most T cells in the periphery have a single TCR that makes them specific to antigen (Brady *et al.*, 2010).

TCRs that are generated during V(D)J-recombination may strongly bind to peptides derived from self-proteins and would thus potentially result in autoimmunity. This problem is accounted for during positive and negative selection in the thymus. Shortly, the TCRs of newly generated T cells are tested based on their affinity for self-peptide/MHC complexes. During this positive selection step, the cells that get selected by binding peptides presented by MHC class I and II become CD8 and CD4 T cells, respectively. If the TCR does not have sufficient affinity to any of these presented self-peptides, they do not get a survival signal and 'die by neglect'. Afterwards, the negative selection process eliminates those T cells that bind self-peptide/MHC complexes with too high affinity. A subset of the CD4 cells, that binds self-peptides with high affinity but survives negative selection, adopts a regulatory T cell phenotype that is characterised by expression of the forkhead box P3 (FOXP3) protein. Regulatory T cells suppress effector T cells to maintain tolerance to self-antigens, and prevent autoimmune disease. In short, the selection process in the thymus selects for TCRs that bind peptides presented in the context of the host's MHC with the right affinity and determines the lineage choice of the newly generated T cells.

The MHC genes are among the most polymorphic in humans and mice (Roy *et al.*, 1989; Trowsdale and Knight, 2013). This means that the subset of the TCRs that can survive thymic selection is different for nearly every individual, unless they are genetically identical such as identical twins or inbred mice. The MHC diversity, the small fraction of the potential TCR diversity that is stochastically produced during a lifetime, and the infection history all contribute to the unique composition of T cells and their TCRs in an individual. The collection of TCRs in an individual is called the TCR repertoire and can function as a personal 'immune fingerprint' (Dupic *et al.*, 2021) that reflects the production, selection, and expansion of T cells in that individual. The insights from characterising the TCR repertoire can thus be of great value to address outstanding questions in T-cell immunology.

TCR sequencing quantitatively characterises the TCR repertoire in a sample

High-throughput sequencing (HTS) is nowadays the most common method for detailed and quantitative characterisation of the TCR repertoire in a sample of cells of interest (Rosati *et al.*, 2017). These techniques have the advantage that the full nucleotide sequence coding for the TCR α and/or β chain can be identified. The complementarity-determining region 3 (CDR3) of these chains is of special interest, as this part covers the V(D)J junctions, which is the most variable domain of the TCR. This means that the CDR3 is most informative about the TCR's specificity, as this region comes into close contact with the antigen (Garcia and Adams, 2005). Alternative methods to assess the TCR repertoire include staining of T cells with monoclonal antibodies, that specifically bind one or more V segments, or by spectratyping, which provides insight into the CDR3-length distribution of the TCRs in the sample. The advances in HTS techniques, especially in the last decade, now enable detailed characterisation of the TCR repertoire of thousands or even millions of cells in a single experiment.

Several methods exist to perform sequencing of the TCR repertoire. Experiments vary for example in starting material, amplification methods and the number of cells that are analysed together. Which method is ideal for a given experiment depends on the research question, and specifically to which extent the TCR information of individual cells needs to be determined. The most detailed characterisation of T cells is achieved by single-cell immune profiling. These techniques generally label the genetic material of a single cell, such that all information can be traced back to the level of individual cells after sequencing. This allows for paired identification of the TCR α and β chains of a T cell, and it is even possible to also measure the gene expression and cell surface protein levels of these individual cells. A disadvantage of single-cell sequencing techniques is the relatively low number of cells that can be analysed in an experiment, such that most of the TCR diversity remains unnoticed (Rosati *et al.*, 2017).

Bulk sequencing methods are an affordable alternative to study the TCR repertoire of larger populations of cells, although they provide a lower level of detail than single-cell profiling. The most relevant difference with respect to the TCR identification is that although both the α and β chain of the TCR can be sequenced, the pairing between each TCR α and β chain remains unknown. The genetic material that is analysed in a bulk HTS experiment can be either the genomic DNA or the RNA coding for the TCR chains. The specific amplification of the DNA coding for the TCR α and β chain requires a multiplex polymerase chain reaction (PCR), which uses a diverse panel of primers binding to the diverse V and J segments (Rosati *et al.*, 2017). Multiplex PCR methods can also be used when RNA is used as starting material, but in general have the disadvantage that they are subject to amplification biases. This means that the V segments with more efficient amplification become over-represented in the resulting products (Benichou *et al.*, 2012). Such biases change the relative abundances compared to the original distribution of cells. An alternative approach overcoming this issue is the use of rapid amplification of 5' complementary DNA ends (5'RACE), that use RNA samples as a starting material (Mamedov *et al.*, 2013). Most TCR sequencing data analysed in this thesis is acquired using this method. Hence, we provide a short overview of the experimental steps that are necessary to sequence the TCR repertoire from a population of cells using the 5'RACE method (Rosati *et al.*, 2017; Heather *et al.*, 2018; Oakes *et al.*, 2017).

A typical TCR repertoire HTS experiment starts with isolation of the cells of interest, usually from a blood sample. These could be general peripheral blood mononuclear cells (PBMCs) but also a specific subpopulation, such as the naive CD8 T cells that are sorted based on subset-specific cell surface markers. Cells are lysed and the 5'RACE method is used for reverse transcription, to obtain a complementary DNA (cDNA) molecule that covers the entire variable region of the TCR sequence. At this stage, a unique molecular identifier (UMI) is incorporated into the sequence. This UMI consists of typically 12 random nucleotides, such that each cDNA molecule is labelled with a genetic barcode. The UMIs are used to trace multiple reads in the eventual sequencing data back to individual cDNA molecules, as they uniquely identify the cDNA molecules prior to amplification. The barcoded cDNA molecules are amplified by a second PCR, and sequencing adaptors are ligated to allow for sequencing of these amplicons. Sequencing is usually done on an Illumina platform that performs sequencing by synthesis with fluorescently labelled nucleotides. This process typically generates millions of sequence reads that should cover enough of the TCR V(D)J region to identify the full nucleotide sequence of the TCR α or β chain. Thus, the sequencing data allows for quantitative characterisation of the TCR repertoire.

TCR repertoire analysis is required to translate data into T-cell immunology insights

Although TCR sequence data contains a wealth of information, its interpretation poses a bioinformatic challenge. It is inevitable the sequence reads contain errors that arose during the PCR amplification and/or sequencing. Thus, small differences in the hypervariable CDR3 region between various reads may arise from truly different TCR sequences, but also from the same TCR sequences in which errors were introduced, or a combination of both. Dedicated TCR-analysis pipelines are designed to overcome this challenge of discriminating between true and error-based TCR diversity. When sequencing UMI-labelled amplicons, the UMI sequences greatly assist such error correction. Typically, reads sharing one particular UMI sequence cover the TCR information of a single cDNA molecule. Since most errors are introduced during the later cycles of the PCR and the sequencing reaction, the base call supported by the largest number of reads is often the true nucleotide at each position (Heather *et al.*, 2018; Oakes *et al.*, 2017). Thus, generating consensus sequences from the reads that share their UMI sequence allows for elimination of many errors in the sequencing data. In addition, since reads that share their UMI sequence are collapsed to a single consensus sequence, the biases introduced by (unequal) PCR amplification are accounted for. However, note that UMIs can also accumulate errors in their sequence (see also Chapter 2).

The various tools that can be used to analyse and error-correct TCR sequencing data vary in their exact algorithms, but generally consist of three major steps. First, as described above, the generation of consensus sequences by merging reads that share their UMI sequence. Second, identification of the V and J gene segments that are used in each of those consensus sequences and subsequently the nucleotide sequence of the CDR3. These three elements together determine the entire variable part of the TCR sequence. The third step involves the comparison of the identified TCR sequences to each other and performing error-correction by merging unexpectedly similar sequences. The resulting output generally consists of a table with each identified TCR sequence expressed in terms of V and J segments and the CDR3 sequence, together with its abundance in the data. The abundance corresponds to the number of reads the TCR was identified in, or the number of distinct UMIs the TCRs was identified with in the case of barcoded sequencing libraries.

Three different TCR analysis pipelines are used in this thesis. In the case of MiGEC and MiXCR (Bolotin *et al.*, 2015), MiGEC is used to extract UMI sequences from sequence reads and generate UMI-based consensus sequences. The coverage of these consensus sequences determines which of them will be accepted, meaning that sequences with support of only a few reads are discarded. MiXCR is then used to identify and error-correct the TCR sequences in this data, to an extent that the user can determine by changing specific parameters. Another pipeline, RecoverTCR (RTCR) (Gerritsen *et al.*, 2016), takes along all (UMI-based consensus) sequences, and estimates the error rate in a data set from the alignment to the V and J gene segment sequences. This data-driven error rate is then used to perform various clustering steps to correct the expected errors in the CDR3 region. Decombinator

(Thomas *et al.*, 2013) also performs the general steps, but in a different order. It starts with identification (decombining) of V and J segments in individual reads, as well as the CDR3. The reads are then collapsed based on the corresponding UMI sequence, allowing for error correction. Thus, although the main steps of various TCR analysis pipelines are similar, the output will differ due to the differences in selection and correction of the reads.

The resulting data can be analysed in many ways, depending on the research question. Basic analyses involve the relative usage of certain V and J segments, or distribution of CDR3 lengths. Note that such analyses are analogous to the older methods of staining with V segment-specific antibodies and spectratyping but provide a higher resolution and more details about the TCR sequences with certain characteristics. The abundance of the TCR sequences in the data set provides information about the distribution of the TCR α and β chains among the cells in the sample. These distributions give insight into the occurrence of cells sharing one or both of their TCR chains, which are often referred to as clones or clonotypes. For example, one can estimate the distribution of clonal sizes in the T-cell pool, and related to this, the diversity of the TCR repertoire. Importantly, a typical sample used for HTS only comprises about 10^6 cells from the total 10^{12} T cells in a human body. This means that small biases in the sample, introduced during the experiment or processing of the data, can have major effects when extrapolated to a pool level. For example, when RNA is used as the starting material for sequencing, single cells may contribute multiple molecules that are each labelled with a distinct UMI (Rosati *et al.*, 2017). As a result, different TCR expression levels between cells, and the stochastic sampling of the RNA molecules from the cells, can easily affect the relative frequency of TCR chains in the data. Thus, the interpretation of TCR abundance in sequencing data requires careful analysis.

Although the V(D)J-recombination mechanism generating the TCR diversity is well studied, one cannot reliably infer the recombination scenario from the TCR chain sequence (Marcou *et al.*, 2018; Murugan *et al.*, 2012). For example, there are many different combinations of deletions and N additions that together generate the exact same sequence. To still obtain insight into the probabilities involved in these stochastic processes, one can use probabilistic modelling to infer recombination models using tools like IGoR (Marcou *et al.*, 2018) and OLGA (Sethna *et al.*, 2019). The abundance of most TCR chains in data is affected by their probability to result from V(D)J recombination, but also by selection in the thymus and the periphery (Elhanati *et al.*, 2014; Sethna *et al.*, 2020). The recombination models are therefore trained on the non-functional sequences that contain a frameshift or a premature stop codon. Such TCR sequences do not code for a productive TCR chain but can still be present if the locus on the other chromosome codes for a TCR that allowed the cell to pass selection. The non-functional TCR sequence is not affecting the selection of the T cell and can thus be used for training of the recombination model, without biases due to selection (Marcou *et al.*, 2018). The resulting models can be used to calculate the generation probability for any given TCR sequence, which quantifies how likely this sequence is the result of V(D)J recombination. TCR sequences differ many orders of magnitude in their generation probability, for example since the insertion of a specific long stretch of N additions is a scenario that is unlikely to happen repeatedly. The trained recombination models are to a large extent similar between people, implying that the generation probability of a given TCR sequence is similar between different individuals. Interestingly, this explains to a large extent the observation of 'public' sequences: TCR chain sequences that are more shared between individuals than others appear to be strongly enriched for high generation probabilities (Elhanati *et al.*, 2018). This example shows that models can help to interpret and extrapolate the findings of TCR repertoire sequencing.

About this thesis

The general aim of the studies described in this thesis is to answer biological questions using analyses of TCR repertoire sequencing data. In **Chapter 2** we devise various models for the distribution of TCR clone sizes in the human naive T cell pool. We characterise the repertoire of TCR α and β chains by HTS of various cell subsets. Comparing the model predictions with the experimental outcomes, we find solid evidence for the presence of very large TCRαβ clones in the naive T cell repertoire. These large naive T-cell clones are only partly explained by their generation probability. **Chapter 3** aims to identify sequence characteristics of such abundant naive T cell clones. We find that the D segment is missing in a substantial fraction of the abundant TCR β sequences in the naive T cell repertoire of young individuals. Such sequences appear to be mostly generated before birth, to persist over a human lifetime, and, as a result, to be excessively shared between individuals. **Chapter 4** is a short report on quantifying the effects of age on the TCR repertoire. We present example analyses and discuss potential pitfalls of assessing ageing effects on the TCR repertoire. **Chapter 5** describes a pilot study on using TCR sequencing to follow the T-cell response upon pneumococcal conjugate vaccination. We define quantitative requirements to classify T-cell expansions but detect these in only a minority of the donors. Our analysis suggests that the vaccine-induced T-cell response is small and/or very broad, and highlights experimental requirements to characterise such responses in future studies.

A key limitation of the studies in humans introduced above is the limited number of cells that can be analysed compared to the total size of the T-cell pool. As a result, many of the TCR chains that make up the diversity of the TCR repertoire will be missed in any analysis. Bulk sequencing methods also do not uncover the full TCR diversity present in an individual as the information on the pairing of α and β chains is not obtained. In **Chapter 6** we overcome both limitations by studying the TCR repertoire of one-TCRa mice that have a strongly reduced receptor diversity. This allows us to study the nearly complete TCR diversity, and compare the repertoire among individual lymph nodes and mice, revealing a spatial organisation of the TCR repertoire. **Chapter 7** concludes the thesis by discussing the main findings and highlighting some promising avenues for future work.

2

The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes

Peter C. de Greef [1]*, Theres Oakes [2]*, Bram Gerritsen [1,3]*, Mazlina Ismail [2], James M. Heather [2], Rutger Hermsen [1], Benjamin Chain [2], and Rob J. de Boer [1]

[1] *Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands*
[2] *Division of Infection and Immunity, University College London, United Kingdom*
[3] *Department of Pathology, Yale School of Medicine, United States*
* *These authors contributed equally to this work*

Elife 9 (2020): e49900.

Abstract

The clone size distribution of the human naive T-cell receptor (TCR) repertoire is an important determinant of adaptive immunity. We estimated the abundance of TCR sequences in samples of naive T cells from blood using an accurate quantitative sequencing protocol. We observe most TCR sequences only once, consistent with the enormous diversity of the repertoire. However, a substantial number of sequences were observed multiple times. We detect abundant TCR sequences even after exclusion of methodological confounders such as sort contamination, and multiple mRNA sampling from the same cell. By combining experimental data with predictions from models we describe two mechanisms contributing to TCR sequence abundance. TCRα abundant sequences can be primarily attributed to many identical recombination events in different cells, while abundant TCRβ sequences are primarily derived from large clones, which make up a small percentage of the naive repertoire, and could be established early in the development of the T-cell repertoire.

Introduction

The human adaptive immune system employs a vast number ($> 10^{11}$ (Clark *et al.*, 1999)) of T lymphocytes, to detect and control pathogens. Most T cells express a single T-cell receptor (TCR) variant, which binds antigen in the form of a short peptide presented 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 ($> 20^9$) a specific TCR is nevertheless expected to bind many different pMHC (i.e., cross-reactivity) (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 $\alpha\beta$ TCRs occurs in the thymus, where thymocytes randomly rearrange and imprecisely recombine gene segments to create a complete receptor (Nikolich-Žugich *et al.*, 2004). This heterodimer is generated by random recombination of Variable, Diversity, and Joining (V, D and J) segments for TCRβ, and V and J segments for TCRα sequences (Davis and Bjorkman, 1988). Most variability arises due to random nucleotide insertions and deletions where the segments are joined (Murugan *et al.*, 2012). Recent estimates of the potential number of TCRs produced by this V(D)J-recombination process range from $> 10^{20}$ (Zarnitsyna *et al.*, 2013) to 10^{61} (Mora and Walczak, 2018), which vastly outnumbers the number of distinct TCRs present in a human body. After generation of the TCR, T cells undergo positive and negative selection, which selects those T cells that have sufficient, but not too high, affinity for any self-pMHC (McDonald *et al.*, 2015). About 3-5% of thymocytes survive selection (Merkenschlager *et al.*, 1997) and enter the periphery as T cells that have not yet encountered foreign cognate antigen, i.e., as naive T cells.

The thymic output of new T cells decreases because of thymic involution, making peripheral division of existing cells the main source of naive T cells from early adulthood onwards in humans (den Braber *et al.*, 2012; Kumar *et al.*, 2018). In the periphery, naive T cells compete for cytokines, such as IL-7, and need to interact with self-pMHC to survive (Tanchot *et al.*, 1997; Takada and Jameson, 2009; Jenkins *et al.*, 2009). Competition between T-cell specificities may reduce repertoire diversity when cells with some TCRs outcompete others (de Boer and Perelson, 1994), resulting in differences in TCR frequencies, and heterogeneous naive T-cell clone sizes. Experimental evidence for large heterogeneity in division and survival rates within the naive T-cell pool has been shown in mice (Hogan *et al.*, 2015; Rane *et al.*, 2018; Reynaldi *et al.*, 2019). Such experiments are not feasible in humans, but mathematical modelling has been used to assess how fitness differences between T-cell clones may affect the frequency of clones in the naive repertoire (Stirk *et al.*, 2008, 2010; Hapuarachchi *et al.*, 2013; Lythe *et al.*, 2016; Desponds *et al.*, 2016, 2021; Dowling and Hodgkin, 2009; Johnson *et al.*, 2012).

Measuring the distribution of TCRα and TCRβ sequences in samples of naive T cells can inform us about the clone-size distribution of the naive T-cell repertoire. Previous studies have reported large heterogeneity in the frequency of TCRβ sequences in naive repertoires from mice (Quigley *et al.*, 2010) and humans (Robins *et al.*, 2009; Venturi *et al.*, 2011; Qi *et al.*, 2014; Pogorelyy *et al.*, 2017). One important factor shaping the abundance of TCR sequences is their likelihood to be produced during VDJ-recombination. Rearrangements with less N-insertions, for example, tend to be more commonly observed (Robins *et al.*, 2009, 2010; Venturi *et al.*, 2011; Pogorelyy *et al.*, 2017). To study this in more detail, the Mora and Walczak groups developed probabilistic models that predict the generation probability of any specific TCRα or TCRβ sequence (Murugan *et al.*, 2012; Marcou *et al.*, 2018). They showed that these sequences (σ) differ by several orders of magnitude in their probability $P(\sigma)$ of being produced by V(D)J recombination in the thymus. Differential generation probabilities do not only impact the abundance of TCRα and TCRβ sequences within an individual, but also contribute to sharing among individuals (Robins *et al.*, 2010; Quigley *et al.*, 2010; Venturi *et al.*, 2011; Qi *et al.*, 2014; Pogorelyy *et al.*, 2017; Elhanati *et al.*, 2018). Hence, it is essential to take the likelihood of generating a sequence into account when interpreting sequencing data of immune repertoires.

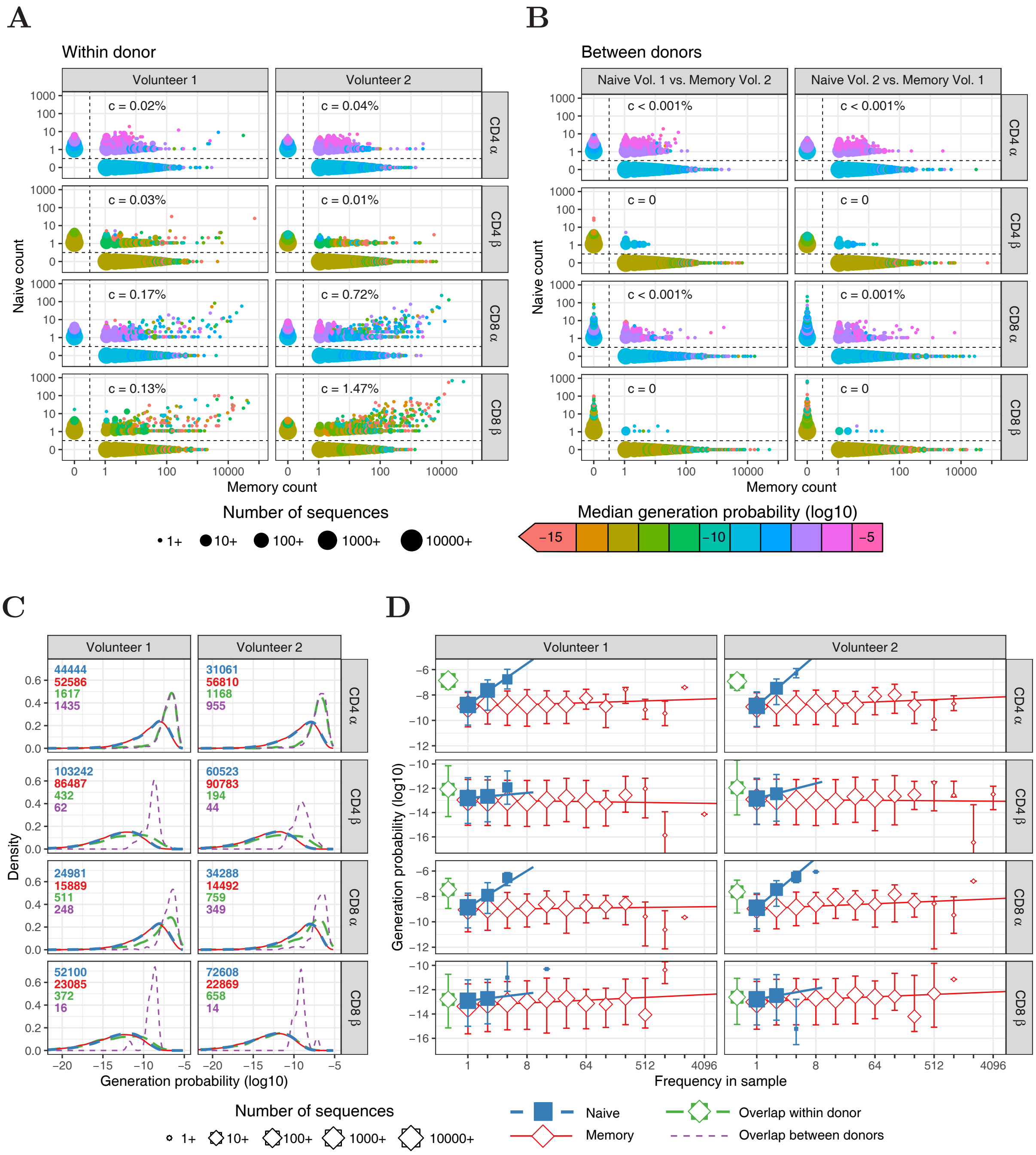
In this study, we characterise the frequency distribution of TCRα and TCRβ sequences in the naive repertoire. We analyse published and new experimental data on both the TCR α and β chain, and combine a quantitative unique molecular identifier (UMI)-based TCR sequencing pipeline with mathematical modelling to consider carefully the contributions of different mechanisms that may lead to observed abundant TCRα and TCRβ sequences in the naive repertoire. Such mechanisms include experimental confounders, such as the purity of the cell populations and repeated sampling of mRNA from the same cell, and diverse biological processes including distinguishing carefully between repeat generation of identical sequences in different cells, and large naive T-cell clones. We show that all these processes are likely to contribute to the observed abundance profile of TCR sequences in samples of naive repertoires. In particular, even after all other mechanisms are accounted for, we find evidence for naive T-cell clone size heterogeneity. Specifically, the results are compatible with an underlying power-law distribution of naive T-cell clone sizes (Desponds *et al.*, 2016), or more generally by models in which 1-5% of naive T cells represent large clones of 10^5 - 10^6 cells. Preferential expansion of some clonotypes, perhaps those occurring early in development of adaptive immunity, therefore plays an important role in shaping the naive T-cell repertoire.

Results

We analysed the frequency distribution of TCR sequences in the naive T-cell compartment, using TCRα and TCRβ sequences published in (Oakes *et al.*, 2017). In brief, peripheral blood mononuclear cells (PBMCs) from two adult volunteers were FACS-sorted into naive (CD27$^+$CD45RA^{high}) and various memory CD4$^+$ and CD8$^+$ populations. TCRα and TCRβ mRNA was reverse transcribed to cDNA molecules to which unique molecular identifiers (UMIs) were attached, followed by PCR-amplification and high-throughput sequencing (HTS) on an Illumina MiSeq platform. We refer to this as experiment 1 below (further details in Materials and Methods). Sequence reads were processed using a customised version of the Decombinator pipeline (Thomas *et al.*, 2013), with an improved error correction on UMIs to more reliably estimate the frequency of nucleotide TCRα and TCRβ sequences in the samples (Figure 2.6). Additionally, we used the RTCR pipeline (Gerritsen *et al.*, 2016) for comparison. The different memory populations were combined for the purpose of the analysis presented below.

Abundant TCR sequences are frequently shared between naive and memory populations, and are enriched for high VDJ recombination probabilities

Within the naive T-cell repertoires, the vast majority of TCRα and TCRβ sequences were observed only once, and most frequencies fall within the range from 1 to 5 (Figure 2.1A). As expected, in the memory repertoires, which contain clonally expanded T cells, much more abundant sequences were present, with a substantial number of α and β chains observed more than 1000 times (Figure 2.1A). The few sequences observed with a frequency higher than 5 in the naive samples were shared in most cases (94.6%) with the corresponding memory subset from the same individual. We examined whether this overlap might arise from imperfect sorting of the T-cell populations, despite the tight non-overlapping sort gates applied (see (Oakes *et al.*, 2017)). A prediction of such sorting contamination is that the abundance of the shared TCR sequences in the naive and memory repertoires should be proportional. Such a linear relationship could be observed clearly for CD8$^+$ TCRα and TCRβ sequences (Figure 2.1A), especially for memory abundances greater than 1000. Correlation measurements suggested that the amount of contamination for CD8$^+$ T cells was 0.1 - 1.5%. As expected, no correlation was observed between the abundance of TCR sequences shared between naive and memory populations of different donors (Figure 2.1B).


A
Within donor
Volunteer 1
Volunteer 2
c = 0.02%
c = 0.04%
c = 0.03%
c = 0.01%
c = 0.17%
c = 0.72%
c = 0.13%
c = 1.47%
CD4 α
CD4 β
CD8 α
CD8 β
Naive count
1000
100
10
1
0
Memory count
0
1
100
10000
B
Between donors
Naive Vol. 1 vs. Memory Vol. 2
Naive Vol. 2 vs. Memory Vol. 1
c < 0.001%
c < 0.001%
c = 0
c = 0
c < 0.001%
c = 0.001%
c = 0
c = 0
Number of sequences
1+
10+
100+
1000+
10000+
Median generation probability (log10)
−15
−10
−5
C
Volunteer 1
Volunteer 2
44444
52586
1617
1435
31061
56810
1168
955
103242
86487
432
62
60523
90783
194
44
24981
15889
511
248
34288
14492
759
349
52100
23085
372
16
72608
22869
658
14
Density
0.6
0.4
0.2
0.0
−20
−15
−10
−5
Generation probability (log10)
D
Volunteer 1
Volunteer 2
Generation probability (log10)
−6
−8
−10
−12
−14
−16
1
8
64
512
4096
Frequency in sample
Number of sequences
1+
10+
100+
1000+
10000+
Naive
Memory
Overlap within donor
Overlap between donors

We next examined the relationship between VDJ recombination probabilities and the overlap between naive and memory repertoires. Using the V(D)J-recombination model (Marcou *et al.*, 2018), we predicted the generation probabilities *P(σ)* of all TCRα and TCRβ sequences in our datasets. As expected, we observed a wide range of *P(σ)* values, which were several orders of magnitude higher for TCRα sequences than TCRβ, due to additional recombination of the D segment. The generation probability distributions of sequences derived from naive and memory T cells were indistinguishable (Figure 2.1C, blue and red, respectively). Thus, our data provide no evidence that the V(D)J-recombination process preferentially produces sequences chains that are more likely to enter the memory pool during an immune response. However, TCRα sequences shared between memory and the corresponding naive samples, were strikingly enriched for high *P(σ)* (Figure 2.1C, green). This enrichment is much less evident for TCRβ sequences. The enrichment for sequences with high *P(σ)* in the population of shared memory/naive TCRα is not compatible with overlap derived from contamination during cell sorting, but rather suggests that the sharing may also arise from T cells which use the same TCRα because of identical VJ recombination events in different T cells. It is important to stress that, since such different T cells are highly unlikely to also share TCRβ sequences, the clonotype, and hence specificity of the T cells in the naive and memory compartments may well be different, despite sharing TCRα sequences.

As a control, we also analysed overlap between the naive sample from one volunteer and the memory sample from the other. In this case, sort contamination of naive repertoires by memory T cells is excluded and a shared sequence can only result from independent identical recombination events, from distinct T-cell clones. For CD4$^+$ cells, we find that

Figure 2.1 ***(preceding page)*** **– Frequencies and generation probabilities of TCRα and TCRβ sequences from memory and naive T cells.** **A.** Frequency of TCRα and TCRβ sequences in naive versus total frequency in memory repertoires sampled from the same volunteer. Symbol sizes represent number of sequences with these frequencies and colour represents their median generation probability *P(σ)*, as determined using IGoR (Marcou *et al.*, 2018). The c value is the slope of linear regression on sequences with a memory count > 100 and indicates the estimated probability that a given TCR sequence from a memory cell appears in the naive sample. **B.** As A., but comparing frequency in naive sample from one volunteer with frequency in memory from the other volunteer. **C.** Distributions of generation probabilities (log10) for TCR α and β sequences from CD4$^+$ and CD8$^+$ from two volunteers. Blue dashed: naive, red solid: memory, green long-dashed: overlap (i.e., sequences observed in both naive and memory within a volunteer), purple dashed: overlap between volunteers (i.e., sequences observed in the naive subset of Volunteer 1 and a memory subset of Volunteer 2, or vice versa). The total number of sequences for each group are indicated in corresponding colours. **D.** The median *P(σ)* is shown for each observed frequency class (log2 bins) of sequences exclusively observed in naive (blue squares) or memory T-cell (red diamonds) samples. *P(σ)* of the overlapping chains is shown in green for reference (irrespective of frequency). Symbol sizes indicate numbers of sequences for each frequency class. Error bars represent the 25% and 75% quartiles, solid lines indicate linear regression between observed frequency and *P(σ)*, weighted by the number of sequences with that frequency.

the number of TCRα sequences shared between naive and memory is similar between and within volunteers, and that the $P(\sigma)$ distribution is nearly identical (Figure 2.1C, purple). For CD8$^+$ cells, the number of sequences shared within an individual is somewhat larger than between individuals, compatible with some degree of sort contamination in this population as discussed above. The small number of TCRβ sequences shared between individuals also had a relatively high $P(\sigma)$, although considerably smaller than for TCRα.

In summary, although contamination with abundant memory T cells may make a small contribution to the TCR sequences which are found in both naive and memory for CD8$^+$ cells, multiple identical recombinations arising from high $P(\sigma)$ values is the dominant mechanism leading to overlap in the TCRα repertoires. Nevertheless, in order to stringently exclude any possible contribution of contamination, we included an analysis which excluded all the shared sequences from the further investigations of the relationship between TCR sequence abundance and $P(\sigma)$ (Figure 2.1D).

The abundances of sequences in all naive repertoires were correlated to $P(\sigma)$ (Figure 2.1D, blue). The median $P(\sigma)$ of the α chains that were observed at least three times was about 154-fold higher than for those that have only been observed once ($p < 10^{-15}$, Wilcoxon test). The enrichment for high $P(\sigma)$ in more abundant TCR sequences was weaker for TCRβ (~2.5-fold, $p < 0.01$, Wilcoxon test), but still stronger than for memory subsets (1.65- and 1.03-fold for TCRα and TCRβ, respectively, $p < 10^{-15}$ and $p = 0.27$). In line with this, the number of N-additions tended to be lower for TCRα and TCRβ sequences abundant in the naive samples (Figure S2.1). These correlations suggest that multiple identical recombination events which occur during formation of the naive T-cell repertoire in the thymus due to high generation probabilities, contributes to the observation of abundant TCR sequences. This is especially evident for TCRα, where the probabilities of producing a given sequence is higher because of the absence of a D region. However, abundant TCR sequences with low $P(\sigma)$ are also observed, especially for TCRβ, leaving open the possibility of large naive T-cell clones.

Frequently observed TCR sequences cannot be attributed only to multiple RNA molecules per cell

T cells contain on average in the order of 100 molecules of TCRα and 300 molecules of TCRβ mRNA (Oakes *et al.*, 2017). Because the TCR sequencing pipeline is not 100% efficient, only a small proportion of these molecules are actually sequenced, but the possibility remains that TCR sequences observed multiple times may be due to repeat sampling from the same cell. Because the variance of this number remains undetermined, it is difficult to computationally determine the contribution of this multiple sampling to the data. Instead, we performed an additional experiment (referred to as experiment 2) in which we sorted naive T cells from an additional volunteer, and split the naive T cells into three subsamples before mRNA extraction. We then carried out library preparations and sequenced TCRα and TCRβ sequences from each subsample independently. In this experiment, sequences observed in more than one subsample must have been derived from different cells, and cannot be a result of sequencing multiple mRNA molecules from a single cell. Repeated

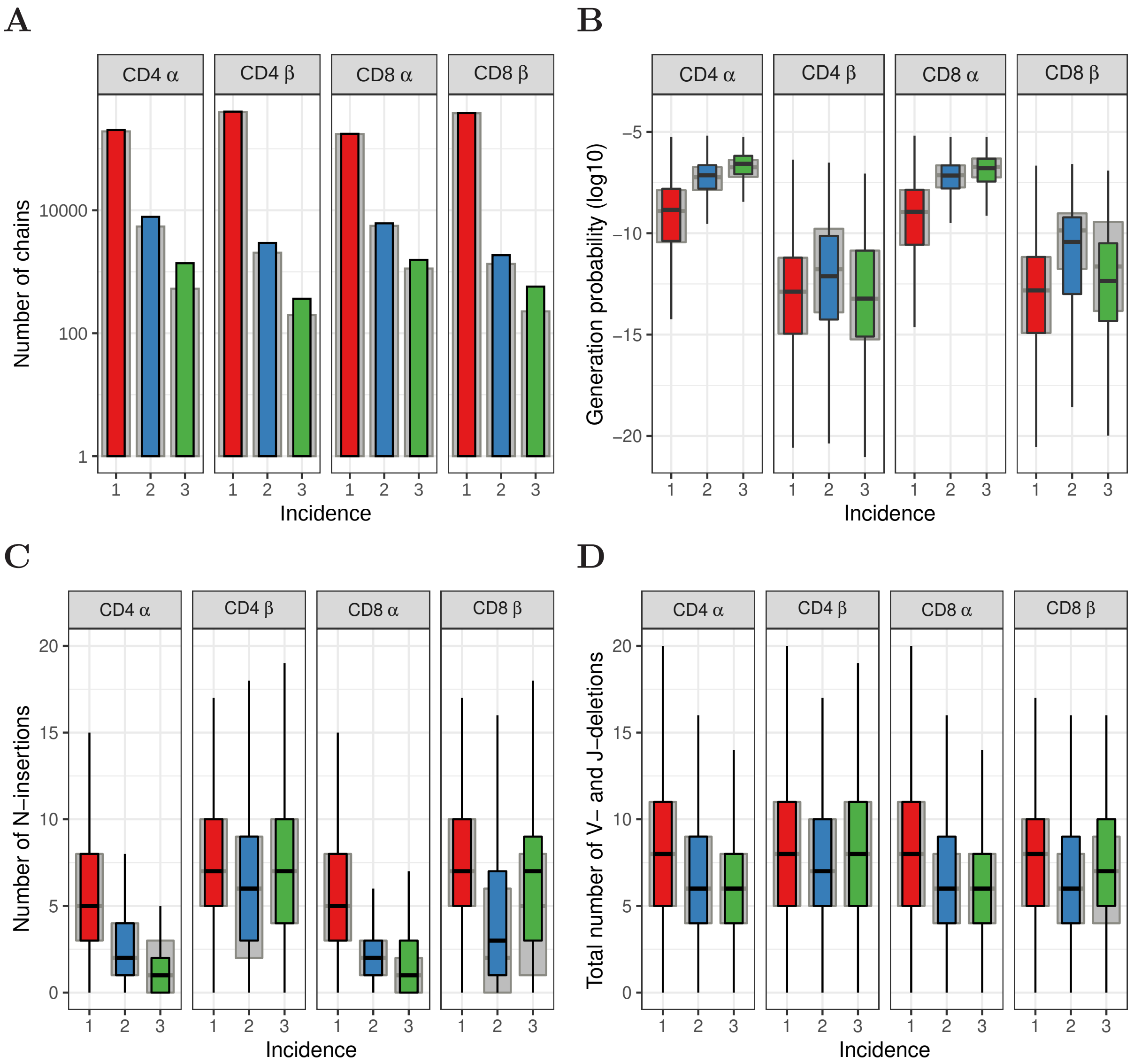


Figure 2.2 – **Subsampling naive T cells confirms that frequently observed TCRα but not TCRβ sequences have high generation probabilities.** **A.** The number of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples (experiment 2). The grey background bars show the results after removing all sequences that were also observed in the corresponding memory samples. **B.** Generation probabilities *P(σ)* (log10) of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. **C.** Minimal number of N-additions of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. **D.** Number of V- and J-deletions of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. The plot shows median (black horizontal line), interquartile range (filled bar) and the range from the bar up to 1.5 times the interquartile range (black vertical range, outliers not shown).

sequences must therefore derive from different cells, and represent abundant sequences.

In total 16913 (3.4%) TCRα sequences, and 5744 (0.61%) TCRβ sequences, were observed in more than one subsample (Figure 2.2A), confirming the existence of a substantial number of frequent TCR α and β chains in the full naive repertoire. In order to exclude any contribution from sort contamination, we also plot the data after removing all TCR sequences found in both memory and naive repertoires (Figure 2.2A, grey bars). A substantial number of α and β chains were still found in multiple subsamples. In order to estimate the impact of multiple sampling on the observed abundances we randomly permuted the TCR sequences between subsamples, and reanalysed the distributions (see Materials and Methods). We estimated that ~ 25% of α and > 75% of β chains with an abundance of greater than 1 in an individual sample may arise from sampling multiple RNA molecules from single cells. The impact is strongest on TCR sequences observed twice (see Figure S2.3). Thus multiple mRNA sampling is an important confounder of estimating TCR sequence abundances in individual repertoires, especially for TCRβ.

Having ruled out the contribution of multiple mRNA sampling experimentally, we examined the relationship between TCR sequence abundance and $P(\sigma)$ in this new data set. The TCRα chains present in more than one naive subsample are dominated by sequences with high $P(\sigma)$. The median generation probability of TCRα sequences observed in two and three subsamples was 56- and 165-fold higher, respectively, than those observed only once (Figure 2.2B). The relationship for TCRβ sequences was remarkably different, however. While TCRβ sequences observed in two subsamples are mildly enriched for high generation probabilities, those observed in three subsamples have hardly any enrichment for high $P(\sigma)$ (Figure 2.2B). Instead, their generation probabilities tend to be lower than those of the sequences observed in two subsamples, and more similar to the generation probabilities of TCRβ sequences seen in only one subsample. We obtained similar results when measuring the number of N-additions and VJ-deletions in the rearrangements: abundant α chains (with incidence 2 or 3) tend be closer to germline rearrangements, while this was only the case for β chains with incidence 2, and not for the most abundant β chains with incidence 3 (Figure 2.2C&D). These trends were observed both with and without removing the sequences that were also observed in memory (Figure 2.2, grey versus coloured bars) and when processing the data with RTCR (Figure S2.5).

We further explored whether the more abundant sequences were also more "public" (found in the repertoires of multiple individuals), which would be predicted if they are more likely products of V(D)J-recombination. We measured the degree of sharing between those TCR sequences observed in 1, 2, or 3 naive subsamples, and the TCRα and TCRβ repertoires of unfractionated blood samples collected from 28 healthy donors. Both TCRα and TCRβ sequences observed in two or three subsamples were found to be significantly more often shared with this independent cohort than those observed once (Figure S2.4A). The most frequent TCRα sequences, which were seen in three subsamples, showed the highest sharing degree, consistent with their strongest enrichment for high generation probabilities. The relatively small number of most frequent TCRβ sequences (i.e., those observed in three subsamples), did not show increased inter-individual sharing compared to the TCRβ sequences observed in two subsamples. Additional comparison with publicly available TCRβ data from a large cohort (Emerson *et al.*, 2017) showed that the most frequently observed β chains, which were observed in all three subsets in experiment 2, were less public than sequences observed in two subsamples (Figure S2.4B). The seemingly paradoxical finding that the most abundant TCRβ sequences (observed in all three subsamples) have lower $P(\sigma)$, and are less public than those found twice, is explored in more detail below.

Computational models of TCR repertoire generation suggest the presence of a small proportion of large T-cell clones in the naive repertoire

In order to more rigorously test our ideas about the frequency distribution of clonotypes in the naive T-cell repertoire, we explored a number of possible computational models of repertoire generation and sampling, and compared model prediction with the experimental data discussed above. The first simplest scenario we considered was a neutral model of repertoire formation, similar to Hubbell's Neutral Community Model (Hubbell, 2001) (Figure 2.3A, details in Materials and Methods). The model assumes that there is no selective advantage of one TCR over another, and therefore the TCR of a naive T cell does not affect its lifespan or division rate. Consider a pool of N naive T cells, from which cells are removed by cell death or by priming with antigen, leading to differentiation into a memory population. A fraction θ of these cells is replaced by thymic production of new clones and the remaining fraction $1-\theta$ gets replaced by division of cells present in the pool. When simulating the naive T-cell pool with this model, the clone-size distribution approaches a "steady state" (not shown). We use this steady-state distribution, for which we have an analytical expression to predict the size of clones in the naive T-cell pool. As the contribution of thymic output decreases during ageing (Steinmann *et al.*, 1985), we evaluated the model for a wide range of values for θ. The clone-size distribution which emerges from the neutral model is approximately geometric for clone sizes larger than the introduction size c (Figure 2.3B). We compared this basic model to models in which we impose other distributions on the underlying clonotype abundances (model details in Materials and Methods). We specifically focused on heavy-tailed distributions such as log-normal and power-law distributions, which have previously been associated with T-cell repertoires (Desponds *et al.*, 2016). The shape of each of these distributions is controlled by a single parameter (as shown in Figure 2.3B), allowing us to compare distributions with different degrees of heterogeneity. In all cases, we normalised the clone-size distribution such that the total number of cells N is constant. Since we had separate experimental data for CD4^{+} and CD8^{+} cells, we considered CD4^{+} and CD8^{+} cells separately, setting $N_{CD4} = 7.5 \times 10^{10}$, and $N_{CD8} = 2.5 \times 10^{10}$.

From all model clone-size distributions we simulate 3 subsamples, so as to compare with the data from the second experiment described above. Each sampled TCR is assigned a TCRα and a TCRβ sequence that were generated with IGoR (Marcou *et al.*, 2018). Previous studies showed that α and β chains with higher generation probabilities tend to have a higher probability to survive selection (Elhanati *et al.*, 2014). Therefore, we train a simple

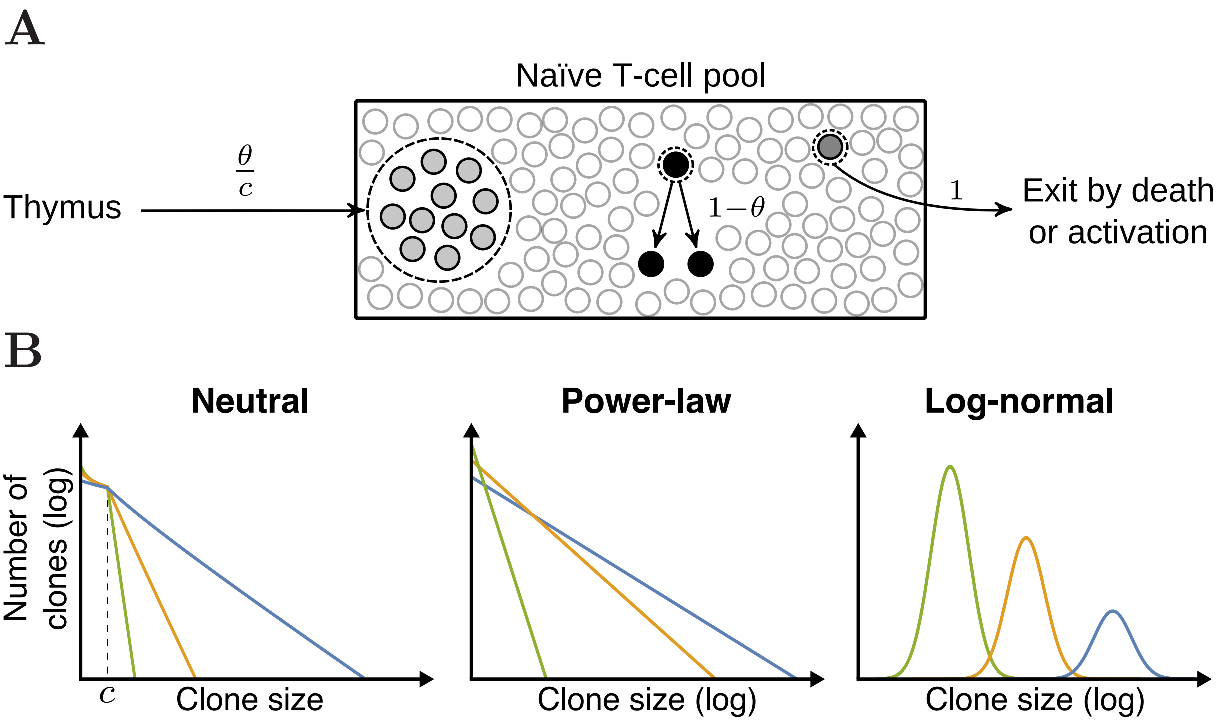
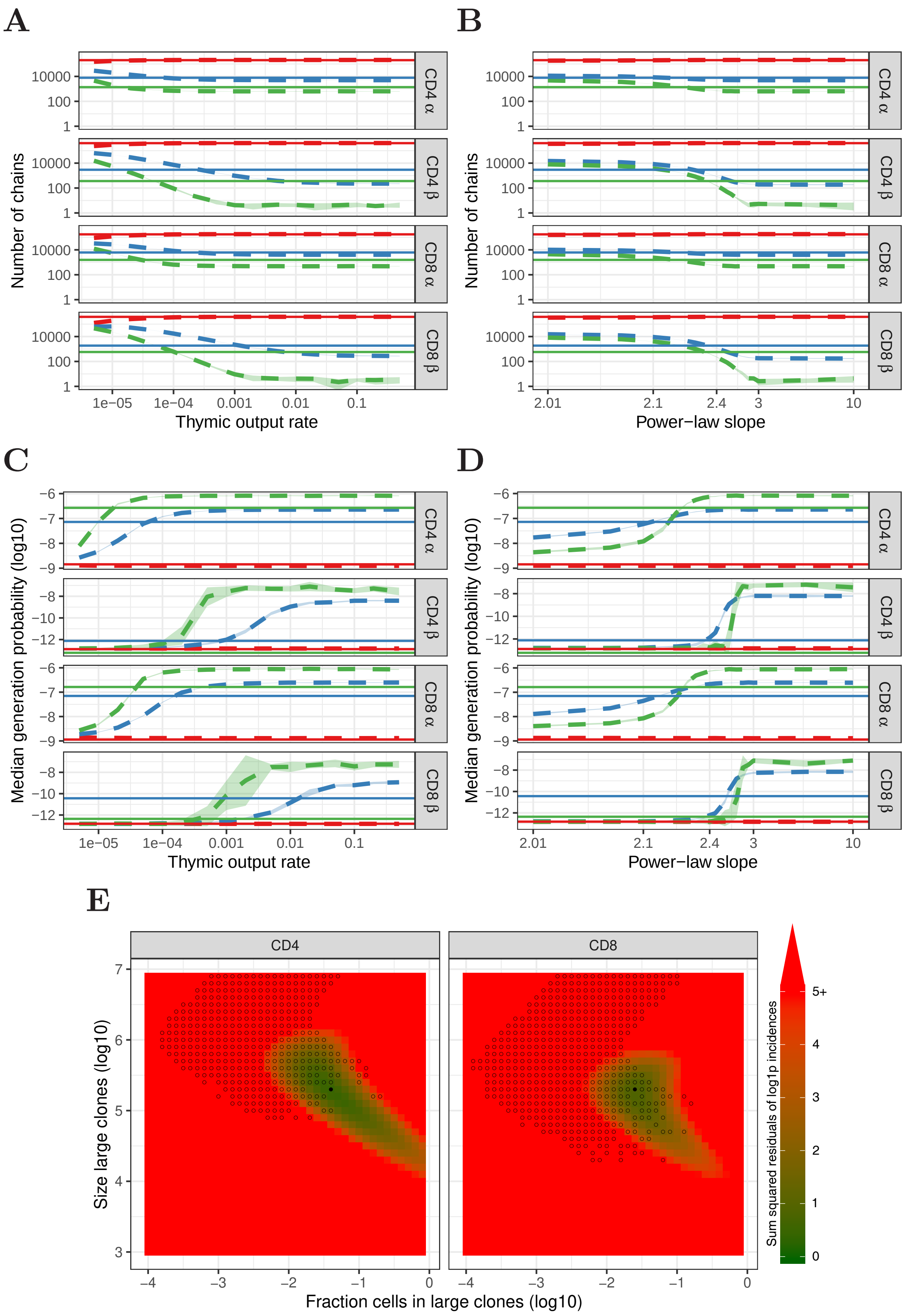


Figure 2.3 – **Schematic representation of the neutral model and various clone-size distributions.** **A.** Schematic representation of the dynamics of the neutral model for the naive T-cell pool. Each event starts with removal of one randomly selected cell from the pool, followed by peripheral division of another cell (with probability $1-\theta$), or a chance for thymic production (probability θ). After c of these thymus events, a clone of c cells is generated and added to the peripheral pool, reflecting the divisions of T cells before entering the periphery. **B.** Schematic representations of the various clone-size distributions that were used to predict the naive repertoire. The green, orange and blue coloured lines depict three parameter choices for each distribution, resulting in a low, medium and high mean clone size, respectively.

$P(\sigma)$-dependent selection model on the data from the single naive T-cell samples shown in Figure 2.1. First, we assume that productively rearranged chains have an overall 1/3 probability to survive thymic selection. Then we bias the probability for bins of sequences based on their $P(\sigma)$, such that the resulting set of α and β chains has the same generation probability distribution as in the experimental repertoire data (Materials and Methods). The models also incorporate the expected number of cells that contribute at least one mRNA molecule. This parameter is also learnt from the data, by setting the number of cells that contributed mRNA such that the predicted diversity of a subsample matches the observed diversity. Taken together, the subsamples we take from the various model clone-size distributions are such that they match the generation probabilities and diversity of the experimental subsamples as closely as possible. We compare the number and median $P(\sigma)$ of the TCR sequences that are predicted to occur in only one, in two or in three subsamples, with the equivalent experimental data from experiment 2 above (Figure 2.4).

We consider first the neutral model (Figure 2.4A&C). For the α chain, a wide range of thymic output rates predict the number of chains occurring in 1, 2 and 3 subsamples


A
B
Number of chains
10000
100
1
CD4 α
CD4 β
CD8 α
CD8 β
1e−05
1e−04
0.001
0.01
0.1
Thymic output rate
2.01
2.1
2.4
3
10
Power−law slope
C
D
Median generation probability (log10)
−6
−7
−8
−9
−10
−12
E
CD4
CD8
Size large clones (log10)
7
6
5
4
3
−4
−3
−2
−1
0
Fraction cells in large clones (log10)
Sum squared residuals of log1p incidences
5+
4
3
2
1
0

reasonably well (Figure 2.4A). The model does not predict the median *P(σ)* of TCRα found in 2 and 3 subsamples well, although qualitatively the model does predict the increasing *P(σ)* with increasing abundance (Figure 2.4C). For the β chains, there is no range of thymic output rates for which the model correctly predicts the number of sequences observed in 2 and 3 subsamples. Moreover, the observation that incidence 2 chains have higher *P(σ)* than incidence 3 chains was not predicted for any value of θ (Figure 2.4C). Thus, although the neutral model captures some features of the observed TCRα sequence abundances, it cannot account for observed TCRβ distributions. A similarly poor match between observed and predicted data is observed for log-normal clonotype model (Figure 2.3B) distributions (not shown).

In contrast, there is a much better fit between observed and predicted data is obtained when the model clonotype frequencies are modelled by a power-law distribution (Figure 2.4B and D). Like the distributions discussed above, in the parameter range where clone-size heterogeneity is limited (i.e., a steep slope), a power-law distribution predicts both the number of TCRα sequences found in 2 and 3 samples, and their larger median *P(σ)*. The number of TCRβ sequences is also predicted well if the slope is close to 2.3 (Figure 2.4B). Remarkably, for this slope the median *P(σ)* of TCRβ sequences found in two samples is higher than the median *P(σ)* of TCRβ sequences found in three samples (Figure 2.4D). Intuitively, we can understand this observation as reflecting the properties of power-law distributions, combined with the lower generation probabilities of TCRβ. Identical TCRβ recombinations occur frequently enough to make a detectable contribution to the TCR sequences observed in two samples, but not to those detected in three samples. Therefore, a significant proportion of TCRβ sequences observed twice are in fact derived from two or more different naive T-cell clones. In contrast, TCR sequences observed three times (or more) must be derived from large naive T-cell clones. Abundant TCRα sequences arise both from large clones and summation of identical TCRα from multiple smaller clones, but due to their higher generation probabilities, the latter dominates the *P(σ)* for TCRα sequences

Figure 2.4 ***(preceding page)*** – **Predictions of the neutral, power-law and two-population model compared with HTS data. A.** Number of TCRα and TCRβ sequences which are predicted to be shared between 1 (red), 2 (blue) and 3 (green) subsamples as a function of the thymic output rate θ for the neutral model. **B.** As A., but as a function of the slope of the power-law distribution. **C.** The median generation probability *P(σ)* of TCRα and TCRβ sequences predicted by the neutral model. Dashed lines depict the mean of 10 model prediction repeats, shaded area indicates the standard deviation, solid lines show observed results in HTS data. **D.** As C., but as a function of the slope of the power-law distribution. **E.** Graphical representation of parameter sweep results for prediction of $CD4^+$ and $CD8^+$ repertoires from αβ clone-size distributions following a mixture model consisting of singleton clones and a small fraction of large clones. The colour represents goodness of fit, with dark green being better predictions for number of sequences per incidence in samples. Empty circles indicate parameter combinations resulting in qualitatively correctly predicted *P(σ)*, i.e., 3 > 2 > 1 for TCRα and 2 > 1 for TCRβ and 2 > 3 for TCRβ. Filled circles indicate parameter combinations with the smallest distance to the incidence data and a correct *P(σ)* prediction.

found twice and three times. Finally, we note that although TCR sequence abundance in the single samples from experiment 1 is likely to incorporate multiple mRNA from single cells, the power-law distribution also predicts abundances in the single samples of experiment 1 reasonably well (Figure S2.6).

The vast majority of TCR sequences in samples of naive T cells are observed only once, and hence we cannot infer anything about their frequency in the whole repertoire, except that it is likely to be below a given abundance threshold. Therefore, we explored whether a more generalised model, which does not make any assumptions about the distribution of the low abundance T cells, would predict our experimental data as well as the power-law model. In this simple mixture model we generate a population in which the majority of cells are present only once, and a minority are present many times. We scanned the parameter space of this model, varying both the proportion of cells in each population, and the size of the larger population. The prediction of the model for each parameter pair was compared to the experimental data from experiment 2, both for the number of TCRs (combining α and β sequences) observed in one, two or three subsamples, and for the median $P(\sigma)$ of these TCR sequences. The best agreement between model and data was observed when 1-5% of the cells were derived from abundant T-cell clones (between 10^5 and 10^6 cells in the whole repertoire) (Figure 2.4E).

Abundant T-cell sequences are enriched for zero insertions and for antigen-association

In human prenatal thymocytes, the enzyme terminal deoxynucleotidyl transferase (TdT) is not expressed, leading to the production of TCR sequences with zero insertions of N-nucleotides. Pogorelyy and colleagues showed that enrichment of zero insertion TCR sequences can be used to detect fetal clones even in adults, and that their contribution to the overall repertoire decays slowly with age (Pogorelyy *et al.*, 2017). Interestingly, the proportion of zero-insertion sequences was strongly enhanced in those sequences observed more than once in the three subsamples examined in experiment 2 (Figure 2.5A). The interpretation of this finding is not straightforward, since zero-insertion TCR sequences have higher median generation probabilities, and this is also a property of abundant sequences as discussed above. Nevertheless, the data are compatible with a model in which the large clones observed in the repertoire are generated preferentially during early prenatal development of the naive T-cell repertoire.

We next examined if the abundant sequences in our data showed characteristics of semi-invariant NKT and MAIT cell populations. Classical NKT cells are characterised by an invariant TRAV24-TRAJ18 α chain and β chains with TRBV11 (Dellabona *et al.*, 1994). MAIT cells are enriched for TCRα rearrangements of TRAV1-2 with TRAJ33, TRAJ12 and TRAJ20 (Reantragoon *et al.*, 2013), and TCRβ sequences predominantly using TRBV20 and TRBV6 (Lepore *et al.*, 2014). Since our HTS data does not contain information on $\alpha\beta$ pairing, we studied both chains separately. A substantial fraction of the observed TCRβ sequences matches the characteristics of MAIT cells, and to a lesser extent NKT cells (Figure 2.5B&C). For both cell types, however, this fraction does not show a clear relation to incidence, and

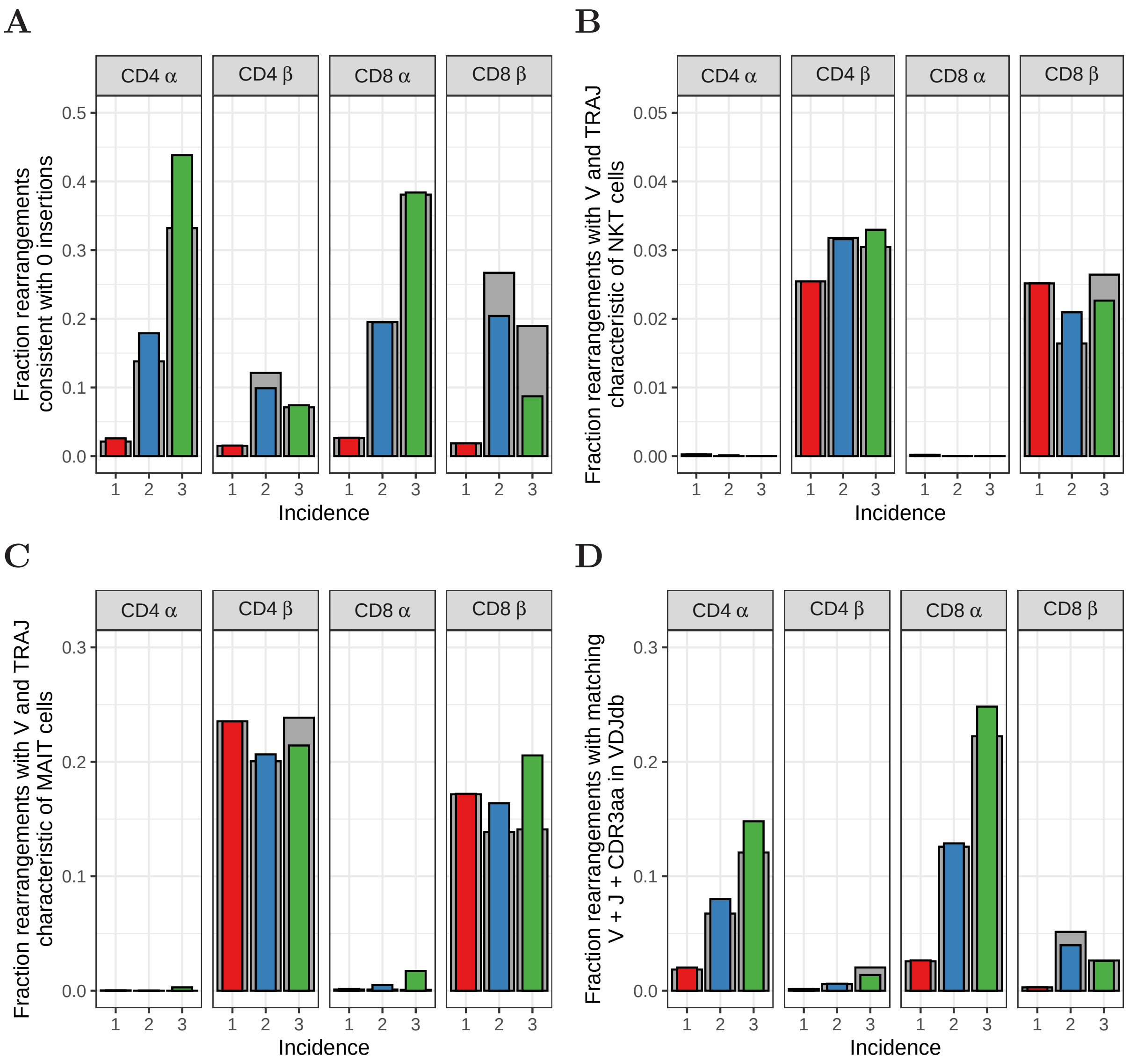


Figure 2.5 – **Characterisation of abundant TCRα and TCRβ sequences. A.** The fraction of rearrangements with zero minimal N-additions for sequences observed in 1, 2 or 3 naive subsamples. Data are shown without (coloured bars) and with cleaning of overlap with memory (grey bars). **B.** Fraction of TCRα and TCRβ sequences with V(J) usage characteristic of NKT cells (TRAV24-TRAJ18 for TCRα; TRBV11 for TCRβ). **C.** Fraction of TCRα and TCRβ sequences with V(J) usage characteristic of MAIT cells (TRAV1-2 with TRAJ33, TRAJ12 or TRAJ20 for TCRα; TRBV20 or TRBV6 for TCRβ). **D.** Fraction of sequences having at least one match (CDR3 amino acid sequence as well as V and J annotation) with the VDJdb (Shugay *et al.*, 2017).

does not suggest enrichment for MAIT or NKT cells among abundant sequences. The most abundant TCRα sequences are enriched for NKT sequences, but these still account for only a small fraction of the total (0.3% and 1.7% for CD4$^+$ and CD8$^+$, respectively, Figure 2.5B). Hence, we conclude that only a small fraction of the abundant sequences are derived from clones with a MAIT or NKT cell phenotype.

Finally we analysed whether the abundant TCR sequences in the naive population could be detected in a database of TCR sequences with known antigen specificity (Shugay *et al.*, 2017). Interestingly, there was a striking enrichment of TCR sequences with known antigen-specific annotation within the high abundance TCRα sequences observed in more than one subsample from experiment 2, and to a lesser extent for TCRβ sequences (Figure 2.5C). Interpretation is again not straightforward, because the high generation probabilities of the abundantly observed chains could lead to these sequences being over-represented in the database (ascertainment bias). Additionally, the observation may also reflect the fact that the naive T-cell populations we sequenced contained some antigen-experienced T cells with a naive phenotype (Pulko *et al.*, 2016). Finally, the observation is also compatible with the hypothesis that TCR recombination has evolved to preferentially generate TCRs specific to common pathogens like CMV or EBV (as discussed in (Thomas and Crawford, 2019)).

Discussion

The diversity and clone size distribution of the naive T-cell repertoire has been the subject of considerable debate, fuelled by the difficulty of obtaining more than a very small sample of the total repertoire, and by a variety of other technical considerations which we address in this study. We use a quantitative UMI-based sequencing protocol, and careful error correction to analyse the naive and memory repertoires from three healthy human volunteers. We convincingly demonstrate that a small proportion of the TCR sequences are present more than once in a sample of naive T cells from blood, corresponding to expected frequencies greater than 1 in 10^5. This number of abundant TCRα sequences is higher than the number of TCRβ sequences.

We carefully considered different mechanisms that could give rise to these abundant TCR sequences. We examined the contribution of potential contamination of the naive population with abundant T cells from the memory compartment during the sorting process, but the extent of such contamination was small (for CD8$^+$ cells) or not detectable (for CD4$^+$ cells). Furthermore, exclusion of all TCR sequences which occurred in both memory and naive populations did not alter the subsequent conclusions of the analysis. We also considered the possibility that abundant TCR sequences were observed due to sampling multiple mRNA molecules from the same cell. In order to exclude this possibility, we carried out an experiment where we divided up a sample of sorted naive T cells into three subsamples prior to lysis, and sequencing. In this experimental paradigm, TCR sequences found in more than one subsample must arise from different T cells. We observed that repeat sampling of mRNA from the same T cell did indeed occur, and might account for as much as 75% of the high abundance TCRβ sequences (for which there are more mRNA molecules per cell, (Oakes *et al.*, 2017)), and as much as 25% of the TCRα sequences. However, this effect was mostly restricted to TCR sequences observed twice, and made little contribution to TCR sequences observed three or more times.

Having excluded methodological causes of high abundance TCR sequences, we examined two biological mechanisms which could explain the data. The first of the mechanisms we consider is that abundant sequences derive from identical TCRα and TCRβ rearrangements occurring in multiple cells. In this model, abundance arises not from multiple sampling from the same large clone of T cells, but from summation over many different clones of T cells, each of which share a α or β chain. The second mechanism is that the naive repertoire clone size distribution is not uniform, but contains many small and some large clones. We combine computational models with experimental data to provide evidence that both mechanisms are required to explain the observed data. The first mechanism dominates the repertoire of TCRα, and is likely to contribute to the majority of observed abundant sequences. Interestingly, the model suggests that those TCRα which have the highest probability of generation are produced hundreds of thousands, or even millions of times within an individual, and must therefore be produced extremely frequently in the thymus. In contrast, the first mechanism has a smaller impact on the TCRβ repertoire, and abundant TCRβ sequences are more likely to arise from large clones in the naive repertoire.

The experimental limitations of sampling small volume of blood which contains only a tiny proportion of the total repertoire has dramatic effects on the observed TCR frequency distribution. One can use the analytical solution of the neutral model (Materials and Methods) with thymic introduction size $c = 1$ to illustrate this extreme sampling effect: $\hat{F}_i \approx F_i(\frac{s}{\theta})^i$, where $\hat{F}_i$ and F_i are the number of clones present with i cells in the sample, and in the pool, respectively, and s is the fraction of the repertoire that was sampled (here $s \sim 10^{-6}$). Since s/θ is of order 10^{-5} and this is raised to the i^{th} power, even very large TCR clones become rare in such a sample. Because of this, it is difficult to be definitive about the exact underlying T-cell clone frequency distribution which gives rise to the abundant TCR sequences we observe. The data are certainly compatible with a power-law distribution, as has been suggested previously (Desponds *et al.*, 2016). But many distributions made up of a mixture of rare clones and a small proportion (1-5%) of large clones (10^5 - 10^6) are compatible with the data we observe.

The demonstration of large clones in the naive repertoire raises the question of what determines the different sizes of different clonotypes. The neutral model already excludes repeated thymic production as explanation for large clones, because the combined probability of repeated αβ-clone production is very low (Dupic *et al.*, 2019). We confirmed that the abundant TCR sequences were not strongly enriched for sequences characteristic of iNKT and MAIT cells (Figure 2.5B&C). An alternative explanation is that the large clones may actually be antigen-experienced, but with a naive phenotype such as memory stem

T cells (Gattinoni *et al.*, 2011; Lugli *et al.*, 2013b,a; Marraco *et al.*, 2015; Pulko *et al.*, 2016). However, a number of alternative explanations for this enrichment exist, as discussed above. Furthermore, antigen-experienced T cells should be present in the memory populations, and large clones could still be observed even after removal of all cells which occur in both memory and naive populations. So, although we cannot exclude that some of the frequent TCR sequences may be derived from T cells that are not truly naive, we believe the data argue for the existence of truly naive large clones.

We speculate that the most likely mechanism for large clones is preferential growth/survival of some clones, presumably due to preferential selection on self-peptide/MHC (Rudd *et al.*, 2011; Lythe *et al.*, 2016). Intriguingly, the abundant TCR sequences we observed were enriched for sequences without N-insertions, a characteristic of TCRs produced prenatally (Pogorelyy *et al.*, 2017). The large clones may therefore be established very early in the development of T-cell adaptive immunity, before homeostasis of the immune system is achieved and when more rapid division and clonal expansion may be favoured.

In conclusion, our study highlights the huge impact of subsampling on correct interpretation of TCR repertoire data. It provides evidence for two different mechanisms which give rise to abundant TCR sequences in the naive human repertoire. The first mechanism, driven by multiple identical recombination events, is frequently overlooked in the analysis of T-cell repertoires, but has important implications in interpretation of observed sharing between different T-cell subpopulations of an individual, and between individuals (public TCR sequences). The second mechanism suggests that the TCR sequence plays a critical role in naive T-cell homeostasis. Further experiments will be required to fully elucidate the cellular and molecular mechanisms which underlie the heterogeneity of the naive T-cell repertoire.

Materials and Methods

Cell sorting and sequencing

Sequence reads came from T cells extracted from blood samples of three healthy volunteers, between 30 and 40 years old. Using CD27 and CD45RA markers, FACS-sorting was performed, identifying naive ($CD27^+CD45RA^+$), CM (central memory, $CD27^+CD45RA^-$), EM (effector memory, $CD27^-CD45RA^-$) and EMRA (effector memory RA, $CD27^-CD45RA^+$) cells. Barcoded TCRα and TCRβ cDNA libraries were obtained by reverse transcription of RNA molecules coding for either the α or β chain, respectively, followed by single strand DNA ligation to attach unique molecular identifiers (UMIs) of 12 nucleotides. These were PCR-amplified and sequenced using the Illumina MiSeq platform. For full description of the sequencing procedure, we refer to (Oakes *et al.*, 2017) and (Uddin *et al.*, 2019). The raw sequence files are available on the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) as experiment SRP109035.

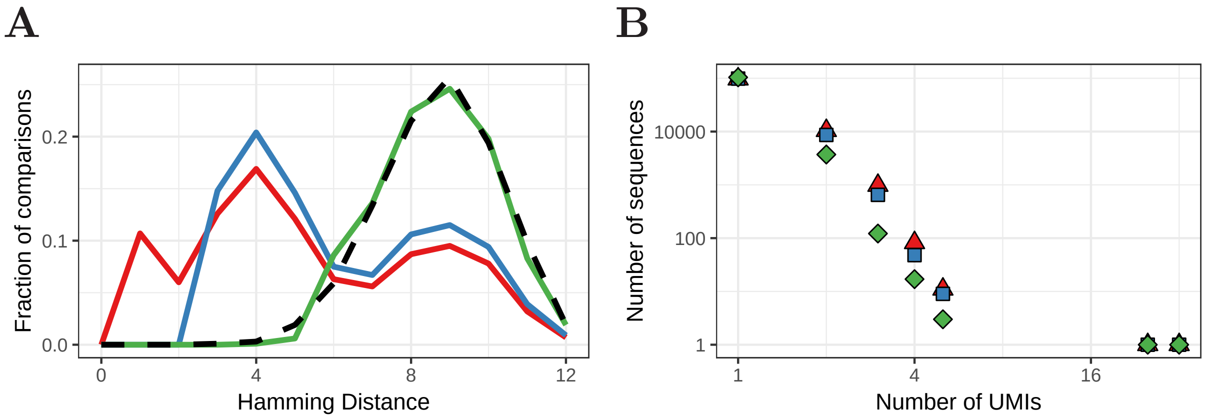


Figure 2.6 – Improved UMI correction leads to more reliable estimation of sequence frequencies. A. Distribution of Hamming Distances of UMIs within TCRβ sequences (naive CD4^{+} sample of volunteer 1) before correction (red), after default correction (blue) and after improved correction (green), in comparison with the distribution of UMIs between sequences (black dashed). **B.** Distributions of the same TCRβ sequences after the different correction strategies. Frequently observed TCRβ sequences remain at the same frequency after correction, whereas the frequency of other sequences tends to be overestimated due to mutated UMIs, which is compensated for by improved UMI correction.

Sequence analysis

We used the Decombinator pipeline (Thomas *et al.*, 2013) (Version 3.1) to demultiplex, annotate, and error-correct the raw sequencing reads. Our reads contain UMIs of 12 base pairs that can be used to identify which TCRα or TCRβ sequences are derived from the same cDNA molecule. Decombinator performs error correction on sequences by collapsing those 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 TCRα or TCRβ 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 sequence, irrespective of its frequency, to contain two UMIs that are nearly identical, concluding the UMIs are different because of PCR or sequencing errors.

We improved this by setting the barcode threshold to 0 and replacing it by an UMI error correction algorithm that takes the number of UMIs into account. Consider a TCRα or TCRβ sequence supported by i different UMIs, i.e., with frequency i. The Hamming distance, H, between two random UMIs of 12 base pairs can be represented by a binomial random variable, $H \sim \mathrm{B}(n,p)$, where $n = 12$ and $p = \frac{3}{4}$ (assuming uniform frequencies of the 4 different bases). There are $\binom{i}{2}$ distinct comparisons between the i UMIs, and assuming that every comparison is independent, the expected distribution of Hamming distances is $n_i(h) = \binom{i}{2} P(H = h)$. To determine whether two UMIs are unexpectedly similar, we define a threshold distance that depends on the frequency of their TCRα or TCRβ sequence (i):

$$D_\alpha = \max(\{d : \sum_{h=1}^{d} n_i(h) \leq \alpha\})\,. \tag{2.1}$$

Our algorithm corrects UMIs for a given sequence as follows: From $d = 1$ to $d = D_\alpha$, for all UMI pairs with $H \leq d$, add the read count of the less frequent UMI to the more frequent UMI and remove the former. We applied this algorithm to every TCRα and TCRβ sequence in our HTS data using $\alpha = 0.05$. The effects of this correction method are shown in Figure 2.6. After the improved correction, the distribution of Hamming Distances within and between distinct TCRα and TCRβ sequences is very similar, indicating that most erroneous UMIs have been removed. Our improved correction decreases the estimated frequency of many sequences at low frequencies, which indicates that many TCRα and TCRβ sequences that were observed two or three times, are actually singletons for which the UMI was mutated once or a few times. In the example given in Figure 2.6, the number of sequences that were observed more than once decreased with 66% by our improved correction (from 11491 to 3855), whereas the default correction estimated 9342 (only 19% reduction) of the sequences to have more than 1 true UMI.

Because our analysis focuses on the naive T-cell repertoire, we combined the different memory populations by adding the abundance of identical TCR sequences (V and J annotation as well as CDR3 nucleotide sequence) in the corresponding CM, EM and EMRA samples. We included for analysis the sequences that were reported as functional by Decombinator and had non-zero $P(\sigma)$. We also processed the HTS reads with RTCR (Gerritsen *et al.*, 2016) (Version 0.4.3). This pipeline determines a sample-based error rate and uses this rate to perform clustering on reads. Compared to Decombinator, RTCR estimates our reads to contain more PCR and sequencing errors and therefore tends to be more conservative in terms of reported diversity. Because RTCR reports fewer distinct rearrangements per sample, the overlap between samples (i.e., the number of chains with incidence 2 and 3) is lower than in Decombinator output. For each of the main-text figures, a supplemental RTCR-based version is provided. Although the quantitative results are not identical, the RTCR results qualitatively match those of the Decombinator output, confirming that our results are not algorithm-dependent.

Subsampling to exclude inflated abundance through multiple RNA contributions by single cells

An important step in our analysis is the additional experiment in which the naive cells were split into three parts before mRNA extraction. The probability for a naive cell to be sampled from the pool is very low ($< 10^{-5}$), but once a cell has been sampled it may likely contribute multiple RNA molecules. These would then be sequenced with different UMIs, inflating the abundance we measure in a sample. Hence, we use subsampling to avoid the noise on TCRα and TCRβ abundance introduced by variable TCR expression between cells. To quantify the possible effect of single cells contributing multiple RNA molecules, we performed a permutation test. We computationally joined the sequences observed in the three independently sequenced replicates, adding the abundance (as measured by UMIs) in each of the three subsamples together. We then randomly assigned the UMIs of these sequences to one of three artificial portions and again scored the incidence of all TCRα and TCRβ sequences. In this setting, RNAs contributed by single cells in a single sample, can be distributed over multiple permuted samples. This was done 10 times for each set of sequences and we found that permutation led to a large increase in the number of sequences occurring in multiple samples (Figure S2.3A). We quantified the number of abundant chains, by counting sequences observed in multiple samples.

When multiple RNA molecules from a single cell can contribute a UMI (i.e., in the permuted set and within a single sample), the number of abundant sequences is greatly overestimated. About 25% of abundant α chains in this setting is actually due to inflated counts. For β chains the effect is much larger, with over 75% of abundance due to RNA content. This difference is consistent with our previous finding that T cells contain in the order of 300 *TCRB* and 100 *TCRA* RNA molecules per cell (Oakes *et al.*, 2017). Moreover, the lower $P(\sigma)$values of β chains readily explains that there are fewer true duplets and triplets than for α chains. Subsampling appears to be very important when obtaining our most surprising result that $P(\sigma)$ values are enriched for β chains with incidence 2, but not incidence 3. After permutation, most duplets are due to RNA content and therefore no longer enriched for high $P(\sigma)$ (Figure S2.3B). These results highlight the importance of our additional step of taking a single blood sample, dividing it into three portions and then analysing all three subsamples separately.

Sharing of TCRα and TCRβ sequences

We sequenced TCRα and TCRβ from whole blood samples taken from 28 healthy volunteers. The study was carried out in accordance with the recommendations of the UK Research Ethics Committee with written informed consent of all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the University College London Hospital Ethics Committee 06/Q0502/92. The raw sequence files are available on the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) as experiments SRP045430 and SRP151125. In order to measure how public the individual sets of sequences were, we measured their degree of sharing between our naive samples and these whole blood repertoires.

As shown in Figure 2.2, we have three sets of sequences, those with incidence 1, 2 and 3. For each set, we measured which fraction is also found in the 28 independent whole blood samples, which delivers 28 estimates of sharing. More precisely, we counted the number of shared TCRα and TCRβ sequences between the sets of sequences observed in two and three naive subsamples, and compared these to sharing with an equal size sample of naive sequences which were only observed in one subsample. Since the number of sequences which occurred more than once was much smaller than the number of sequences which only occurred once, we subsampled the set of unique sequences 10 times. The results are shown as the number of shared TCRα or TCRβ for each whole blood repertoire, as a proportion of their number of sequences in the samples being tested (Figure S2.4A). In order to study the sharing of the β chains in our data with higher resolution, we also analysed overlap of the sets of sequences with the TCRβ data from a large cohort of 786 people published in (Emerson *et al.*, 2017) (Figure S2.4B).

Neutral model for dynamics of naive T cells

To model naive T-cell dynamics in the absence of peripheral selection, we developed a model that is similar to the Neutral Community Model (NCM) of Hubbell (Hubbell, 2001). Naive T cells, viewed through an ecological lens, are individuals, and all naive T cells sharing the same TCRα and TCRβ sequence are part of the same species ($\alpha\beta$-clone). Neutrality, as defined by Hubbell, means that all species have the same per capita probability of birth (peripheral division) and death. When considering the model, we ignore the very small chance that an existing $\alpha\beta$-clone is produced again by the thymus. Hence, in our simulations we assume that the thymus produces T-cell clones that are unique and novel.

Consider a pool of N naive T cells belonging to clones, each consisting of i cells, which changes by thymic production, cell division and cells leaving the naive pool (as a result of cell death or activation). During each event, one randomly selected cell exits the pool, causing the corresponding clone to decrease in size from i to $i-1$ cells. With probability $1-\theta$, another randomly selected cell will divide, causing the corresponding clone to increase its size from i to $i+1$ cells. Alternatively, with probability θ, thymic production *can* occur: every c events in which no peripheral division occurred, the thymus will release c cells of a newly produced clone. So, the pool size N only fluctuates by c cells, and because $N \gg c$, the total number of cells stays almost constant during the entire simulation. The per capita birth rate ($(1-\theta)/N$) and death rate ($1/N$) are equal for all T-cell clones, which makes this a neutral model. In this discrete-time model, exit and production are coupled, but its dynamics can be approximated by a continuous-time model, in which thymic production, cell division, and deaths are uncoupled Poisson processes. This is illustrated by the following Markov chain, in which the states are clone sizes and the rates show the probabilities of clones moving to another state:

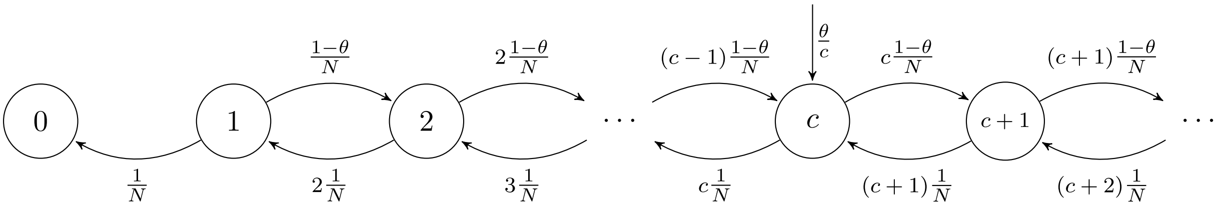


Figure 2.7 – **Markov chain representation of the neutral model with thymic introduction size** c.

This Markov process describes the dynamics of the clone-size distribution F, i.e., the total number of clones F_i consisting of i cells. After many birth and death events, individual clones still change in clone size over time, but the clone-size distribution approaches equilibrium. At this steady state, the rate at which new clones enter the naive pool, θ/c, equals the rate at which clones leave the pool, i.e., $F_1(1/N)$. Hence, in equilibrium, the number of singletons, clones with only one cell, approaches $F_1 = \theta N/c$. The total rate at which the cells of clones with i cells divide and die depends on the total number of cells belonging to F_i clones: iF_i. For clone sizes up to c cells, the rate at which the cells of the F_i clones die, (iF_i/N), balances the division the cells of F_{i-1} clones $((i-1)F_{i-1}(1-\theta)/N)$ and the rate at which new clones enter the pool (θ/c). The analytical solution to this recurrence relation $iF_i/N = (i-1)F_{i-1}(1-\theta)/N + \theta/c$ is:

$$F_i = \frac{N - N(1-\theta)^i}{ic}, \quad \text{for } 1 \le i \le c\,. \tag{2.2}$$

For states with $i > c$, only birth and death of cells need to balance between states $i-1$ and i (as there is no net flux from clones introduced by the thymus): $iF_i/N = (i-1)F_{i-1}(1-\theta)/N$. This recurrence relation has the following analytical solution:

$$F_i = \frac{cF_c(1-\theta)^{i-c}}{i}, \quad \text{for } c \le i \le N\,. \tag{2.3}$$

When predicting the full clone-size distribution, we use Equation 2.2 and Equation 2.3 to calculate the steady-state distribution. The total number of all distinct clones (i.e., the richness) in the steady-state repertoire is simply the sum over all their frequencies F_i, $R = \sum_{i=1}^{\infty} F_i$, which has a simple closed-form solution for $c = 1$,

$$R = \sum_{i=1}^{\infty} F_i = \frac{\theta N \ln\theta}{\theta - 1} \quad \text{for } c = 1\,. \tag{2.4}$$

The Simpson's diversity of the steady state repertoire also has a simple form,

$$S = 1/\sum_{i=1}^{\infty} F_i \left(\frac{i}{N}\right)^2 = \frac{2\theta N}{2 + (c-1)\theta}\,, \tag{2.5}$$

which equals $F_1 = \theta N$ for $c = 1$, and is a saturated function of θ if $c > 1$.

We consider the sampling process of a small fraction s from a naive T-cell pool of N cells, which clones follow the distribution F in Equation 2.2 and Equation 2.3. Assuming the naive pool is large and well-mixed, the number of T cells, X, sampled from the j cells belonging to a particular clone, can be approximately represented by a binomial random variable, $X_j \sim \mathrm{B}(n = j, p = s)$. The expected clone-size distribution of the sample, $\hat{F}$, is then given by

$$\hat{F}_i = \sum_{j=i}^{N} F_j P(X_j = i)\,. \tag{2.6}$$

The strong distortion of sampling from clone-size distributions can be illustrated using the analytical solution of Equation 2.6 for the neutral model for $c = 1$:

$$\hat{F}_i = F_i \left(\frac{s}{s + (1-s)\theta}\right)^i\,. \tag{2.7}$$

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 that even very abundant clones will become rare or absent in a small sample.

Clone-size distributions of the naive T-cell pools

Since our data contains separate data on both CD4$^+$ and CD8$^+$ T cells, we predicted the clone-size distributions of both subsets separately. To account for the larger CD4$^+$ pool

(Wertheimer *et al.*, 2014; Westera *et al.*, 2015), we set its pool size $N = 7.5 \times 10^{10}$ cells, while we used $N = 2.5 \times 10^{10}$ for the naive CD8⁺ pool.

When analysing the neutral model, we used its steady-state distribution (Equation 2.2 and Equation 2.3). Since the β chain rearranges first, followed by a few divisions before rearrangement of the α chain (Gonçalves *et al.*, 2017), we use $c = 100$ for TCRβ and $c = 10$ for TCRα. We also used various phenomenological clone-size distributions that are not based on a mechanistic model. To allow for exploration of a wide range of distributions, we chose mathematical functions which form can be changed by a single parameter, such as the slope of the power-law distribution.

The power-law distribution with form $F_i = F_1 \times i^{-k}$ shows a straight line on a log-log plot. Since all F_i are written as a function of F_1, the total number of cells $N = F_1(1 + 2 \times 2^{-k} + 3 \times 3^{-k} + \ldots) = F_1 \sum_{i=1}^{\infty} i^{1-k}$. This sum is convergent for $k > 2$ and gives

$$F_i = \frac{N i^{-k}}{\zeta(k-1)}, \quad \text{for } k > 2 \tag{2.8}$$

for the power-law clone-size distribution, in which ζ is the Riemann zeta function.

We also studied repertoires with log-normal distributions of clone-sizes by drawing from a normal distribution and raising 10 to the power of these numbers for clone sizes. For this we used varying μ and $\sigma = \mu/10$. These distributions yielded results that were qualitatively similar to those from the neutral model (not shown). For the simple mixture model (Figure 2.4E), we defined two populations of clones: (1) singletons (clones of just one cell that can only contribute to high TCRα or TCRβ abundances by sharing a chain with many other clones) and (2) large clones of equal size. We varied the fractions of both populations as well as the size of the large clones to find which fraction of the cells in the naive repertoire is expected to belong to large clones. A similar analysis, combining the aforementioned distribution following from the neutral model with a log-normal distribution for the population of large clones, produced very similar results (not shown).

In silico samples from modelled clone-size distributions

To compare the clone-size distributions with the HTS data of the blood samples, we generated TCRα and TCRβ repertoires using IGoR (Marcou *et al.*, 2018). We generated 10^8 TCRα and TCRβ sequences using IGoR's default recombination model and parameters. We selected the rearrangements which CDR3 nucleotide sequence consisted of a multiple of 3 nucleotides (in frame) and did not contain in-frame stop codons, in line with the inclusion criteria of productive rearrangements in our HTS samples (~ 28%). Next, we calculated generation probabilities $P(\sigma)$ for all these rearrangements. This may seem a detour, but this is needed as many different scenarios can lead to the same TCRα or TCRβ rearrangement.

Only a small percentage of thymocytes that undergo rearrangements in the thymus will eventually be exported as a naive T cell. This is due to out-of-frame rearrangements, but also as a result of both positive and negative selection. Moreover, the generation probability distributions of pre- and post-selection TCRα and TCRβ repertoires are markedly different (Elhanati *et al.*, 2014). To account for these observations, we train a $P(\sigma)$-dependent selection model to account for the effects of thymic selection on our IGoR-produced TCRα and TCRβ sequences. Note that this selection method is based on single chains rather than on αβ-TCRs. This is because recombination of β and α chains occurs at different points in T-cell differentiation. The first step in selection, after formation of the β chain, is based on correct folding and expression, using a pre-α pseudochain for pairing. If the T cell survives this step, it undergoes multiple rounds of divisions, by which its β chain can pair with many different α chains. The second step is positive and negative selection based on MHC-peptide interactions, which is likely to operate on a joint αβ pair. It is unknown how much each of these two steps contributes to the overall selection process.

For TCRβ selection, we reason that selection on pairing with the invariant pre-α chain acts exclusively on the level of single β chains, and once the T cell survives this first step, it is expected to survive with at least one of the many α chains it can pair with during the second step. The absence of strong structural constraints on αβ pairing supports this idea (Tanno *et al.*, 2020). Additionally, the large $P(\sigma)$ shift between pre- and post-selection TCRβ repertoires is indicative of selection acting on the level of single β chains (i.e., the probability for a β chain to be selected is largely irrespective of the α chain). For α chains this shift is less pronounced, and a newly generated α chain only pairs with a single β chain. We therefore also tested the effect of an alternative selection model in which a given α chain survives selection with a given probability for repeated production events, reflecting different selection outcomes when pairing with different β chains. This approach decreases the average frequency of α chains in the post-selection repertoire (since in this case they will on average survive only in a fraction of selection events, instead of our default all-or-nothing model). This did not affect our results in a qualitative manner and we proceeded with selection on the level of single chains for both TCRα and TCRβ.

We use each of the HTS data sets from the single sample experiment (shown in Figure 2.1) to calculate the relative enrichment or depletion of 100 log10 $P(\sigma)$ bins (ranging from -50 to 0) compared to 100 equally sized samples of the IGoR output, for TCRα and TCRβ separately. If the HTS data contained few rearrangements for a given bin, we joined adjacent bins in such a way that the bin-specific selection factor was always based on at least 1% of the experimental observations (Figure 2.8). This approach yielded $P(\sigma)$-specific selection factors $f_{P(\sigma)}$ ranging from 0.6 to 1.15 (i.e., our data suggests that sequences with a preferable $P(\sigma)$ are about 2 times as likely to be selected as those in the least preferable $P(\sigma)$ domain). We assumed an overall selection factor of 1/3, meaning that one out of 3 productive TCRα or TCRβ rearrangements would survive selection. We then allowed sequences to be part of the post-selection repertoire with probability

$$p_{selected} = f_{P(\sigma)}/3 \tag{2.9}$$

and stored the outcome to make a consistent decision when multiple copies of the same TCRα or TCRβ sequence were present in the pre-selection repertoire. This approach yielded a post-selection repertoires with $P(\sigma)$ distributions similar to the single sample HTS data. Other values for the overall selection probability, ranging from 1/10 to 1, were also tested, but yielded similar qualitative results (not shown).

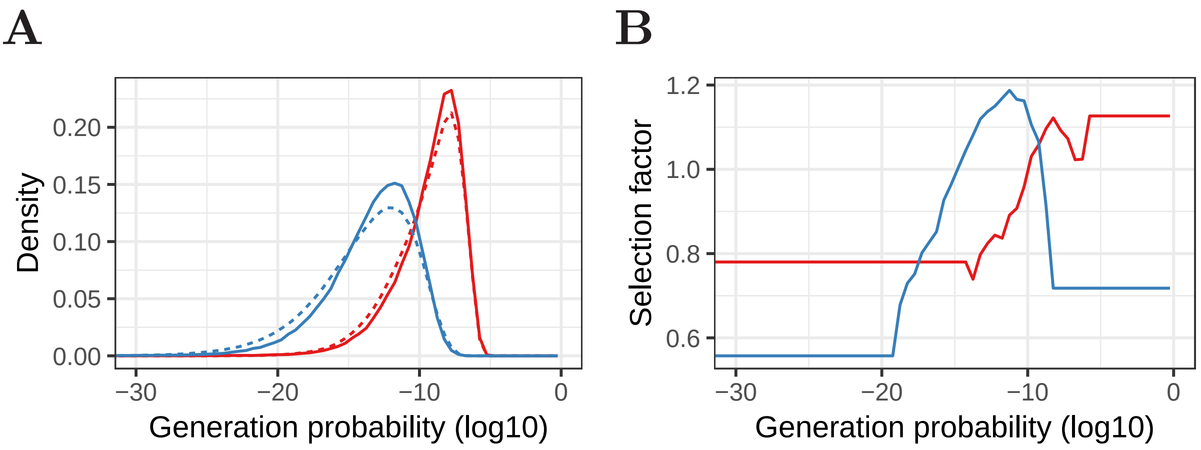


Figure 2.8 – **Pre- and post-selection *P(σ)* densities and *P(σ)*-dependent selection factors for α and β chains. A.** Relative frequency of generation probabilities of TCRα (red) and TCRβ (blue) sequences in the combined HTS data (solid) and IGoR output (dashed). **B.** The bin-specific selection factors $f_{P(\sigma)}$ are determined by division of the density of a given bin in the HTS data by the density in the pre-selection IGoR output. A value of 1 means that a sequence with this *P(σ)* has an average probability to be selected in the thymus, whereas lower values indicate stronger selection and higher values weaker selection (i.e., a higher probability to pass selection).

We could have assigned all clones in the clone-size distribution an α and β chain with this approach. However, since only a very small part of the repertoire is sampled, we chose to only assign an identity to those clones present in the samples. Hence, we started with predicting the presence of all clones, as a function of their size, in each of the samples. The probability that a clone with i cells is represented by at least one cell in a sample of n cells from a pool of N cells is

$$p_i = 1 - (1 - \frac{i}{N})^n \tag{2.10}$$

Given F_i, which is the number of clones in the pool with clone size i, the number of these clones present in the sample of n cells can be approximately represented by a binomial random variable, $X_i \sim B(n = F_i, p = p_i)$. We evaluate this for the entire clone-size distribution F. N and F are known from the model but one cannot directly determine the number of sampled cells n. This is because individual cells may contribute multiple mRNA molecules and many cells may have been present in the FACS-sorted sample without contributing mRNA to the eventual sequenced fraction. Therefore, we learn the sample size by assigning α or β to sampled clones and choosing n such that the predicted diversity (i.e., number of distinct chains) matches the experimental observations. We took the number of distinct TCRα or TCRβ sequences as lower bound for the sample size, since in this model individual cells are assumed to express one functional α or β chain. The total number of cells reported by the FACS-sorter was used as upper bound. We also checked the implications of the observation that some T cells contain two functional α and/or β chains, but this did not qualitatively change our results (not shown).

Thus, we adjusted the generation probability distribution by training a $P(\sigma)$-dependent selection model on independent HTS data and based the sample size on the corresponding subsamples. Hence, the predicted individual subsamples reflect the experimental observations in terms of diversity and generation probabilities. We use the chains occurring in multiple samples (i.e., those with incidence 2 and 3) to assess the agreement between model predictions and the HTS data. We repeated the sampling process and assignment of α and β chains 10 times for each model-parameter combination to account for the stochastic nature of sampling and V(D)J recombination.

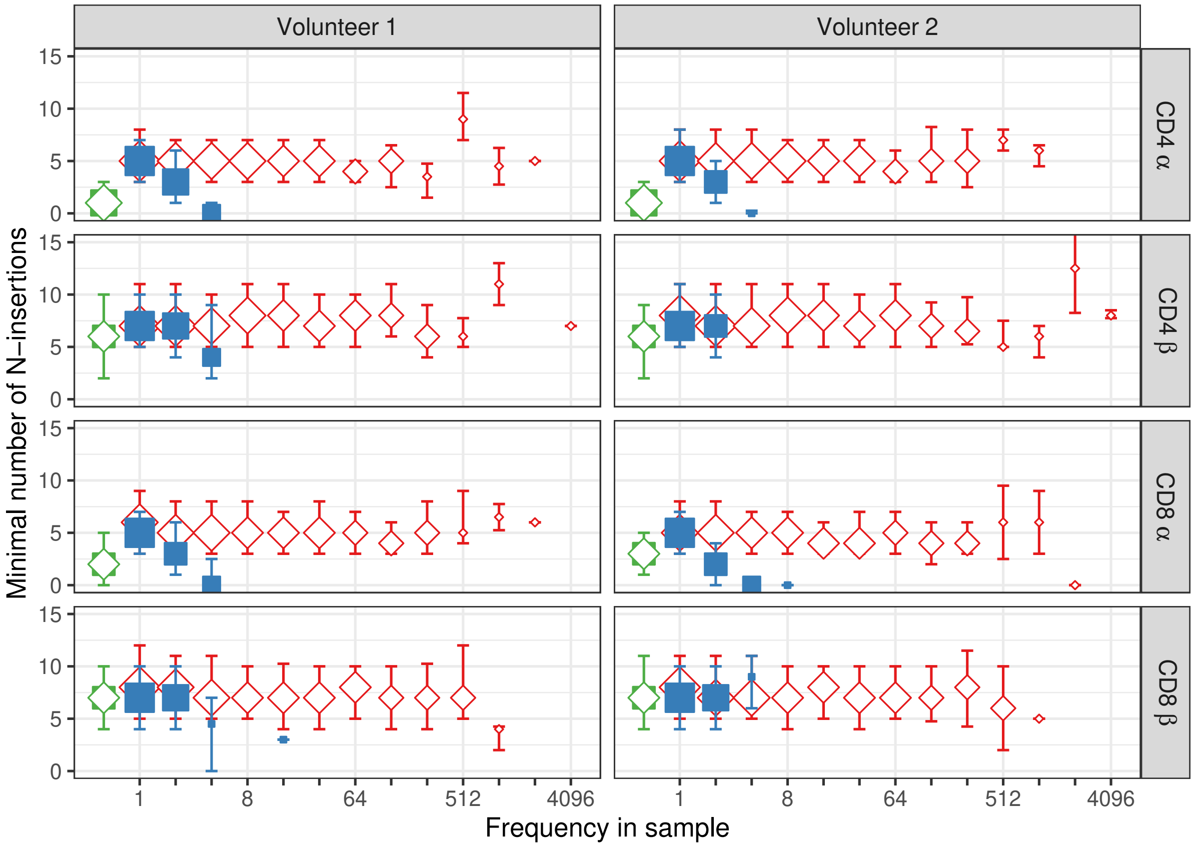
Author contributions

P.C.d.G. - Conceptualisation, Formal analysis, Visualisation, Methodology, Writing - original draft, Writing - review and editing; T.O. - Validation, Investigation, Writing - review and editing, Experimentation - T cell sorting and receptor sequencing; B.G. - Conceptualisation, Software, Formal analysis, Visualisation, Methodology, Writing - original draft, Writing - review and editing; M.I. - Software, Formal analysis; J.M.H. - Software, Methodology, Writing - review and editing; R.H. - Formal analysis, Writing - review and editing; B.C. - Conceptualisation, Supervision, Writing - review and editing; R.J.d.B - Conceptualisation, Supervision, Writing - review and editing.

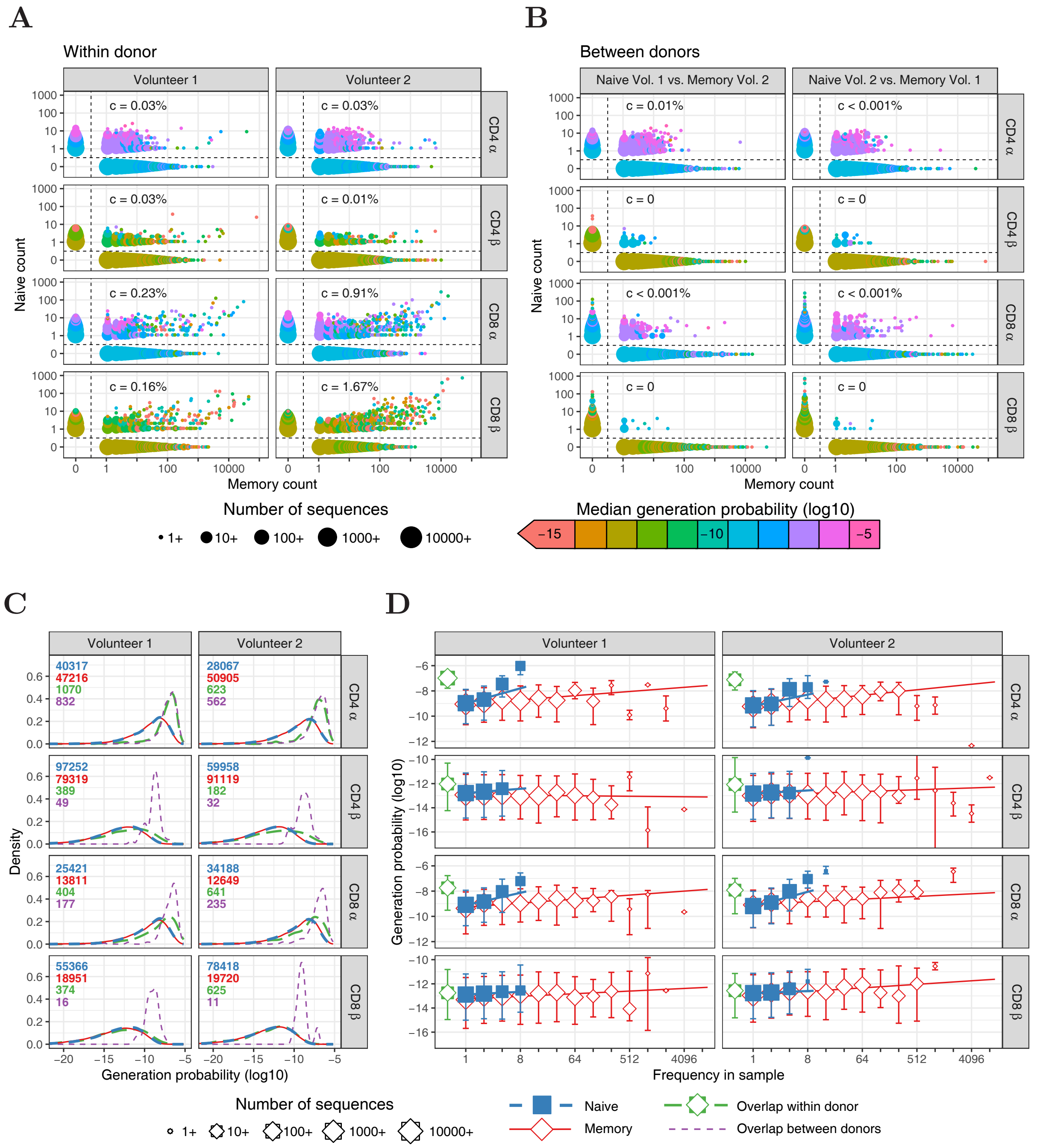
Acknowledgements

We thank Laurens Krah for mathematical advice and helpful discussions. This work was supported by The Netherlands Organization for Scientific Research (NWO) Graduate Program 022.005.023 (to P.C.d.G.), the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch government (FES0908) (to B.G.), by a grant to B.C. from Unilever PLC and supported by the National Institute for Health Research UCL Hospitals Biomedical Research. J.M.H. was supported by an MRC studentship.

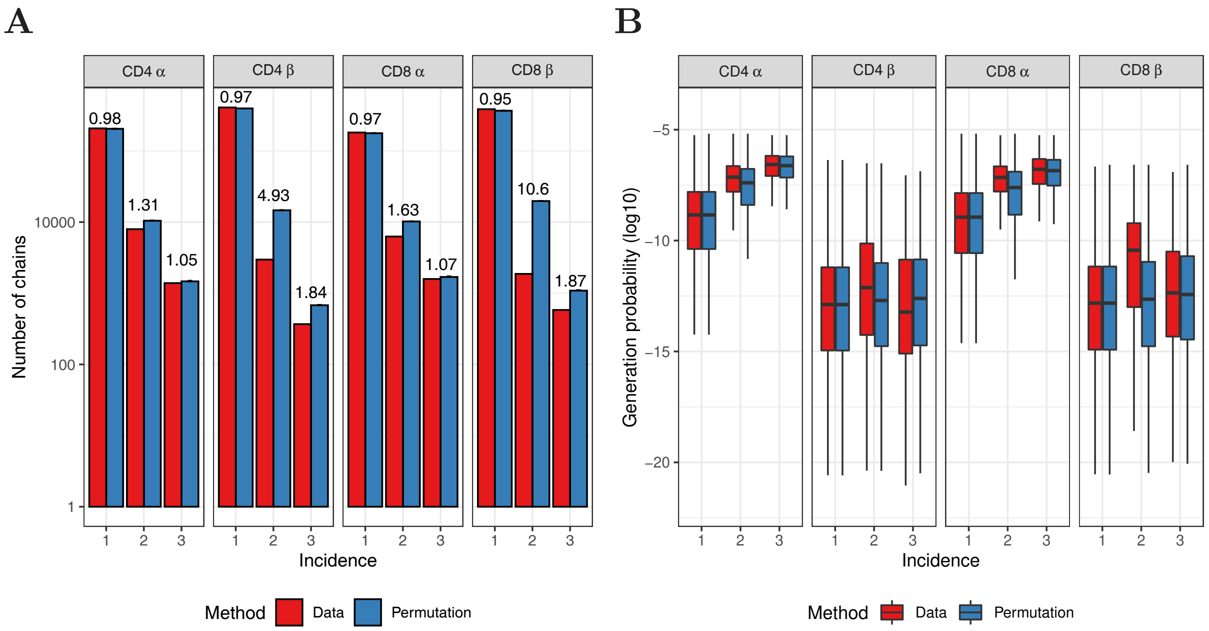
Supplemental Figures



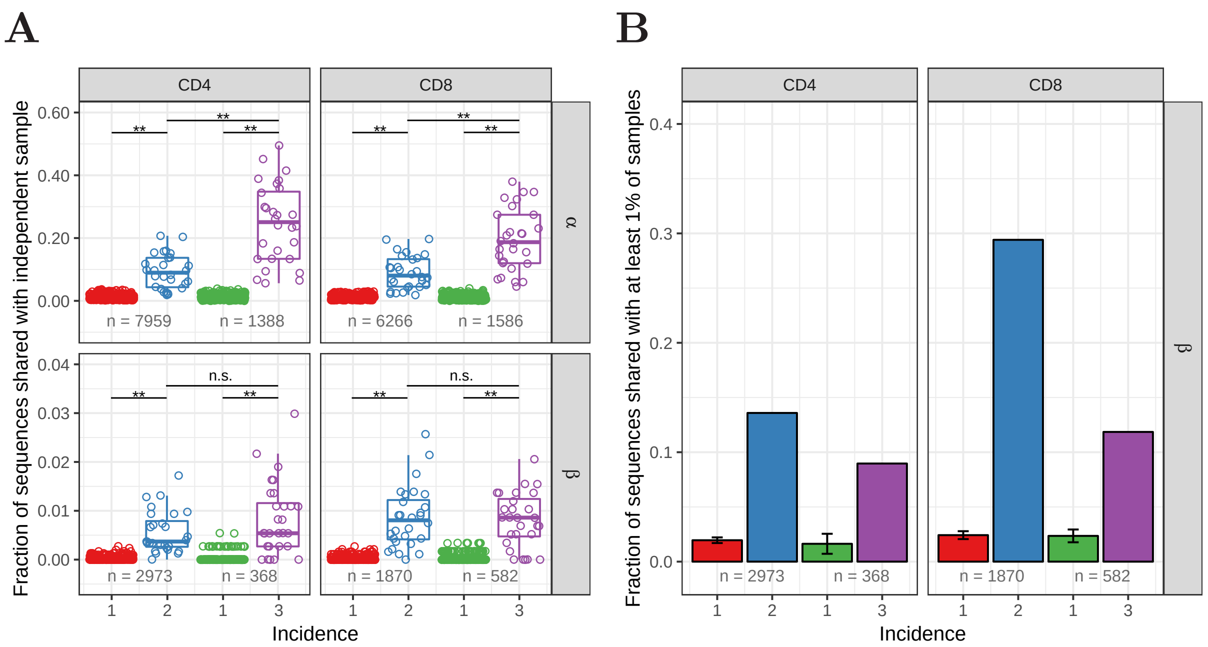
Supplementary Figure S2.1 – **TCRα and TCRβ sequences abundant in naive tend to have less N-insertions.** For each sequence σ in our dataset of single samples, the minimal number of N-insertions was determined by the length of the CDR3 nucleotide sequence not matching germline TRAV and TRAJ (for TCRα), or TRBV, TRBD and TRBJ (for TCRβ). The median insertion length is shown for each observed frequency class (log2 bins) in naive (blue squares) and memory T-cell (red diamonds) samples. Insertion length of the overlapping TCR sequences is shown in green for reference (irrespective of frequency). Symbol sizes indicate numbers of sequences for each frequency class. Error bars represent the 25% and 75% quartiles.



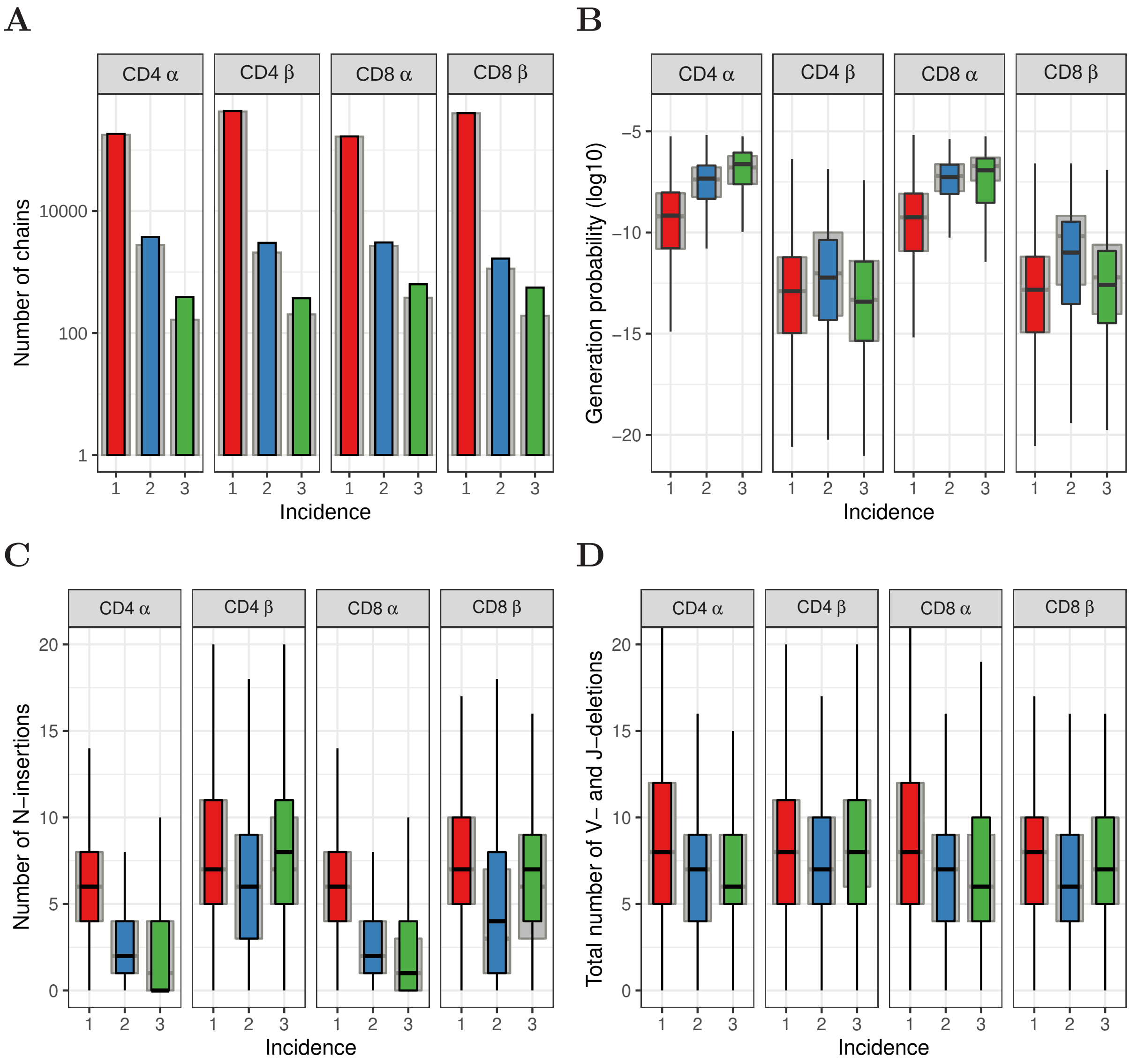
Supplementary Figure S2.2 – Similar to Figure 2.1, but for HTS data processed with RTCR. A. Frequency in of TCRα and TCRβ sequences in naive versus total frequency in memory samples of the same volunteer. **B.** As A., but comparing frequency in naive sample from one volunteer with frequency in memory from the other volunteer. **C.** Distributions of generation probabilities (log10) for TCR α and β sequences from CD4$^+$ and CD8$^+$ from two volunteers. **D.** The median *P(σ)* is shown for each observed frequency class (log2 bins) of sequences exclusively observed in naive (blue squares) or memory T-cell (red diamonds) samples. *P(σ)* of the overlapping chains is shown in green for reference (irrespective of frequency). Further details are provided in the legend of Figure 2.1.



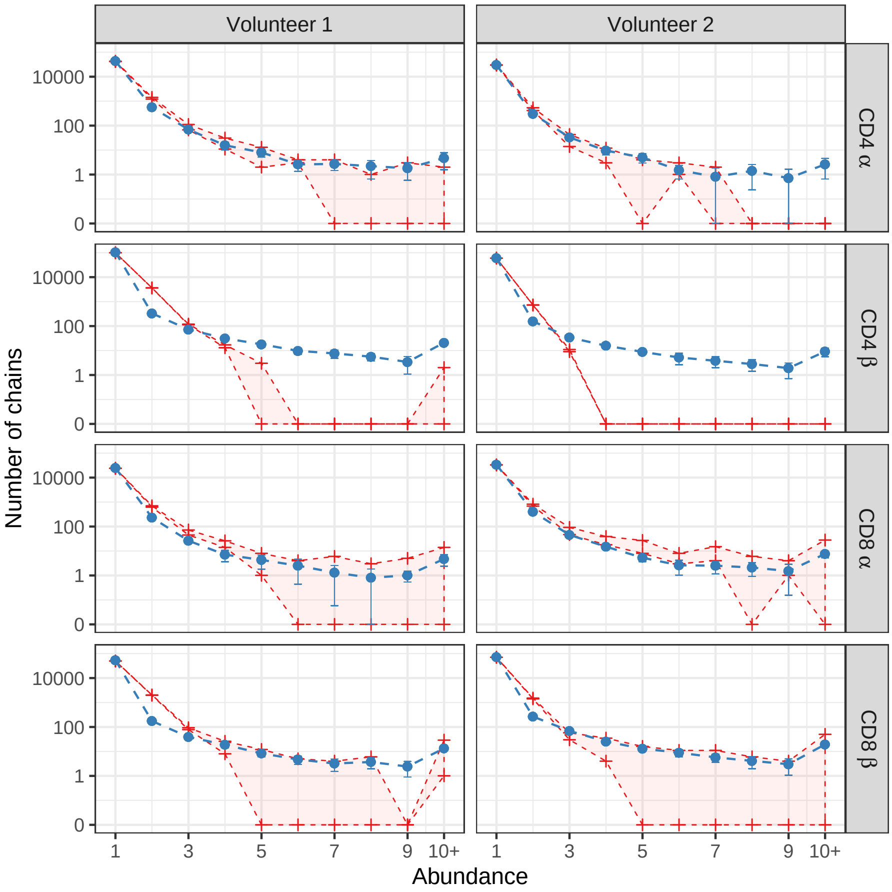
Supplementary Figure S2.3 – **Permutation of subsampling experiment. A.** Number of chains observed in 1, 2 or 3 subsamples (red) and after redistributing the sequences over the samples (blue). For the permutation test, mean values of 10 iterations are shown, with error bars indicating 1 standard deviation. The fold-change between data and permutation is indicated on top of the bars. **B.** Generation probabilities of the sequences in A, as determined with IGoR (Marcou *et al.*, 2018). The plot shows median (black horizontal line), interquartile range (filled bar) and the range from the bar up to 1.5 times the interquartile range (black vertical range, outliers not shown).



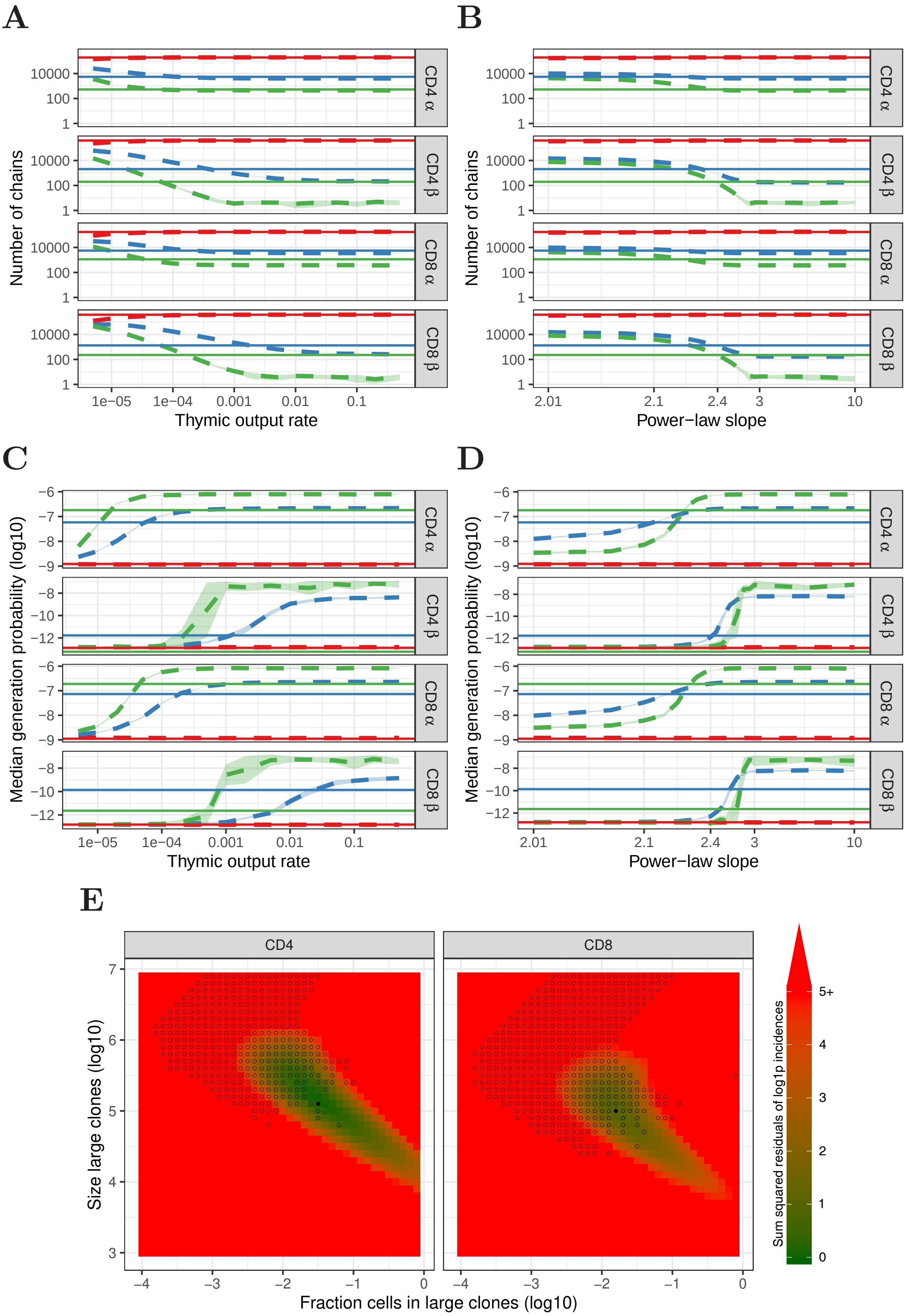
Supplementary Figure S2.4 – **Observed frequency predicts sharing for TCRα but not TCRβ sequences.** **A.** We compared the occurrence of TCRα and TCRβ sequences observed in two or three subsamples (incidence 2 or 3, respectively), and equal-size samples of sequences observed in one subsample (incidence 1), in unfractionated blood samples collected from 28 healthy donors. Symbols depict the number of shared TCRα or TCRβ sequences for each whole blood repertoire, as a proportion of the total number in the samples being tested (the latter is indicated at the bottom). The boxplot depicts the median value and 25th and 75th percentiles. Shared fractions were compared by Wilcoxon-Mann-Whitney test, **: $p < 0.01$, n.s.: not significant ($p > 0.05$). **B.** Fraction of each set of sequences from A that was observed in at least 1% of the samples from a large cohort of 786 individuals (Emerson *et al.*, 2017). Error bars show the standard deviation for the multiple sets of sequences with incidence 1. A smaller fraction of the most frequently observed β chains (incidence 3) are shared than those with incidence 2, which is in line with the *P(σ)* observations using IGoR.



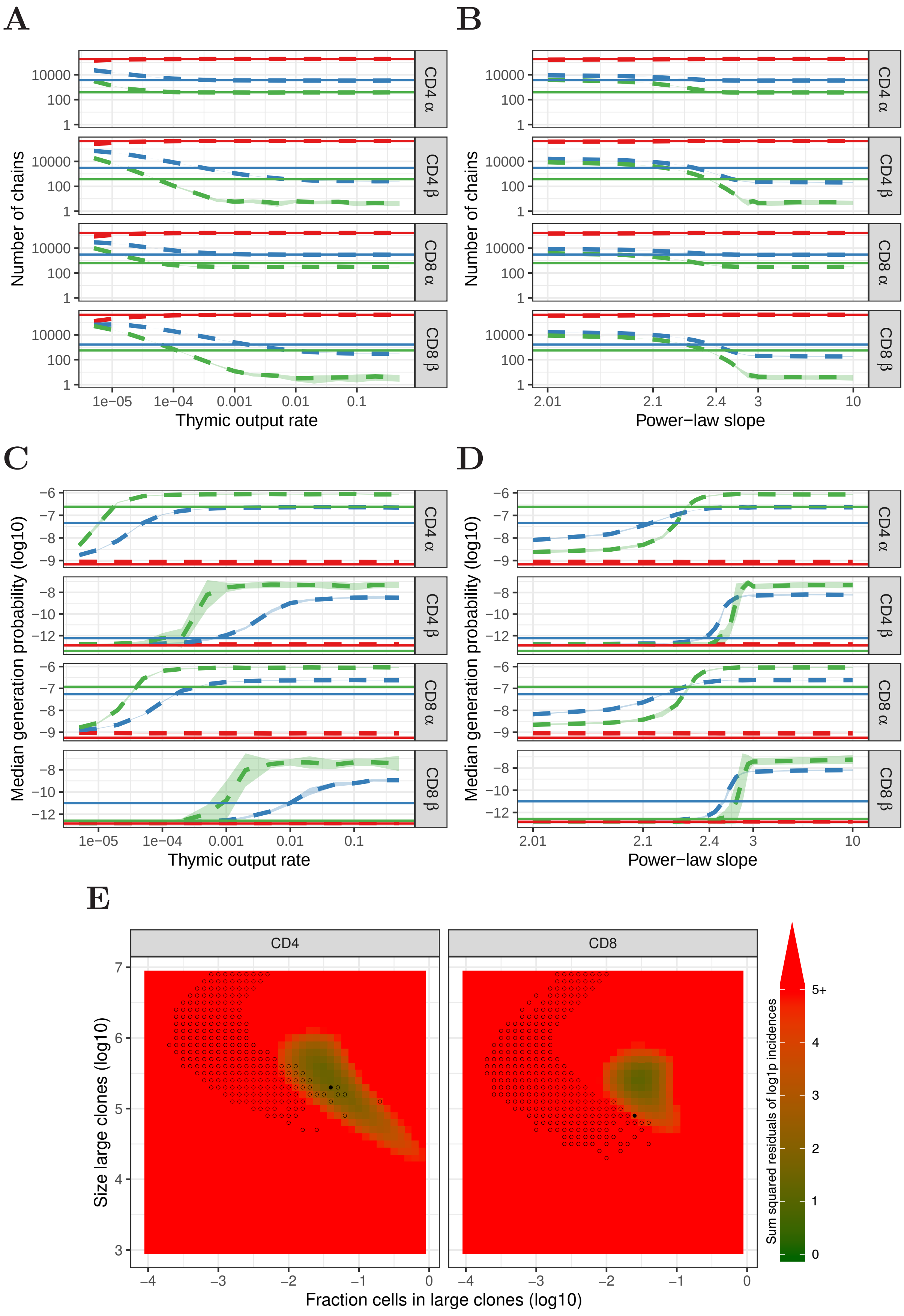
Supplementary Figure S2.5 – **Similar to Figure 2.2, but for HTS data processed with RTCR. A.** The number of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. The grey background bars show the results after removing all sequences that were also observed in the corresponding memory samples. **B.** Generation probabilities *P(σ)* (log10) of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. **C.** Minimal number of N-additions of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. **D.** Number of V- and J-deletions of TCRα and TCRβ sequences observed in 1, 2 or 3 subsamples. The plot shows median (black horizontal line), interquartile range (filled bar) and the range from the bar up to 1.5 times the interquartile range (black vertical range, outliers not shown).



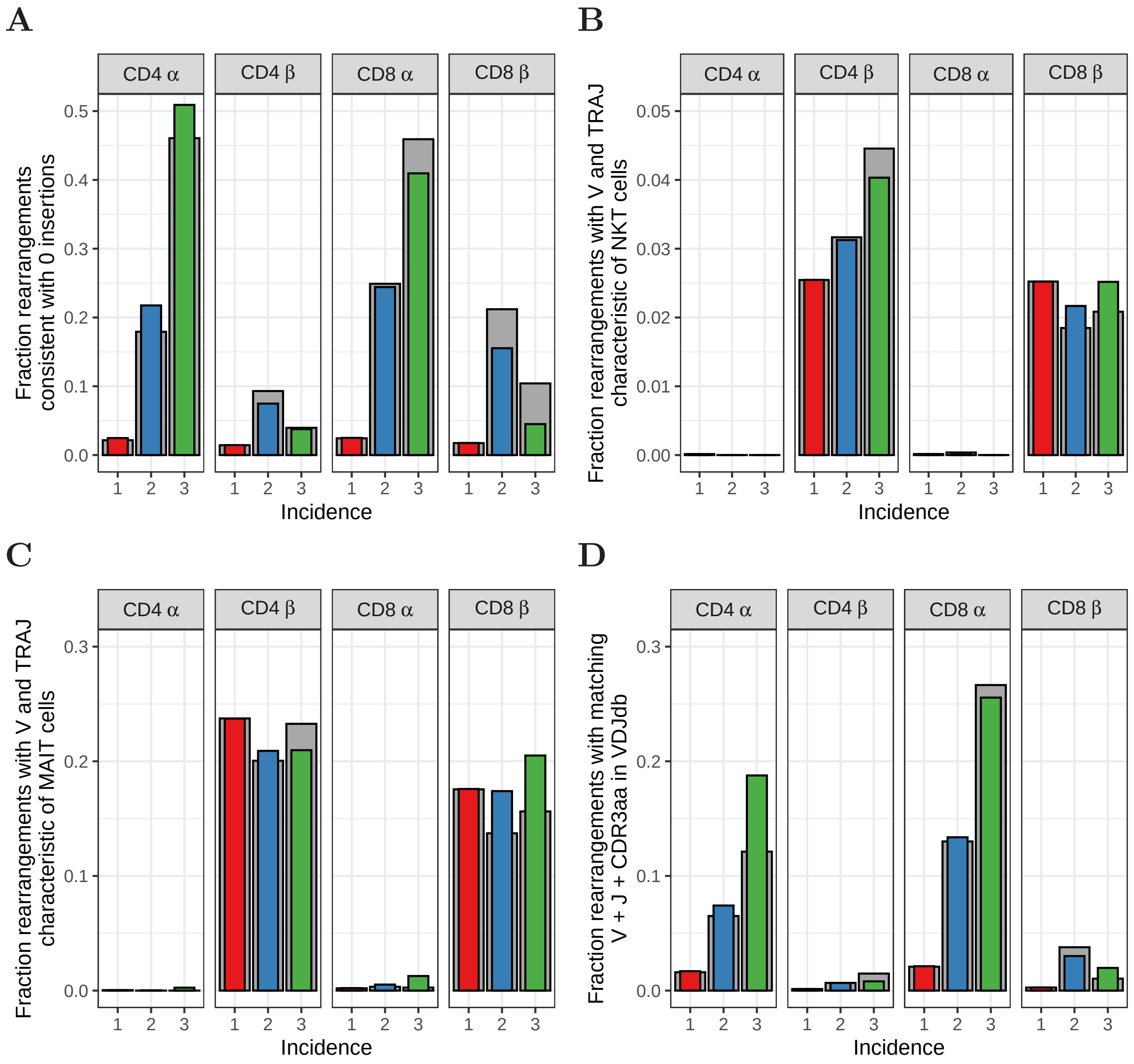
Supplementary Figure S2.6 – **Prediction of power-law model (exponent 2.3) for single sample data.** Red lines indicate abundance of TCRα and TCRβ sequences, both without (top) and with cleaning of overlap with memory (bottom). Blue lines represent model prediction for the abundance of α and β chains (blue dashed line, error bars indicate standard deviation over 10 simulations, note the different predictions between the volunteers due to different sample sizes). Sequences represented with 10 or more UMIs are grouped ("10+"). Due to the impact of sampling multiple RNAs from the same cell, the predictions under-predict the number of doublets, especially for β chains. Apart from this, the model fits well to the “uncleaned” data in most cases, but the cleaned data under-represents abundant clones, suggesting that several clones shared between memory and naive represent genuine abundant naive clones. The only exception is for CD4$^+$ TCRβ, where we observe much fewer large clones than the model predicts.



Supplementary Figure S2.7 – Similar to Figure 2.4, but for HTS data from which TCRα and TCRβ sequences were removed that also occurred in the corresponding memory samples.



Supplementary Figure S2.8 – Similar to Figure 2.4, but for HTS data processed with RTCR.



Supplementary Figure S2.9 – Similar to Figure 2.5, but for HTS data processed with RTCR. A. The fraction of rearrangements with zero minimal N-additions for sequences observed in 1, 2 or 3 naive subsamples. Data are shown without (coloured bars) and with cleaning of overlap with memory (grey bars). **B.** Fraction of TCRα and TCRβ sequences with V(J) usage characteristic of NKT cells (TRAV24-TRAJ18 for TCRα; TRBV11 for TCRβ). **C.** Fraction of TCRα and TCRβ sequences with V(J) usage characteristic of MAIT cells (TRAV1-2 with TRAJ33, TRAJ12 or TRAJ20 for TCRα; TRBV20 or TRBV6 for TCRβ). **D.** Fraction of sequences having at least one match (CDR3 amino acid sequence as well as V and J annotation) with the VDJdb (Shugay *et al.*, 2017).

3

TCRβ rearrangements without a D segment are common, abundant, and public

Peter C. de Greef and Rob J. de Boer

Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands

Proceedings of the National Academy of Sciences 118.39 (2021): e2104367118.

Abstract

T cells play an important role in adaptive immunity. An enormous clonal diversity of T-cells with a different specificity, encoded by the T-cell receptor (TCR), protect the body against infection. Most TCRβ chains are generated from a V-, D-, and J-segment during recombination in the thymus. Although complete absence of the D-segment is not easily detectable from sequencing data, we find convincing evidence for a substantial proportion of TCRβ rearrangements lacking a D-segment. Additionally, sequences without a D-segment are more likely to be abundant within individuals and/or shared between individuals. Our analysis indicates that such sequences are preferentially generated during fetal development and persist within the elderly. Summarizing, TCRβ rearrangements without a D-segment are not uncommon, and tend to allow for TCRβ chains with a high abundance in the naive repertoire.

Introduction

The adaptive immune system relies on large and diverse repertoires of B- and T-lymphocytes. When encountering antigen, cognate lymphocytes start proliferating to clear the pathogen. Many of the cells die after clearance, but others are maintained and form a memory that can be recalled after repeated antigen exposure. The specificity of αβ T-cells is determined by the α and β chain of the T-cell receptor (TCR). These are generated by recombination of variable (V), diversity (D) and joining (J) regions for the TCRβ and V and J for the TCRα chain. During V(D)J-recombination in the thymus, one variant of each of these segments is recombined in a semi-random manner, with deletions and non-templated additions occurring at the junction(s). The combination of the generated β and α chains of the TCR yield an enormous potential diversity ($> 10^{20}$ (Zarnitsyna *et al.*, 2013; Mora and Walczak, 2018)), of which only a small subset is realised in the actual TCR repertoire with a diversity estimated to be around 10^8 (Qi *et al.*, 2014).

The recombination process is guided by recombination signal sequences (RSSs) flanking the V, D and J segments. The RSSs contain spacers of 12 or 23 base pairs (bp), and two gene segments can only be recombined when they have different spacer lengths, a principle that is known as the 12/23 rule. In the TCRB locus, the 3' ends of V and D segments have 23-bp spacer RSSs, while the 5' end of D and J segments have 12-bp spacer RSSs. Following the 12/23 rule, it is therefore possible to have direct V-to-J rearrangements, not including a D-segment. Ma *et al.* studied TCRβ sequencing data in which no D-segment could be identified and estimated that this occurs in about 0.7% of rearrangements in humans (Ma *et al.*, 2016). Previous studies in human cell lines and mice reported V-to-J rearrangements to be rare due to the so called 'beyond 12/23 restriction' (Bassing *et al.*, 2000; Tillman *et al.*, 2003). Another scenario that would lead to the complete absence of the D-segment is a large number of deletions, which may happen before and/or after Terminal deoxynucleotidyl transferase (TdT)-mediated N-additions. It is not possible to uniquely infer the underlying mechanism from TCR sequencing data as different recombination scenarios lead to identical TCRβ rearrangements (Venturi *et al.*, 2011). Moreover, the measured fraction of V-J rearrangements also critically depends on the method used for estimating which nucleotides are derived from the D-segment.

When sequencing TCR α or β chains from samples of T cells, large differences in frequency are observed, even within samples of naive T cells (Quigley *et al.*, 2010; Venturi *et al.*, 2011; Qi *et al.*, 2014; Pogorelyy *et al.*, 2017). Several factors contributing to abundance have been identified in previous studies. TCR chains differ orders of magnitude in their likelihood to be generated, i.e., their generation probability, which can be estimated using generative models (Murugan *et al.*, 2012; Marcou *et al.*, 2018; Sethna *et al.*, 2019). The generation probabilities of TCRα chains correlate well with abundance in the naive repertoire (de Greef *et al.*, 2020), which implies that early or repeated generation of single TCR chains contributes to their abundance. Another factor that contributes to abundance of TCR chains is generation before birth, when N-additions are less likely inserted due to down-regulation of TdT. Such TCR sequences have limited diversity, and were shown to maintain high abundance for decades, while being excessively shared among individuals (Pogorelyy *et al.*, 2017). It should be noted that abundant α or β chains do not provide direct evidence for the existence of large αβ clones in the naive compartment, as one α-chain may pair with many different β-chains (and vice versa) (de Greef *et al.*, 2020). Still, αβ clones could become large as a result of increased division rates in the periphery (Gaimann *et al.*, 2020), which may be due to TCR interactions with self-peptide MHC complexes.

Here we study characteristics of TCRβ sequences that are abundant in the naive T-cell compartment. We find that TCRβ rearrangements without a D-segment are a likely outcome of V(D)J-recombination, but not easily identified. TCRβ chains that are abundant among naive T cells are strongly enriched for having no D-segment. We performed a meta-analysis of TCRβ sequence data, providing evidence for fetal origin of many of such sequences, which may explain why they are shared between so many individuals. Together, this shows that absence of a D-segment is not uncommon in TCRβ rearrangements and that it is an import factor explaining TCRβ abundance in the naive repertoire.

Results

The naive T-cell repertoire contains abundant TCR sequences that lack glycine in their CDR3

The naive T-cell repertoire consists of a huge clonal diversity, of which just a small fraction can be observed in a typical sample of cells. In addition, when RNA is used to sequence the TCRβ chains in T cells, differential TCR expression levels may overestimate the measured abundance of a given T-cell clone (de Greef *et al.*, 2020). We therefore re-analyse the data from (Qi *et al.*, 2014), who sequenced the TCRβ repertoires of memory and naive T-cells from young and aged healthy individuals using five replicates per subset. Measuring the number of samples a given TCRβ appears in, i.e. the incidence, classifies the abundance of sequences without biases due to multiple RNA contributions by single cells. We processed the subsamples independently using RTCR (Gerritsen *et al.*, 2016), which performs clustering of likely erroneous sequences using sample-specific estimates of error rates, while maintaining as much as possible of the diversity.

In line with the results presented in Qi *et al.*, we find that the vast majority of the sequences in the naive T-cell samples appears only in a single subsample, underlining the enormous diversity of the naive repertoire. However, there is also a substantial proportion of TCRβ sequences that are found in two or more subsamples of naive T-cells (Figure 3.1A). The median fraction of sequences with an incidence > 1 was 8.0 times higher in aged than in young individuals, confirming the earlier finding that naive T-cell diversity is lost with age (Qi *et al.*, 2014). We reasoned that some, and in particular the more abundant sequences may be derived from missorted memory T-cell clones. Therefore, we also analysed the effect of discarding all sequences that were also observed in at least one of the corresponding memory samples. Although this correction did remove a larger fraction of the abundant sequences than of those with incidence 1, the incidence of most abundant rearrangements remained unchanged (Figure 3.1A). This confirms that the naive T-cell receptor repertoire of both young and aged individuals contains abundant TCRβ sequences (de Greef *et al.*, 2020).

TCRβ sequences differ several orders of magnitude in their probability of being generated during V(D)J-recombination. To investigate to what extent this relates to abundance in the naive T-cell repertoire, we estimated generation probabilities of the sequences using OLGA (Sethna *et al.*, 2019). The average generation probability of infrequent TCRβ sequences (observed in a single subsample) was very similar among all individuals. The abundant TCRβ sequences, however, were enriched for having a high generation probability in young individuals, albeit to a different extent (Figure 3.1B). This confirms that the likelihood of TCRβ generation, which could reflect repeated thymus production, plays a role in the abundance of TCRβ sequences within the naive repertoire of young adults. Samples from aged individuals, that have much lower (Westera *et al.*, 2015) or even no thymus T-cell production (Thome *et al.*, 2016), contained many more abundant sequences, but showed a much smaller enrichment of high generation probability (Figure 3.1B). These results remained qualitatively similar after cleaning potential contamination by removing sequences overlapping with the memory compartment (dashed lines in Figure 3.1B). Together, these results indicate that likelihood of V(D)J-recombination affects TCRβ abundance in young individuals, and that this effect dilutes with age.

One of the main determinants of the generation probability is the number of N-additions in the rearrangement, since a specific long stretch of N-additions is not a likely outcome of the V(D)J-recombination process. Hence, the observation that abundant sequences in young individuals tend to have high generation probability predicts that they may have shorter CDR3 lengths. Indeed, when we analysed the number of CDR3 amino acids as a function of abundance, we observed on average shorter CDR3s among

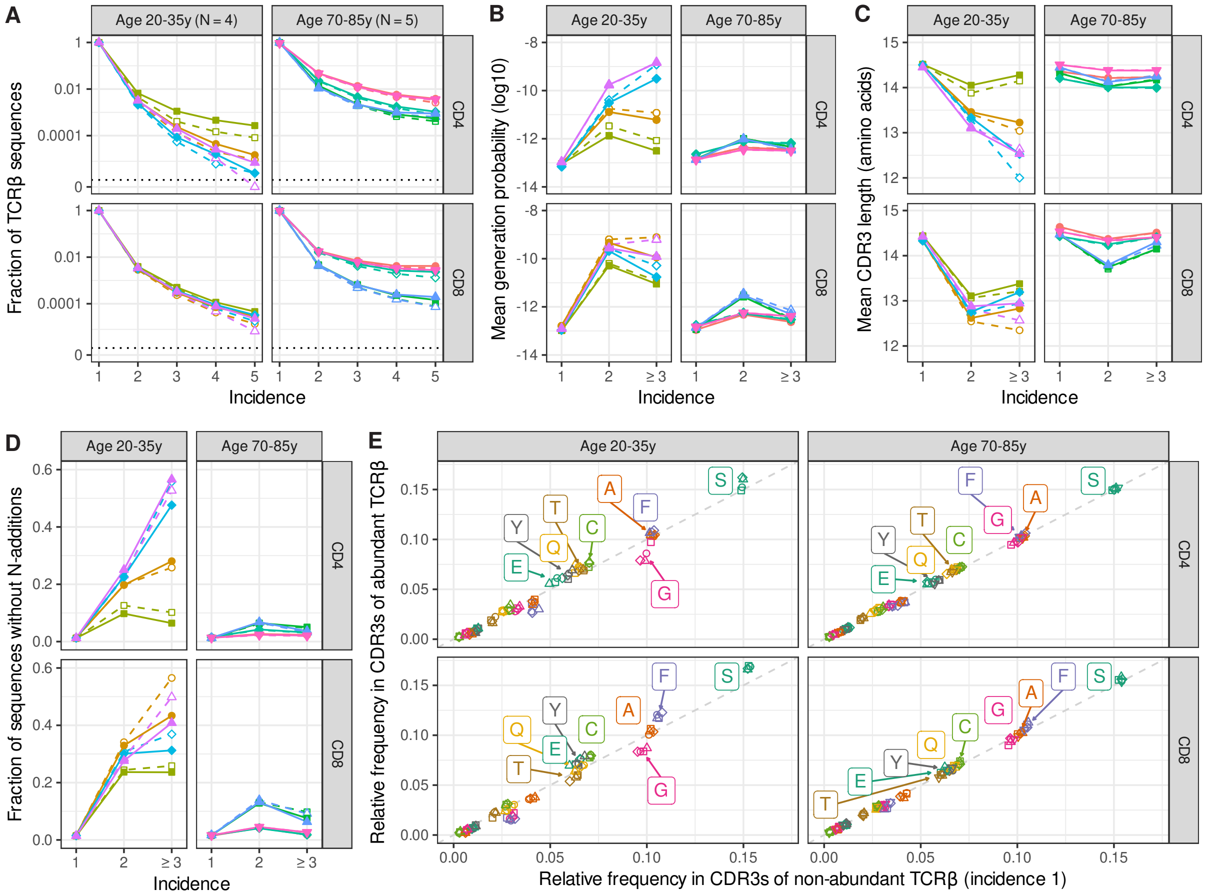


Figure 3.1 – Features of abundant sequences in the naive T-cell repertoire of young and aged individuals. A. Fraction of TCRβ sequences occurring in one or multiple subsamples (i.e., incidence). The vertical axis is log-scaled with 0 added at the bottom. The solid lines and closed symbols are based on all productive sequences, the dashed lines with open symbols show results after removing sequences that were also present in the corresponding sample(s) of memory T cells. Colours and symbols represent the individuals in the Qi *et al.* dataset and are used consistently in the other figures. **B.** Geometric mean of the TCRβ sequence generation probabilities, as a function of their abundance. The most abundant sequences, i.e., those shared across 3-5 subsamples, are grouped together due to the relatively small number of observations. **C.** Mean number of amino acids in the CDR3s of TCRβ sequences, as a function of their incidence. **D.** Fraction of TCRβ sequences without detectable N-additions, i.e., which CDR3 nucleotide sequence can be fully aligned to the identified germline V-, (D-,) and J-segments. **E.** Mean amino acid frequencies among CDR3s of non-abundant (incidence 1) versus abundant (incidence > 1) TCRβ sequences. The amino acid frequency of each CDR3 is calculated as the CDR3 amino acid counts divided by its total length, to account for differential CDR3 lengths. The grey dashed line represents the identity line, and the amino acid identity is shown for those that exceed a relative frequency of 0.05 in any sample. Colours represent the amino acids, plot symbols indicate individuals, the latter being consistent with the other figures.

abundant TCRβ sequences, as compared to the sequences found in only a single subsample (Figure 3.1C). Another factor that is related to N-additions and is reported to play a role in abundance is the generation of TCR sequences before birth (Pogorelyy *et al.*, 2017). The enzyme inserting N-additions (TdT) is downregulated during early ontogeny, making rearrangements without N-additions much more likely during early fetal development. The absence of N-additions cannot be proven for a given rearrangement, as many different recombination scenarios, with and without N-additions, lead to the same TCRβ nucleotide sequence. To still obtain a proxy for the relation between absence of N-additions and abundance, we counted the sequences that are consistent with having no N-additions (i.e., their full CDR3 nucleotide sequence can be mapped to the identified V-, (D-,) and J-segments. Among the TCRβ sequences from young individuals, we found that this was much more common for the abundant sequences than for the sequences with incidence 1 (Figure 3.1D). In aged individuals, both the effects on CDR3 length and absence of detectable N-additions were considerably less pronounced (Figure 3.1C&D).

As the TCRβ sequences are coding for TCR-specificity, we translated them to obtain CDR3 amino acid sequences. We compared the relative amino acid usage in the CDR3 of abundant TCRβ sequences with those observed in only a single subsample (Figure 3.1E). In the samples taken from the aged individuals, we found no relation between amino acid usage and abundance (the points in Figure 3.1E are very close to the diagonal). For the samples from young individuals, however, there were considerable differences between abundant and other TCRβ sequences, especially among the more commonly used amino acids (Figure 3.1E and Figure S3.1A). Within the CDR3s of abundant sequences, there was an over-representation of serine, phenylalanine and cysteine. These amino acids are particularly found at both ends of the CDR3, because they are encoded by either all germline V-segments (S and C) or J-segments (F). The observed enrichment of germline-encoded amino acids is thus to be expected given the observation that CDR3s of abundant TCRβ sequences tend to be shorter (Figure 3.1C). Glycine, in contrast, is rarely encoded by germline V- and/or J-segments. While being the fourth most common acids in CDR3s, it was consistently underrepresented within the abundant sequences of the young individuals, as compared to their sequences with incidence 1. By focusing the analysis on the middle five amino acids of the CDR3, which are most likely to contact the peptide epitope, we also found fewer glycine residues in abundant sequences from young individuals ($p < 0.01$, Wilcoxon signed-rank test; Figure S3.1B). Although glycine residues in the CDR3 could also arise from N-additions, they are generally encoded by the guanine-rich parts of the germline D-segments (Figure 3.2A). Hence, the under-representation of glycine among the CDR3s of abundant TCRβ sequences may reflect the absence of nucleotides derived from the D-segment.

Many abundant TCRβ sequences are V-J rearrangements without a D-segment

We therefore investigated if abundant TCRβ rearrangements in young individuals indeed contain fewer nucleotides that are encoded by the D-segment. It is not straightforward to analyse the D-segment length, as there is no way to reliably tell from the CDR3 sequence

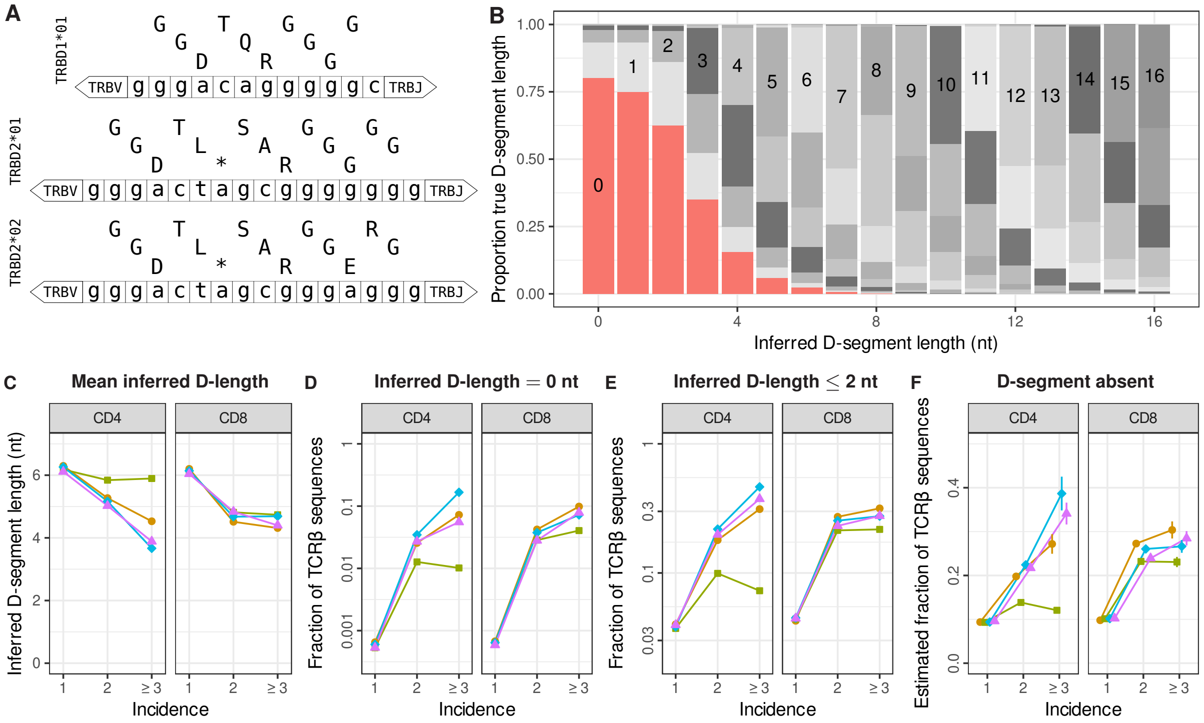


Figure 3.2 – TCRβ sequences occur without a D-segment and are enriched among abundant sequences. **A.** Nucleotide sequences with amino acid translation in the three reading frames of the human TRBD alleles, as listed in IMGT (Lefranc, 2001). **B.** Comparison between inferred and true lengths of D-segments in an *in silico* repertoire of 10^7 productive rearrangements generated using IGoR (Marcou *et al.*, 2018). The proportion of true D-segment lengths is plotted as a function of the inferred D-segment length, which is the maximum region in the non-V/J encoded part of the CDR3 nucleotide sequence matching any D-allele. The bar graph segments are coloured by true D-segment length, with inserted numbers indicating identical true and inferred values. **C.** Mean inferred D-segment length as a function of incidence in the naive repertoires of young individuals (colours matching Figure 3.1). **D.** Fraction of sequences with an inferred D-segment length of 0 nucleotides as a function of incidence. **E.** Fraction of sequences with an inferred D-segment length of 2 or fewer nucleotides, most of which likely representing rearrangements without a D-segment. **F.** Population-based estimate on the fraction of sequences without a D-segment (see Supplement). The expected value is shown with closed symbols, the vertical bars indicate the confidence range (standard deviation).

which nucleotides originated from V/D/J-segments and which from non-templated additions. We thus removed the nucleotides at the 3' and 5' end of the CDR3 that perfectly matched the germline sequence of the annotated TRBV and TRBJ sequence, respectively. The longest match of the remaining sequence with any of the TRBD alleles was taken as a conservative proxy for the D-segment length. We observed a negative relation with incidence in our samples, i.e., sequences shared between samples had on average fewer nucleotides matching a D-segment (Figure 3.2C). This indicates that D-deletions may have a positive effect on the abundance in the naive T-cell repertoire.

As described above, there is no method to accurately measure the D-segment length of any given TCRβ rearrangement. We evaluated the performance of our conservative method by applying it to an *in silico* repertoire generated with IGoR (Marcou *et al.*, 2018). This tool allows one to generate TCRβ sequences with probabilities for gene choices, deletions and additions that are trained on sequence data. The advantage of such generated sequences is that the true D-segment length is known for each recombination scenario. Overall, only 38% of the predictions of our conservative method was correct, which was mainly due to an overestimation of the D-segment length, especially for short D-segments (Figure 3.2B). Intuitively these results can be understood because any N-addition will match at least one nucleotide in any of the TRBD alleles. This means that it is very unlikely to observe the complete absence of the D-segment, implying that potential absence of D-segments in TCRβ sequences is likely overlooked.

We further studied the role of potential D-segment absence on the abundance of TCRβ sequences in the naive T-cell repertoires of young individuals. We started with a very strict threshold, by counting the number of rearrangements with an inferred D-segment length of 0 nucleotides. Note that this requirement only includes sequences in which neither N-additions nor a D-segment is identified. Overall, this feature was very rare in our samples ($< 0.1\%$), but much more common among sequences with higher abundance in the naive repertoire (Figure 3.2D). We also counted the number of sequences with 2 or fewer nucleotides matching a D-segment, accounting for the observation that the majority of rearrangements with an inferred D-length of 1 or 2 nucleotides in the *in silico* repertoire does not have a D-segment (Figure 3.2B). Such sequences were found much more often and made up an even larger fraction of the abundant sequences (Figure 3.2E).

In addition to the classification of individual sequences, we established a quantitative method to make a population-based estimate on the fraction of sequences without D-segment. We first split the *in silico* repertoire into sequences with and without detectable N-additions. For both sets, we calculated the probability of D-segment absence as a function of the inferred D-segment length, using the known true D-segment length of each of these sequences (Figure S3.2A). We then weighted our D-length measurements by these probabilities to estimate the fraction of sequences without D-segment in the TCRβ sequencing data (Supplemental Information). In general, about 10% of the sequences was estimated to not have a D-segment, but this fraction was much higher among the abundant sequences (Figure 3.2F). Together, both methods confirm that there is a substantial fraction of the TCRβ repertoire of naive T cells that does not contain a D-segment.

TCRβ sequences without a D-segment are abundant, functional, and public

The abundant TCRβ sequences in the naive repertoire are enriched for having high generation probabilities (Figure 3.1B), short CDR3s (Figure 3.1C) and no N-additions (Figure 3.1D), but also for having no D-segment (Figure 3.2D-F). We investigated the relative contribution of these factors in more detail by studying TCRβ sequences that were abundant (i.e., those that were shared between subsamples). Generation probabilities appeared to be highest for abundant TCRβ sequences without a D-segment (Figure S3.3A). This reveals that the recombination model, of which individual probabilities were trained on large samples of TCR sequencing data, predicts that (almost) complete absence of the D-segment is likely to occur during TCRβ rearrangement. Importantly, the fraction of TCRβ sequences without detectable N-additions is enriched among the abundant sequences with few or without detectable D-segment nucleotides (Figure 3.3A). All sequences with an inferred D-segment length of 0 nucleotides cannot have N-additions, as any detectable N-addition would be counted as at least one TRBD-derived nucleotide. This means that multiple sequence features, characteristic of abundant TCRβ sequences, are not independent of each other. Hence, these factors may be confounding the analysis of how D-segment absence impacts abundance in the naive repertoire.

To discriminate between the effect of each of these factors, we focused on the sequences in which no N-additions could be identified. We calculated which fraction of these sequences with an inferred D-segment length of 0 nucleotides was abundant, i.e., shared between subsamples of naive T cells. This fraction was relatively consistent between individuals and between CD4$^+$ and CD8$^+$ samples (~15% and ~19%, respectively; Figure 3.3B). The sequences without detectable N-additions but most likely with D-segment (i.e., > 2 nucleotides inferred D-segment length) were not nearly as often abundant (Figure 3.3B). From this set, we also selected sequences with similar generation probabilities, CDR3 lengths or both, to control for a confounding role of these factors. Still, these sequences were much less often abundant than those sequences without a D-segment (Figure 3.3B). So, in addition to absence of N-additions, high generation probabilities and short CDR3s, absence of a D-segment is on its own an important factor affecting the abundance of TCRβ sequences in the naive repertoire.

The ubiquity of TCRβ rearrangements in the naive repertoire lacking a D-segment raises the question if such receptors are functional. We therefore assessed their presence in the memory samples from the same dataset and correlated this with incidence among these samples. The fraction of sequences with an inferred D-segment length of 0 or ≤ 2 nucleotides appeared similar between naive and memory samples (Figure S3.2B-E), suggesting that absence of the D-segment does not affect the probability of participating in an immune response. The strong relation with incidence that was observed for the naive samples, was however absent for samples of memory T cells. We performed a similar D-segment inference method on the human entries in the VDJdb of reported antigen-specific TCR amino acid sequences (Shugay *et al.*, 2017) and found over 1% of sequences to not have

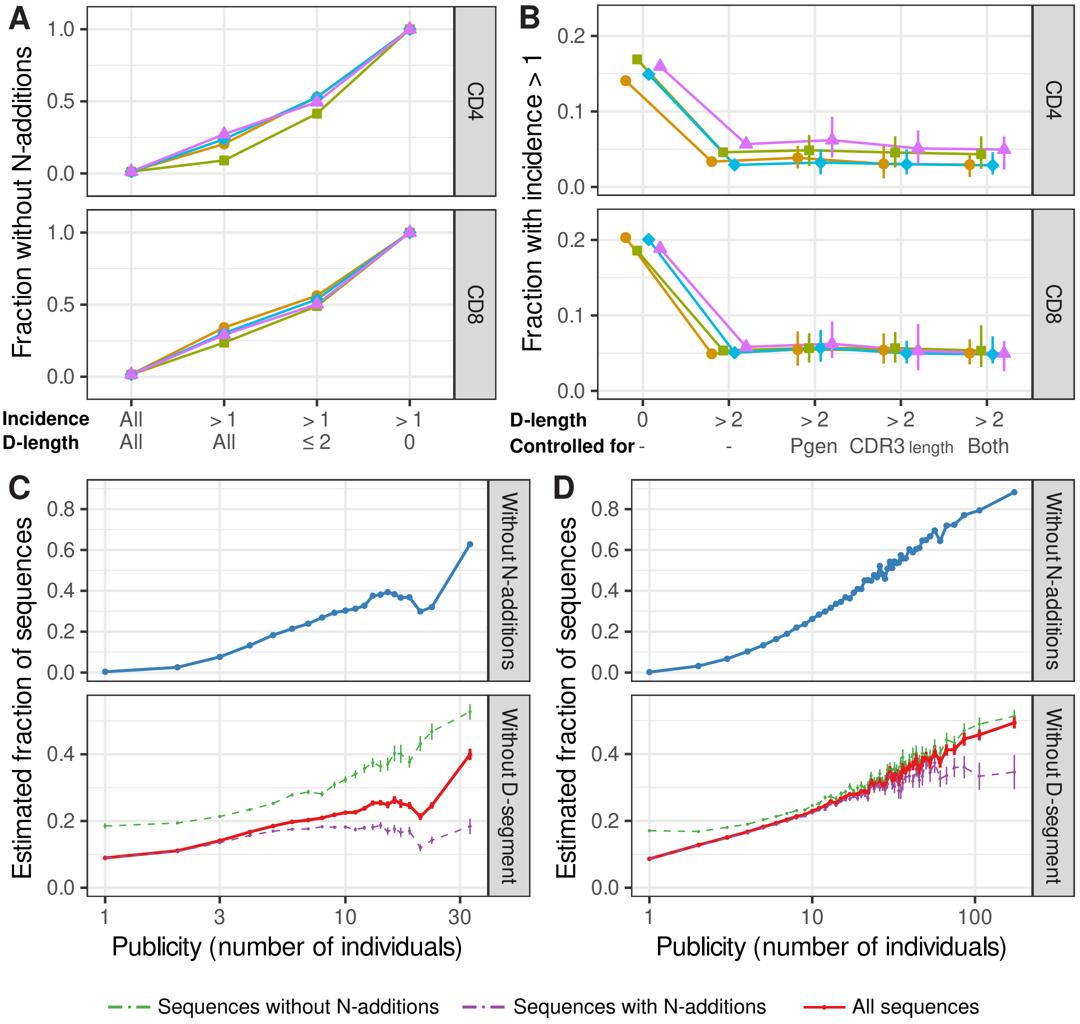


Figure 3.3 – **TCRβ sequences without a D-segment are more often abundant and shared between individuals. A.** Fraction of TCRβ sequences without detectable N-additions, as a function of the incidence and the number of inferred D-segment nucleotides. **B.** Fraction of sequences occurring in multiple subsamples of naive T cells, among the sequences in which no N-additions are identified. Sequences with an inferred D-segment length of 0 nucleotides are compared to other sequences with an inferred D-segment length of more than 2 nucleotides, which indicates presence of a D-segment. The other points show results after selecting sequences that have a similar distribution of generation probabilities (Pgen), CDR3 lengths or both as the sequences without a D-segment (see Supplement). Closed symbols show mean of 100 iterations, total range is indicated with vertical bars. **C&D.** Estimated fraction of sequences without N-additions or without a D-segment, as a function of publicity. Publicity values are measured as the number of samples the TCRβ nucleotide sequence appears in and are binned such that every data point is based on at least 500 TCRβ sequences and shown as a weighted average. Top: fraction of sequences in which no N-additions are identified. Bottom: estimated fraction of sequences without a D-segment among TCRβ sequences with or without detectable N-additions (green and purple, respectively) or all sequences together (red). The expected values are shown as a line, with the vertical bars indicating the confidence range (standard deviation). Data from Britanova *et al.* (C, 73 individuals), and Emerson *et al.* (D, 666 individuals).

D-matching amino acids. Interestingly, common pathogens, like InfluenzaA, EBV, and CMV seem to evoke more responses lacking a D-segment than HIV-1, a more rare pathogen that is typically encountered later in life (Fisher's exact test, $p = 0.002$, Figure S3.3B). Together, these results indicate that TCRβ sequences without a D-segment are not functionally impaired.

The observation that absence of the D-segment causes TCRβ sequences to be abundant within individuals, predicts that such sequences may also be more often shared between individuals. We tested this by analysing inter-individual sharing of productive TCRβ sequences from two published TCRβ datasets of 73 (Britanova *et al.*, 2016) and 666 (Emerson *et al.*, 2017) individuals. In both cohorts, the fraction of sequences without detectable N-additions was much higher among the sequences that were shared between many individuals (Figure 3.3C&D; top). As explained above, this could act as a possible confounder, which we took into account by separately estimating the absence of D-segments among sequences with and without N-additions. In both sets, and also in general, there was a striking relation between publicity and the inferred absence of the D-segment (Figure 3.3C&D; bottom). While absence of the D-segment was not very common among private sequences (~ 10%), this was the case for > 40% of the most public sequences. Together, this confirms that TCRβ sequences without a D-segment are not only abundant within the naive repertoire, but also more likely shared between individuals.

TCRβ sequences without a D-segment are preferentially generated before birth and still present at old age

We wondered why especially TCRβ sequences without a D-segment are abundant in the naive repertoire of young individuals. An explanation could be that they were generated prenatally, when clonal competition may be less restrictive. To test this idea, we studied the samples previously described by (Carey *et al.*, 2017). They sorted $CD8^+$ naive T cells from samples of cord blood from extremely preterm and term neonates and peripheral blood from infants and adults. For these samples, it is even more important to take the effect of N-additions into account, as the enzyme inserting N-additions (TdT) is down-regulated during early ontogeny. In line with this, we found the largest fraction of sequences without N-additions among the preterm cord blood samples (Figure 3.4A; top). This was much lower for the term cord blood samples, indicating that TdT-downregulation already stops long before birth. In the peripheral blood of infants and adults, we found only ~2% sequences without detectable N-additions. We then analysed the inferred D-segment lengths like before and found that the fraction of sequences without D-segment was highest among the preterm cord blood samples (Figure 3.4A; bottom). Interestingly, this was mostly the case due to the sequences that lacked both N-additions and a D-segment (Figure 3.4A; dark purple). This indicates that generation of sequences without N-additions and without a D-segment is most likely during early fetal development and that these rapidly dilute, even before birth.

We tested the persistence of sequences without a D-segment by correlating the estimated fraction of sequences lacking a D-segment with age in the Britanova *et al.* dataset

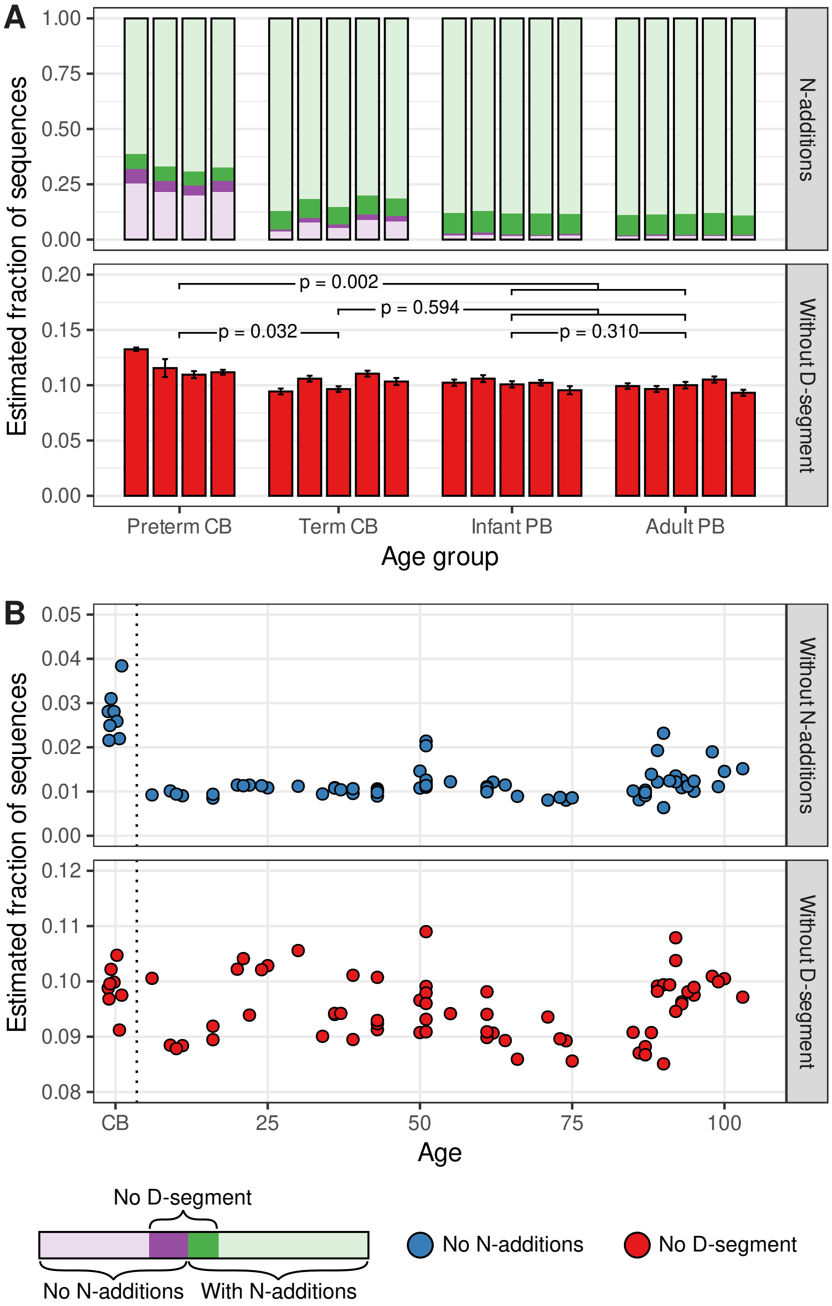


Figure 3.4 – **TCRβs without a D-segment are preferentially generated before birth.** **A.** Absence of N-additions and D-segment in samples of naive CD8$^+$ T cells from cord blood (CB) and peripheral blood (PB), as described in (Carey *et al.*, 2017). Top: Fraction of the sequences without detectable N-additions (purple). The estimated fraction of sequences without a D-segment among sequences with and without N-additions is indicated in dark green and dark purple, respectively. Bottom: Total estimated fraction of sequences without a D-segment, with error bars indicating the confidence range (standard deviation). The p-values are determined with the Mann-Whitney U test. **B.** Absence of N-additions and D-segment, measured in samples of unsorted T cells from cord blood (CB) or peripheral blood, as a function of the individual's age. Due to the large number of sequences per individual, the confidence range (standard deviation) in the bottom panel is too small to be visible. Data from (Britanova *et al.*, 2014).

(Britanova *et al.*, 2016), which includes 8 cord blood samples. It should be noted that these samples contained PBMCs that were not sorted to only include naive T cells. We found an increased fraction of sequences without N-additions in the cord blood samples, in line with the previous results (Figure 3.4B). The estimated fraction of sequences without D-segment, which takes this confounding factor into account, was only slightly higher in cord blood samples compared to the peripheral blood samples (Figure 3.4B; $p = 0.049$, Mann-Whitney U test). This corresponds to the previous observation that the over-representation of sequences without D-segment is mostly limited to the very preterm cord blood samples (Figure 3.4A). Interestingly, we noted that the fraction of sequences without N-additions and also the fraction of sequences without D-segment do not decrease in the elderly of > 80 years (Figure 3.4B). Since there is not much thymic production of new T-cell clones at old age, this indicates that TCRβ sequences without a D-segment may persist longer than other sequences. Together, these results suggest that TCRβ sequences without a D-segment are preferentially generated prenatally, dilute before and after birth, but are maintained until very old age.

Discussion

Here we analysed TCRβ sequencing data from naive, memory and unsorted repertoires to identify sequence characteristics that correlate with abundance. We first confirm that abundant TCRβ sequences in naive T-cell samples of young individuals are characterised by high generation probabilities, short CDR3s and absence of N-additions (Robins *et al.*, 2010; Venturi *et al.*, 2011; Pogorelyy *et al.*, 2017; de Greef *et al.*, 2020). In the aged individuals, there are more abundant sequences, which are less often characterised by these factors. This may be partly explained by the decreased thymus production in the elderly, but also indicates that some T-cell clones are preferentially selected in the periphery due to other factors. In young individuals, where the latter mechanism likely plays a smaller role, we found a relative depletion of glycine in abundant naive TCRβ sequences, indicating a role of the D-segment in abundance. Although it is not possible to reliably measure the number of CDR3 nucleotides originating from the D-segment, we used a conservative method and evaluated its performance on an *in silico* repertoire. The D-segment length inference of individual sequences is not very reliable, but by splitting the data, the quantitative population-based estimates show a substantial population of TCRβ sequences with complete absence of the D-segment in the naive T-cell repertoire. We show that such sequences tend to be much more often abundant in the repertoire and, as a result, more often shared between individuals than other TCRβ sequences.

From our sequencing data, we cannot infer the recombination scenario by which sequences without a D-segment were generated. Following the 12/23 rule, direct V-to-J recombination is possible during TCRβ rearrangement, although several studies reported this to be rare due to the beyond 12/23 restriction. Still, such a scenario cannot be excluded given the enormous number of TCRβ rearrangement events during a human lifetime.

Alternatively, a large number of deletions at the 3' and/or 5' end could remove all D-segment nucleotides. A possible scenario is that N-additions 'protect' the D-segment against excessive deletion. The TdT enzyme, that is responsible for inserting N-additions during V(D)J-rearrangement, is downregulated during early ontogeny (Pogorelyy *et al.*, 2017), which could make complete deletion of the D-segment more likely, and would explain the large fraction of sequences without a D-segment in the cord blood samples from extremely preterm neonates. As a result, complete deletion of the D-segment would become less likely once TdT is activated, causing the rapid dilution of TCRβ rearrangements without a D-segment even before birth. By this time, the TdT-independently generated clones may have undergone multiple rounds of division, increasing their abundance in the naive repertoire (Gaimann *et al.*, 2020), which may be one of the reasons why such rearrangements persist over a human lifetime and even tend to increase in relative frequency in the elderly.

Although our main goal was to describe which sequence characteristics explain abundance in the naive T-cell repertoire, the memory T-cell samples also contain TCRβ sequences without a D-segment. The abundance in the memory repertoire is not affected by absence of the D-segment, however. Abundant TCRβ sequences in the memory compartment likely reflect large clonal expansions rather than the more subtle differences within the naive repertoire. Still, the existence of sequences without a D-segment in samples of memory T cells indicates that such rearrangements are functional and participate in immune responses. Moreover, we find that about 1% of the reported TCRβ sequences specific for common pathogens does not have any D-matching amino acid in the CDR3 (Figure S3.3B). The observation that this percentage is higher for common viral pathogens than for the more rare HIV-1 makes it tempting to speculate about the effect of age at which individuals get exposed to the pathogen. Most people get exposed to common pathogens at young age, when a relatively large fraction of naive T cells originates from prenatally generated clones. Exposure to HIV-1 is much more likely when these sequences are strongly diluted already. If this were the case, it would explain the higher generation probabilities of TCRs specific for common antigens without needing to invoke the previously suggested evolution of the recombination machinery towards TCRs specific for common pathogens (Thomas and Crawford, 2019).

Together, our study highlights absence of the D-segment as an important determinant for TCRβ abundance in the naive T-cell repertoire. Many of them are likely generated long before birth, when TdT is still down-regulated. Such sequences are often shared and present at very old age, indicating that the TCR repertoire maintains TCRβ chains that resemble TCRα chains.

Materials and Methods

All sequencing data described in this study was collected in previous studies and downloaded from the NCBI and Adaptive Biotechnologies servers. Raw sequencing data was processed using RTCR (Gerritsen *et al.*, 2016). All analyses were restricted to productive rearrangements, and individual sequences were defined by the combination of V gene, CDR3 nucleotide sequence, and J gene. The D-segment length was inferred as the maximum match of the inferred inter-V-J sequence with any of the TRBD alleles. Detailed information on all analyses is given in the Supplemental Information.

Author contributions

P.C.d.G. and R.J.d.B. conceived the study. P.C.d.G. analysed the data with input from R.J.d.B. P.C.d.G. wrote the manuscript, which was edited and approved by both authors.

Acknowledgements

This work was supported by The Netherlands Organization for Scientific Research Graduate Program 022.005.023 (to P.C.d.G.). We gratefully acknowledge all authors involved in generating the TCR sequencing datasets we studied, for providing access to their data.

Supplemental information

Data sources

The dataset by Qi *et al.* (Qi *et al.*, 2014) was obtained from dbGaP found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi- bin/study.cgi?study_id=phs000787.v1.p1 through dbGaP study accession number PRJNA258304. These data (project "Immunosenescence: Immunity in the Young and Aged") were provided by Jorg Goronzy on behalf of his collaborators at PAVIR and Stanford University. In this study, five replicates with each 10^6 cells per aliquot of naive and memory CD4 T cells were collected. For CD8 T cells, 0.25×10^6 T cells were collected per replicate, except for the naive CD8 T cells from young individuals, from which 10^6 cells per aliquot were analysed. The TCRβ repertoire sequence data of 666 individuals was downloaded from the Adaptive Biotechnologies website (originally published as (Emerson *et al.*, 2017)). The data of cord blood and peripheral blood TCRβ repertoires (Carey *et al.*, 2017) were downloaded from the same website. The results presented in Figure 3.3C and Figure 3.4B are based on data from (Britanova *et al.*, 2016) downloaded from the NCBI SRA archive Bioproject accession PRJNA316572.

Sequence analysis

After pairing reads with Paired-End reAd mergeR (PEAR) (Zhang *et al.*, 2014), the reads from Qi *et al.* were processed using Recover TCR (RTCR) (Gerritsen *et al.*, 2016). Each sample of naive or memory cells was split into five subsamples that were sequenced separately. RTCR estimates a per-sample error rate to account for the inevitable inaccuracies that occur during PCR and sequencing errors. To prevent the occurrence of high-incidence reads by the error-correction clustering algorithm, we processed each subsample separately. When running RTCR, we did not perform Unique Molecular Identifier (UMI)-guided error correction. This was because the incorporated UMI sequences were composed of only 4 nucleotides. Thus there are only 256 unique combinations possible, which does not allow for collapsing of PCR duplicates into reliable consensus sequences. As we did not use within-sample read counts, but only used the incidence in multiple samples as a measure for abundance, our results should not be influenced much by uneven PCR amplification. After error-correction, reads were filtered following default RTCR settings, i.e., if V and J were in-frame and the CDR3 did not contain in-frame stop codons or ambiguous bases. TCRβ sequences were defined by the combination of CDR3 nucleotide sequence, TRBV gene and TRBJ gene.

The data by Britanova *et al.* was processed with RTCR using the barcode files given at https://github.com/ mikessh/aging-study. First, UMIs were extracted in forward and reverse reads using the Checkout algorithm of RTCR. UMI-guided consensus sequences were generated using the umi_group_ec algorithm, which were processed with the main pipeline of RTCR using default settings.

The data from Emerson *et al.* was already processed. TCRβ sequences were extracted from the column "rearrangement", V and J genes from the columns "v_gene" and "j_gene", respectively. Out-of-frame sequences, those with an unresolved V or J gene or containing an in-frame stop codon, were filtered out and not used for analysis. Publicity was measured as the number of samples that contained the combination of CDR3 nucleotide sequence, V gene and J gene.

For analysis of the Carey *et al.* data, TCRβ CDR3 nucleotide sequences were taken from the column "cdr3_rearrangement", V and J genes from the columns "v_gene" and "j_gene", respectively. Out-of-frame sequences, those with an unresolved V or J gene or containing an in-frame stop codon, were filtered out and not used for analysis.

Inference of D-segment length

Many different recombination scenarios can lead to the exact same sequence, e.g., after deletion of nucleotides they can be added again as N addition. As these differences are not visible in the sequence, it is impossible to uniquely tell which nucleotides are encoded by V, D or J segment, and which by N additions. To still estimate the number of nucleotides originating from the D segment, we started with a conservative approach, by taking the CDR3 nucleotide sequence and matching the nucleotides at the 5' end to the germline sequence of the identified V gene segment. The first mismatch position is assumed to be the end of the V-segment, although technically this mismatch could also occur due to e.g. a sequencing error. The same procedure is followed by matching the 3' end of the remaining sequence to the identified J gene segment. Any remaining nucleotides could be a mixture of a D gene segment, N additions and P additions. We inferred the length of the D segment by taking the longest exact match of any of the three germline TRBD allele sequences (as listed in IMGT (Lefranc, 2001): TRBD1*01: GGGACAGGGGGC; TRBD2*01: GGGACTAGCGGGGGGG; TRBD2*02: GGGACTAGCGGGAGGG) with the inter-V-J sequence.

To evaluate the performance of the D-segment length inference method, we generated 10^8 TCRβ sequences using the default generation model of IGoR without sequencing errors.

We randomly selected 10^7 sequences that were productive, i.e. in-frame and not containing a stop codon, and which V- and J-segments were in the RTCR germline reference set (e.g., excluding pseudogenes). For each sequence, we inferred the D-segment length as described above. We compared this to the true D-segment length, by subtracting d_5_del and d_3_del (if positive) from the total length of the selected D segment. The agreement between true and inferred D-segment lengths is shown in Figure 3.2B.

We also split the *in silico* repertoire of Figure 3.2B based on the presence of detectable N additions, i.e., nucleotides in the inferred inter-VJ sequence that did not match any of the TRBD allele sequences (Figure S3.2A). This improves the performance of our inference method, e.g. since there are no sequences with N additions and an inferred D-segment length of 0 nucleotides, as at least one N addition would be counted as derived from the D segment. The probability for a sequence in the *in silico* repertoire to not contain a D segment, given its inferred D-segment length and the presence/absence of detectable N additions, is shown in red in Figure S3.2A. We used these probabilities to obtain a population-based estimate on the fraction of rearrangements without D segment in the sequencing data: the expected value was calculated by weighing the inferred number of D-segment-derived nucleotides with these probabilities. We also determined the confidence range on this estimate by re-sampling 100 times with these probabilities, reporting the standard deviation on the estimated fraction of sequences without a D segment (e.g., in Figure 3.2F and Figure 3.3C&D).

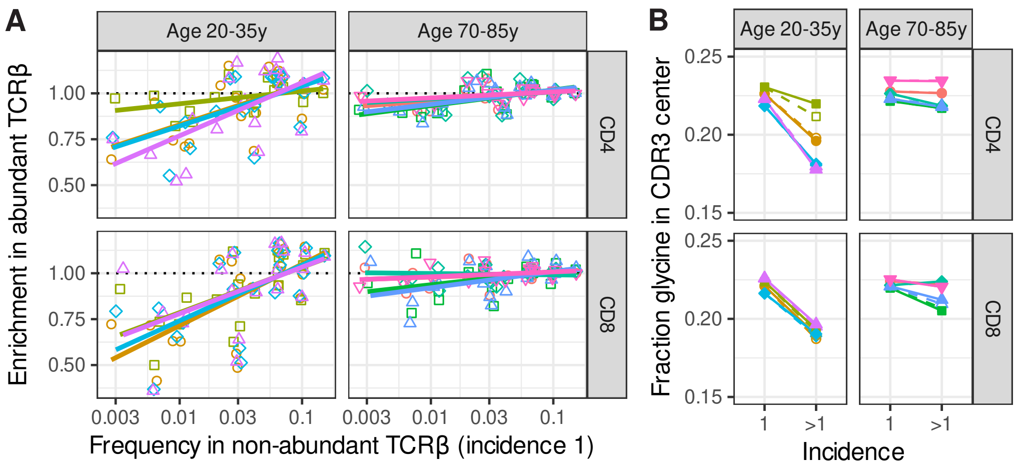
Controlling for confounders

Various measures correlate with abundance and are not independent of each other (e.g., absence of N additions and inference of 0 D-derived nucleotides). In Figure 3.3B we therefore analyse the subset of sequences that has no detectable N additions. Those with an inferred D-segment length of 0 nucleotides (Set A) are compared with sequences with more than 2 nucleotides inferred D-segment length but also no detectable N additions (Set B). When controlling for generation probabilities, CDR3 lengths or both, we selected sequences from Set B that were matching the characteristics of the sequences in Set A. This was possible for on average 96.2% of the sequences of Set A (range: 95.2-97.3%) and was done 100 times. We also performed a logistic regression analysis on all data, predicting the probability to be abundant as a function of generation probability (log10), CDR3 length, presence of detectable N additions and presence of D-derived nucleotides, the latter both with the threshold of 0 and 2 nucleotides. The results of this analysis are summarised in Table S3.1 and confirm the effect of D-segment absence on abundance in the naive repertoire.

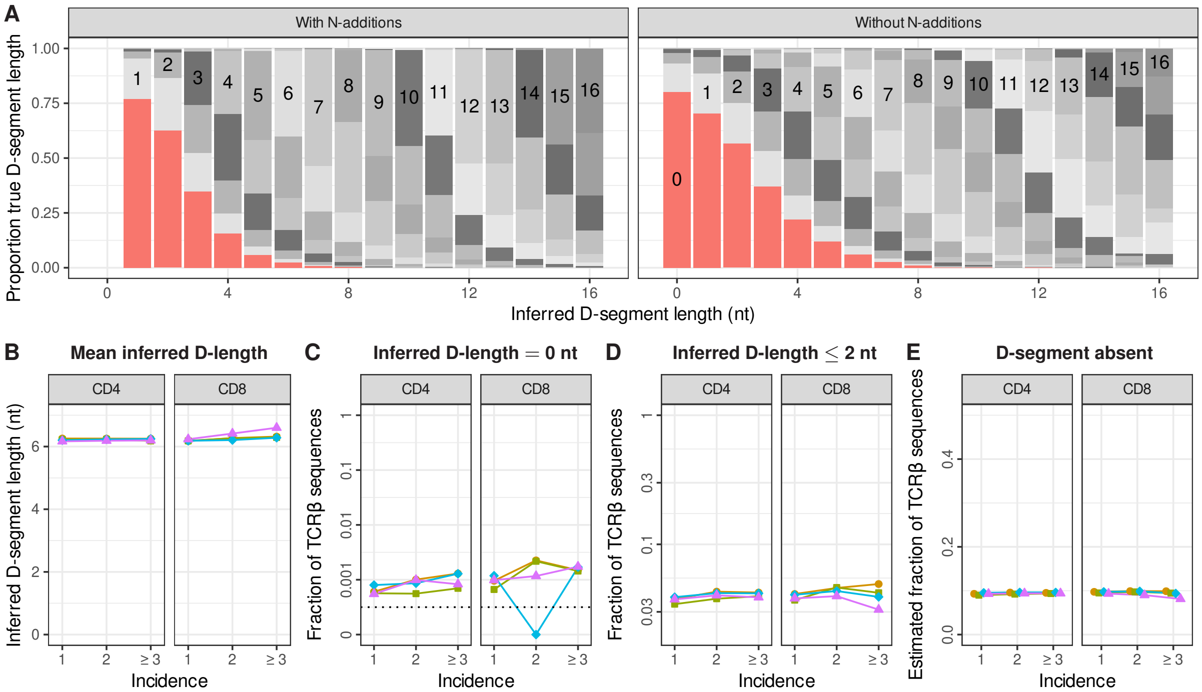
Supplementary Table S3.1 – **Logistic regression analysis**

	Inferred D-segment length = 0 nt	Inferred D-segment length <= 2 nt
	β (SE)	β (SE)
Generation probability (log10)	0.345 (0.004)	0.341 (0.004)
CDR3 length (nt)	-0.045 (0.004)	-0.001 (0.004)
No N additions	1.679 (0.017)	1.673 (0.017)
No D-segment	1.118 (0.039)	0.796 (0.016)

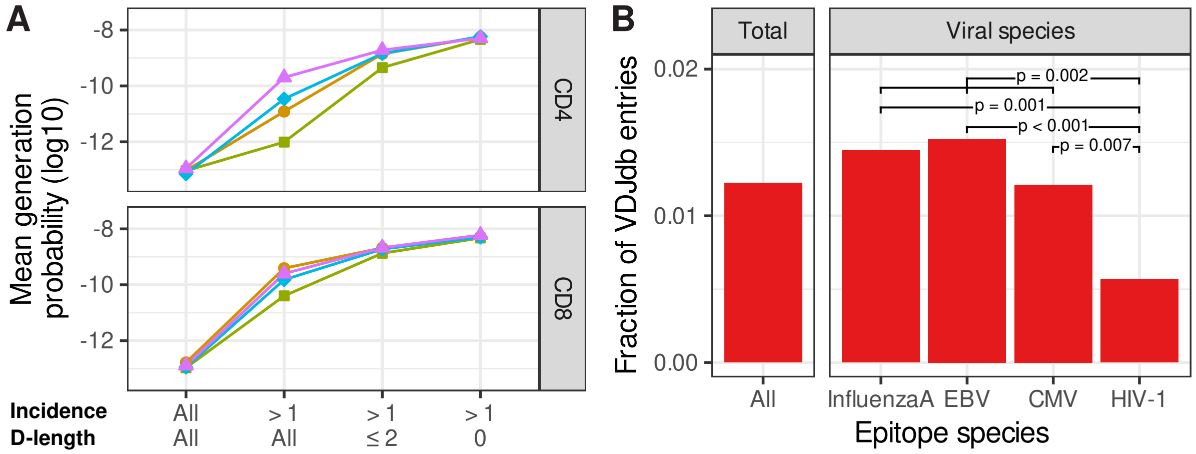
Supplementary Figures



Supplementary Figure S3.1 – **Supplemental to Figure 3.1.** **A.** Enrichment of relative CDR3 amino acid values as a function of their proportion among sequences with incidence 1 (log-scaled). Enrichment is calculated by dividing the relative usage among sequences with incidence 1 by the relative usage among sequences with incidence > 1. Each colour represents a different individual, matching the colours in the main text figures. The positive slopes of the linear regression lines indicate that common amino acids tend to be enriched in abundant sequences. **B.** The mean proportion of glycine residues among the centre 5 amino acids of the CDR3 from TCRβ sequences with incidence 1 or higher. The solid lines and closed symbols are based on all productive sequences, the dashed lines with open symbols show results after removing sequences that were also present in the corresponding sample(s) of memory T cells.



Supplementary Figure S3.2 – **Supplemental to Figure 3.2.** **A.** Comparison between inferred and true lengths of D segments in an *in silico* repertoire of 10^7 productive rearrangements generated using IGoR (Marcou *et al.*, 2018), like in Figure 3.2B. Sequences are split based on presence (left) or absence (right) of detectable N additions. The proportion of true D-segment lengths is plotted as a function of the inferred D-segment length, which is the maximum region in the non-V/J encoded part of the CDR3 nucleotide sequence matching any D-allele. The bar graph segments are coloured by true D-segment length, with inserted numbers indicating identical true and inferred values. The red bars show the true fraction of rearrangements without a D segment (i.e., having a D-segment length of 0 nucleotides), as a function of the inferred D-segment length. Note that there are no sequences with N additions and an inferred D-segment length of 0 nucleotides, as at least one N addition would be counted as derived from the D segment. **B.** Mean inferred D-segment length as a function of incidence in the memory repertoires of young individuals (colours matching Figure 3.11). **C.** Fraction of sequences with an inferred D-segment length of 0 nucleotides as a function of incidence among memory samples. **D.** Fraction of sequences with an inferred D-segment length of 2 or fewer nucleotides, most of which likely representing rearrangements without a D segment. **E.** Population-based estimate on the fraction of sequences without a D segment. The expected value is shown with closed symbols, the vertical bars indicate the confidence range (standard deviation), which is very small due to the large number of observations.



Supplementary Figure S3.3 – **Supplemental to Figure 3.3.** **A.** Median TCRβ generation probabilities as a function of the incidence and the inferred D-segment length. **B.** Inference of the D-segment length on CDR3 amino acid sequences in the VDJdb, retrieved on 8 January 2021 (Shugay *et al.*, 2017). Like for nucleotide sequences, we assigned matching CDR3 amino acids to the translated germline V and J sequences. In the remaining amino acids, we used the maximum match with any of the reading frames of translated TRBD alleles (Figure 3.2A) as a proxy for the number of D-encoded amino acids in the CDR3. Shown are the fractions of unique sequences having 0 D-encoded amino acids among all human TCRβ records (left), and those specific for the viral epitope species with at least 1000 unique V-CDR3-J combinations (right). P-values are determined with Fisher's exact test.

4

Towards a robust comparison of diversity between sampled TCR repertoires

Peter C. de Greef and Rob J. de Boer

Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands

Preprint at bioRxiv (2023): 10.1101/2023.02.10.528010.

Abstract

T-cell receptor (TCR) repertoire sequencing data provides quantitative insight into the distribution of T-cell clones. The diversity of the TCR repertoire in humans tends do decrease with age, which may be a key determinant explaining immune senescence in older individuals. To address this, we first analyse how the diversity of a potential T-cell response against an unseen pathogen changes with age. Next, we discuss the complications with interpreting the outcomes of such an analysis. Specifically, the changes in T-cell subset sizes confound analyses of TCR diversity, and typical sample sizes do not easily allow for a robust quantification of this diversity. Thus, explaining immune senescence as a result of decreasing TCR diversity is far from straightforward and requires a detailed, robust, and quantitative analysis.

Introduction

The human TCR repertoire has a unique composition that results from thymic production and selection, as well as exposure to antigens in the periphery. The extreme diversity of TCR sequences makes comparison of the TCR repertoire between individuals a challenging task. In addition, the Human Leukocyte Antigens (HLAs) are highly polymorphic in the human population, implying that a TCR may bind different antigens in other individuals. A measure that summarises the entirety of a TCR repertoire is its diversity. In ecological studies, diversity is often measured considering the richness, which is defined as the total number of species in a system, and the evenness, which quantifies to which extent these species differ among each other in frequency (Chao *et al.*, 2020). In a TCR repertoire context, the species are the distinct TCR sequences, and their diversity can be estimated using high-throughput TCR sequencing.

The generation of new T-cell clones decreases with age, to an extent that naive T-cell production by the thymus is diminished (Westera *et al.*, 2015) or even absent (Thome *et al.*, 2016) in older individuals. This decreasing source of new diversity may lead to reduced T-cell immunity since TCR diversity is a key feature of a functional T-cell pool. The scenario that the TCR repertoire lacks T-cells that are specific for a given foreign antigen has been described as 'holes in the repertoire' (Yager *et al.*, 2008). Such an absence of required T-cell specificities may be the result of reduced TCR richness, illustrating the need for accurate estimates of this richness. Here we present an intuitive analysis in which we estimate how many T-cell clones would be recruited against an unseen pathogen. This case study provides quantitative insights into the potential losses of responses with age, but also highlights key caveats of such an analysis. We discuss important challenges of estimating TCR diversity based on sequencing data of sampled T cells. These insights will help to refine future experiments and analyses to better compare the TCR diversity between sampled repertoires.

Results and Discussion

The estimated richness of a putative response and the total repertoire decreases with age

A reduced TCR repertoire diversity, leading to a failure to mount a T-cell response, should be reflected in a strongly reduced number of TCR sequences specific for a given pathogen. We tested this by using the VDJdb, which is a database that lists TCR sequences that are found to be specific for certain epitopes (Shugay *et al.*, 2017). We checked the occurrence of such sequences specific for a particular pathogen across a large cohort of individuals up to an age of about 70 years (Emerson *et al.*, 2017). Since the vast majority of the population in Western countries is HIV-negative, HIV-1 will be an unseen pathogen to most individuals in this dataset. So, the number of TCRβ sequences specific for an HIV-1 epitope serves as a proxy for a putative T-cell response against a pathogen without previous exposure. This is important, as both the VDJdb and the TCR repertoires in this dataset will be enriched for specificities towards common pathogens. We counted the number of matches between the HIV-1 entries in the VDJdb and the TCRβ repertoires that were size-normalised to exclude heterogeneous sample sizes as a confounding factor (see Methods). Interestingly, although the data only covers a tiny portion of each total T-cell repertoire, we found such sequences in all individuals, across the entire age range (Figure 4.1A).

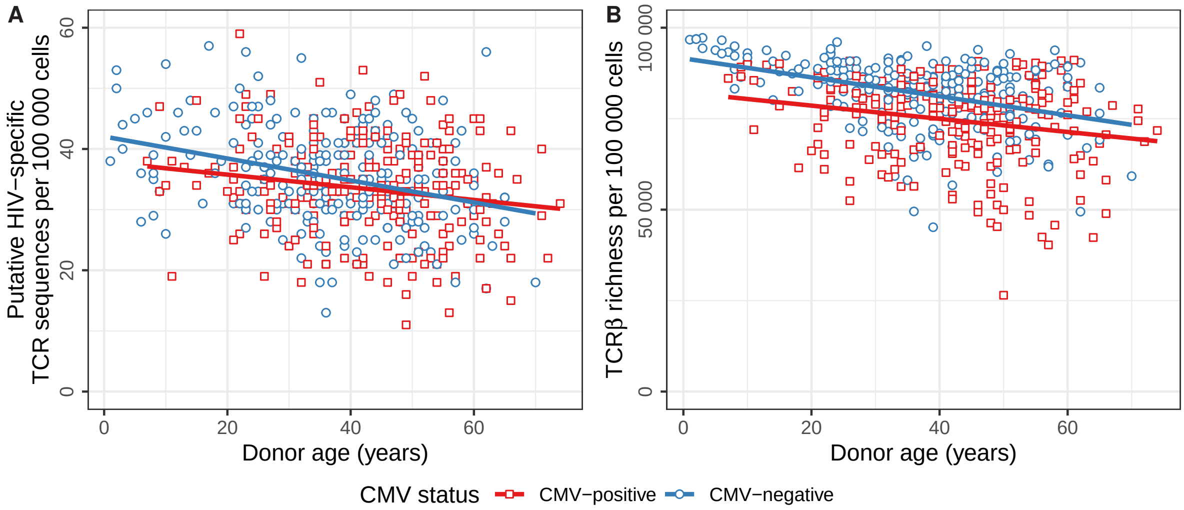


Figure 4.1 – **The richness of unsorted T-cell repertoires tends to decrease with age and CMV. A.** Total number of distinct TCRβ nucleotide sequences from (Emerson *et al.*, 2017) that match the VDJdb (Shugay *et al.*, 2017) as being specific for an HIV-1 epitope in CMV-positive (red boxes) and CMV-negative (blue circles) individuals. TCRβ sequences are counted as a match with the database if their CDR3 amino acid sequence as well as their V- and J-gene families are identical. Differences in sample size are normalised for by randomly sampling a 100 000 templates from each complete sample (see Methods). **B.** Total number of distinct TCRβ nucleotide sequences among 100 000 unsorted T cells. Linear regression lines are shown for CMV-positive (red) and CMV-negative (blue) individuals.

The richness of the putative HIV-1 response varies considerably between donors, even of the same age. This reflects a large individual heterogeneity, for example due to different HLA compositions within the cohort. A linear regression analysis, with age as the independent variable, showed that the number of putative HIV-1-responsive clones tends to decline, both with age and CMV-infection (Figure 4.1A). This implies that the number of responding clones tends to decrease with age, and that at an age of 60 years about a quarter of the diversity found in children is lost. The data also suggest that CMV-infection reduces the number of HIV-1-responsive clones, probably due to the repertoire being skewed towards a limited number of expanded CMV-specific T-cell clones. Note that the TCRα chain was not sequenced but also determines the TCR specificity, and that being reported as binding to epitopes may be restricted to HLA alleles that are absent in many donors. Our analysis thus does not exactly quantify the T-cell response against HIV-1 epitopes. However, the decrease in reported TCR hits is expected to reflect an actual decrease in richness of the T-cell response against an unseen pathogen. To place these pathogen-specific results into a more general context, we also quantified the overall changes in repertoire richness with age. In line with previous studies (Britanova *et al.*, 2014; Krishna *et al.*, 2020; Qi *et al.*, 2014; Yoshida *et al.*, 2017), we found a moderate decline in richness per normalised number of sequences with both age and CMV-infection (Figure 4.1B).

TCR repertoires are dominated by naive T cells in young and by memory T cells in older individuals

Importantly, the TCRβ dataset we used consists of sequences that were observed in a peripheral blood sample, that was not sorted to only contain a specific T-cell subpopulation. Naive T-cell frequencies are a major determinant of TCR repertoire richness because encounter with antigen leads to proliferation, and the resulting effector and/or memory clones will mostly persist at a higher frequency than the initial frequency of the naive T-cell clone. It is important to make a distinction between the number of naive T cells per unit of blood, and the percentage of naive T cells among other subsets. For example, when the memory compartment grows with age, the naive T-cell diversity does not have to decrease if its total pool size remains stable, while the percentage of naive T cells would decrease. In line with this, Wertheimer *et al.* reported that the percentage of naive T-cells decreases significantly with age, while the absolute number of naive CD4 T cells per μl of blood is rather similar between young and older individuals, if they are CMV-negative (Wertheimer *et al.*, 2014).

We consulted various studies reporting the number of naive T cells per volume of blood to estimate how the naive T-cell pool size changes with age (Figure 4.2A). The reported counts vary widely between individuals and different studies. Part of this variation may be explained by the different markers that are used to sort the naive sub-population from blood, some being more stringent than others. The studies mostly agree on the observation that naive CD8 T-cell numbers in blood decrease strongly with age (Figure 4.2A; right), implying that the total number of naive CD8 T cells is much smaller in older than in young individuals. However, some of the naive CD8 T cells may have undergone phenotypic changes without having responded to foreign antigen, for example becoming virtual memory cells. The effects of age on the naive CD4 T-cell numbers in blood are less pronounced (Figure 4.2A; left), suggesting that the absolute size of the naive CD4 T-cell pool does not change dramatically with age, while the relative frequency of naive CD4 T cells may decrease substantially.

To assess to which extent the observed changes in richness as described above may reflect relative changes in subset sizes, we defined a simple subset classifier. We split each TCR repertoire into a putative naive and effector/memory fraction (see Methods). We again plotted the size-normalised richness of each repertoire, like in Figure 4.1B, but now

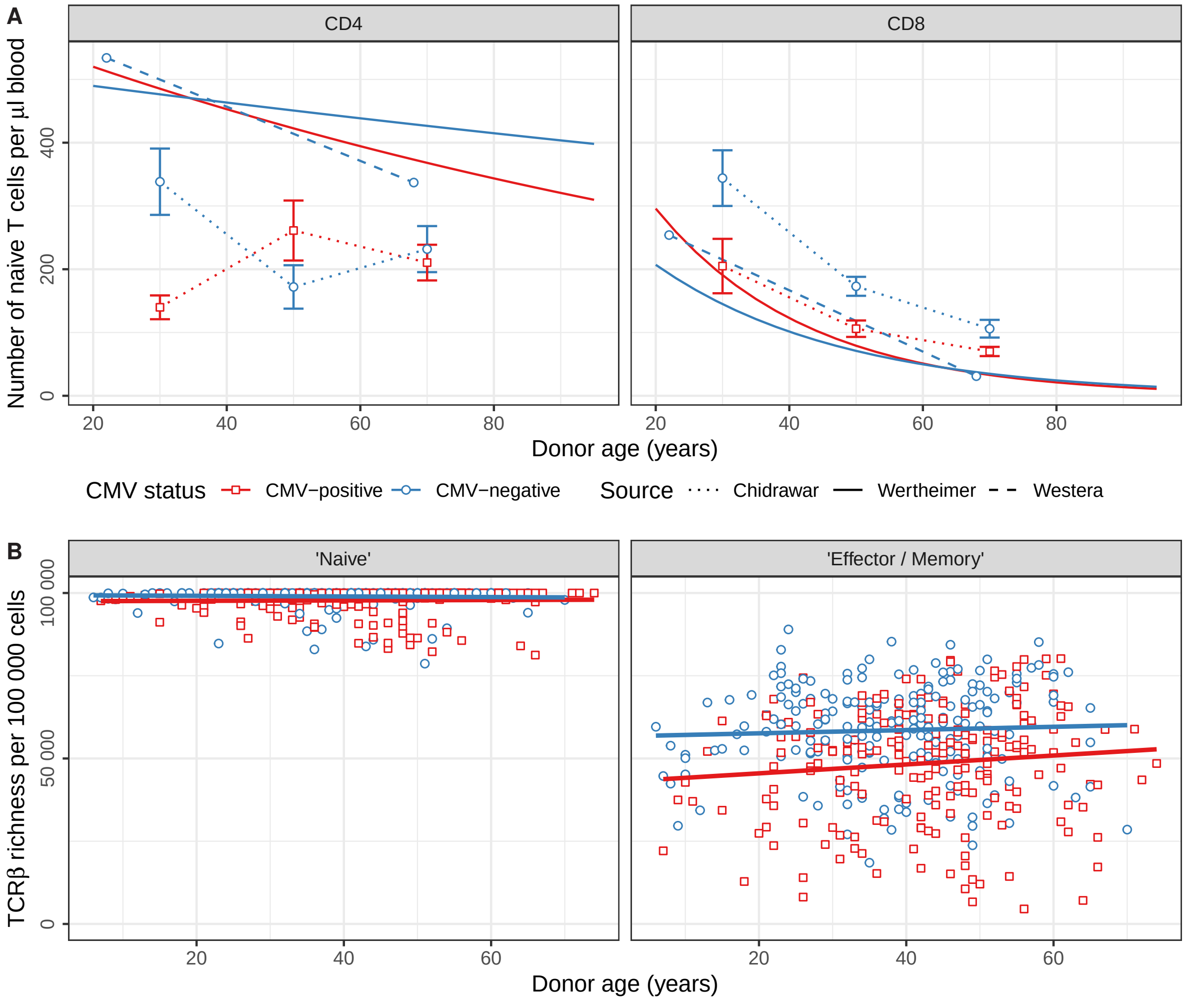


Figure 4.2 – **The observed T-cell diversity largely depends on the relative frequency of naive T cells. A.** Estimated number of naive CD4 and CD8 T cells per µl of blood as reported in three studies. Solid lines show the results of a regression analysis (Wertheimer *et al.*, 2014), the dashed line indicates the median values of a young and aged age group (Westera *et al.*, 2015), and the dotted lines connect the median values per age group, with the error bars indicating the reported standard deviation (Chidrawar *et al.*, 2009). **B.** Size-normalised richness (similar to Figure 4.1B) after splitting the TCRβ repertoires into a 'naive' and 'effector/memory' fraction (see Methods). Linear regression lines are shown for CMV-positive (red) and CMV-negative (blue) individuals.

separately for the two inferred subpopulations (Figure 4.2B). The absence of a decrease in TCR repertoire richness in both compartments suggests that the diversity of each subpopulation may be rather stable with age. Although the separation between naive and effector/memory will be far from perfect with this approach, it reveals a key determinant of any TCR diversity analysis. By a relative increase of expanded subpopulations, the richness of an unsorted repertoire will decrease, while the richness of the individual subpopulations can remain stable. This means that it is crucial to analyse such subpopulations separately to allow for conclusions on diversity loss within for example the naive T-cell compartment.

The total richness of the naive T cell repertoire can only be estimated by combining multiple subsamples

To assess the changes in naive TCR repertoire richness with age we re-analysed the TCRβ sequencing data studied before in (Qi *et al.*, 2014). They estimated the richness of naive and memory populations from four young (20-35y) and five aged (70-85y) individuals. Importantly, the cells were split into multiple subsamples before mRNA extraction, enabling the use of the Chao2 estimator to impute the richness of each total T-cell pool. They estimated naive TCRβ repertoire richness in the order of 10^7 to 10^8, with a two- to fivefold richness decrease in older healthy donors when compared to the young individuals (Qi *et al.*, 2014). We revisited this naive T-cell data by performing additional analyses on the changes with age. First, we estimated the relative contribution of sequence reads by single cells and used this to impute the distribution of cells from the read counts in each sample (see Methods). We then plotted the measured and extrapolated richness using rarefaction curves to account for differences in sampling depth (Figure 4.3). In line with the results in Figure 4.2B, the TCR richness of a given number of naive T cells is very similar between the age groups, especially for lower numbers of cells (Figure 4.3A). At a depth of 50 000 cells, which was reached in all subsamples, the maximum richness difference between any sample pair was only 13% for naive CD8 T cells, and less than 5% for CD4. Although the curves diverge somewhat more at a higher depth, the observed differences in measured richness between the age groups remain rather limited.

Since typical samples of T cells only comprise a minor fraction of the entire T-cell repertoire, it can be useful to extrapolate the richness to the level of the entire T-cell pool. The non-parametric Chao1 estimator accounts for the unobserved diversity, based on the number of TCR sequences observed once and twice in a sample (Chao *et al.*, 2009). For each subsample, the total richness estimate using the Chao1 estimator typically exceeds the observed richness by orders of magnitude (Figure 4.3B). Notably, although multiple estimates of the naive TCR repertoire richness in a single individual can be quite heterogeneous, the estimated TCR richness is clearly distinct between both age groups. In addition, the rarefaction curves are mostly flat at the highest sampling depth, suggesting that the richness estimate is quite robust. We also integrated the information from multiple subsamples using the Chao2 estimator (Figure 4.3C). The Chao2 estimates are based on occurrence in one or multiple samples rather than on the abundance in an individual sample (Chao *et al.*, 2009). The estimates did not saturate completely at the current sampling depth

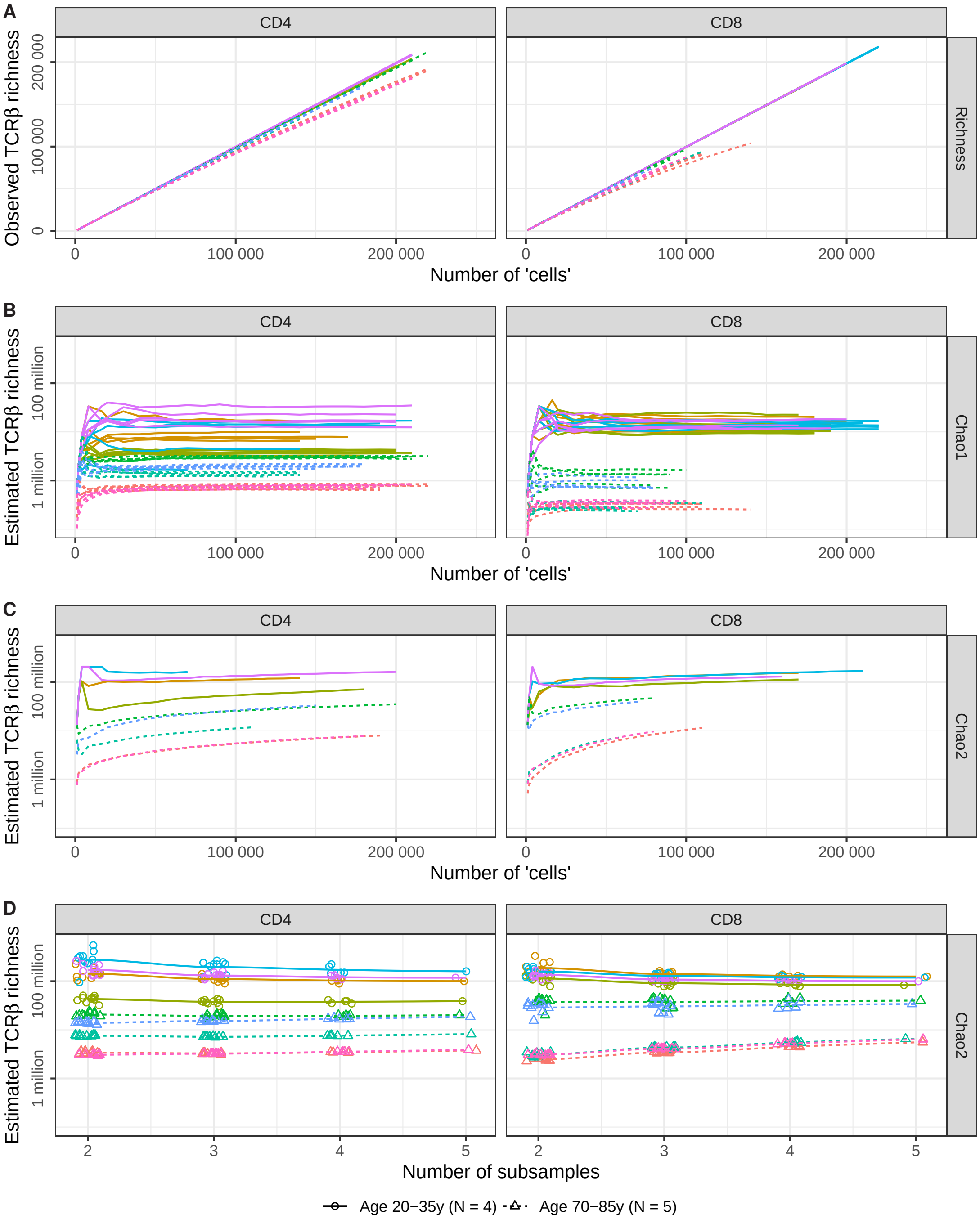


Figure 4.3 – **Naive TCRβ richness in young and aged individuals as a function of sample size and number. A-C.** Rarefaction curves for the observed TCRβ richness (A) and estimated total richness using the Chao1 (B) and Chao2 (C) estimators. Shown are results based on CD4 (left) and CD8 (right) naive T-cell repertoires from young (solid lines) and aged individuals (dashed lines). The colours indicate individual donors, and the horizontal axis depicts the inferred number of cells (see Methods). The rarefaction curve of the Chao2 estimator runs until the sampling depth of the smallest subsample of each individual. **D.** Estimated total TCRβ richness using the Chao2 estimator based on 50 000 inferred 'cells' from all (5) or a subset of the subsamples. Shown are the loess regression lines with colours and line style as in A-C, with the symbols showing individual Chao2 estimates for repertoires of young (circles) and aged (triangles) individuals.

but were clearly distinct between the age groups, in line with the previous analyses (Qi *et al.*, 2014). In comparison with the Chao1 estimates, we arrived at much higher estimates for the total repertoire richness using the Chao2 estimator. This may indicate that the inferred cell abundance in individual samples is far from perfect, casting doubt on the accuracy of the Chao1 estimates in Figure 4.3B. The total richness predicted using the Chao2 estimator appeared surprisingly consistent when based on a subset of the subsamples at a given sampling depth (Figure 4.3D). So, the estimated richness based on multiple subsamples allows to discriminate between the naive TCR repertoires of young and aged individuals, even at a very limited coverage of the entire TCR diversity.

Towards a robust comparison of diversity between sampled TCR repertoires

Here we analysed the observed and extrapolated richness of multiple existing TCR repertoire sequencing datasets. While identifying 'holes in the repertoire' using TCR sequencing may be an attractive idea, the extremely limited coverage of typical samples weakens the outcomes of such an analysis. While estimating the richness of specific T-cell responses is problematic, even estimating the total richness of a diverse T-cell repertoire appears far from straightforward. Differences in richness and evenness may reflect differences in subset frequency rather than true differences in richness within a given T-cell pool. Addressing the changes in diversity during healthy ageing requires sequencing the TCR repertoire of multiple samples from sorted T-cell populations. It remains to be determined to which extent the observed or estimated TCR diversity can functionally explain or predict clinical outcomes in health and disease. This report illustrates that careful experiments and analyses are necessary to obtain robust signals from sampled immune repertoires.

Methods

TCRβ sequencing data

The processed TCRβ repertoires that were used for the analysis presented in Figure 4.1 and Figure 4.2, originally published in (Emerson *et al.*, 2017), were downloaded from the Adaptive Biotechnologies website. We only used the repertoires of donors with a known age and CMV-infection status, for which 'counting method v2' was applied. To eliminate the heterogeneity in sample size in Figure 4.1, we used the 482 repertoires with a template count of at least 100 000, and down-sampled these without replacement to contain 100 000 templates. The TCR richness was quantified in these size-normalised repertoires, based on the combination of the identified V-gene, J-gene, and CDR3 nucleotide sequence. The dataset that was used for the analysis presented in Figure 4.3, originally published in (Qi *et al.*, 2014), was obtained from dbGaP found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000787.v1.p1 through dbGaP study accession number PRJNA258304. These data (project "Immunosenescence: Immunity in the Young and Aged") were provided by Jorg Goronzy on behalf of his collaborators at PAVIR and Stanford University. In this study, five replicates with each 10^6 cells per aliquot of naive and memory

CD4 T cells were collected. For CD8 T cells, 0.25×10^6 T cells were collected per replicate, except for the naive CD8 T cells from young individuals, from which 10^6 cells per aliquot were used for sequencing. The sequencing data was processed using RTCR (Gerritsen *et al.*, 2016) as described before (de Greef and de Boer, 2021).

Inference of a T-cell response against an unseen pathogen

The VDJdb (Shugay *et al.*, 2017) was downloaded from https://vdjdb.cdr3.net on 15 December 2022, by selecting human TRB sequences that were found to be specific for epitopes derived from HIV-1, with a confidence score of at least 1. When matching the size-normalised sequencing data with this database, we required the translated CDR3 sequence, as well as the V- and J-gene families to be identical.

Reported naive T-cell counts

We searched the literature for studies reporting naive CD4 and CD8 T-cell counts as a function of age. We only selected studies in which individuals were stratified by CMV-status. We used the regression formulas presented in Table S1 in (Wertheimer *et al.*, 2014), the median ages and T-cell counts for both age groups reported in Tables 1 and 2 in (Westera *et al.*, 2015), and the median T-cell counts plus standard deviation for the age groups reported in Tables 2 and 3 in (Chidrawar *et al.*, 2009).

Computational classification of total TCR repertoires into naive and effector/memory subpopulations

The dataset reported in (Emerson *et al.*, 2017) is based on unsorted T-cell repertoires. To address the potential changes in richness for the underlying T-cell subpopulations, we performed a TCR sequences by subset. We separated TCR sequences that are expected to be derived from naive T cells from those that were expected to be derived from effector/memory T cells. Specifically, we assumed the percentage of naive T cells in these samples to decrease from 90% at birth to 50% at 20 years of age (2 percent point decrease per year), and a further decrease of 0.25 percent point per year. We used the same selection criteria described above and inferred the total number of naive and effector/memory T cells in these samples. For the 420 TCR repertoires in which both numbers exceeded 100 000 cells (which in total thus contained at least 200 000 cells), we sorted the TCR repertoire by TCR sequence frequency. The least abundant sequences were classified as naive, until their relative frequency was equal to the estimated fraction of naive T cells. The remaining, more abundant, sequences were classified as being derived from effector/memory cells. Both fractions were down-sampled without replacement to a template count of 100 000 to eliminate the heterogeneity in sample size. The richness of each down-sampled TCR repertoire is shown in Figure 4.2B.

Inference of T-cell distributions from mRNA-based read distributions

The sequencing data obtained in (Qi *et al.*, 2014) is based on PCR-amplified TCRβ mRNA transcripts from samples containing variable numbers of cells. The TCR frequencies in the data will thus be affected by differences in cell count, TCRβ expression level and amplification efficiency. These factors may influence the estimated richness in each of the samples, complicating reliable comparison between samples. To reduce the effect of these factors we inferred a distribution of cells from the distribution of reads in each sample. We reasoned that the majority of the TCR sequences observed in only one subsample will be derived from a single T cell. This allowed us to infer the distribution of how many reads are contributed by each single cell, for each subsample individually. Importantly, this distribution only describes the cells that contributed any TCR read, since cells without any contribution remain unnoticed. Using the average of the inferred number of reads that was derived from each contributing cell, we estimated how many cells contributed all reads together in each sample. Reassuringly, the subsamples containing naive CD8 T cells from aged individuals, that were known to contain fewer cells than the other samples, also contained fewer cells according to our TCR repertoire-driven estimate. We then assigned TCR sequences that were supported by a single read to single cells, since these reads cannot be contributed by multiple cells. The remaining TCR sequences were then assigned to inferred 'cells' based on the estimated read-contribution distribution, starting from the TCR sequence supported by the largest number of reads. This resulted in a table of TCR sequences, each supported by an inferred number of cells that was smaller than or equal to the observed number of reads supporting that TCR sequence (see example table below). The analyses presented in Figure 4.3 are based on these inferred distributions of cells instead of the potentially more biased frequency of reads. Rarefaction curves were obtained by sampling without replacement from these inferred TCR repertoires.

TCRβ sequence			Observed number of reads			Inferred number of 'cells'		
V	J	CDR3	Subsample 1	…	Subsample 5	Subsample 1	…	Subsample 5
7-9	2-7	TGTGCCA…GTACTTC	0	…	20	0	…	2
7-9	2-3	TGTGCCA…GTATTTT	10	…	7	1	…	1
20-1	2-7	TGCTGTA…GTACTTC	0	…	2	0	…	1
…	…	…	…	…	…	…	…	…
20-1	2-7	TGCGGTG…GTACTTC	1	…	0	1	…	0
9	1-1	TGTGCCA…TTTCTTT	38	…	25	5	…	3
7-9	2-3	TGTGCCA…GTATTTT	0	…	20	0	…	3

Author contributions

P.C.d.G. and R.J.d.B. conceived the study. P.C.d.G. analysed the data with input from R.J.d.B. P.C.d.G. wrote the manuscript, which was edited and approved by both authors.

Acknowledgements

This work was supported by The Netherlands Organization for Scientific Research Graduate Program 022.005.023 (to P.C.d.G.). We gratefully acknowledge all authors involved in generating the TCR sequencing datasets we studied, for providing access to their data.

5

On the feasibility of using TCR sequencing to follow the vaccination response - lessons learned

Peter C. de Greef [1]*, Josien Lanfermeijer [2,3]*, Marion Hendriks [2], Alper Cevirgel [2,4], Martijn Vos [2], José A.M. Borghans [3], Debbie van Baarle [2,4], and Rob J. de Boer [1]

[1] *Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands*
[2] *Center for Infectious Disease Control, National Institute for Public Health and the Environment, The Netherlands*
[3] *Center for Translational Immunology, University Medical Center Utrecht, The Netherlands*
[4] *Current affiliation: Department of Medical Microbiology and Infection Prevention, University Medical Center Groningen, The Netherlands*
* *These authors share first-authorship*

Submitted.

Abstract

T cells recognise pathogens by their highly specific T-cell receptor (TCR), which can bind small fragments of an antigen presented on the Major Histocompatibility Complex (MHC). Antigens that are provided through vaccination cause specific T cells to respond by expanding and forming specific memory to combat a future infection. Quantification of this T-cell response could improve vaccine monitoring or identify individuals with reduced ability to respond to a vaccination. In this proof-of-concept study we use longitudinal sequencing of the TCRβ repertoire to quantify the response in the CD4 memory T-cell pool upon pneumococcal conjugate vaccination. This comes with several challenges owing to the enormous size and diversity of the T-cell pool, the limited frequency of vaccine-specific TCRs in the total repertoire, and the variation in sample size and quality. We defined quantitative requirements to classify T-cell expansions and identified critical parameters that aid in reliable analysis of the data. In the context of pneumococcal conjugate vaccination, we were able to detect robust T-cell expansions in a minority of the donors, which suggests that the T-cell response against the conjugate in the pneumococcal vaccine is small and/or very broad. These results indicate that there is still a long way to go before TCR sequencing can be reliably used as a personal biomarker for vaccine-induced protection. Nevertheless, this study highlights the importance of having multiple samples containing sufficient T-cell numbers, which will support future studies that characterise T-cell responses using longitudinal TCR sequencing.

Introduction

Vaccination has proven to be a safe and effective method for immunisation, limiting the spread of numerous infectious diseases. Exposure of a pathogen or its subunits to the adaptive immune system provides immunity that can potentially last a lifetime. Neutralising antibody titres typically serve as a correlate of protection in an individual (Katzelnick *et al.*, 2016; Khoury *et al.*, 2021; Tsang *et al.*, 2014) but do not cover the immunity provided by T cells, which is often crucial to prevent infection (Pizzolla *et al.*, 2017; Wilkinson *et al.*, 2012). Quantitative characterisation of the T-cell response induced by vaccination thus has the potential to provide an important additional measure of protection in an individual (Fink, 2019). T cells recognise antigens by their highly specific T-cell receptor (TCR) presented as peptides on the Major Histocompatibility Complex (pMHC). Activation through the TCR is followed by clonal expansion and maintenance at increased frequencies as memory T cells, resulting in an enhanced immune response at a next encounter with a similar pathogen. In the context of vaccination, providing an antigen induces a T-cell response. The TCR repertoire dynamics reflecting this response can be followed using high-throughput TCR repertoire sequencing (Dykema *et al.*, 2022; Fink, 2019; Miyasaka *et al.*, 2019; Pogorelyy *et al.*, 2018; Sycheva *et al.*, 2018).

Previous studies have used TCR repertoire sequencing to characterise the T-cell response after yellow fever vaccination (YFV) (DeWitt *et al.*, 2015; Pogorelyy *et al.*, 2018). This live-attenuated virus vaccine induces a large CD8+ T-cell response, which could be quantified by measuring T-cell expansion and contraction after vaccination by longitudinally sequencing the TCR repertoire. This allowed the identification of YFV-specific TCR sequences, which occupied up to 8% of the total CD8+ T-cell repertoire two weeks after vaccination (Pogorelyy *et al.*, 2018). Other vaccine-induced T-cell responses have been characterised by sequencing the TCR repertoire of cells that were sorted for binding known influenza epitopes (Dash *et al.*, 2017; Glanville *et al.*, 2017). For many other vaccines, however, the epitopes that induce a T-cell response remain unknown. In those cases, following vaccine-specific T-cell clones requires characterisation of the total TCR repertoire. It remains to be determined whether or not TCR sequencing of the overall T-cell repertoire can serve as a suitable biomarker to quantify T-cell responses induced by such vaccinations.

In the present proof-of-concept study we aimed to identify specific expansion of T-cell clones in the CD4 memory T-cell pool after pneumococcal vaccination. Although this vaccine mainly induces pneumococcal serotype-specific antibodies, T cells are activated by the CRM197 conjugate, a carrier protein which is a non-toxic mutant of diphtheria toxin. The activated CD4 T cells provide additional help to B cells to produce specific antibodies (Sterrett *et al.*, 2020). As CRM197 is also the main antigen of the diphtheria vaccine given in early childhood, the vaccine is anticipated to boost existing T-cell memory. However, the height of the T-cell response may be lower compared to the T-cell response against YFV and immunodominant epitopes are less well described. We performed longitudinal TCR sequencing of the CD4 memory T-cell pool before and after pneumococcal conjugate vaccination. By taking replicate samples we defined quantitative requirements to classify expansions and we identified critical parameters that aid in reliable analysis of the data. The absence of detected robust T-cell expansions in many of the vaccinated individuals illustrates the challenges of using TCR sequencing to quantify specific T-cell responses after vaccination. We conclude that the T-cell response induced by the conjugate in the pneumococcal vaccine is often too small or too diverse to allow for reliable quantification using TCR sequencing. Finally, our analysis identified specific requirements for monitoring T-cell responses using longitudinal TCR sequence data.

Results

Study design

We tested the application of TCRβ sequencing using samples from a human cohort that was part of a vaccination study with Prevanar 13, a conjugated vaccine targeting 13 pneumococcal strains. Blood samples were taken from 13 adult individuals before vaccination (day 0) and at day 7, day 28, and between 4 to 8 months after vaccination (Table S5.1, Figure 5.1A). The antibody response was quantified by measuring diphtheria-specific IgG antibodies, which showed a clear response in 10 out of 13 individuals (Figure 5.1B). Typically, they showed IgG

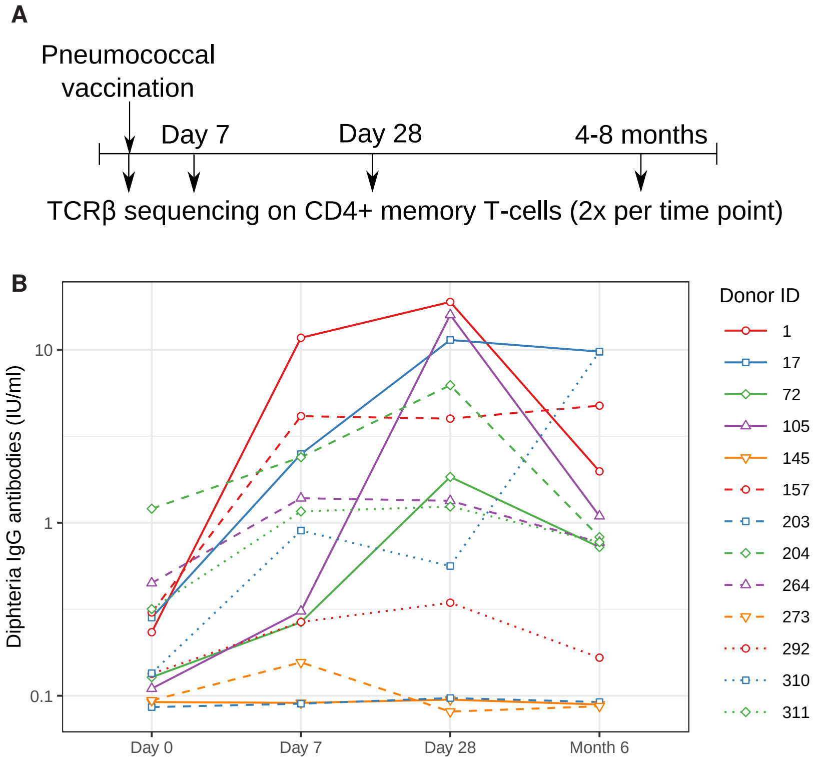


Figure 5.1 – **Study overview and measured antibody response. A.** Schematic overview of the vaccination and sampling time course. Individuals were vaccinated at day 0, blood samples were drawn before vaccination (day 0) and at three follow-up time points. **B.** Quantification of the antibody response to the diphtheria toxin. Values above 0.1 IU/ml are considered protective. Solid lines indicate older individuals (over 65 years of age).

levels that were considered protective already before vaccination, which increased about an order of magnitude at day 7 and/or day 28.

The T-cell response can be characterised using longitudinal TCR repertoire sequencing

The presence of a clear antibody response after vaccination in most individuals suggests effective T-cell help, most likely provided by CD4 memory T cells. We characterised this T-cell response by isolating and combining three subsets of CD4 memory T cells ($CD27^{+}CD45RO^{+}$, $CD27^{-}CD45RO^{+}$, and $CD27^{-}CD45RO^{-}$). The cells were split in two portions, yielding sorted populations containing in the order of 10^5 CD4 memory T cells per subsample, per time point per individual (Figure S5.1A). mRNA was extracted from the cells in each sample for TCRβ cDNA library preparation (see Methods). The libraries were barcoded with Unique Molecular Identifiers (UMIs) to overcome biases in PCR amplification and to allow for error correction of the sequence reads.

Without prior knowledge which TCRs are induced by the conjugate of the pneumococcal vaccine, we relied on detection of expansion of TCRβ chains upon vaccination. One would expect the frequency of specific TCRβs to have increased at day 7 and/or 28 with respect to

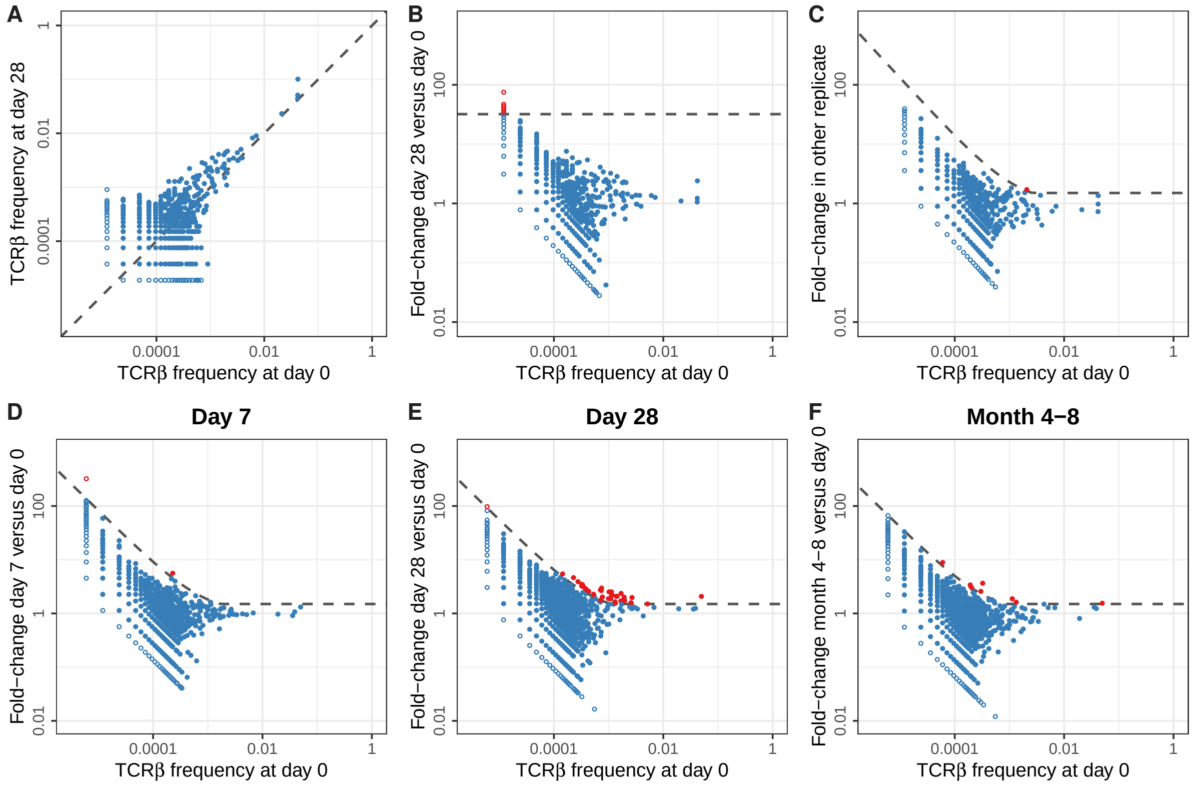


Figure 5.2 – Classification of expansion between time points for TCRβ repertoires of example donor 292. A. Relative frequency of TCRβ sequences found at day 0 and/or day 28 in samples taken from donor 292. Sequences observed at only one of the two time points are plotted as open symbols, at half the frequency corresponding to a single UMI in the sample they were not present. The dashed line is the diagonal, representing equal frequencies between both samples. **B.** Fold-change in TCRβ frequencies at day 28 versus day 0. The dashed line indicates a general fold-change threshold of 32x, as was used in (Pogorelyy *et al.*, 2018). **C.** Fold-change in TCRβ frequencies between the two replicate samples taken at day 0 from donor 292. The dashed line represents the combined threshold for quantification of expansion (see Methods). Red dots represent sequences that meet both requirements to be classified as expanded. In this case, comparing two samples from the same TCR repertoire, only a single sequence was classified as expanded (false-positive). **D-F.** Similar to C, but now quantifying the fold-change between different time points. Replicate samples from each time point are joined, after which expansion at day 7 (**D**), day 28 (**E**), and month 4-8 (**F**) is quantified versus day 0.

the pre-vaccination sample and potentially to be lower in the samples of the last time point. We thus measured the frequency of TCRβ sequences post-vaccination and compared these to the corresponding pre-vaccination frequencies (Figure 5.2A). TCRβ frequencies appeared highly correlated between time points, confirming the persistence of many T-cell clones at similar frequency during the study period. We quantified the fold-change of each observed TCRβ sequence between pre-vaccination and post-vaccination time points, revealing the highest fold changes for the least abundant sequences (Figure 5.2B). Naturally these small clones give the strongest signal, as the fold-change results from dividing by a small pre-vaccination frequency. As a result, applying a general fold-change threshold to classify TCR sequences as being expanded would focus the analysis on those sequences of which the dynamics are estimated with the highest uncertainty (red points in Figure 5.2B). A similar pattern even occurs when comparing two replicates from the same time point (Figure 5.2C), although these are samples containing cells from the exact same TCR repertoire.

As an alternative to a generic fold-change threshold, we used the many replicates within our dataset to estimate the effects of sampling noise during the generation, sequencing, and annotation of the TCRβ libraries (see Methods). We identified two requirements that together identify expanded TCRβs in our dataset: (1) a fold-change of at least 1.5, and (2) an absolute TCRβ-UMI count exceeding the relative pre-vaccination frequency by at least 30 UMIs. These thresholds were calibrated by balancing specificity, removing false-positive 'expansions' between samples from the same time point, and sensitivity, to allow for detection of expanded clones between time points (Figure S5.2). The combination of these requirements provides a fold-change threshold that is dependent on the pre-vaccination frequency and the sizes of the samples that are being compared. We classified few false-positive 'expansions' between samples from the same time point (red point in Figure 5.2C), and a variable number of expanded TCRβs at the three post-vaccination time points (Figure 5.2D-F). To reduce the number of comparisons, while increasing the size of the samples being compared, we pooled replicates from the same time point when classifying expansion between time points before and after vaccination.

Most repertoires allow for detection of few TCRβs that expand upon vaccination

We quantified TCRβ expansion after vaccination in each donor by applying the two requirements defined above when making pairwise comparisons between TCRβ frequencies at pre-vaccination and the corresponding post-vaccination time points. In donor 203, for which we retrieved the largest number of TCRβ sequences (Figure S5.1B), this allowed us to identify > 100 TCRβ sequences that expanded at day 7 and/or day 28 after vaccination (Figure 5.3A). These sequences together increased from about 5% of the repertoire to over 13% at the peak (day 7), followed by a decline to about 8% of the CD4 memory pool by day 28 and month 4-8 after vaccination (Figure 5.3B – blue dashed line). The expansion of these sequences was reflected by a decrease in the size-normalised TCRβ diversity in the samples at day 7 post vaccination (blue dashed lines in Figure S5.3D-F). Strikingly, there was only little overlap between the TCRβs classified as expanded at day 7 and at day 28. In addition, while showing the largest number of expanded TCRβ sequences, donor 203 did not show a Diphtheria-specific IgG response after vaccination (blue dashed line in Figure 5.1B). These two observations together raise the question whether the detected expansions in this donor truly reflect the dynamics of a T-cell response induced by the vaccine. As we rely on the dynamics of the overall CD4 memory T-cell repertoire, we cannot exclude the possibility that these expansions may have been caused by other ongoing immune responses. A potential scenario is a response against CMV, as this donor turned out the one of the two CMV-positive individuals in our study (Table S5.1). Thus, although our findings suggest that longitudinal TCRβ sequencing can be used to detect T-cell clones that change in abundance after vaccination, they are not guaranteed to be specifically activated by the vaccine.

When investigating the smaller samples of the other donors we were unable to detect responses of a similar magnitude as in donor 203 using the same classification method. In all but one donor we detected TCRβ expansions at day 7 and/or day 28, although the expansions were again rarely detected at consecutive time points (Figure 5.3A). Donors 1 and 311 showed the largest number of expanded TCRβ sequences at day 7, while most expansions for donors 17 and 292 were detected at day 28 post vaccination. Our study did not involve TCRβ sequencing of individuals that did not receive the vaccination, which

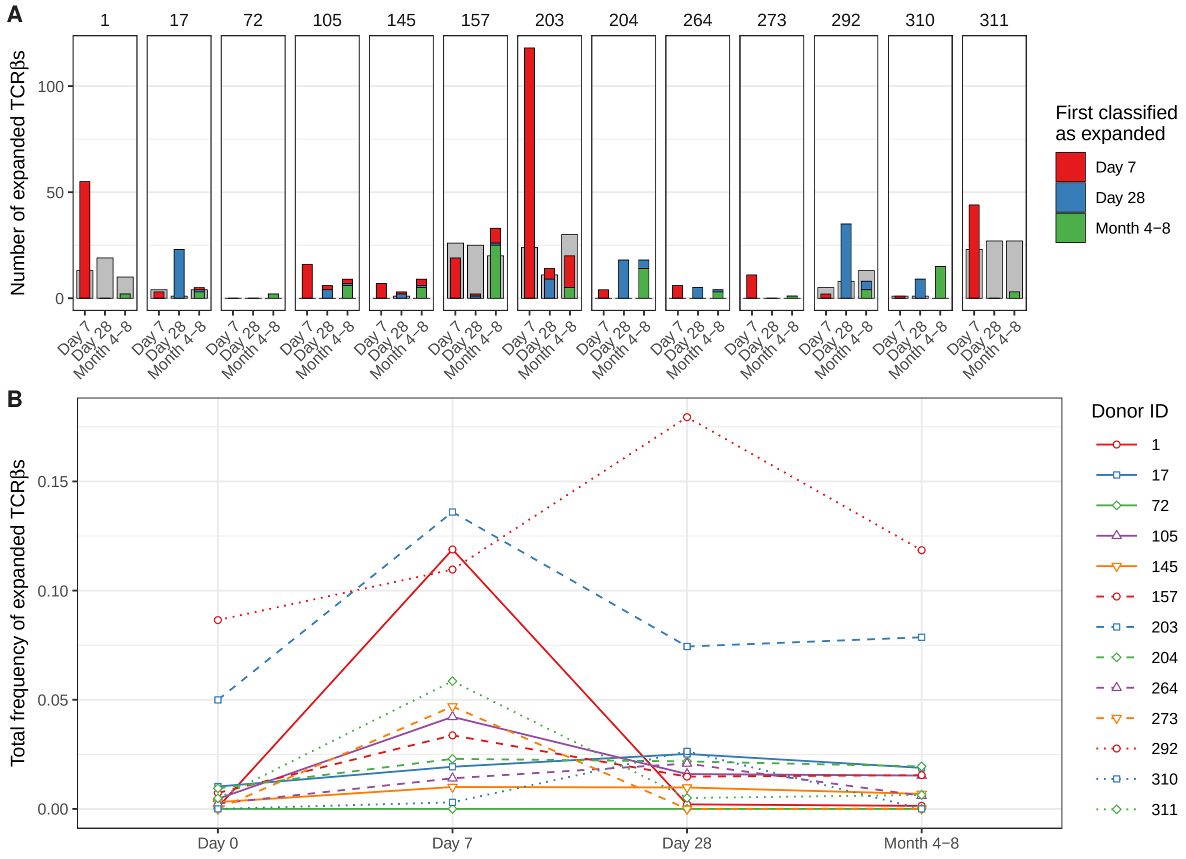


Figure 5.3 – **Number and dynamics of TCRβs sequences classified as expanded. A.** Number of expanded TCRβs at timepoints after vaccination compared to day 0. Colours indicate the first time point at which the specific sequence was classified as expanded (red: day 7, blue: day 28, green: month 4-8). The classification of expansion was performed after pooling the replicates per time point (see Figure S5.4A for the classification based on comparisons between individual replicates between time points). The grey bars serve as a proxy for dynamics that are not induced by the vaccination, by classifying 'expansion' while permuting the time points. Specifically, we classified how many sequences would be considered 'expanded' in the pooled pre-vaccination samples, when compared to the indicated post-vaccination time points. **B.** Dynamics of the sequences classified as expanded at day 7 and/or day 28: the total relative frequency of these sequences is shown.

would have allowed us to estimate how many of the observed expansions are induced by the vaccination. To still validate our findings to some extent, we performed a permutation analysis by switching the order of the time points in each comparison. The number of expanded TCRβs often did not exceed the number of 'expansions' after permutation, which in fact reflect contractions of a similar magnitude, for example by non-specific dilution (Figure 5.3A - grey bars). This indicates that many of the detected expansions may have occurred independently of the vaccination. The total proportion of expanded TCRβs varied considerably between individuals, mostly owing to the different number of detected expansions for each donor (Figure 5.3B). The largest contraction of expanded TCRβs happened between day 7 and day 28 post-vaccination in most individuals, most clearly pronounced in donors 1 and 203. Notably, detecting TCRβ expansions by making comparisons between individual replicates of different time points yielded similar numbers of expanded TCRβ sequences (Figure S5.4A). Thus, in most donors we only detected a few TCRβs that expanded upon vaccination at a single time point when compared to their pre-vaccination frequency.

The sample size dominates the number of detected TCRβ expansions

The large differences that we observed in the number of expanded TCRβs between donor 203 and the other donors could be caused by a biological effect, but also by technical variation, e.g., the number of identified TCRβ sequences. In order to distinguish between biological variation regarding the vaccination response in different donors, and sources of technical variation between samples, we computationally down-sampled the TCR repertoires from donor 203 to the sample sizes of the corresponding samples from the other donors. In each of these down-sampled sets we could still detect expansions, but the number of identified TCRβ expansions was at least a 5-fold lower for each down-sampled repertoire (Figure S5.4B). This emphasises that the total number of identified TCRβ sequences is a critical parameter for longitudinal characterisation of T-cell dynamics and a large determinant for the ability to detect expansions. This analysis also suggests that we could have identified even more TCRβ expansions in the other donors than in donor 203, if their sample sizes would have been as large as those of donor 203 (Figure S5.4B).

While the analysis so far focused on the identification of individual TCRβ sequences that expand upon vaccination, we also checked if we could detect more general signatures of TCR repertoire dynamics after vaccination. We quantified the changes in overall diversity of the TCRβ repertoire using various estimates (see Methods). The overall repertoire diversity varied considerably, but not consistently, between donors and time points (Figure S5.3A-C). Since the diversity measures are strongly affected by the sample size, we also normalised the estimates by down-sampling each repertoire to the same number of UMIs (Figure S5.33D-F). Even after size-normalisation, we observed increases as well as decreases in TCRβ upon vaccination. So, although the expansion of specific TCRβ sequences could be reflected in a decreased diversity after vaccination, we did not detect such dynamics consistently in most the donors. An alternative scenario that has been proposed is that the recruitment of (naive) vaccine-specific clones could increase the estimated diversity (Miyasaka *et al.*, 2019), perhaps reducing our ability to detect consistent diversity dynamics between vaccinated individuals.

Another approach to follow the dynamics of many T-cell clones together is by quantifying the changes in TCRβ V-gene usage. The observation that TRBV usage varies considerably between samples, both from the same and from different time points (Figure S5.5) confirms that sampling effects have a profound effect on TCRβ frequencies, partly masking clonal dynamics that allow for quantification of the T-cell response induced by vaccination. Together, these results identify the size of T-cell samples as a key factor that determines to which extent T-cell responses can be quantified and compared using TCR sequencing.

Discussion

In this proof-of-concept study, we applied longitudinal TCRβ sequencing on CD4 memory T cells from individuals before and after pneumococcal conjugate vaccination. We developed specific criteria to classify clonal expansions from longitudinal TCR sequencing data, aiming to discriminate between biological and technical variation in the results. Doing so, we identified some TCRβs that expanded after vaccination, although these were mostly limited to a few individuals and a single time point. The absence of detectable and persistent T-cell expansions in most individuals illustrates the complications of longitudinal TCR sequencing when there is a small and/or diverse T-cell response. The sample size appears a crucial factor for detection of TCRβ expansion in the overall T-cell repertoire. An overview of critical technical requirements for a robust longitudinal TCRβ repertoire characterisation are detailed in Box 1.

The complications with detecting considerable T-cell expansions upon pneumococcal conjugate vaccination does not indicate absence of a substantial T-cell response, or lack of a protective effect of vaccination. Firstly, the total size of the T-cell response in these donors is unknown and may well be below the 1% range, meaning that all vaccine-specific sequences together are still relatively rare and not easily distinguishable from all sequences with different specificities. Secondly, the total response is expected to be composed of many individual TCRαβ clonotypes with different TCRβ sequences. Many of their frequencies will fall below our limit of detection (Figure 5.4B) especially when we track cells at the level of individual TCRβ sequences. Moreover, even when TCRβ sequences are present at multiple time points, their dynamics cannot always be distinguished from noise. Thirdly, it remains to be determined whether the size and/or diversity of the T-cell response correlates with protection against infection. The sharp increase in Diphtheria-specific IgG antibodies indicates a substantial response upon vaccination, while the size and breadth of the T-cell response currently remains an open question as a functional characterisation of the T-cell response was not included in this study.

Box 1: Challenges of following a T-cell response with TCR sequencing

Its enormous diversity is one of the key features of the TCR repertoire, but also poses a major challenge to measure T-cell responses using longitudinal TCRβ sequencing. Without *a priori* knowledge about which TCRs are antigen-specific, the responders must be distinguished from the T-cell clones with different specificities, just based on the changes in their abundance. Thus, a sufficient increase in frequency is required to identify the potentially many TCRβ chains that are expressed by the antigen-specific cells mounting the T-cell response. For each involved T-cell clone, this requires: (1) an abundance in the repertoire that is sufficient to be present in the sample, (2) a TCR sequencing protocol that is sensitive enough to detect changes in frequency, and (3) a careful analysis to distinguish between technical variation and true clonal dynamics.

1. Sufficient abundance of cells of interest in the sorted cell population

There are about 10^{12} T cells in the human body, so even samples of millions of cells will only constitute a tiny proportion of the total pool. Combined with the large diversity of the TCR repertoire, this results in limited TCR overlap between samples of naive and memory T cells, even between replicates of the same time point (Warren *et al.*, 2011). The measured overlap correlates strongly with the sequencing depth of the sample, which depends on the starting number of cells and the sequencing protocol (Figure 5.4A). These observations follow from the probability of clonal presence in a sample, which, for small samples, scales roughly linearly with the sample size. A central question is which proportion of the cells in the sample are expected to be participating in the response towards the vaccine. A previous study estimated the response after yellow fever vaccination (YFV) to comprise 2-8% of the total T-cell pool in blood, which is composed of many clones which frequencies differ by several orders of magnitude (Pogorelyy *et al.*, 2018). Since most vaccines are expected to induce a much smaller response than YFV, the frequency of most vaccine-specific TCRβs will be very low, even at the peak of the response. It may thus be useful to enrich the sample for T cells that participate in the response, to obtain enough signal. In this study, we sorted the CD4 memory T-cell population because a response was anticipated to occur within this cell population. Although we were limited by the availability and size of the samples, further enrichment may be possible by sorting for activation markers and antigen-specificity using available tetramers when possible. While enrichment for cells with specific characteristics allows for quantitative estimates of the total T-cell response, the identification of individual clonal T-cell dynamics will become more complicated.

2. Sensitive TCR sequencing protocols

When calculating the expected effect size in a sample for a given number of sorted cells, it is important to take the loss of information during the TCR sequencing protocol into account. We estimated that probably less than 10% of the cells contributed one or more mRNA molecules to the eventual dataset after amplification, sequencing, and processing of the data (Figure S5.6). Moreover, some of the cells perhaps contributed multiple mRNA molecules, each labelled with a separate UMI sequence, adding to the uncertainty of estimating clonal abundance before and after vaccination. The frequencies of TCRβs in the data is distorted by many stochastic processes, including the sampling of cells and mRNA molecules, as well as the amplification and sequencing of transcripts. We quantified the contribution of these factors by comparing replicate samples from the same TCR repertoire, which revealed that typically a frequency of at least 0.1% of the memory T-cell repertoire is required to be stably present in multiple samples at our average sequencing depth (Figure 5.4B). We also found considerable differences in TCRβ abundance between replicates, requiring us to set strict thresholds to discriminate T-cell expansion from technical variation. Note that sequencing the same TCRβ library twice yielded much more similar results, indicating that most uncertainty is introduced before the sequencing (Figure 5.4A - red points). Having multiple samples from the same TCR repertoire is essential to estimate the contribution of technical variation to the measured abundances. Specific algorithms exist to model the noise introduced during TCR sequencing and to discriminate this from true TCR dynamics (Koraichi *et al.*, 2022; Touzel *et al.*, 2020). Another factor to consider is the contribution of uneven PCR amplification. While UMIs are used to factor this out, the UMI-based error correction of the sequences requires multiple reads sharing their UMI. The wide distribution of the number of reads per UMI results from the uneven amplification by PCR and identifies a sufficient sequencing depth as a key requirement to allow enough sequences to reach the threshold for error correction (Figure 5.4C).

3. Processing the samples and quantification of expansion

During the steps outlined above, from reverse transcription, via amplification, to sequencing the TCRβ libraries, it is inevitable that errors are introduced. As UMIs label cDNA molecules before amplification, they greatly assist error-correction of the reads. Dedicated pipelines exist to perform these steps, which can also correct other errors by clustering of low-quality or nearby sequences (Bolotin *et al.*, 2015; Gerritsen *et al.*, 2016). The resulting data provides a way to estimate the changes in frequency of each TCRβ sequence. Careful interpretation is necessary for the reasons explained above, mainly distinguishing between real biological effects and the technical variation arising during the entire TCR sequencing process. This requires a robust classification of expansion, which is ideally calibrated using on samples from the same and different time points.

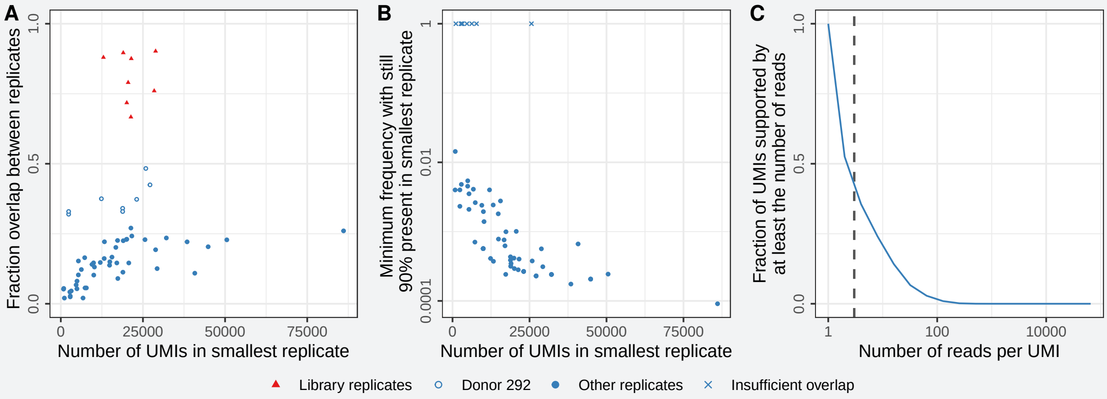


Figure 5.4 – **Sample overlap and coverage. A.** Fraction of the sample overlapping between two replicates from the same time point (see Methods). Comparisons between multiple samples of sorted cells (blue circles) and replicates generated by sequencing the same TCRβ library twice (red triangles). The open blue circles indicate samples from donor 292, which have a relatively high overlap due to a low TCRβ diversity (see also Figure S5.3). **B.** TCRβ frequency in the largest replicate at which sequences start missing in the smallest replicate (see Methods). X marks indicate comparisons in which the most abundant TCRβ sequence did not overlap between both samples. **C.** Cumulative frequency of the UMI-coverage, plotted as the fraction of UMIs supported by at least a given number of reads (horizontal axis). The vertical dashed line indicates a coverage of 3 reads per UMI, which was used as the minimum support to take a sequence into account in the analysis (see Methods).

TCR repertoire characterisation is usually done by sequencing of the mRNA coding for the α- and/or β-chain of the TCR. While single-cell techniques exist to perform paired sequencing of both TCR chains, their drawback is the limited number of cells that can currently be profiled. Instead, many studies focus on the TCRβ-chain, for which high-throughput methods allow characterisation of millions of cells. Our choice to sequence the bulk TCRβ repertoire instead of paired TCRαβ single-cell sequencing has both advantages and drawbacks. Doing so, we could characterise the TCR repertoire from many cells per time point, which is necessary to detect clonal expansion. Although missing information on the TCRα, a substantial expansion of a TCRαβ clonotype will likely be reflected by an increased frequency of its TCRβ sequence. More detailed identification of the expanding TCRs in donor 203 would have required single-cell analysis of many cells. This could have allowed us to further characterise the expanded clones, for example by their transcriptional profiles. In addition, the technical variation stemming from the fact that single cells can contribute multiple mRNA molecules in a bulk analysis is excluded when sequencing the repertoire at the single-cell level. Currently, however, considering the limited expansion observed in most bulk repertoires, we do not expect that we could have captured a larger response using single-cell sequencing, because sample sizes would have been even smaller.

A key challenge remains to distinguish between technical variation and true dynamics of the TCR repertoire. This discrimination requires a sufficient sample size, which in turn requires large numbers of input cells and minimisation of information loss during the TCR-sequencing procedure (see also Box 1). This reduces the relative contribution of sampling noise and technical biases, which allows setting less strict thresholds to quantify expansion. The TCR dynamics result from a combination of vaccine-induced expansions and other ongoing immune responses. Thus, functional assays are crucial to verify the specificity of the expanded T-cell clones. Such information will also help to interpret the dynamics functionally, such as the changes in TCR diversity after vaccination. It could be that the ongoing immune response is not easily detected in blood samples, if the response mostly occurs in lymphoid organs or specific tissues. The large variation in MHC-genes across individuals causes immune responses to be mostly private. Still, finding motifs in vaccine-specific TCR sequences would enable more direct identification of the vaccine-induced T-cell response (Mudd *et al.*, 2022), perhaps even without the need for data from consecutive time points.

Some vaccines, like YFV, may elicit large T-cell responses that can be accurately quantified by longitudinal TCR sequencing. Vaccines that activate fewer T cells, or a wide diversity of T-cell clones will be much more challenging to characterise with sequencing of the TCR repertoire. Translating parameters of the T-cell response into a personalised biomarker of vaccine efficacy involves several other challenges. A first step would be to relate these to other correlates of protection such as (neutralising) antibody titres. For example, this may give insight into the importance of the breadth and depth of the T-cell response, which can be estimated using TCR sequencing. The relevance of such features beyond the currently known risk factors and serological assays will require extensive clinical studies.

While currently perhaps not yet feasible, technological advances may enable this in the future. This study should be considered as one of the first steps on the way to personalised vaccination strategies that will further protect people at risk from infectious diseases.

Materials and Methods

Study cohort

Samples used in this study were selected from The Vaccines and InfecTious disease in the Ageing PopuLation (VITAL) cohort (Baarle *et al.*, 2020), which was started in 2019 in the Netherlands. For this study healthy individuals were recruited who did not use immune-modulatory drugs and who were not immunocompromised due to a medical condition. This study was approved by the acknowledged ethical committee METC Noord Holland and carried out in accordance with the recommendations of Good Clinical Practice with written informed consent from all subjects, in accordance with the Declaration of Helsinki.

Sample selection

For this study, the samples were selected from the VITAL cohort based on age. We selected 8 donors with an age between 25 and 40 years (young adults), as well as 5 adults that were over 65 years (older adults). Individuals were vaccinated with the pneumococcal vaccine Prevenar 13. Blood samples were collected from all individuals at day 0 (before vaccination), at day 7, at day 28 and between 4 or 8 months post-vaccination (see Table S5.1).

PBMC and serum isolation

Peripheral blood mononuclear cells were obtained by Lymphoprep (Progen) density gradient centrifugation from heparinised blood, according to the manufacturer's instructions. PBMCs were frozen in 90% fetal calf serum and 10% dimethyl sulfoxide at -135°C until further use. Serum was isolated out of tubes with clot-activation factor and stored at -80°C until further use. Blood withdrawals were postponed if participants received other vaccinations or had elevated body temperatures (> 38°C).

Cytomegalovirus (CMV)-specific antibodies

Anti-CMV IgG antibody concentrations were measured by an in-house-developed multiplex immunoassay (Tcherniaeva *et al.*, 2018). Cut-off values were based on previous calculations: Individuals with a CMV-specific antibody level of ≤ 4 arbitrary units (RU)/ml were considered CMV-negative, individuals with an antibody level > 7.5 RU/ml were considered CMV-positive, and those with a level between 4 and 7.5 RU/ml were considered inconclusive and hence excluded from further analysis (Samson *et al.*, 2020).

Determination of diphtheria-specific antibody concentrations

Nunc MaxiSorp ELISA plates were coated with 2.5μg/ml diphtheria toxoid (Statens Serum Institute) and blocked with 0.01M Glycin. Plasma samples were analysed in duplicates. Bound antibodies were detected with HRP-conjugated secondary Rabbit Anti-Human

IgG Antibody (Sigma-Aldrich) and TMB one component Substrate Solution (Diarect). IgG antibodies were quantified in IU/ml using Standard Diphtheria Antitoxin Human Serum (NIBSC). The detection limit of the assays used was 0.015 IU/ml and antibody concentrations above 0.1 IU/ml were considered as protective.

Isolation of CD4 memory T cells for TCR repertoire analysis

Approximately 5 million PBMCs were labelled at 4°C for 30 min with the following mAbs mix: CD8(RPTA-T8)-FITC, CD3(UCHT1)-PerCP, CD4(RPA-T4)-APC/Cyanine7, CD27(O323)-Brilliant Violet 765, CD45RO(UCHL1)-PE (All Biolegend). After the staining was finished, samples were split in two portions and sorted separately to obtain duplicates. CD4$^+$ memory T cells were defined as CD27$^+$CD45RO$^+$, CD27$^-$CD45RO$^+$, and CD27$^-$CD45RO$^-$ (thus, all CD4$^+$ cells, except the naive T cells (CD27$^+$/CD45RO$^-$)) and sorted using a FACS Melody (BD). CD4$^+$ memory T cells were sorted directly into PBS, spun down and resuspended in RNA Later (Ambion Inc. Applied Biosystems). Sorted samples were stored at -80°C for subsequent TCRβ clonotype analysis.

Preparation of TCRβ cDNA libraries for sequencing

mRNA was isolated with the RNA microkit (Qiagen) according to the manufacturer's protocol. Isolated mRNA was used in the 5' RACE-based SMARTer Human TCR a/b Profiling Kit v2 (Takara Bio USA, Inc.) to perform sequencing of TCRs, following the manufacturer's protocol using only the TCRβ-specific primers. Clean-up was performed with AMPURE XP clean-up beads (BD). The resulting TCRβ libraries were sequenced on an Illumina MiSeq (paired-end 2x300nt). The reproducibility of the sequencing was analysed by sequencing the libraries of donor 145 twice. To check the effects of a larger number of shorter reads, we sequenced the samples from donor 204 and 292 on an Illumina NextSeq (paired-end 2x150nt) instead.

Processing of TCRβ sequencing data

TCRβ sequencing data was processed using the CogentTM NGS Immune Profiler pipeline (version 1.0), as provided by Takara Bio. We set the overseq-threshold to 3, meaning that UMI-TCR pairs supported by at least 3 reads were taken into account. We defined a TCRβ sequence as the combination of V-segment, CDR3 amino acid sequence and J-segment. For the analyses presented in Figure 5.2D-F, Figure 5.3, and Figure S5.3 we joined the counts from replicates of the same time point to arrive at a single TCR repertoire per time point. The equivalent analysis of Figure 5.3A using the individual replicates is shown in Figure S5.4A.

A robust classification of expansion

TCRβ sequences have frequencies that can differ by multiple orders of magnitude. The most abundant sequences are often measured with a sufficient number of UMIs to reliably estimate their frequency in the repertoire. The frequency of many rare sequences is much less certain, as their proportion in the data is relatively more affected by sampling noise. While classifying expansion between timepoints, we accounted for these differences using a two-step approach. We used replicates of the same time point of the same donor (which are thus samples from the same T-cell repertoire) to estimate the sampling noise. The requirements for expansion were optimised to be both specific and sensitive: they should result in little or no expansion between samples from the same time point, while allowing for detecting expansions between time points.

Firstly, we determined a general fold-change threshold based on the abundant TCRβ sequences. Specifically, we analysed the sequences present at a relative frequency of more than 0.5% in a sample and quantified their fold-change when comparing with another sample taken. To still obtain a fold-change in cases where a sequence was present in one sample, but absent in the other, we replaced the frequency in the second sample with half the frequency of being represented by a single UMI in that sample. We determined the optimal fold-change threshold by comparing replicates of the same time point, and samples from different time points. Since we know that by definition there should be no expansion between the repertoires sampled at the same time point, we used 1.5 as the optimal fold-change threshold for our samples (Figure S5.2A).

Secondly, we quantified the sampling noise, which is expected to have a larger effect on the fold-change of rare sequences. Based on the relative frequency of a sequence in a reference sample, we calculated its expected number of UMIs in another sample. If no further threshold would be added, this would result in many sequences to be classified as expanded, even between samples from the same time point (Figure S5.2B). We therefore added a threshold to this number (accounting for the contribution of sampling noise to the absolute UMI count), to obtain the minimum UMI count to be classified as expanded compared to the reference sample. By setting an absolute UMI count threshold of 30 UMIs, we decreased the number of expanded sequences between samples from the same time point, while still allowing the detection of expansions between time points (Figure S5.2B).

Thus, we classify a sequence as expanded between sample 1 and 2, if (1) the relative frequency in sample 2 is at least 1.5 times higher than in sample 1, and (2) the absolute UMI count in sample 2 exceeds the relative frequency in sample 1 with at least 30 UMIs. These two requirements together result in the dashed lines shown in Figure 5.2C-F.

Quantification of overlap between samples

We quantified the TCRβ overlap between sample pairs using Bray-Curtis dissimilarity, because it takes abundance into account and its value can be intuitively understood. For a collection of TCRβ X with proportions X_i and X_j in sample i and j, respectively, the Bray-Curtis dissimilarity is calculated as $BC = 1 - \sum \min(X_i, X_j)$. The relative overlap, $1 - BC$, can thus be understood as the proportion that is identical between two samples, in terms of identity and abundance, such that no overlap remains if this part is removed from both samples.

To obtain a quantitative estimate of the minimum resolution of the TCRβ sequencing assay, we compared replicates from the same time point from the same individual with each other. Since these are obtained from the same TCR repertoire, they provide the opportunity to estimate the minimum TCRβ sequence frequency that guarantees overlap between samples. We then sorted the TCRβ sequences based on abundance in the largest sample. Starting from the sequence with the highest frequency, we kept track which fraction of the TCRβs was also observed in the smaller replicate. Continuing this until less than 90% of the most abundant sequences was overlapped, we obtained an estimate on the minimum TCRβ frequency that is required to guarantee overlap between samples from the same TCR repertoire.

Diversity estimates

Many measures exist to quantify diversity of a sample, which mostly differ in the relative contribution of richness and evenness. Richness relates to the distinct number of TCRβ sequences in a sample, while evenness quantifies the differences in abundance between TCRβ sequences. We used three distinct measures to estimate the TCRβ diversity in our samples. The richness is the total number of distinct TCRβ sequences in the sample. Given a collection of TCRβ X with proportions X_i in a sample, the Shannon index is $H = -\sum X_i \ln X_i$, which can be expressed as the effective number of species by e^H. The Simpson index is given by $\lambda = \sum X_i^2$, of which the inverse $1/\lambda$ is the effective number of species. These three measures were evaluated for the TCRβ repertoires at each time point, and plotted in Figure S5.3A-C. To compare diversity between donors and timepoints while accounting for the different sample sizes, we computationally down-sampled all samples to have a total number of UMIs equal to the smallest sample in the set. The normalised diversity measures calculated from these down-sampled repertoires are provided in Figure S5.3D-F.

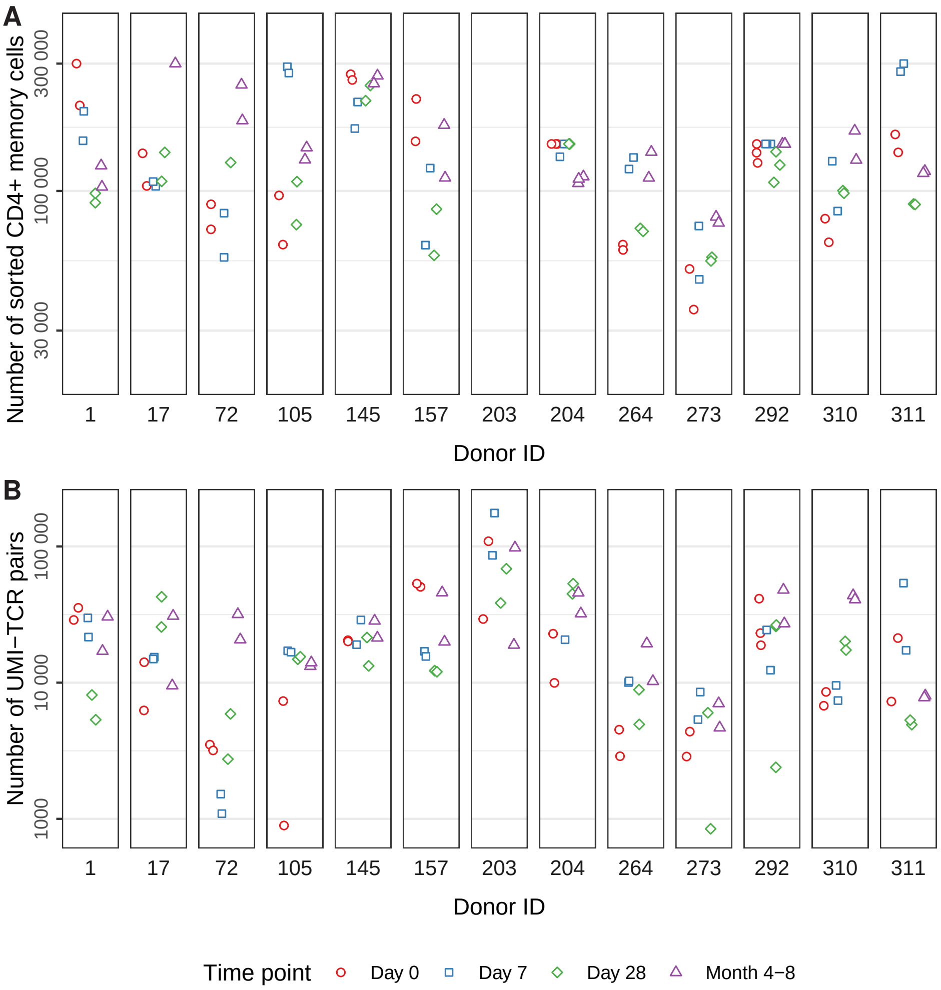
Author contributions

P.C.d.G., J.L., J.A.M.B., D.v.B., and R.J.d.B. conceived the study. The experimental procedures were carried out by J.L., M.H., A.C., and M.V.. P.C.d.G. performed the data analysis in consultation with the other authors. P.C.d.G. and J.L. wrote the manuscript with input from J.A.M.B., D.v.B., and R.J.d.B.. The manuscript was edited and approved by all authors.

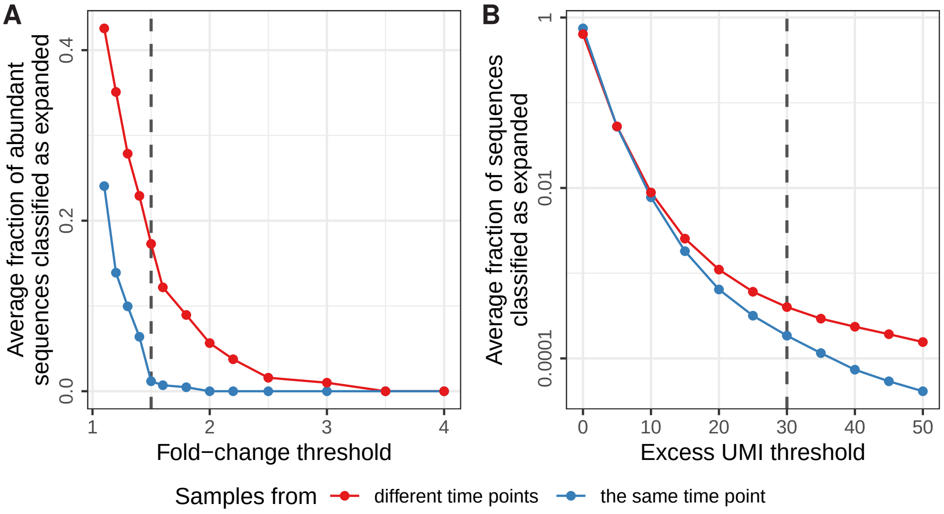
Acknowledgments

We thank Birgit Weinberger and Magdalena Hagen for providing the DT antibody data, Anne Marie Buisman for providing the CMV data, Josine van Beek for sample selection, and Martijn van de Garde for critically reading the manuscript. This work was supported by The Netherlands Organization for Scientific Research Graduate Program 022.005.023 (to P.C.d.G.). The VITAL project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 806776 and the Dutch Ministry of Health, Welfare and Sport. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA-members.

Supplementary Material



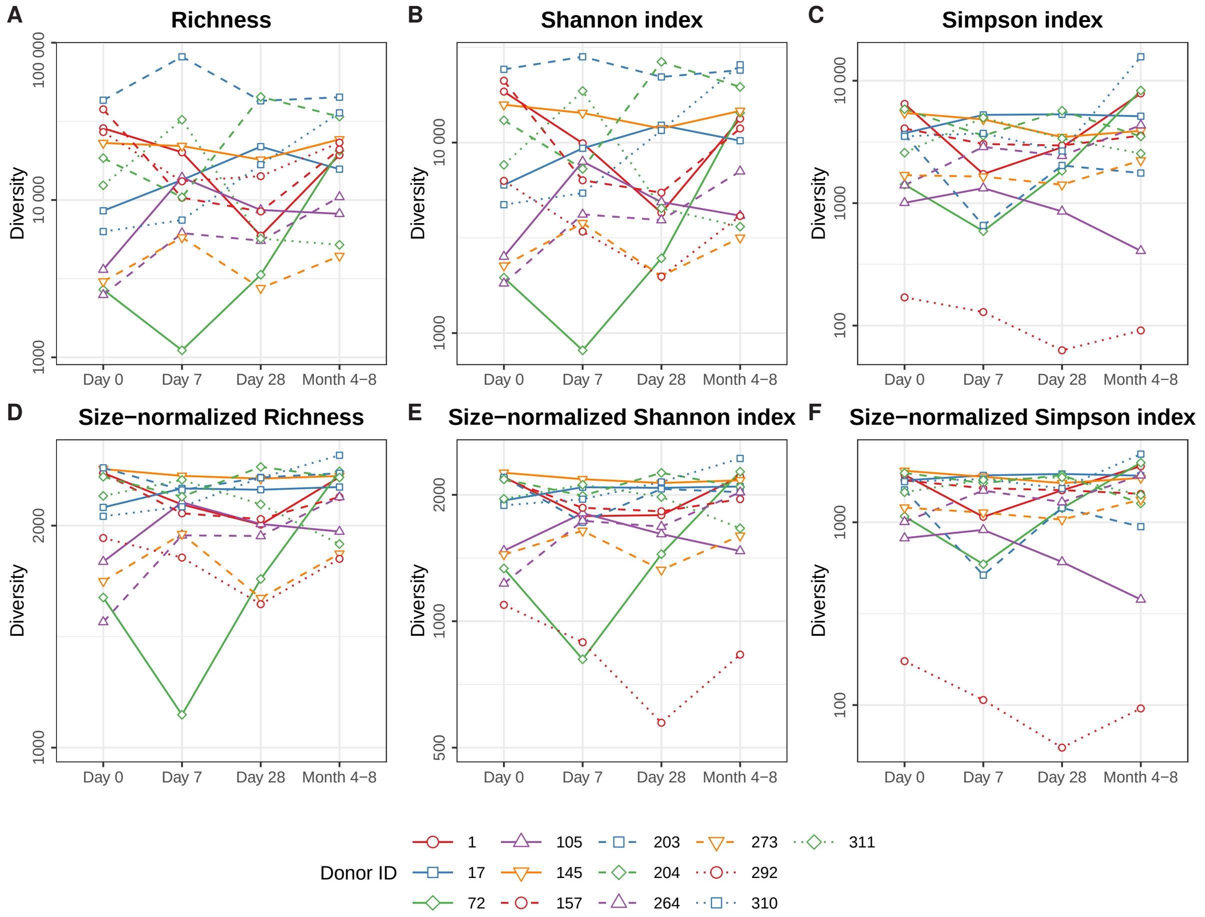
Supplementary Figure S5.1 – **Number of sorted cells and identified TCRβ sequences. A.** Number of sorted CD4 memory T cells per sample. While the numbers of sorted cells for donor 203 were in the same range as the other samples, the exact numbers could not be retrieved. **B.** Total number of TCRβ sequences retrieved by TCR sequencing per sample.



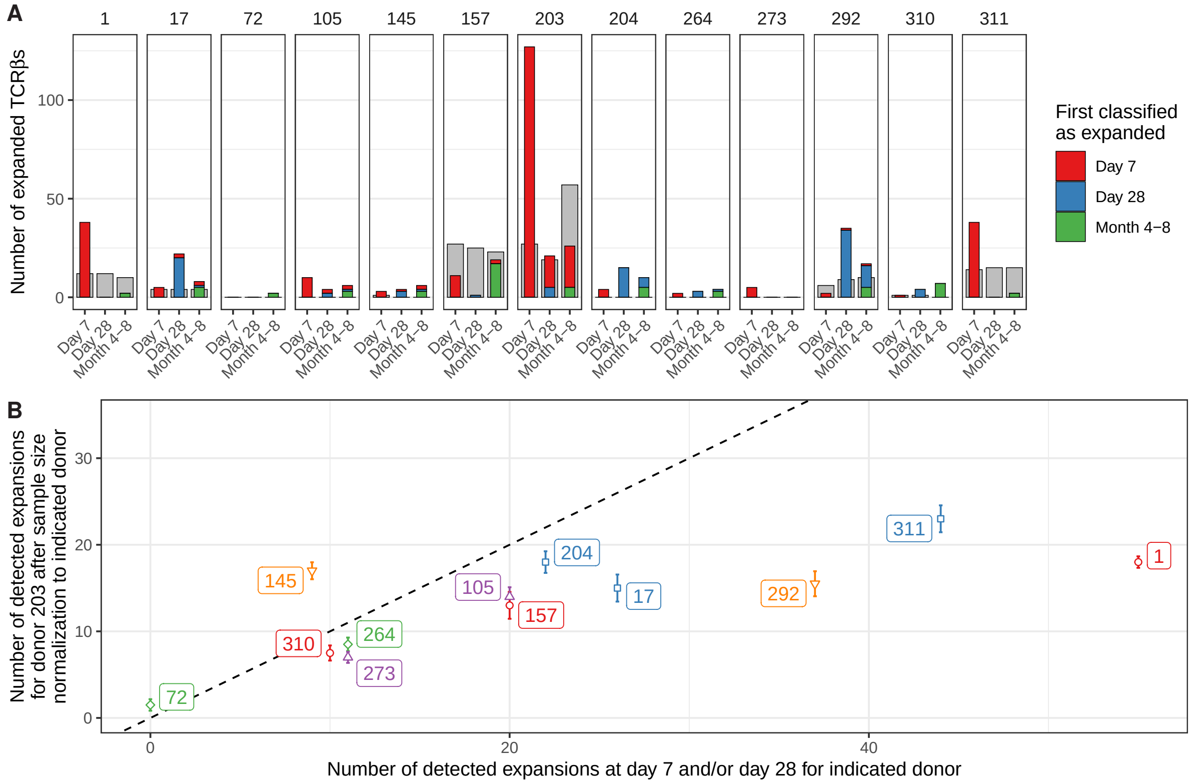
Supplementary Figure S5.2 – **Thresholds for classification of expanded sequences.** **A.** The fraction of abundant sequences (> 0.5% in the reference sample) that would be classified as expanded because their fold-change exceeds the threshold (horizontal axis). Comparisons were made between samples from the same time point (blue) and between a reference sample and a sample from a later time point (red). The vertical dashed line indicates a fold-change of 1.5, which was used in the analysis to classify expansion. **B.** Similar to A, but now for all sequences and additionally requiring an absolute difference in UMI count. A sequence is classified as expanded if (1) the fold-change is larger than 1.5, and (2) the number of UMIs for that sequence exceeds the reference relative frequency with at least the excess UMI threshold (that varies on the horizontal axis). The vertical dashed line indicates the threshold of 30 UMIs, which was used in the analysis to classify expansion.

Supplementary Table S5.1 – **Donor characteristics**

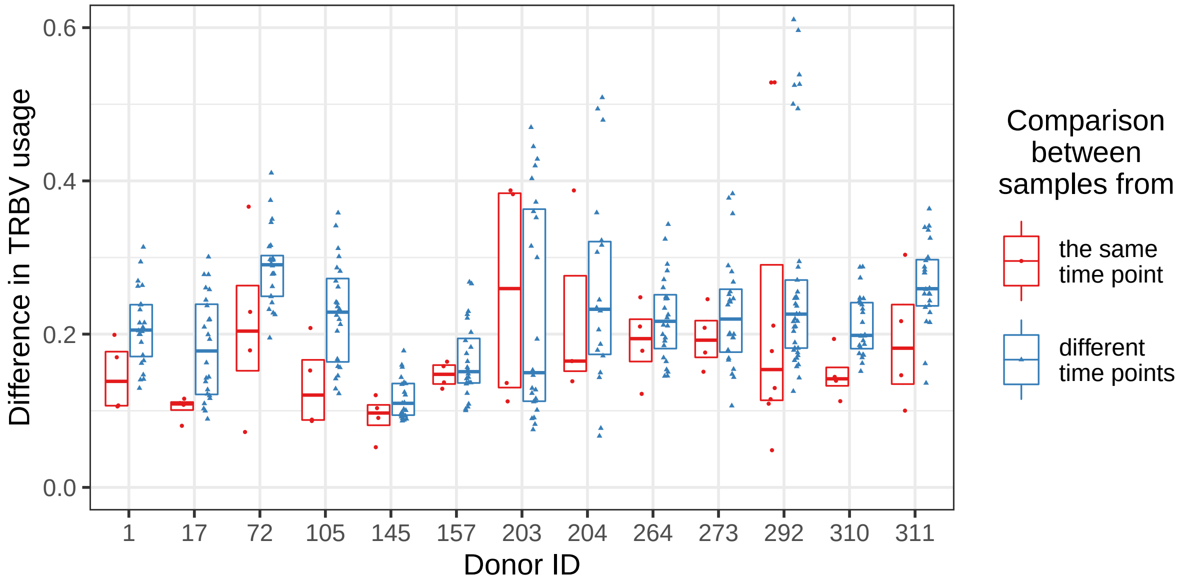
Donor ID	Age (years)	Sex	CMV status	Timing follow-up sample	Comments
1	70	M	CMV-	6.5 months	
17	73	F	CMV-	4.4 months	
72	71	F	CMV-	6.5 months	
105	68	M	CMV-	5.5 monts	
145	69	F	CMV-	6.2 months	Sequenced twice
157	27	M	CMV-	6.9 months	
203	38	F	CMV+	7.7 months	
204	25	F	CMV-	6.0 months	Illumina NextSeq; 3 samples per time point
264	27	F	CMV-	6.8 months	
273	31	M	CMV-	5.7 months	
292	28	F	CMV+	6.5 months	Illumina NextSeq; 3 samples per time point
310	30	F	CMV-	5.4 months	
311	25	F	CMV-	5.5 months	



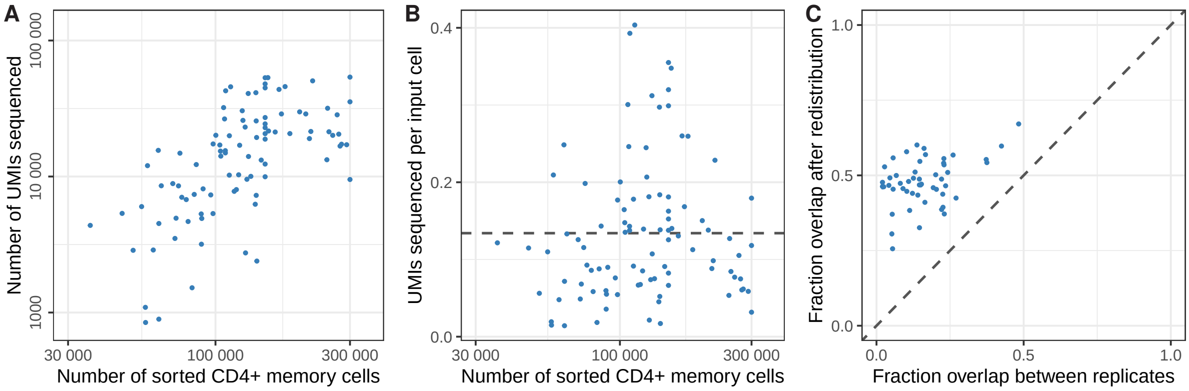
Supplementary Figure S5.3 – **Estimated TCRβ diversity in pooled replicates before and after vaccination. A.** Number of distinct TCRβ sequences per time point, after pooling the sequence counts of the corresponding replicates. The vertical axes have a logarithmic scale. **B.** Effective number of species (TCR sequences) as quantified with the Shannon index (see Methods). **C.** Effective number of species (TCR sequences) as quantified with the Simpson index (see Methods). **D-F.** Similar to A-C, but after normalising the size of the samples by down-sampling to the smallest sample (2613 UMIs) to allow for comparison between time points and individuals.



Supplementary Figure S5.4 – **Expansion based on individual replicates and down-sampling of pooled samples.** **A.** Similar to Figure 5.3A, for which samples from the same time point were pooled before classification of expansion. Here, the comparisons were made between individual replicates, meaning that a sequence will be classified as expanded if it satisfies both requirements in any of the post-vaccination replicates compared to any of the pre-vaccination replicates. The number of TCRβs meeting both requirements of expansion in any of those comparisons is plotted. Colours indicate the first time point at which the specific sequence was classified as expanded (red: day 7, blue: day 28, green: month 4-8). The grey bars serve as a proxy for dynamics that are not induced by the vaccination, by classifying 'expansion' while reversing the order of the time points. **B.** Comparison of the number of detected expansions in donor 203 versus the other donors, after correcting the data for the different sample sizes by down-sampling. We down-sampled the data of donor 203, pooled per time point like in Figure 5.3, at each time point 10 times to the number of UMIs that were identified for the corresponding pooled samples from each of the other donors. Plotted on the vertical axis is the number of detected expansions for donor 203 that remain after down-sampling (median ± standard deviation), which is compared to the number of detected expansions of the indicated donor on the horizontal axis. The dashed line indicates the identity line, which indicates an identical number of detected expansions in both cases. The low numbers of expansions that remain after down-sampling indicate that the sample sizes of most of the donors were insufficient to detect the breadth of the response that was detected in donor 203. The observation that the other donors often showed more expanding clones than observed in the equally sized samples of donor 203, suggests that they may in fact have experienced a more diverse T-cell response than donor 203.



Supplementary Figure S5.5 – **TRBV usage differences between samples from the same TCR repertoire and another time point.** Comparison of TRBV usage between samples from the same individual. The total difference in TRBV usage between samples is quantified by summing the differences in relative frequency for each TRBV gene. Comparisons are performed per donor, between samples from the same time point (red circles) and different time points (blue triangles). The result of each comparison is plotted, with the boxes summarising the median difference (thick line inside), as well as the first and third quantiles (bottom and top of the box, respectively).



Supplementary Figure S5.6 – TCRβ contribution per input cell. **A.** Total number of TCRβ mRNA molecules (uniquely labelled with UMIs) retrieved by TCR sequencing as a function of the number of sorted CD4 memory T cells for the corresponding sample. **B.** Average mRNA contribution per input cell, as measured by dividing the total number of TCRβ sequences by the number of input cells. The horizontal dashed line shows the mean value. This can be considered an upper bound of the probability a given cell in the sample will contribute an mRNA molecule, as cells can contribute multiple molecules. **C.** Fraction overlap (see Methods) between replicates before and after redistributing sequences over the samples. The relative fraction of TCR sequences that overlap between the two replicates is shown on the horizontal axis. We then combined the TCRβ counts of both samples and randomly redistributed the sequences to arrive at two artificial samples with total counts identical to the original samples. We performed this redistribution 100 times, arriving at 100 estimates for the overlap after redistribution. The median of these values is plotted on the vertical axis, with error bars indicating the standard deviation (often invisible due to the range of the error bars being smaller than the plot symbols). The dashed line indicates identical overlap between two samples for both comparisons, which is the expectation if every cell in the sample would have contributed maximally one mRNA molecule (de Greef *et al.*, 2020). The increase of overlap for all sample pairs after redistribution indicates that a substantial fraction of the cells contributed multiple mRNA molecules. Hence, the probability for a given cell to contribute a TCRβ mRNA is expected to be considerably lower than the upper bound shown in B.

6

The TCR directs naive T cells to a preferred lymph node

Peter C. de Greef [1], Sospeter Njeru [2*], Claudia Benz [2*], Jörg Kirberg [2+], and Rob J. de Boer[1+]

[1] *Theoretical Biology and Bioinformatics, Utrecht University, The Netherlands*
[2] *Paul-Ehrlich-Institut, Div. Immunology (3/3), Langen, Germany*
* *These authors contributed equally*
[+] *These authors share last-authorship*

In preparation.

Abstract

The diversity of the T-cell receptor (TCR) repertoire is an essential hallmark of the adaptive immune system. However, this diversity also complicates analysis of its spatial organisation. Here we characterise the TCR repertoire of individual lymph nodes from mice with a reduced receptor diversity. This approach allows for a reproducible comparison of repertoires within and between mice. We find evidence for a deterministic CD4/CD8 lineage choice, and a consistent spatial structure in the repertoire. Specifically, we identify a small subset of T cells with a TCR-driven preference for one or multiple lymph nodes. This shows that the spatial organisation of a part of the naive T-cell repertoire is non-random and may be affected by localised self-antigens.

Introduction

αβ T cells are one of the major cell types in adaptive immunity. The specificity of each T cell is determined by its heterodimeric T-cell receptor (TCR). The TCR genes of each T cell are formed in the thymus by a process termed V(D)J recombination. This process involves the recombination of germline gene segments, that are fused in an imprecise manner, giving rise to a huge diversity of different TCRs (Hayday *et al.*, 1985; Izraelson *et al.*, 2018; Zarnitsyna *et al.*, 2013). The α- and β-chain sequences of the TCR can be uniquely described by the combination of their V and J segments and the hypervariable CDR3 region between these segments. Resting, i.e., naive T cells are activated through binding of their TCR to cognate antigen, which are complexes of peptides with Major Histocompatibility Complex (MHC)-encoded molecules (peptide/MHC). In the absence of cognate foreign antigen, naive T cells remain resting, and their maintenance requires 'tickling' of their TCR by low-affinity binding to self-peptide/MHC molecule complexes (Brocker, 1997; Kirberg *et al.*, 1997; Tanchot *et al.*, 1997). Differences in the expression of self-antigens between anatomical locations may thus lead to a spatial organisation of the TCR repertoire.

T cells used to characterise the TCR repertoire are often sampled from the blood, as this is easily accessible and can be sampled longitudinally. Still, only a small fraction of T cells appears in the blood, as the vast majority occurs at other locations throughout the body, including the spleen, lymph nodes, and gut (Ganusov and Boer, 2007). Previous studies in mice and humans showed that there are considerable differences between the TCR repertoires sampled from distinct anatomical locations, even for naive T cells that have not been activated by cognate antigen and expanded by proliferation (Bergot *et al.*, 2015; Thome *et al.*, 2016; Zhang *et al.*, 2021). However, such analyses are always hampered by the combination of a high diversity and limited sample size and/or sequencing depth, which means that only a limited fraction of the local repertoire could be analysed. This directly affects any measure of repertoire overlap, as missing or overrepresented clonotypes could arise due to these sampling effects (Ferrarini *et al.*, 2017; Madi *et al.*, 2014) and need not reflect a true spatially structured organisation of the repertoire.

Thus, it remains an open question whether or not the TCR repertoire of naive T cells is uniformly distributed across tissues. This can be addressed in a more reproducible manner, reducing biases arising from sampling effects, when the TCR diversity is reduced in mice. In the most extreme case, this involves studying the spread and cell-division history of T cells harbouring monoclonal, transgene-encoded, TCRs after transfer into T-cell deficient recipients, such as Rag2$^{-/-}$ mice. This reveals that for several of such TCR-transgenic cells, these distribute evenly across the lymph nodes, while other TCRs show cell division, increased residence, and/or activation-marker expression in a limited but reproducible set of lymph nodes (data Kirberg lab). This indicates that even naive T cells may have a spatial niche determined by their TCR, in which their TCR may bind to localised (self-)peptide/MHC complexes that, in case of a T-cell depleted environment, even allows for cell division, i.e., homeostatic proliferation. Currently it is not known to which extent this spatial organisation was the result of excessive proliferation by these transgenic cells in a lymphopenic environment, and/or whether these rules also apply to the normal situation.

Here, we characterised the spatial structure of the TCR repertoire by Vα staining and by TCR sequencing of T cells from individual lymph nodes taken from 'one-TCRα' mice. The animals were raised using gnotobiotic technique (Heine, 1998) and only had a minimal, strictly anaerobic flora to exclude variation caused by external stimuli from newly acquired bacteria or pathogens, or changes in the microbiome composition. While their TCRα repertoire originates from V to J recombination, as in wild type mice, it has a strongly reduced diversity as all T cells must use one specific TCRβ chain. This provides a unique opportunity to perform high-throughput TCR sequencing on millions of T cells, covering their nearly complete diversity. Here, this information is used to reproducibly compare repertoires across lymph nodes and mice. By enriching the lymph nodes for resident T cells, we increase the repertoire differences between lymph nodes, suggesting that there are resident cells which are spatially structured. This reveals a layered organisation of the T-cell repertoire, in which many cells are circulating, while a small fraction shows a TCR-driven preference for a specific set of lymph nodes.

Results

Vα staining in one-TCRα mice reveals spatial differences in the TCR repertoire

The spatial structure of the TCR repertoire may be reflected by differential usage of germline TCR segments that may get evident by comparing T cells from individual lymph nodes for Vα-family usage. We performed such analyses in one-TCRα mice (see Methods) that have a strongly reduced TCR diversity. This is achieved by a transgene-encoded TCRβ chain on a homozygous *TCR*β*-/-* background. In addition, one-TCRα thymocytes can only use the TCRα recombination product of one of their chromosomes, since the other TCRα allele is not functional. Thus, every T cell must use the identical transgenic TCRβ chain together with a recombined TCRα chain of the functional allele. In this way, there are no in-frame secondary TCRα chains that, in the normal situation, a third of all T cells may

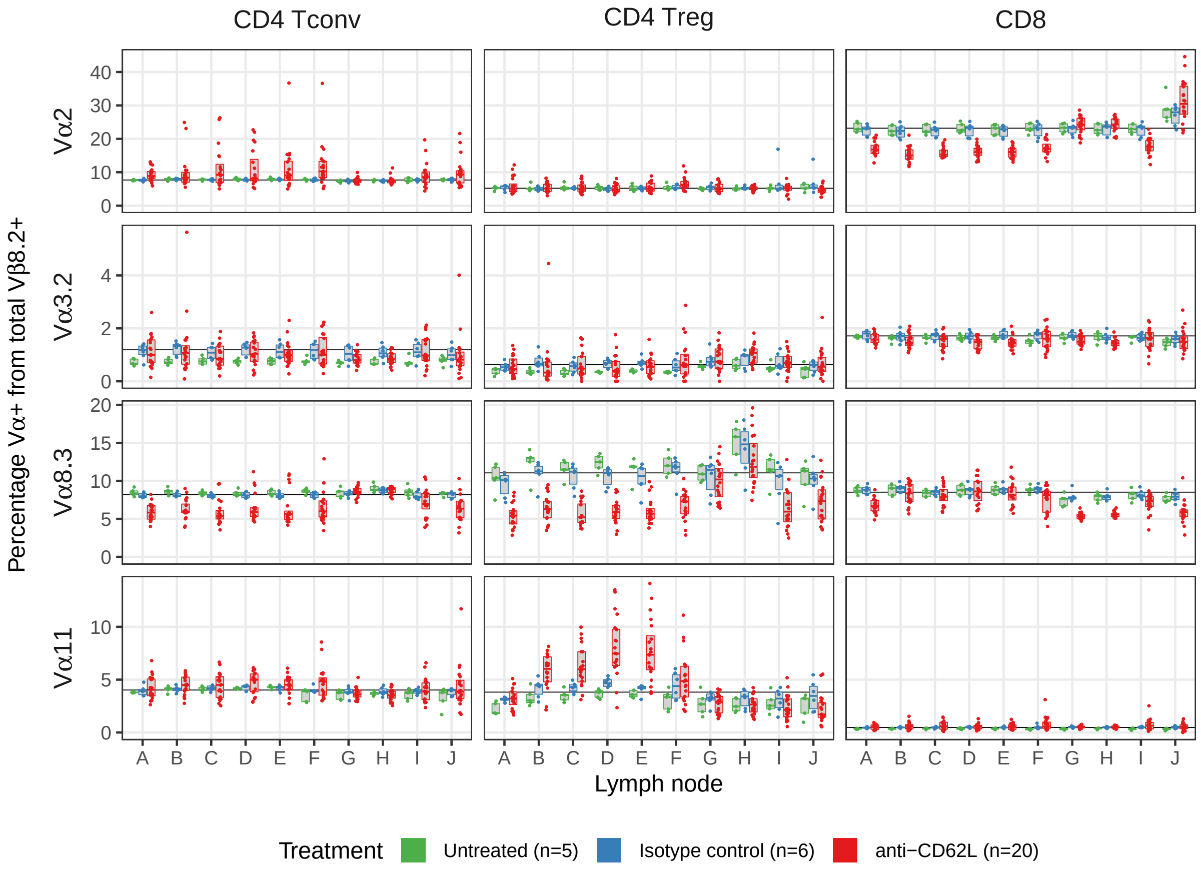


Figure 6.1 – Vα staining of T-cell subsets in lymph nodes. Shown is the fraction of T cells staining positive for the indicated Vα-family specific antibodies as a percentage of all αβ T cells (αβ T cells being identified as Vβ8.2$^+$ since all αβ T cells in one-TCRα mice must use the transgenic TCRβ chain). Note that the scales of the vertical axes differ between rows. Top and bottom of boxes indicate the first and third quantiles, respectively, and the line inside represents the median value for each lymph node. Individual values are shown as small dots. The horizontal black line indicates the median value of the isotype control group across all lymph nodes per Vα family for each T-cell subset. The anatomical location of each lymph node is shown in Figure S6.1A.

have (Casanova *et al.*, 1991). The diversity of the peripheral TCR repertoire in one-TCRα mice is strongly constrained, due to the lack of TCRβ chain diversity and because only the TCRα chains that are compatible with the fixed TCRβ chain to survive thymic selection can be present. In these mice we can study the distribution of the TCR repertoire by just analysing the TCRα repertoire.

We thus stained T cells derived from various lymph nodes of individual one-TCRα mice (Figure S6.1A) maintained by gnotobiotic technique and subdivided the T cells by gating into the CD4 Tconv, CD4 Treg and CD8 populations (see Methods). For each of these, we determined the proportion of cells using one of the four Vα-families for which specific monoclonal antibodies are available. This revealed Vα-usage patterns that were mostly population-specific, indicating that the TCR repertoire of CD4 Tconv cells is clearly distinct from that of CD8 T cells (Fig. 1 - blue and green). The CD4 Tconv and CD8 subset each showed a very reproducible Vα usage across mice and lymph nodes. Vα usage of CD4 Treg overall resembled that of CD4 Tconv cells, although being less consistent between mice and showing more variation for some Vα-family usage in one or more lymph nodes. Together, this shows that Vα staining does not reveal if there is a spatial distribution within the TCR repertoire, even with the reduced TCR diversity of one-TCRα mice.

Differences in the TCR repertoire between lymph nodes may be too subtle to be picked up by Vα-family staining. For example, there could be a small fraction of cells with a clear preference for a given lymph node, whose signal gets diluted by circulating cells with other TCR specificities but using a V segment belonging to the same Vα family. We enriched the repertoire for 'resident' T cells by treating mice with antibodies to block CD62L (L-selectin). This treatment prevents lymphocytes from entering the lymph nodes from the blood during their normal recirculation, since CD62L is needed to interact with ligands present on the endothelium of the high endothelial venules (HEV). Thus, T cells that have TCRs without a specificity for self-antigens presented in the given lymph node will likely leave faster than those T cells that have a TCR that, in the given lymph node, will bind to local (self-)peptides.

Indeed, after treatment with anti-CD62L, the Vα usage of the remaining cells showed increased inter-individual variation and was clearly distinct to that in mice that were not treated or received control antibodies (Figure 6.1 - red *vs.* blue and green). The increased inter-individual variability may be partly the result of the lower number of remaining cells after treatment. Still, there were also consistent differences that were induced by the treatment. For example, the relative increase for Vα2$^+$ cells among CD8 T cells in the renal lymph nodes (J) co-occurred with a relative decrease of such cells in lymph nodes draining the skin (A-F) (Figure 6.1). Also, the fraction of Vα8.3$^+$ CD4 Tconv cells was reduced in all lymph nodes except for those located close to the gut (G&H) (Figure 6.1). These results indicate that the drainage of cells out of lymph nodes is not a random process that would on average affect all cells equally, since in that case the frequency in the use of specific Vα-families should not change upon blocking CD62L. Rather, it appears that lymph nodes may contain a subset of cells that are (relatively more) resident and that carry a distinct TCR repertoire.

The TCR repertoire is reproducible and the result of deterministic lineage choice

Although the Vα staining on mice treated with anti-CD62L antibodies showed some spatial patterning, these results are based on the summation of TCRs that share Vα-family use while still having a diverse TCR repertoire. We then characterised the distribution of individual TCRs, thus T cells having the same TCR specificity because they share their TCRα amino acid sequence, across lymph nodes and between individual mice by TCRα repertoire sequencing. As many cells as possible of the CD4 Tconv, CD4 Treg and CD8 subsets were sorted from each of the lymph nodes and their TCR repertoire was characterised by targeted NGS sequencing of TCRα cDNA libraries with an average coverage of ~ 30 reads

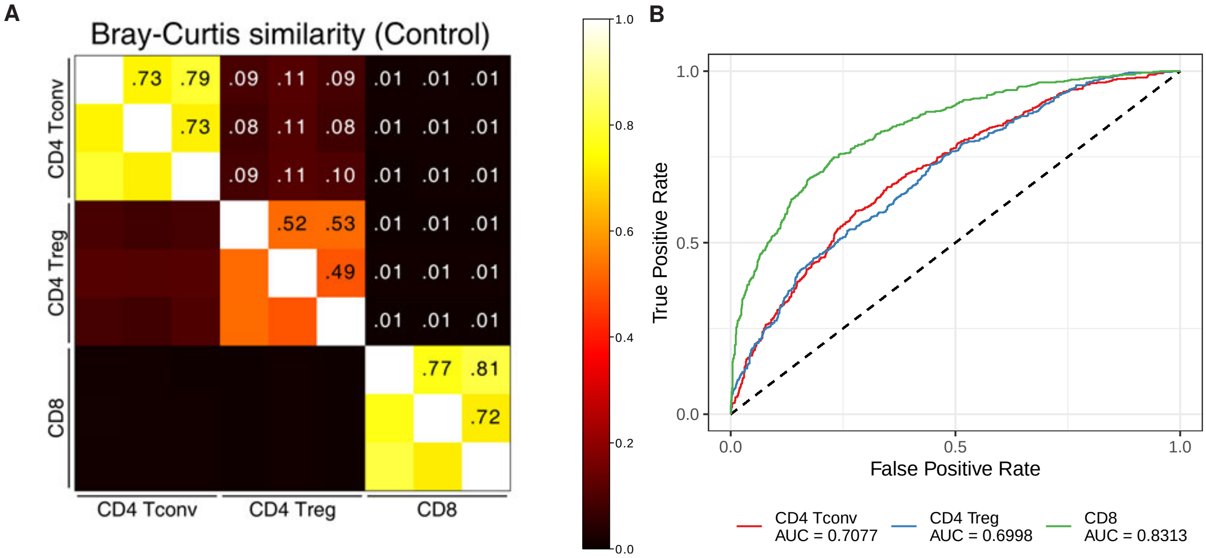


Figure 6.2 – **TCR repertoires are reproducible and result from deterministic lineage choice in one-TCRα mice. A.** Heatmap showing Bray-Curtis similarity (see Methods) between repertoires across cell subsets of three mice. Data from individually sequenced lymph nodes were merged to obtain subset-specific repertoires for each mouse. On the diagonal the score is maximal, since comparing a sample against itself yields a similarity score of 1. Above the diagonal, values are shown to allow for quantitative comparison. **B.** Receiver operating characteristic (ROC) curves resulting from a neural network that predicts cell subset based on TCR sequence (see Methods). The area under the curve (AUC) values are indicated and can be compared to a random prediction (dashed line, AUC = 0.5).

per cell. We performed error correction using Unique Molecular Identifiers (UMIs) and the RTCR pipeline (Gerritsen *et al.*, 2016) (see Methods), which resulted in a dataset of > 25 million error-corrected TCRα sequences. Since T cells in one-TCRα mice have a fixed TCRβ chain, each T cell's specificity is entirely encoded by its single functional TCRα chain. TCRα chains were differentiated according to their individual combination of V family, J gene and CDR3 amino acid sequence. We compared the similarity for TCRα-chain usage between any two samples using the Bray-Curtis index (see Methods). This measure can be interpreted as the fraction of the sample comprising identical TCRs at the same frequency in both samples. Thus, this measure takes both identity and abundance of the TCRs into account to derive a numerical value of the repertoire overlap. In other words, it yields 1 for identical repertoires, when all TCRα sequences occur at the same relative frequency, and yields 0 for repertoires having no overlapping sequences.

The TCR repertoire of each subset shows a high overlap between individual mice, especially for the CD4 Tconv and CD8 subsets (Figure 6.2A). Apparently, this is a direct result of the fixed TCRβ chain which forces the peripheral TCRα repertoire, downstream to thymic selection, to obtain a less diverse TCR repertoire for which TCR frequencies are shared, even among individual mice. Although the CD4 Treg repertoires have a smaller overlap between mice than CD4 Tconv and CD8 cells, they still contain many TCRs that are present in all mice and at similar frequencies.

In contrast to the overlap of each subset between individual mice, the TCRs appear extremely subset-specific such that almost no TCRs are found to occur both in CD4 and CD8 cells. Only a small fraction of the TCR repertoire overlaps between the CD4 Tconv and Treg subsets and these show virtually no overlap with the CD8 subset, both within individual mice and comparing across individual animals (Figure 6.2A). Technically, this implies a high purity of the samples, which is essential to characterise TCR repertoires. Biologically, it means not only that the TCR repertoire of CD4 and CD8 cells are different (Bergot *et al.*, 2015; Camaglia *et al.*, 2023; Mark *et al.*, 2022), but that the CD4/CD8 lineage choice of T cells is completely TCR-driven, and that this choice is made consistently within and between mice. In other words, it shows that CD4/CD8 fate choice is largely deterministic, although it is based on the affinity of the TCR for a large variety of self-peptide/MHC complexes. We illustrate the deterministic nature of these lineage choices further, by training a neural network on the data (see Methods). The prediction of a CD8 fate based on the TCRα sequence achieved a high AUC of 0.8, indicating that the TCR indeed determines which phenotype the cell adopts (Figure 6.2B). While the model cannot capture the full decision process, the thymus apparently makes a consistent CD4/CD8 choice given a specific TCR. Note that these observations depend on the fixed TCRβ chain and that TCRα-based predictions would not be possible when paired with variable TCRβs (Camaglia *et al.*, 2023; Carter *et al.*, 2019). Also, the absence of secondary TCRα chains, that arise due to lack of allelic exclusion for TCRα, will have contributed during the training of the neural network. In fact, since the combination of TCRα and TCRβ determines the affinity for self-peptide/MHCs, one would have to perform paired TCR$\alpha\beta$ sequencing on many more cells to capture the reproducibility of the selection process in wild type mice with full TCR diversity. Thus, one-TCRα mice provide a unique opportunity to characterise the rules governing lineage choice during thymic selection.

Resident T cells have a lymph node-specific TCR repertoire

The TCRα repertoire of one-TCRα mice demonstrates considerable overlap across various lymph nodes within and between untreated mice. Similarity scores were highest for CD8 cells, followed by CD4 Tconv cells, while CD4 Treg cells only showed a limited overall degree of overlap (Figs. 3A and S3A&B). Thus, without impeding lymphocyte recirculation, the TCRα repertoire of conventional T cells (CD4 Tconv and CD8) is mostly shared among the various lymph nodes, both within and across individuals. The high overlap within individuals is expected to be caused by a continuous T-cell redistribution due to lymphocyte recirculation (Textor *et al.*, 2014). For CD4 Treg, the overlap between different lymph nodes is less complete, demonstrating that a significant fraction of the repertoire is either not shared or shared with different frequencies among the different lymph nodes of an individual (Figure 6.3A). Interestingly, the CD4 Treg subsets showed a clear spatial structure with the TCR repertoire of external lymph nodes (B-E) being more similar to each other as compared to the other nodes.

The lymph nodes from mice that were treated with anti-CD62L contained far fewer

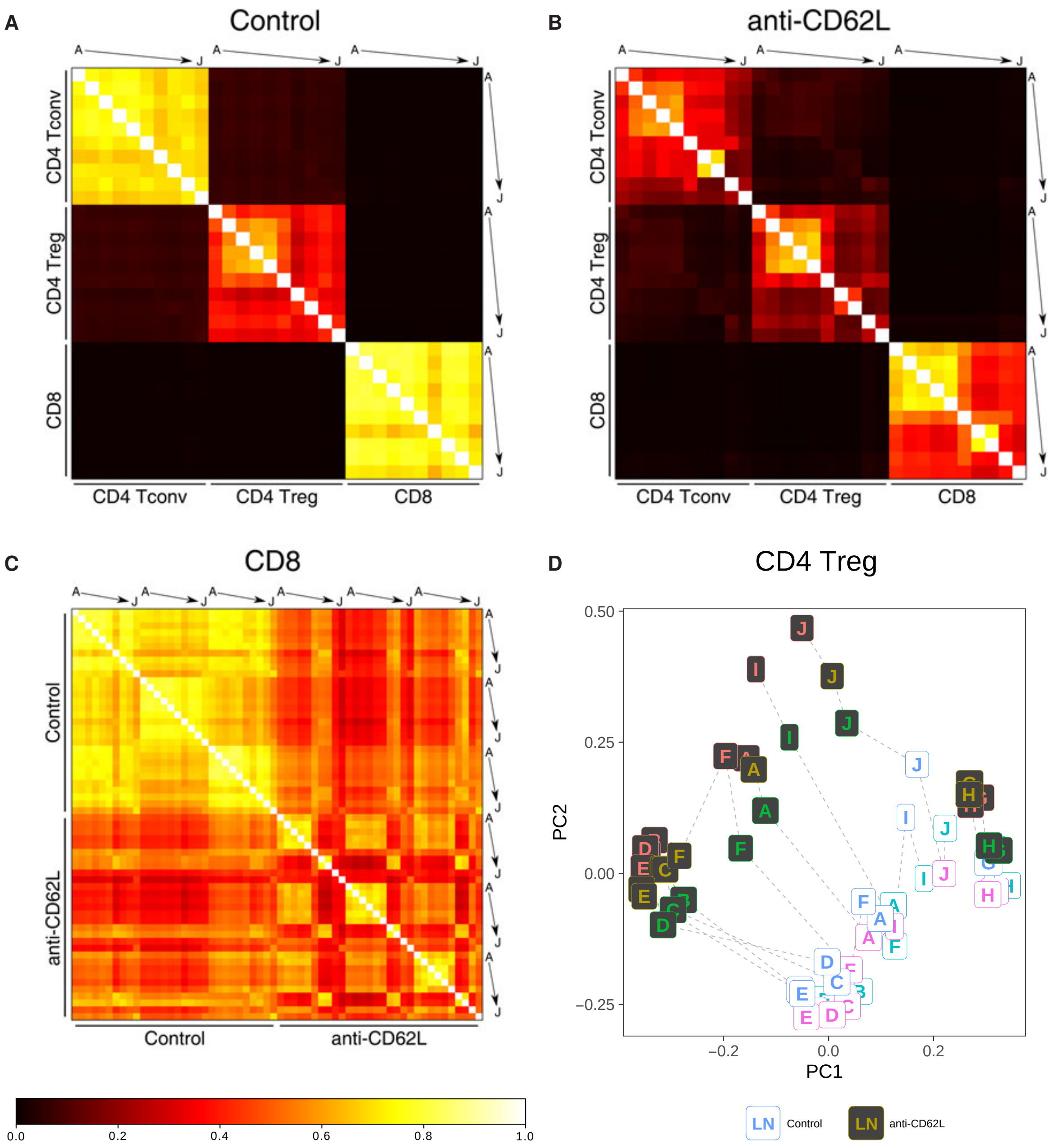


Figure 6.3 – **Lymph node-specific changes to the TCR repertoire upon blocking CD62L. A&B.** Heatmaps showing Bray-Curtis similarity between TCR repertoires of the indicated T cell subsets from individual lymph nodes of a control (A) and an anti-CD62L-treated mouse (B). Further mice are shown in Figure S6.3. **C.** Heatmap showing Bray-Curtis similarity between the TCR repertoire of CD8 T cells from individual lymph nodes of control and anti-CD62L-treated mice. The other subsets are shown in Figure S6.4A&B. **D.** Visual representation of the TCR repertoire structure by Principal Coordinate Analysis (PCoA) on CD4 Treg cells. Lymph nodes are depicted as letters, individual mice are shown as colours, with the background distinguishing control (white) or anti-CD62L-treated (black) mice. Identical lymph nodes across mice are connected by dashed lines. Other subsets shown in Figure S6.4C&D.

cells (Figure S6.1B) and generally showed less overlap between lymph nodes (Figs. 3B and S3C&D). The pattern that external lymph nodes of untreated mice have a similar CD4 Treg TCR repertoire was, however, enhanced by anti-CD62L treatment. Moreover, a similar pattern became apparent for CD4 Tconv and CD8 cells. In addition, the mesenteric lymph nodes (G&H) of treated mice showed high overlap within and between mice, indicating a consistent preference of some TCRs towards these nodes. Altogether, the TCR repertoire of T cells retained within specific lymph nodes is to a large extent reproducible, as shown by the overlap between identical lymph nodes of several treated mice (Figure 6.3C and Figure S6.4A&B). By performing a down-sampling analysis we could show that the observed effects of treatment are to a large extent consistent across mice and not simply the result of sequencing fewer cells (see Methods and Figure S6.2). A PCoA analysis on the dissimilarities between repertoires showed that the differences between TCR repertoires of individual lymph nodes are mostly due to lymph node identity and treatment, rather than mouse identity (Figs. 3D and S4C&D). Even on the level of individual TCRα sequences we find consistent spatial patterns that increase in magnitude upon anti-CD62L treatment (Figure 6.4). This figure also illustrates that there are abundant TCRs that are consistently enriched or depleted in certain lymph nodes after treatment. Specifically, 8 out of 9 abundant TCRs showed a clear preference for just one or a few lymph nodes.

A T cell's preference for a lymph node is TCR driven

The strong and reproducible effects of anti-CD62L treatment on the repertoire distribution makes it possible to classify TCRα sequences for each lymph node, based on their relative enrichment or depletion upon treatment. Since we are comparing across different mice, such an analysis is limited to the TCRs that are sufficiently shared between treated and/or control mice, which account for 50-90% of the total TCRα-repertoire in most samples, depending on the subset and lymph node location (Figure 6.5A - coloured bars). We found many TCRs to be depleted upon treatment, indicating that those T cells are leaving a lymph node relatively quickly, most likely since their TCR does not (strongly) interact with (self)-peptide/MHC as present in the given lymph node. Such T cells could be interpreted as circulating cells (Figure 6.5A - orange bars). They accounted for a large fraction of T cells in control mice with up to 60% of such T cells present in the external lymph nodes (B-E). A much smaller set was found to be consistently enriched upon treatment with anti-CD62L, indicating that these T cells are to some extent specific to that lymph node and are longer retained in there. While consistently and often strongly enriched in treated samples, we found that such 'resident' T cells only make up a minor fraction in the control lymph nodes, typically 1-3% for CD4 Tconv and Treg, and 10-15% of CD8 cells (Figure 6.5A - red bars in control samples).

We then compared the amino acid sequences of TCRs to each other based on their enrichment or depletion in specific lymph nodes upon anti-CD62L treatment. Interestingly, pairs of TCR sequences for which their CDR3 amino acid sequences are similar were much more often enriched in the same lymph nodes than those that were less similar in CDR3 amino acid sequence (Figure 6.5B). Sequences with only one amino acid difference were

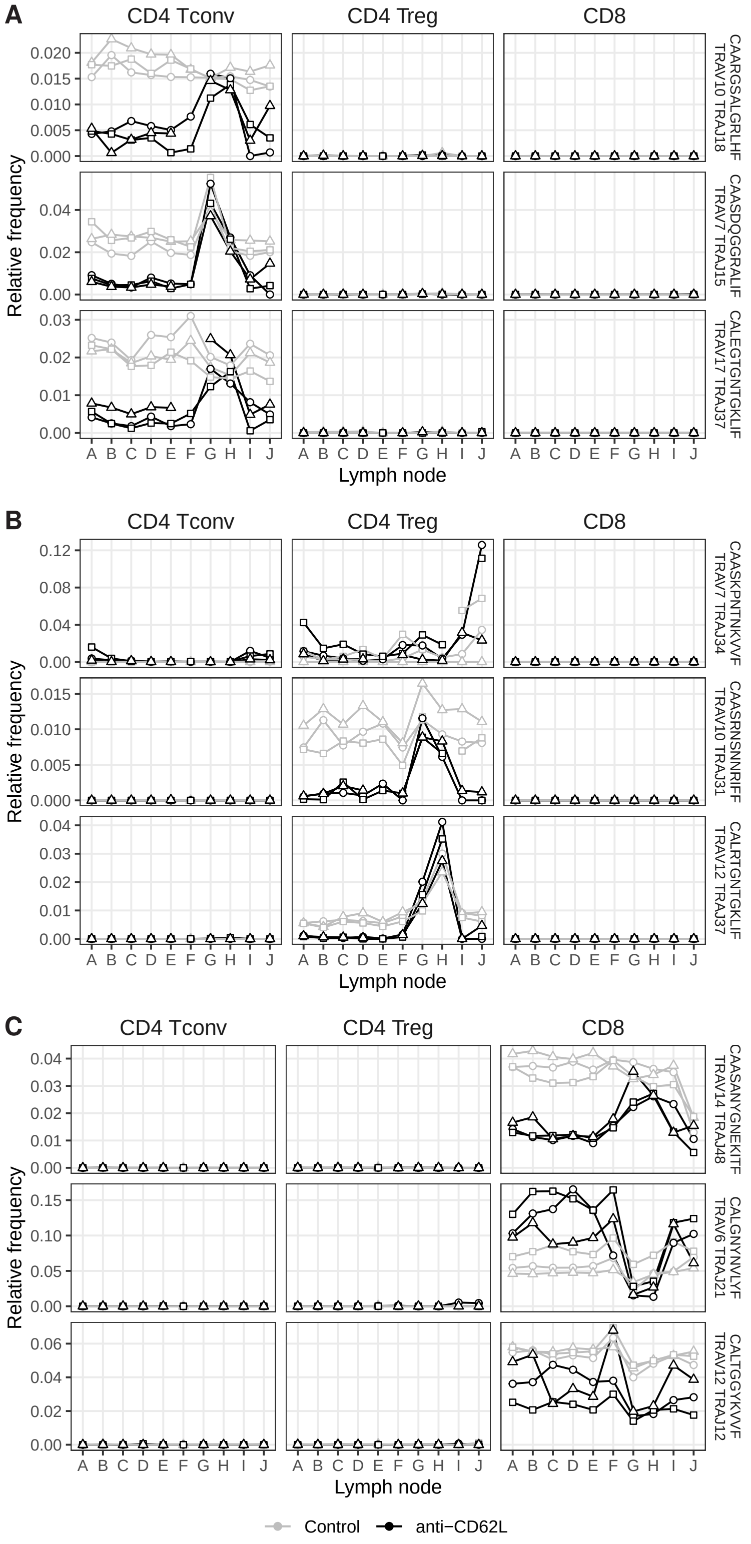


Figure 6.4 – **Frequency patterns of individual abundant TCRs.** Relative frequencies of the three TCR sequences that are most abundant in control CD4 Tconv (**A**), CD4 Treg (**B**) and CD8 (**C**) samples. Samples from the same mice share the grey or black plotting symbol and are connected with lines. Note that some data points are missing due to excluded samples (see Methods). Control mice are shown in grey, mice treated with anti-CD62L in black.

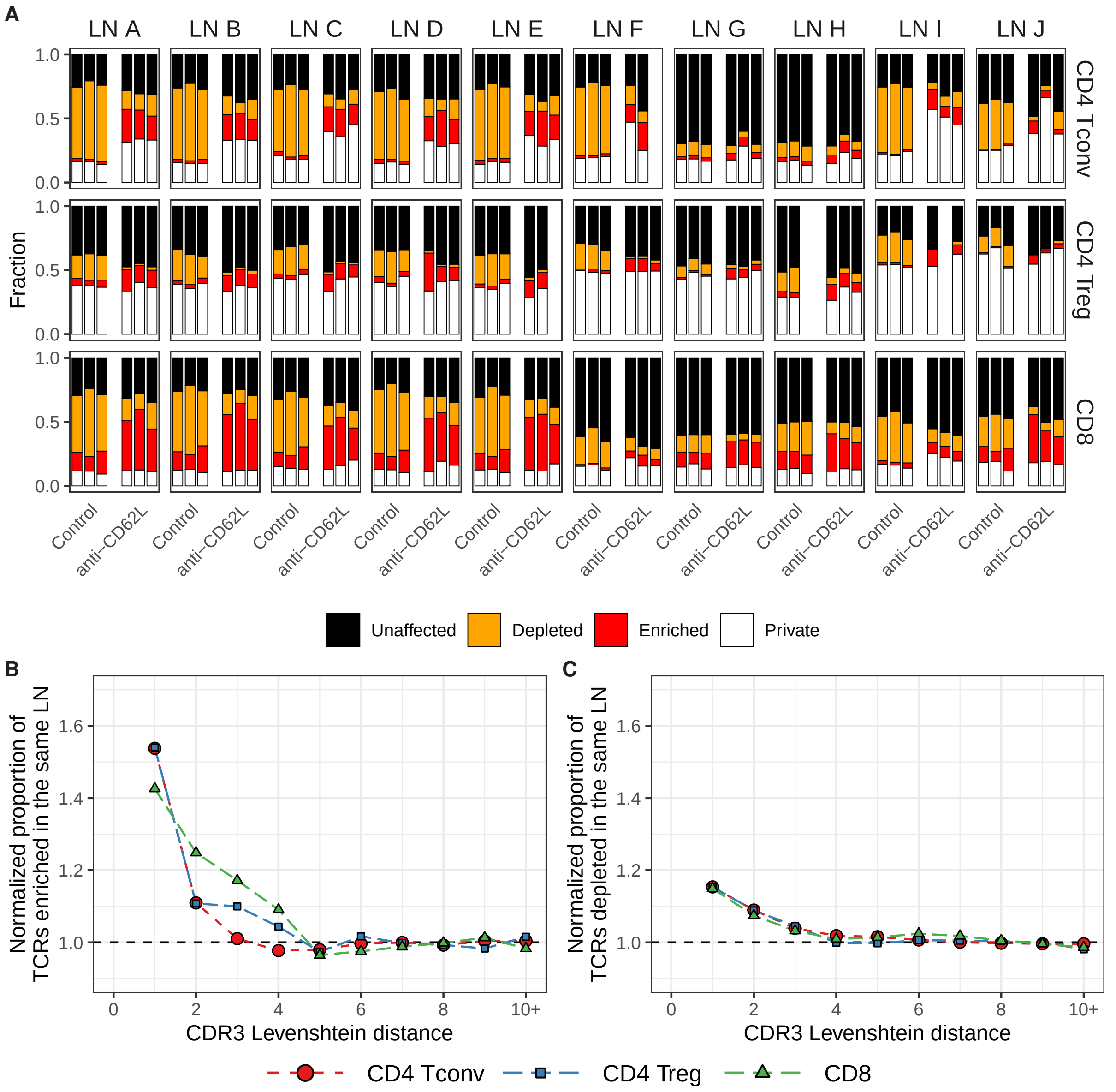


Figure 6.5 – **Lymph node residence is determined by TCR amino acid sequence.** **A.** Cumulative proportion of TCRs, classified by their frequency change upon anti-CD62L treatment in each lymph node (LN). White: TCRs that are absent in at least one control sample and at least one treated sample ('private'). Red: TCRs of 'resident' T cells that consistently showed a higher frequency in the specific lymph nodes from treated mice than those from control mice ('enriched'). Orange: TCRs of 'non-resident' T cells that consistently showed a lower frequency in the specific lymph nodes from treated mice than those from control mice ('depleted'). Black: TCRs that are sufficiently shared across mice but do not consistently differ in frequency between lymph nodes from control and treated mice ('unaffected'). **B&C.** Relation between lymph node residence behaviour upon CD62L treatment (being either 'enriched' or 'depleted') and the number of amino acid positions that differ for the TCRs being compared. Plotted is the proportion of TCR-pairs, at each Levenshtein distance between their CDR3 sequences, that share at least one lymph node in which they are classified as enriched (B) or depleted (C). The plotted values are normalised by dividing by the overall matching fraction irrespective of Levenshtein distance.

most often enriched in the same lymph node, with a difference of 3 or 4 amino acids still showing a trend towards sharing such behaviour. This suggests that there are groups of similar TCRs that are specific to one or more (self-)antigens present in specific lymph nodes, which would serve as a niche for T cells expressing these TCRs. Sequences with a lower similarity did not show a shared enrichment towards the same lymph nodes, consistent with a different TCR-specificity. Interestingly, we observed a similar but much weaker trend for the depletion of TCR sequences upon treatment (Figure 6.5C). This would be consistent with having a TCR for which the relevant (self-)peptide/MHC is not presented within the specific lymph nodes. Together, this shows that the TCR sequence is the major determinant for the residence of T cells in specific lymph nodes.

Discussion

In this study, we characterised the spatial distribution of the TCR repertoire in mice. Robust analysis of such spatial differences is challenging due to the enormous diversity of the TCR repertoire, secondary TCRα chains arising by TCRα allelic inclusion, and the limited throughput of single-cell sequencing techniques that are necessary to capture paired TCR$\alpha\beta$ sequence information. We solved these issues by studying one-TCRα mice, for which high-throughput TCRα sequencing captures the full diversity of the TCR repertoire as present in these mice. Doing so, we could show that CD4/CD8 lineage choice is a process that is largely deterministic given a TCR sequence. Moreover, we could identify TCR-abundance patterns across lymph nodes that were consistent between mice. Enriching for resident T cells by treating with anti-CD62L to block lymphocyte recirculation, we could realise a distinction between TCRs that are either retained or remain circulatory with respect to a given lymph node. This revealed that the vast majority of naive T cells in a lymph node is circulating, while a small fraction of cells is more resident, reflecting a TCR-directed preference for one or more specific lymph nodes.

The mice we studied were kept using gnotobiotic technique and only had a minimal, strictly anaerobic flora. Thus, the spatial distributions we observed are unlikely to be imposed by stimulation with foreign antigens apart from those taken up as food. Although we did not specifically sort the cells to express characteristic naive T-cell surface markers, this implies that the observed preference for certain lymph nodes is driven by interactions with locally presented self-peptides. The 4-day period between anti-CD62L treatment and cell isolation leaves the opportunity that these interactions result in both residence and local proliferation. Our study focused on the TCR repertoire, leaving the physiologic function of such lymph node-specific preferences as an open question. We speculate that specific lymph nodes function as a niche for certain TCR specificities, where these cells are more likely to be retained, receive survival signals by the ‘tickling’ of their TCR, and/or undergo cell division. Such localised fitness differences may play a role in the maintenance of naive T-cell diversity, in contrast to a scenario in which competition between TCR clones of similar specificity happens throughout the lymphoid system.

Although we could impute a TCR-driven preference for some of the analysed T cells by our strict classification, this was not possible for all sequences. Firstly, a comparative TCR profiling requires harvesting of several different lymph nodes and needs inter-individual comparisons to be able to quantify changes in TCR frequency as induced by treatment. As a result, any analysis must be limited to those TCR sequences that are sufficiently shared across individuals. Secondly, although we harvested lymph nodes throughout the body, there are many more locations that contain T cells, including other lymph nodes, spleen, and gut (Cose *et al.*, 2006). Hence, we would predict that TCRs that were consistently reduced in all studied lymph nodes may still have a specific niche at a location that we did not include in the current analysis. Alternatively, a low TCR affinity to a self-peptide/MHC ligand may not cause a significant increase in the lymph node retention time to be noticed in the time frame used here. Thirdly, the effects of treatment were not as prominent for all TCRs in all lymph nodes. For example, the mesenteric lymph nodes had a distinct TCR repertoire even before treatment (Figure 6.3D). Instead of being enriched in these lymph nodes after treatment, the TCRs with a preference for the mesenteric lymph nodes were mostly characterised by consistent depletion in all other lymph nodes. All in all, if a full characterisation of TCR-driven preferences for anatomical locations would be possible, it would require more samples from many more mice, which is difficult to realise even by current advances in high throughput sequencing.

While the use of one-TCRα mice provides the opportunity of nearly complete TCR repertoire profiling, it remains an open question to which extent the resulting insights translate to mice with a fully diverse αβ TCR repertoire. The observation that a fixed TCRβ chain allows for the generation of both CD4 and CD8 T cells supports the consensus that the lineage choice of double positive T cells is not restricted by one of the TCR chains, α or β, but rather depends on the combination of TCRα and TCRβ (Carter *et al.*, 2019; Springer *et al.*, 2021; Tanno *et al.*, 2020; Uematsu *et al.*, 1988). This would predict that the TCRα repertoire of CD4 and CD8 T cells could be very distinct in the context of another TCRβ chain or in the presence of a different MHC allele. Thus, characterising the TCR repertoire of T cell subsets while fixing different TCRβ chains or MHC haplotypes would be a promising avenue for further characterisation of the deterministic nature of lineage choice in the thymus.

Even though limited-diversity T-cell repertoires may not be as protective against foreign pathogens (Messaoudi *et al.*, 2002), they provide feasible opportunities that can be generalised to, but not directly tested in fully diverse TCR repertoires. Taken together, one-TCRα mice offer the ability to reduce the complexity of the T-cell repertoire to a level that allowed us to shed a light on the spatial organisation of TCR repertoires. This study revealed the deterministic nature of T-cell lineage choice, and the non-uniform distributions of TCRs, which are shaped by TCR-driven lymph node preferences.

Methods

Animals

One-TCRα mice were bred by intercrossing the following alleles to yield $TCR\alpha^{+/-}$ $TCR\beta^{-/-}$ TCRβ-Kaa-tg^{+} $Foxp3^{gfp/gfp}$ ♀ or $Foxp3^{gfp/Y}$ ♂ animals used for the experiments:

- TCRα (Mombaerts *et al.*, 1991),
- TCRβ (Mombaerts *et al.*, 1992),
- Foxp3-IRES-eGFP (Wang *et al.*, 2008), and
- TCRβ-Kaa (Kieback *et al.*, 2016).

Initially, animals for breeding were obtained by rederivation into isolators using hysterectomy and dunk-tank passage, placing the foetuses with MAS/A foster females that have a minimal, strictly anaerobic flora. Progeny were genotyped and the line maintained by continual intercrosses using gnotobiotic technique in isolators. Both female ♀ and ♂ male animals were used for experimentation.

Treatment with anti-CD62L

Mel-14 hybridoma cells were used to produce batches of anti-CD62L mAb (Gallatin *et al.*, 1983). 511.7D12 was used to produce an isotype-matched control mAb (anti-KLF3, (Alles *et al.*, 2014)). Antibodies were purified by Protein G affinity chromatography. To block lymphocyte recirculation *in vivo*, animals were transferred into sterile IVC-caging and received 1 mg of Mel-14 mAb in PBS i.v. on day 0 and were analysed on day 4 (Lepault *et al.*, 1994). As indicated, control animals instead received anti-KLF3 (same dose), PBS (carrier), or were untreated but similarly manipulated as injected ones. Three different batches of Mel-14 gave similar results (two produced in-house, one from low endotoxin contract production at InVivo Biotech Service, Hennigsdorf, Germany).

Cell sorting and FACS analysis

Lymph nodes were isolated, ruptured in-between nylon-meshes (40 μm pore size), filtered (40 μm pore size), and the cells obtained were stained for TCR-Vβ8.2 (F23.2-Alexa700, prepared in-house), CD4 (Pe, BD), CD8 (Pe-Cy7, BD), and CD24 (HSA, eFluor450, BD) in the presence of 2.4G2 mAb (Fc-block) tissue culture supernatant at 1/5 to 1/10 dilution. Sytox-blue dye was added and all dye-excluding cells were sorted on a 4 laser (405, 488, 561, 633 nm) BD AriaFusion cell sorter. Four subpopulations were collected, in this order from far left to far right:

1. control cells (mostly B cells: $V\beta8.2^{-}$ $CD4^{-}$ $CD8^{-}$ $CD24^{+}$, not further analysed),
2. CD4 Tconv ($V\beta8.2^{+}$ $CD4^{+}$ CD8- $CD24^{-}$ Foxp3-gfp^{-}),
3. CD4 Treg ($V\beta8.2^{+}$ $CD4^{+}$ $CD8^{-}$ $CD24^{-}$ Foxp3-gfp^{+}), and
4. CD8 ($V\beta8.2^{+}$ $CD4^{-}$ $CD8^{+}$ $CD24^{-}$ Foxp3-gfp^{-}) cells,

using ‘4-way purity’ sort-mode and at sorting speeds to generally have > 90% sort efficiency. In this way, up to ~ 700 000 cells were collected into 1.5 ml Eppendorf tubes; when necessary, additional tubes were used to collect all cells from a given origin (subpopulations 1-4 and lymph nodes A-J, see Figure S6.1A). When more than ~ 700 000 cells could be obtained, these would be processed independently and only the final NGS sequencing data were to be aggregated computationally. Reassuringly, the resulting repertoires showed high overlap before aggregation.

Following sorting, cells were pelleted in a swing-out rotor, the supernatant aspirated, and the pellet was resuspended by vortexing while adding TriZol (Sigma) to which 1/500 volume of 2-mercaptoethanol had been added, using up to 600 µl per tube. Lysates prepared in this way were long-term stored at -75°C. Where appropriate, graded numbers (10 or 40 cells) of monoclonal T cells (spike-in cells) were sorted directly on top of the lysate. To this end, the T cells were isolated from pooled lymph nodes of Rag1- or Rag2-deficient OT-2, P14 (327), OT1, DO11.10 or HA-TCR transgenic mice and stained for CD4, CD8 and TCR to identify the nominal T cells for sorting.

For FACS analyses, mAbs either to Vα2 and Vα11 or to Vα3.2 and Vα8.3 were combined with Vβ8.2, CD4, and CD8 specific reagents for staining in the presence of 2.4G2 mAb (Fc-block) tissue culture supernatant. For analysis on a LSRII-SORP 4 laser (405, 488, 561, 633 nm) instrument, cells were washed, adding Sytox-blue dye before analysis to exclude dead cells by gating.

TCRα library preparation and sequencing

Briefly, total RNA was isolated from cells lysed and stored in TriZol, ensuring complete phenol removal by an additional chloroform extraction, and adding GlycoBlue (Invitrogen) co-precipitant to ease quantitative recovery. Precipitated RNA was dissolved in citrate buffer and stored at -75°C until used. For cDNA preparation, a maximum RNA amount corresponding to ~ 400 000 cells was processed (lysates containing more sorted cells were split, processed in parallel, pooling the final material before PCR) using a gene specific primer corresponding to the B6 TRA-C 3'UTR allele polymorphism in order to limit any contribution from the 129-derived TCRα-KO allele in the presence of a 5'- and 3'-blocked Template-Switch-oligonucleotide (TSO, containing 20 nt of the Illumina Read2 sequencing primer, a UMI [N6S or N7S], and rGrGrG) at predetermined optimal conditions for first strand synthesis using any SuperScriptII (RNaseH-) equivalent commercial reverse transcriptase (usually ProtoScript II, NEB). Remaining RNA was digested, salts were removed by precipitation, and a first PCR was performed using all of the cDNA with Phusion polymerase (Finnzyme or NEB, (Wang *et al.*, 2004), a forward primer completing the Illumina Read2 sequence (introduced by the TSO during first strand cDNA synthesis), and a reverse TRA-C specific primer that overlaps the neo-insertion cassette of the TCRα-KO allele in order to exclude any PCR-products arising from the inactivated allele. Following purification by precipitation, a semi-nested second PCR used one of a series of 8 reverse primers that (from its 5' to 3') add the Illumina Read1 primer sequence, each having one of a balanced and phased internal index sequence (3 to 10 nt long, similar to (Glenn *et al.*, 2019)), and 23 nt complementary to the 5' end of the TRA-C coding region.

For cDNA and PCR, lysates were grouped and processed according to input cell number. The number of PCR cycles for each group was adjusted such that in theory (all mRNA molecules being reverse transcribed and each PCR cycle leading to a doubling of product copy number) an assumed number of 5 TCRα mRNA copies per cell (Inaba *et al.*, 1991; Ma *et al.*, 2018) would give a detectable but not overamplified band (actually a smear since derived from a multitude of TCRs) on an agarose gel. Either our assumptions were correct or T cells may have a higher TCRα mRNA copy number and our methodology is sub-optimal, but we had no lysates that failed. The resulting bands, derived from lysates containing roughly similar numbers of sorted cells, were cut out of the gel and the DNA was purified on commercial glass-milk columns (SMARTPURE kit, Eurogentec), quantified by Nanodrop, and similar amounts of the isolated DNA, each derived by the use of one of the 8 reverse primers, were pooled to yield a sample to be processed for NGS.

NGS was performed on an Illumina NovaSeq instrument using XP flowcells to obtain 2 x 250 nt paired-end reads at the DRESDEN-concept Genome Center (DcGC) / CRTD / CMCB Deep Sequencing Facility, Dresden, Germany. To this end, the DNA quantity in each sample was re-determined on a Qbit (PicoGreen fluorometry) and 8 ng DNA was subjected to a 6 cycle extension PCR with NEBNext Q5 II polymerase and primers to complete the Illumina Read2 sequencing primer region, adding unique dual indexes specific for each sample (to be read by IndexRead1 and IndexRead2 sequencing), and to incorporate the Illumina p5 and p7 regions for on-chip immobilisation and bridging PCR amplification (Bentley *et al.*, 2008). The resulting material was subjected to XP beads purification, quantified on Qbit and fragment analyser, and finally pooled, balancing the molar input according to the number of cells that contributed in order to achieve a read-depth of ~ 30-40-fold per input cell.

TCRα sequencing data processing

Paired-end reads were merged with PEAR (Zhang *et al.*, 2014) using default settings. For each read that could not be paired, for example because the product was too long to have any overlap, we added the last 15 nucleotides of Read2 to the Read1 to combine the UMI and TCR information in a single read. We then split the reads based on the 8 different internal index sequences, by selecting those merged reads that started with an exact match to one of the internal indices, followed by CAGCAGGT (which corresponds to the primers' 3' TRA-C sequence region that the series of 8 reverse primers contain to bind to TRA-C). We extracted the UMI sequences by taking the last nucleotides from the merged reads (7 or 8 bases, depending on the N6S or N7S UMI design of the actual TSO used), and only accepted those that had CCC following the UMI sequence, matching the TSO design with rGrGrG at its 3' end. RTCR (Gerritsen *et al.*, 2016) was used to generate UMI-based TCRα consensus sequences (umi_group_ec), and we proceeded with all sequences in the output that were based on at least three reads. We processed this data using the main RTCR module (run) with a B6-specific TRA reference set that corresponds to the single functional TCRα allele in one-TCRα mice (see below).

Table 6.1 – **Selected alleles from IMGT for one-TCRα repertoire analyses.**

TRAV					**TRAJ**	
1*02	6-5*04	7N-6*01	12-2*02	14D-1*01	2*01	32*01
2*01	6-6*04	8-1*03	12-3*05	14D-2*01	3*01	33*01
3-1*01	6-7/DV9*04	8-2*01	12D-1*02	14D-3/DV8*01	4*02	34*02
3-3*01	6D-3*03	8D-1*01	12D-2*02	14N-1*01	5*01	35*01
3-4*01	6D-4*01	8D-2*03	12D-3*04	14N-2*01	6*01	37*01
3D-3*02	6D-5*01	8N-2*01	12N-1*01	14N-3*01	7*01	38*01
3N-3*01	6D-6*05	9-1*02	12N-2*01	15-1/DV6-1*01	9*02	39*01
4-2*02	6D-7*04	9-2*02	12N-3*01	15-2/DV6-2*02	11*01	40*01
4-3*03	6N-5*01	9-3*01	13-1*02	15D-1/DV6D-1*07	12*01	42*01
4-4/DV10*02	6N-6*01	9-4*02	13-2*03	15D-2/DV6D-2*03	13*01	43*01
4D-2*01	6N-7*01	9D-1*03	13-3*01	15N-1*01	15*01	44*01
4D-3*06	7-1*02	9D-2*05	13-4/DV7*02	15N-2*01	16*01	45*01
4D-4*03	7-2*02	9D-3*03	13-5*01	16*01	17*01	46*01
4N-3*01	7-3*04	9D-4*01	13D-1*04	16D/DV11*01	18*01	47*02
4N-4*01	7-4*01	9N-2*01	13D-2*03	16N*01	21*01	48*01
5-1*02	7-5*03	9N-3*01	13D-3*01	17*02	22*01	49*01
5-2*01	7-6*02	9N-4*01	13D-4*03	18*02	23*01	50*01
5-4*02	7D-2*01	10*02	13N-1*01	19*03	24*02	52*01
5D-4*01	7D-3*03	10D*01	13N-2*01	20*01	25*02	53*01
5N-4*01	7D-4*02	10N*01	13N-3*01	21/DV12*03	26*01	54*01
6-1*03	7D-5*02	11*02	13N-4*01		27*01	56*01
6-2*02	7D-6*02	11D*01	14-1*01		28*01	57*01
6-3*02	7N-4*01	11N*01	14-2*02		30*01	58*01
6-4*03	7N-5*01	12-1*06	14-3*01		31*01	61*01

A customised set of one-TCRα-specific TRAV and TRAJ germline reference sequences

The TCRα locus contains many gene segments, as multiple duplications led to nearly identical V gene paralogues present in the germline as allelic forms, by which strains may differ. As a result, it is not always possible to uniquely identify a V gene in TCR sequencing data, especially when using relatively short reads and when the reference set of sequences derives from a cumulative database on all mouse strains, without clear assignment from which strain a given allele is derived. Indeed, the TCRα locus is polymorphic and structurally different between several inbred mouse strains indicating a long history that is at evolutionary scale (Rupp *et al.*, 2016). In one-TCRα mice, the single functional TCRα allele is derived from C57BL6/J, while the default mouse TRA germline reference set of RTCR is based on a combination of this and several other mouse strains, such as 129Sv, BALB/c, and many others. Because RTCR estimates the error rate of the data based on errors with respect to the germline sequence, it is crucial to minimise mismatches between the true TCRα germline-corresponding V-segment, as found during experimental sequencing, and the germline reference used for alignment. To obtain the best possible germline reference set for one-TCRα mice, we downloaded all available Mus musculus TRAV and TRAJ gene segment sequences from IMGT (Lefranc *et al.*, 2009). By running MiGMAP (https://github.com/mikessh/migmap) on one million consensus sequences supported by at least 10 reads that were randomly selected from the data, we identified the best-matching allele for each gene segment. The selected 116 TRAV and 48 TRAJ alleles that were then used as germline reference are provided in Table 6.1.

Data analysis

Cells of the same subset and derived from the same lymph node that were divided over more than one tube because they exceeded ~ 700 000 cells were processed as individual samples. Afterwards, the resulting TCR tables were joined to obtain a single TCR dataset. TCRα sequences matching those of the spike-in monoclonal TCR-transgenic cells were removed before analysis.

A total of four erroneous samples were excluded from analysis and are shown as grey bars in the heatmaps. Two samples were excluded because they were accidentally mixed during the experimental protocol. Two other samples showed a substantial repertoire overlap with another T cell subset. One of these had a TCR diversity that far exceeded the number of input cells. As to these observations these samples are apparently dominated by contamination, motivating their exclusion during further analyses.

Since the TCRβ sequence is fixed, the TCR-specificity of each T cell in one-TCRα mice is entirely conveyed by its TCRα amino acid sequence. To avoid missing overlap between samples due to ambiguous V-gene assignment, we used a combination of V family, CDR3 amino acid sequence, and J gene to identify a TCR and compare its frequency between samples.

Many different methods are used in ecology and biology to compare the composition of (TCR-)species between sites or samples. Here, we chose to use the Bray-Curtis similarity index (Bray and Curtis, 1957) to compare TCR repertoires across pairs of samples, because it takes abundance into account and its value can be intuitively understood. Specifically, for a collection of TCRs X with proportions X_i and X_j in sample i and j, respectively, Bray-Curtis similarity can be calculated as $BC = \sum \min(X_i, X_j)$. The Bray-Curtis similarity can thus be understood as the proportion that is identical between two samples, in terms of identity and abundance, such that no overlap remains if this part of the sample is removed. The Principal Coordinate Analysis (PCoA; Figure 6.3D and Figure S6.4) was performed using the scikit-bio library in Python 3, by interpreting the Bray-Curtis dissimilarity ($1 - BC$) as a distance.

Subsampling

Typically, treatment with anti-CD62L reduces the cell number in a lymph node and enriches for T cells that remain longer ('resident' T cells). To check to which extent the set of remaining TCRs, as compared to the control situation, is reproducible across mice, we used a subsampling approach (Figure S6.2A). For each lymph node, we calculated the Bray-Curtis similarities between a 'central' treated sample and another treated sample (BC_{TT}) as well as a control sample (BC_{TC}). The latter two samples were diversity-matched by subsampling. Specifically, we randomly selected TCRs from the most diverse sample, until the selection reached a TCR diversity identical to the other sample. Thus, if treatment leads to a random subset of the TCR diversity remaining in the lymph node, BC_{TT} and BC_{TC} are expected to be similar. The observation that in most of such iterations BC_{TT}, the similarity between treated samples, exceeds BC_{TC}, the similarity between a control and treated sample, confirms the non-random effect of the experimental treatment (Figure S6.2B). In other words, the 'resident' T cells express a, to some extent, reproducible subset of the TCR diversity.

Subset prediction from TCR sequences

The deterministic nature of lineage choice is illustrated using a neural network. The absence of significant repertoire overlaps between TCRs of the different subsets, especially between CD4 (both CD4 Tconv and Treg) and CD8 cells indicates that the thymus makes a consistent lineage choice based on TCR specificity. This decision process *in vivo* is determined by the binding affinity to many different (self-)peptide/MHC complexes, with CD4 T cells (both CD4 Tconv and Treg) having a TCR binding to complexes of class II and CD8 T cells a TCR binding to complexes of class I MHC molecules. Here we illustrate the deterministic role of the TCR in this process *in silico* by training a neural network that predicts the likelihood that a particular TCR sequence leads the T cell to adopt the CD4 Tconv, CD4 Treg, or CD8 phenotype. Specifically, for each subset, we took the 5000 most abundant TCR sequences (in terms of V family, CDR3 amino acid sequence and J gene) across all control mice and lymph nodes analysed. We then trained a neural network using the DeepTCR architecture (Sidhom *et al.*, 2021) using 80% of the data as input and using the default parameter settings for supervised training. The predictive performance of the model was then tested using the remaining 20% of the sequences.

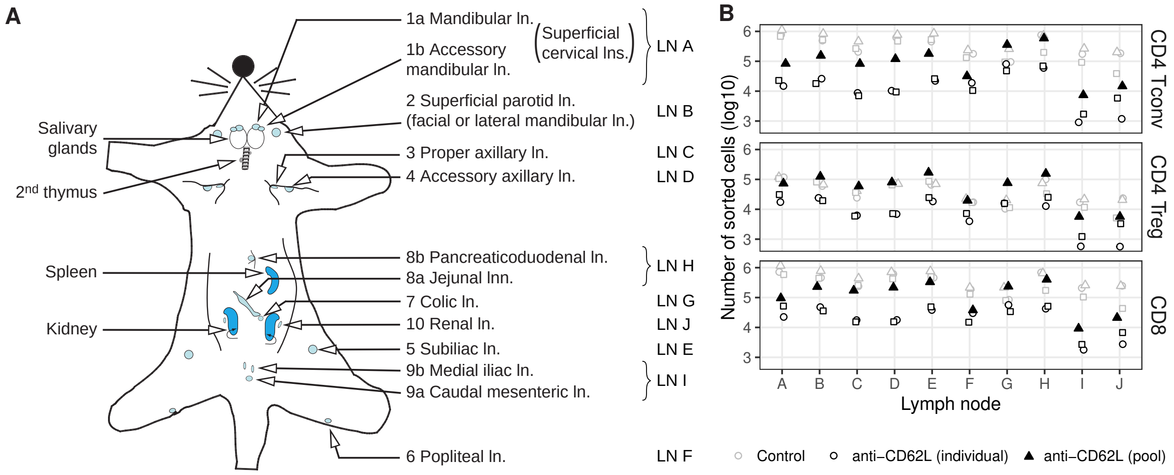
Author contributions

J.K. conceived the study. The experimental procedures were carried out by S.N., C.B. and J.K.. P.C.d.G. performed the data analysis with input from J.K and R.J.d.B.. P.C.d.G. and J.K. wrote the manuscript, which was edited and approved by all authors.

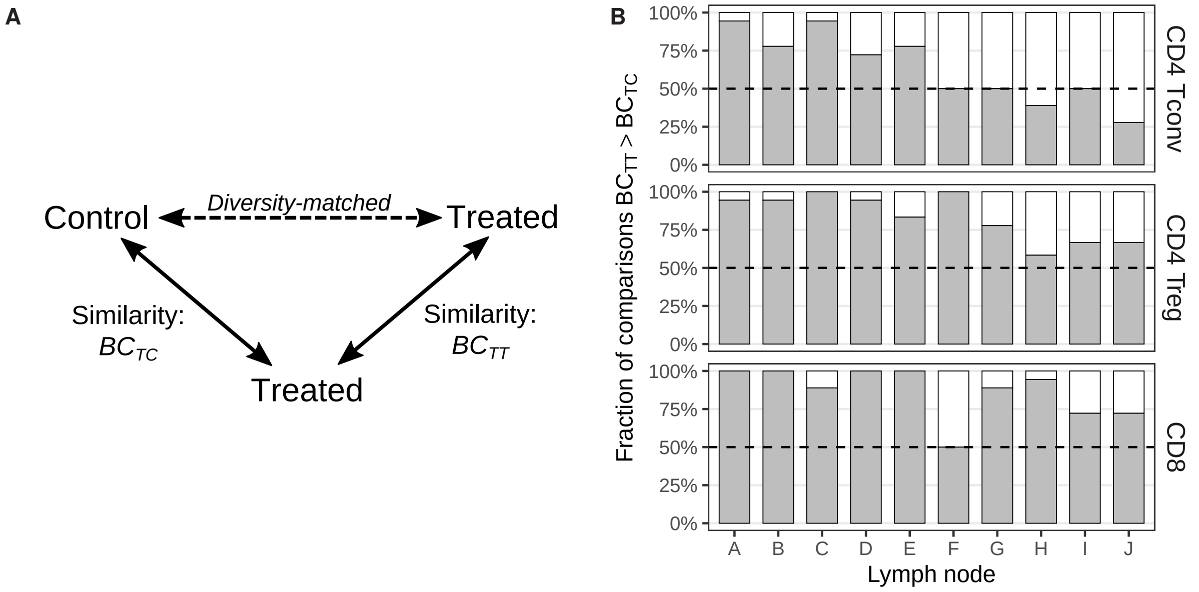
Acknowledgments

We thank Irene ter Horst and Ramona Mirke for their help with mice and cell sorting. This study was supported by The Netherlands Organization for Scientific Research Graduate Program 022.005.023 (to P.C.d.G.) and by the German Science Foundation (DFG) supporting the Forschergruppe CONTROL-T (FOR 1961) with grants KI611/2-1 and KI611/2-2 (to J.K.).

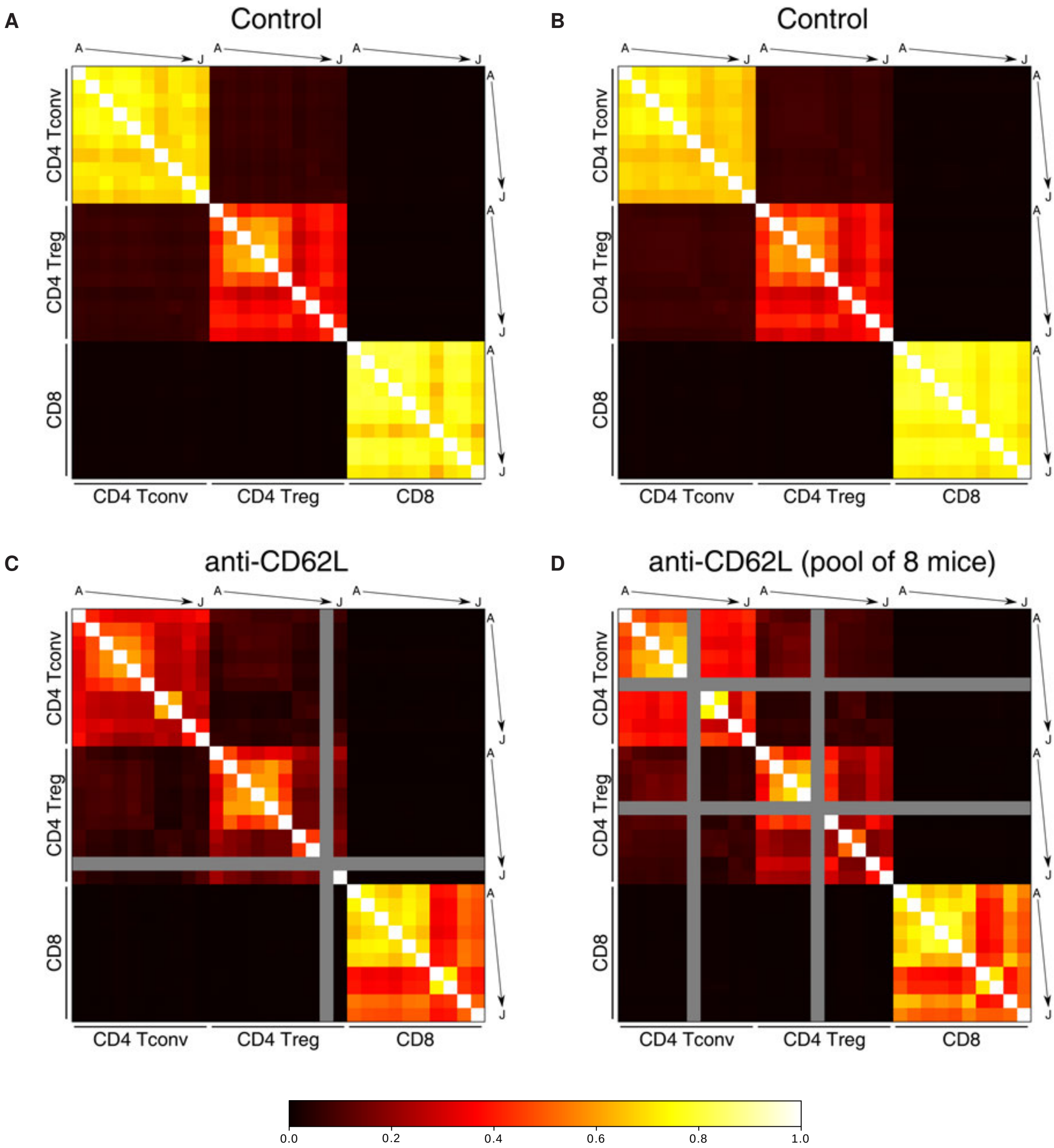
Supplemental Figures



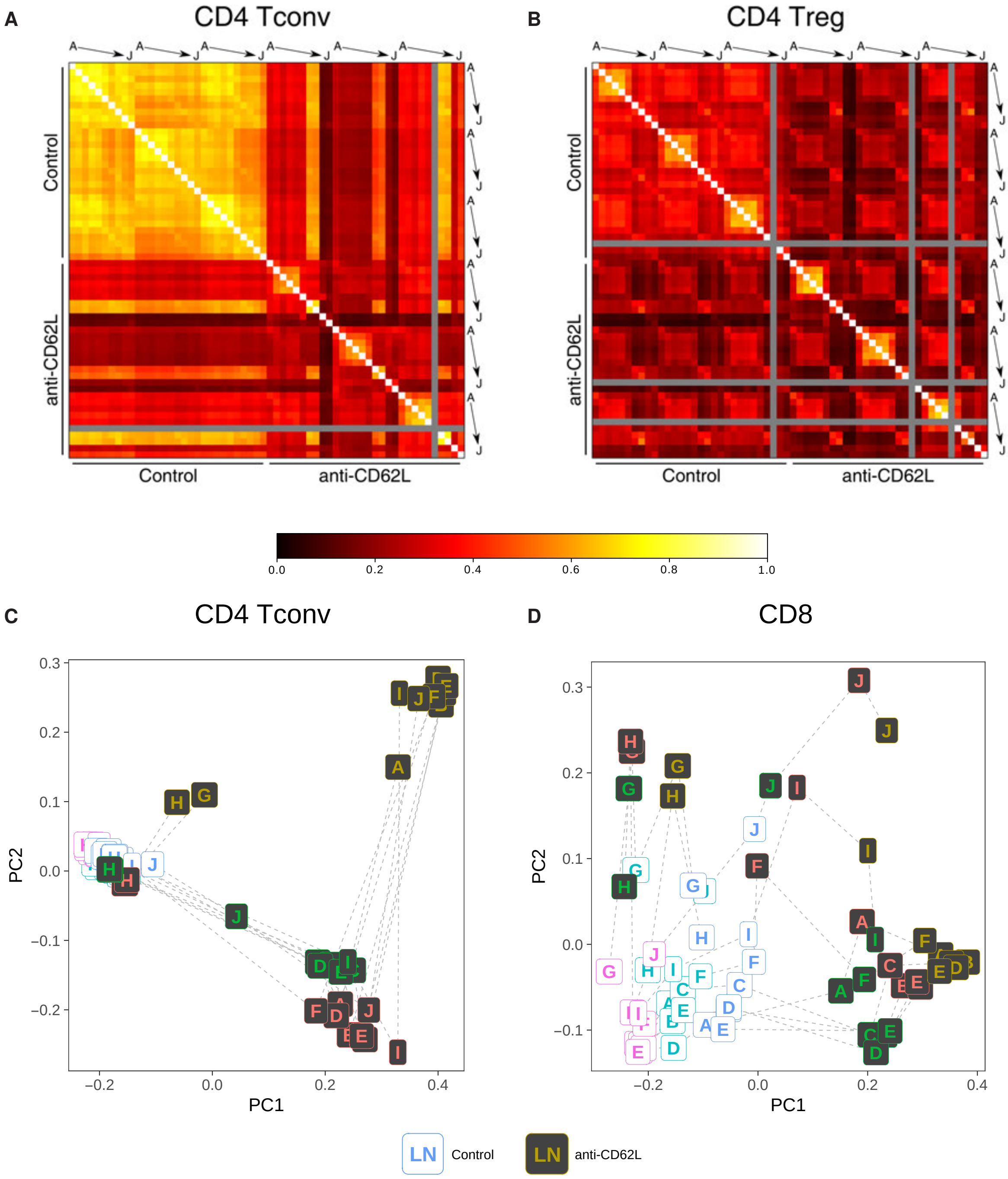
Supplementary Figure S6.1 – **Cell sampling from mice. A.** Schematic representation of mouse anatomy and sampled lymph nodes. The letters indicate how the lymph nodes are referred to throughout the text and figures. **B.** Number of cells that were sorted for each subset per lymph node. The filled black triangles represent the samples that were composed of lymph nodes pooled from 8 individual mice that were treated with anti-CD62L.



Supplementary Figure S6.2 – **Inter-mouse similarity is not just the result of sequencing fewer cells. A.** Schematic representation of sample comparison (see Methods for details). **B.** Fraction of comparisons, n=18 if all samples were available, resulting in a Bray-Curtis similarity that is higher between treated samples than between control and treated samples. The horizontal dashed line indicates 0.5 which would be consistent with a TCR-independent effect of anti-CD62L treatment.



Supplementary Figure S6.3 – Supplemental to Figure 6.3A&B. Heatmaps showing Bray-Curtis similarity between TCR repertoires of different cell subsets from individual lymph nodes in control mice (**A** and **B**) as well as an anti-CD62L treated mouse (**C**) and a pool of 8 anti-CD62L treated mice (**D**). Grey values indicate excluded samples (see Methods).



Supplementary Figure S6.4 – **Supplemental to Figure 6.3C&D. A&B.** Heatmaps showing Bray-Curtis similarity between TCR repertoires of CD4 Tconv (**A**) and Treg (**B**) cells from individual lymph nodes of control and anti-CD62L-treated mice. Grey values indicate excluded samples (see Methods). **C&D.** Visual representation of the repertoire structure by Principal Coordinate Analysis (PCoA) on CD4 Tconv (**C**) and (**D**) CD8 samples. Lymph nodes are depicted as letters, individual mice are shown as colours, with the background distinguishing between control (white) and anti-CD62L-treated (black) mice. Identical lymph nodes across mice are connected by dashed lines.

7

General discussion

The advances in high-throughput sequencing techniques, in particular during the past decade, enabled the collection of enormous amounts of TCR repertoire sequencing data. Experimental procedures, like the use of UMIs, and bioinformatic solutions in the form of dedicated software pipelines address some of the challenges with reliable processing of this data. Both are necessary to obtain a quantitative and reproducible characterisation of the TCR repertoire. In this thesis, we went a step further by analysing the TCR repertoire of humans and mice along several axes, to address outstanding immunological questions. This general discussion chapter summarises the main findings, discusses the interpretation of the bigger picture and poses outstanding questions that require follow-up studies.

Multiple samples from the same TCR repertoire

Each sample of T cells that is used to characterise the TCR repertoire only covers a tiny subset of the total T-cell pool. The resulting TCR sequence frequencies in the data reflect the true abundance in the repertoire but are also shaped by variation between cells, for example as a result of differential expression levels and amplification efficiencies. A common approach in this thesis is that we compared TCR repertoires derived from multiple samples. Doing so, we studied the TCR repertoire in several dimensions: the differences in abundance between TCR sequences, the TCR dynamics after antigen exposure and during healthy ageing, and the spatial organisation of the TCR repertoire in an individual. The integration of the many samples we analysed provided the opportunity to address immunological questions from TCR sequencing data.

Multiple samples to reliably classify TCR sequences by abundance

In **Chapter 2 – 4** we integrated the information obtained from multiple subsamples to discriminate between abundant and rare TCR sequences in the repertoire. We used the inter-sample incidence rather than the intra-sample abundance to classify TCR sequences by their frequency in the repertoire. In fact, single-cell sequencing techniques take this approach to the extreme, by capturing only a single cell in each subsample. Our approach combines the advantages of bulk and single-cell TCR sequencing: it allows for affordable characterisation of large numbers of cells, while reducing the impact of variable mRNA contributions by individual cells. This method enabled us to show in **Chapter 2** that T-cell clones in the naive repertoire are extremely heterogeneous with respect to their abundance, which is only partly explained by their generation probability. In **Chapter 3** we identified that, in addition to previously identified features, absence of a D-segment is a characteristic of many abundant TCRβ sequences in the naive repertoire of young individuals. The short **Chapter 4** discussed the analysis of changes in the TCRβ repertoire during healthy ageing, for which having multiple subsamples appears crucial to extrapolate findings from small samples to the size of the entire repertoire.

In each of these studies, we processed the data of the subsamples individually before joining the data to determine the incidence of each TCR sequence. This has practical reasons, as the output of the TCR sequence data processing pipelines we used just provides a count for each identified TCR sequence. This information does not allow for determining the incidence of a TCR sequence, because the subsample from which each individual TCR transcript originated remains unknown. The error correction, merging similar and/or low-quality TCR sequences, was thus performed on the level of individual subsamples. With respect to the identification of abundant TCR sequences this can be considered a conservative approach. If error correction would have been performed by clustering TCR sequences over multiple subsamples, a high incidence could result from incorrectly clustering similar TCR sequences found in separate subsamples. In our approach, TCR sequences can only have a high incidence by independent identification in multiple subsamples. At the same time, we may have missed TCR sequences overlapping between subsamples, for example due to sequencing errors in the CDR3 region and a failure to correct these by clustering. Such cases could be accounted for by future TCR processing pipelines that can process multiple samples from the same repertoire together. This approach would allow for error-correction of low-quality sequences in a sample, based on a similar high-confidence sequence in another subsample. Corrections like this should be performed with care, to prevent the overestimation of abundant sequences by excessive clustering. Another future pipeline feature that would benefit the analysis is the inclusion of the UMI sequences and sample indices in the output. Such information allows one to verify that TCR sequences with high incidence are not due to errors during demultiplexing, i.e., the separation of reads from a sequencing run based on sample-specific indices. The combination of TCR and UMI sequences in the output would also allow one to perform error-correction on the UMIs, which improves the estimates of TCR sequence frequencies in a single sample.

Multiple samples to distinguish T-cell dynamics from technical variation

In **Chapter 5** we aimed to characterise the T-cell response after pneumococcal vaccination using TCR sequencing. Without prior information about vaccine-specific TCR sequences, this requires quantification of the frequency differences between samples from different time points, which we compared against the differences between replicates, i.e., samples from the same time point. By having multiple samples for individual time points, we could quantify how reproducibly TCR sequence frequencies are estimated between samples from the exact same repertoire. The observed differences in frequency were for many sequences as large between samples from the same time point as between different time points, illustrating the need for a robust classification to obtain solid evidence for changing TCR frequencies with time. In addition to requiring a minimum relative fold-change, we therefore only classified a TCR sequence as expanded if the absolute number of UMIs supporting the sequence exceeded the pre-vaccination frequency by another threshold. Using this combination of two thresholds, only few sequences were classified as expanded in most individuals. Importantly, the size of the different samples seems to dominate the estimated number of expanded TCR sequences, complicating quantitative comparison between donors. We thus showed the complications in using TCR sequencing to identify a response upon vaccination and identified the sample size as a limiting parameter to obtain a biological signal that can be distinguished from technical variation.

The challenge of discriminating between technical and biological effects is not unique to following TCR frequencies longitudinally. The quantification of differences in gene expression of cells between different experimental conditions poses in fact a similar question. Dedicated software packages, such as EdgeR (Robinson *et al.*, 2010), are used in such cases to model both the biological and technical variability. At the same time, the number of distinct TCR sequences in humans (in the order of 10^8 (Qi *et al.*, 2014)) largely exceeds the number of protein-coding genes on the human genome (about 20 000 (Willyard, 2018)). As a result, technical variation potentially has an even larger effect on most of the observed TCR frequencies. It could therefore be that only the dynamics of the most abundant TCR sequences can be reliably quantified. A tool that is specifically developed to fulfil this task is NoisET (Koraichi *et al.*, 2022; Touzel *et al.*, 2020), which can be used to estimate the noise from multiple replicates and then classify sequences based on changes that exceed the estimated noise levels. Both EdgeR and NoisET use statistical models to account for the noise. A more mechanistic understanding of the sources of technical variation will require dedicated benchmarking experiments. The resulting insights should be used to optimise the experimental TCR sequencing pipeline such that the noise is reduced and the remaining technical variation is accurately estimated. Only in such a way reliable conclusions can be drawn from comparisons between TCR repertoires, for example from different time points after infection or vaccination.

Multiple samples to uncover a multi-level organisation of the TCR repertoire

In **Chapter 6** we studied the TCR repertoire of one-TCRα mice that have a strongly restricted TCR diversity. By comparing samples from different T-cell subsets against each other, such as CD4 versus CD8 T cells, we obtained evidence for a lineage choice that is deterministic given the TCR sequence. Moreover, by sampling the TCR repertoire from individual lymph nodes across genetically identical mice, we revealed a reproducible, TCR-driven preference of some T cells to a limited set of lymph nodes. These results indicate that at least not all T-cell clonotypes are randomly distributed throughout the body, but are organised based on their TCR specificity. So, the comparison between multiple samples from diverse anatomical locations in genetically identical mice revealed a non-random organisation of the TCR repertoire. The expected biological mechanism behind the reproducible CD4/CD8 lineage choice and the lymph node preference is the affinity of certain TCR specificities for self-peptide/MHC complexes. Although we did not study these interactions in more detail, our analysis suggests that even the naive TCR repertoire has a spatial organisation. The TCR repertoire as measured in the blood does not allow for addressing research questions about the mechanisms behind this spatial distribution, as well as its function. An example of such a question would be what changes the TCR repertoire distribution undergoes after presentation of one specific peptide/MHC complex in a constrained anatomical location.

The analysis of spatial differences in the TCR repertoire is much more complicated to perform in humans. This is not only because of the very sparse availability of human tissue and lymph node samples, but also due to the wide diversity of TCR specificities and (self-)

peptides, as well as the MHC polymorphism in the human population. These complicating factors play a much smaller role when studying one-TCRα mice since they have a reduced TCR diversity, a limited impact of foreign peptides when kept using gnotobiotic technique, and no MHC diversity between individuals. Hence, the animal model allowed us to address biological questions that are not straightforward to study in the more complex human context.

Another method to reduce the complexity of the immune system is using computational models. Examples range from statistical models, which infer the underlying probabilities during TCR generation and selection, to mechanistic models of repertoire maintenance (**Chapter 2**). In the latter case, we hypothesised that the dynamics of naive T-cell clonotypes are neutral, which we tested against NGS data using a simple model. This comparison was only possible by reducing the actual complexity of the T-cell biology. One example of a simplification in this study is that the *in silico* repertoires of CD4 and CD8 T cells were created from the same set of TCR sequences. In **Chapter 6**, however, we showed that the TCR repertoire of CD4 and CD8 T cells is clearly distinct, at least in the one-TCRα mice. This means that the simplification of generating *in silico* repertoires from the same set of TCR sequences may not be completely accurate, but it still allowed us to reject neutral dynamics as a mechanism for the maintenance of the naive TCR repertoire. Thus, addressing fundamental questions about the biology of adaptive immune repertoires often requires simplification of the system. The art is to find a balance that aims to reduce the complexity of the system experimentally and/or computationally to an appropriate level.

From the TCR as a barcode towards a functional interpretation of the repertoire

The analyses described in this thesis generally focus on the repertoire of TCR sequences, without studying them functionally, for example in terms of the antigen specificities. In other words, we interpreted the TCR sequence often like a ‘neutral barcode’. This barcode is shared between the T cells that belong to the same clonotype, that have the same antigen specificity. Even without knowing the specificity, this allows quantitative insight into the dynamics of the clonotypes. Importantly, the sequence of a single TCR chain is also shared between clonotypes that differ in the other TCR chain. This means that the sequence of a single TCR chain does not uniquely label a single antigen specificity. From this perspective, having only single-chain data is not optimal, as during the analysis different TCR specificities are joined as one entity whenever they share one of their TCR chains. This challenge does not occur in single-cell data or when one of the TCR chains is fixed, like in one-TCRα mice. Still, even in individuals with a fully diverse TCR repertoire, following single chains allows one to measure the cumulative frequency of groups of T cells sharing one of their TCR chains.

The TCR sequence as a 'neutral barcode' to study clonal dynamics

Longitudinal samples of the TCR repertoire allow for a quantitative analysis of T-cell dynamics at different time scales. An example is the quantification of the response after infection or vaccination, like we performed on the TCRβ repertoire in **Chapter 5**. In such an analysis, an expanding TCRβ sequence may actually represent multiple clonotypes. This data therefore cannot provide a complete picture of the breadth and depth of the T-cell response. Although this is a limitation, one can still compare the number and total frequency of the expanding TCRβ sequences to approximate the level of the response. In addition, the durability of the response can be estimated by studying the frequencies of these sequences in follow-up samples. Thus, although it does not cover the full diversity of clonotypes in a sample, single-chain TCR sequencing does provide valuable insights into T-cell responses.

Another question that can be addressed using longitudinal TCR sequencing relates to the T-cell dynamics on a longer time scale, which are expected to be less dominated by the effect of individual stimuli. Bensouda Koraichi *et al.* used Bayesian inference to estimate the parameters of clonal dynamics in the T-cell pool based on longitudinal TCR sequencing data spanning up to three years (Bensouda Koraichi *et al.*, 2023). Interestingly, they found a strong negative correlation between the turnover rate and the age of an individual, suggesting that repertoires of older people have slower dynamics than those of young individuals. Longitudinal TCR repertoire studies spanning longer time spans are very rare, understandably because the current sequencing techniques have been developed quite recently. An exception is the study by Yoshida *et al.* which involved TCRβ repertoire sequencing of samples that were collected approximately 20 years apart and then cryopreserved (Yoshida *et al.*, 2017). They describe an overall decrease in CD8 T cell diversity with age, while the diversity of the CD4 T cell repertoire retained a fairly high diversity. Note that the interpretation of TCR repertoire diversity changes with age is challenging and does not always yield robust results (**Chapter 4**). They also found considerable TCRβ overlap between samples taken about two decades apart (Yoshida *et al.*, 2017), indicating that many T cell clonotypes are maintained at a similar frequency for long periods of time. Biologically, these findings show that the T-cell pool can be very dynamic in response to specific antigen, but also quite stable over a long time span. Technically, it means that much can be learned about the dynamics of T-cell clonotypes, even if the data is far from complete. TCR sequences of single chains from small samples are very informative about the dynamics of clonotypes on a pool level.

Heterogeneity between and within T-cell subsets

When studying T-cell biology by analysing TCR sequences, it may be an oversimplification to study the TCR repertoire in broadly defined subsets of circulating T cells. In fact, there is no such thing as the TCR repertoire. The T cells in a sample that harbour the observed TCRs actually have different phenotypes and fulfil many different functions. Single-cell studies show that T cells have an enormous diversity in gene expression profiles, even within the well-established T-cell subsets (Guo *et al.*, 2018; Schattgen *et al.*, 2021; Zemmour *et al.*, 2018;

Zheng *et al.*, 2017). These differences likely relate to T-cell function, which implies that each TCR repertoire analysis actually studies a collection of TCR repertoires. Each of these may have different dynamics and inferring the dynamics from TCR repertoire data only provides an average of the actual production and turnover rates. The differences between the actual rates are reflected in the TCR frequency distribution, which can be very broad (Desponds *et al.*, 2016; Gaimann *et al.*, 2020; Qi *et al.*, 2014). For the total T-cell pool this is natural as a memory T-cell clone specific for a persistent virus is expected to occur at a much higher frequency than an unactivated T-cell clone that has been recently produced by the thymus. Moreover, even in repertoires of naive T cells we found a large heterogeneity with respect to their clone size (**Chapter 2**). This implies that survival and division rates are heterogeneous, even within the naive T-cell pool. Note that these observations were based on samples from the blood and even larger heterogeneity may exist between different tissues (**Chapter 6**). So, while current studies already indicate large differences between T-cell clones, even more variability is to be expected when zooming in into specific T-cell subsets at particular anatomical locations.

An interesting example of a relevant difference between T-cell clones is whether their TCR was recombined in the presence or absence of TdT. This enzyme is necessary for the non-templated insertions in the TCR sequence that contribute to the diversity of the TCR repertoire. In both humans and mice, TdT expression is suppressed during early repertoire development before birth (Benedict *et al.*, 2000; Gregoire *et al.*, 1979). This means that the repertoire starts with a population of T-cell clones with limited diversity. Direct analysis of this specific set of sequences is not straightforward, as sequences without detectable N-additions can also be generated in the presence of TdT (Marcou *et al.*, 2018; Murugan *et al.*, 2012). Still, these sequences are found to be abundant, excessively shared between identical twins, and persistent over multiple decades (Pogorelyy *et al.*, 2017). This implies that the early presence of such clones may have allowed them to increase in abundance before experiencing much clonal competition, simply because they were generated first (Gaimann *et al.*, 2020). In addition, it is tempting to speculate that TCRs without N-additions play a role explaining their early production. For example, they may have an increased probability to survive thymic selection, to boost a quick generation of an early T-cell repertoire. Alternatively, they could have a broad antigen specificity, allowing them to provide protection, even with a relatively small TCR repertoire. Such a hypothesis can be explored by studying biochemical properties of sets of TCR sequences with and without N-additions. An example of a potential mechanism is the stickiness of certain amino acids in the CDR3, which was reported to be characteristic for the TCR repertoire of regulatory T cells (Lagattuta *et al.*, 2022) and other CD4 T-cell subsets (Kasatskaya *et al.*, 2020). These features may help to also interpret the early production of TdT-independent TCR sequences functionally and thus explain this level of heterogeneity in the TCR repertoire.

Classification of sequences in the TCR repertoire

One of the largest challenges when studying the adaptive immune system is a classification task: to tell which TCR sequences will and will not bind a given a peptide/MHC complex. The opposite question may be even more complicated to address: given a TCR sequence, which of the many possible antigens will it bind to? Central to these challenges is the immense diversity of both TCRs and peptide/MHC complexes, and the complications with studying their interactions at a large scale. Antigen-specific T cells can be sorted using tetramers or dextramers, which are complexes of MHC molecules that are associated with a specific peptide and bound to a fluorochrome. By performing TCR sequencing on these sorted populations, one obtains TCR sequences that bind one given peptide/MHC combination. The resulting set of TCR sequences is sometimes characterised by a motif in the TCR sequence, such as an over-represented V/J combination or a specific stretch of amino acids in the CDR3 sequence (Cukalac *et al.*, 2015; Dash *et al.*, 2017; Pogorelyy *et al.*, 2022; Sant *et al.*, 2018). Such features help generalise predictions of antigen binding, based on the presence of these motifs in other TCR sequences. Multiple databases, including VDJdb (Shugay *et al.*, 2017), McPAS-TCR (Tickotsky *et al.*, 2017) and IEDB (Vita *et al.*, 2019), collect sets of TCR sequences that have been found to bind to specific peptide/MHC complexes. These are interesting resources to address the challenge of annotating TCR sequences with their antigen specificity.

Even though these databases exist, there are several reasons that this challenge is far from solved. First, the data is extremely sparse, since currently only a tiny fraction of the potential peptide/MHC complexes has been studied using tetramers. Second, because most of the TCR sequences found to bind to specific antigens have been sequenced in bulk, which does not provide information on the pairing between TCRα and TCRβ chains. Since the antigen specificity of a TCR is defined by the combination of both chains, the information on antigen-specific TCRs is often incomplete. Third, since studies vary in their approach to assess the antigen-specific TCR repertoire, for example using tetramers or dextramers, the confidence on TCR-peptide/MHC interactions is variable. Even though these challenges make it difficult to generalise predictions for TCR specificity, various machine learning approaches have been developed to fulfil this task. The predictive success of tools like TCRex (Gielis *et al.*, 2019), DeepTCR (Sidhom *et al.*, 2021), and pMTnet (Lu *et al.*, 2021) indicates that such approaches support the prediction of TCR specificity. With more data becoming available, especially at the single-cell level, this seems a promising development. It remains to be determined if the TCR amino acid sequences themselves provide enough information. Alternatively, dedicated structural predictions with tools similar to AlphaFold (Jumper *et al.*, 2021) may be necessary for accurate annotation with antigen specificity. It should be noted that reported binding of a TCR to a peptide/MHC complex is not a guarantee for T-cell activation (Vazquez-Lombardi *et al.*, 2022), imposing another level of complexity for functional predictions of T-cell reactivity.

Towards a functional interpretation of the TCR repertoire

The insights and developments discussed so far mostly relate to fundamental questions about the biology of T cells. The importance of T cells in health and disease also implies that this knowledge has practical applications. Obvious examples include novel treatments against cancer, such as boosting the action of tumour infiltrating lymphocytes, or immunotherapy with chimeric antigen receptors (CARs) that are expressed by engineered T cells. An application of T-cell repertoire data would be to assess the T-cell immunity of an individual qualitatively or quantitatively. A proof of principle of this potential is the possibility to infer past exposure to an antigen based on the TCR repertoire. This approach relies on TCR sequencing of a cohort, of which for each individual is known whether or not they were exposed to a virus, for example. By training machine learning models on the repertoire data it appears possible to infer cytomegalovirus (CMV) seropositivity (Emerson *et al.*, 2017) and presence or absence of a T-cell response against SARS-CoV-2 (Dalai and Baldo, 2021; Gittelman *et al.*, 2022) based on the TCRβ repertoire. Such approaches do not require a mechanistic understanding of which TCR sequences are part of the T-cell response. Moreover, this was possible even though large MHC polymorphism exists in the human population, which means that TCR sequences that are specific to a virus in one individual are not guaranteed to be specific for the same virus in another individual. These results suggest that inferring antigen exposure based on a TCR repertoire seems feasible for at least some pathogens, provided that there is enough TCR sequencing data to train a classification model.

The TCR repertoire is shaped by pathogen exposure, and defines to which extent T cells can protect against future challenges. This contains a wealth of information that cannot only be used to uniquely identify an individual (Dupic *et al.*, 2021), but also to make quantitative predictions about the current state of T-cell immunity in an individual. It is therefore crucial to learn which parameters of the TCR repertoire are most relevant to infer the level of protection. An example is that decreasing TCR diversity with age has been proposed as one of the mechanisms behind immune senescence (Amoriello *et al.*, 2021; Xu *et al.*, 2020; Yager *et al.*, 2008). It remains to be determined which measure of TCR diversity is most informative about such an effect. A substantial decrease in repertoire richness may result in 'holes in the repertoire', but absence of a response could as well be the result of functionally exhausted T cells. Technically, it is notoriously difficult to reliably estimate the actual richness of the TCR repertoire from a small sample (**Chapter 4**). Functionally, the actual number of distinct clonotypes may not be very informative, as a high number of different but similar TCR sequences does not guarantee coverage of the entire antigenic space. Similar questions arise when assessing the response upon vaccination with TCR sequencing. It will require large cohort studies to determine which aspects of the T-cell response are most informative to estimate the level of protection. If this information cannot be inferred from TCR repertoire data alone, a more complete picture can be achieved by a multi-omics approach using single-cell sequencing, combining TCR repertoire analysis with a functional characterisation of T-cell subsets (Schattgen *et al.*, 2021).

Outlook

The huge diversity of the TCR repertoire and the constraints regarding the number of cells that can be analysed in an experiment pose a fundamental challenge for TCR repertoire studies. In addition, like in any experiment, inevitable errors that are introduced during the experimental procedures make the analysis of the resulting data difficult. Instead of striving for excluding these complications completely, our ambition should be to account for them. Dedicated control experiments, replicate sampling, and quantitative modelling appear promising methods to obtain immunological insights, given the experimental limitations.

With the advances in high-throughput and single-cell sequencing, it is more feasible and affordable than ever to acquire large amounts of TCR sequencing data. Now this progress is continuing, challenges arise at new levels. It will require a large effort to organise TCR repertoire data in such a way that it becomes useful to train machine learning models to annotate repertoires functionally. Such interpretation is promising, as the TCR repertoire is even more personal than the genome, because it provides information on exposure history and future protection to pathogens. The resulting immunological insights are expected to support the further development of personalised medicine, which will benefit both science and society.

References

Alam, S. M. and Gascoigne, N. R. J. (1998). Posttranslational regulation of TCR Vα allelic exclusion during T cell differentiation. *The Journal of Immunology*, 160:3883–3890.

Alles, M., Turchinovich, G., Zhang, P., Schuh, W., Agenès, F., and Kirberg, J. (2014). Leukocyte β 7 integrin targeted by Krüppel-like factors. *The Journal of Immunology*, 193:1737–1746.

Amoriello, R., Mariottini, A., and Ballerini, C. (2021). Immunosenescence and autoimmunity: Exploiting the T-cell receptor repertoire to investigate the impact of aging on multiple sclerosis. *Frontiers in Immunology*, 12.

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human αβ T cell receptor diversity. *Science*, 286:958–961.

Baarle, D. V., Bollaerts, K., Giudice, G. D., Lockhart, S., Luxemburger, C., Postma, M. J., Timen, A., and Standaert, B. (2020). Preventing infectious diseases for healthy ageing: The VITAL public-private partnership project. *Vaccine*, 38:5896–5904.

Bassing, C. H., Alt, F. W., Hughes, M. M., D'Auteuil, M., Wehrly, T. D., Woodman, B. B., Gärtner, F., White, J. M., Davidson, L., and Sleckman, B. P. (2000). Recombination signal sequences restrict chromosomal V (D) J recombination beyond the 12/23 rule. *Nature*, 405(6786):583–586.

Benedict, C. L., Gilfillan, S., Thai, T.-H., and Kearney, J. F. (2000). Terminal deoxynucleotidyl transferase and repertoire development. *Immunological Reviews*, 175:150–157.

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.

Bensouda Koraichi, M., Ferri, S., Walczak, A. M., and Mora, T. (2023). Inferring the T cell repertoire dynamics of healthy individuals. *Proceedings of the National Academy of Sciences*, 120(4):e2207516120.

Bentley, D. R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59.

Bergot, A. S., Chaara, W., Ruggiero, E., Mariotti-Ferrandiz, E., Dulauroy, S., Schmidt, M., von Kalle, C., Six, A., and Klatzmann, D. (2015). TCR sequences and tissue distribution discriminate the subsets of naive and activated/memory Treg cells in mice. *European Journal of Immunology*, 45:1524–1534.

de Boer, R. J. and Perelson, A. S. (1994). T cell repertoires and competitive exclusion. *Journal of Theoretical Biology*, 169:375–390.

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 2015 12:5*, 12:380–381.

Borghans, J. A., Noest, A. J., and Boer, R. J. D. (1999). How specific should immunological memory be? *Journal of Immunology*, 163:569–75.

den Braber, I., *et al.* (2012). Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36:288–297.

Brady, B. L., Steinel, N. C., and Bassing, C. H. (2010). Antigen receptor allelic exclusion: an update and reappraisal. *Journal of immunology (Baltimore, Md. : 1950)*, 185:3801–3808.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, 27(4):326–349.

Britanova, O. V., *et al.* (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *The Journal of Immunology*, 192:2689–2698.

Britanova, O. V., *et al.* (2016). Dynamics of individual T cell repertoires: From cord blood to centenarians. *The Journal of Immunology*, 196:5005–5013.

Brocker, T. (1997). Survival of mature CD4 T lymphocytes is dependent on Major Histocompatibility Complex class II–expressing dendritic cells. *Journal of Experimental Medicine*, 186:1223–1232.

Camaglia, F., Ryvkin, A., Greenstein, E., Reich-Zeliger, S., Chain, B., Mora, T., Walczak, A. M., and Friedman, N. (2023). Quantifying changes in the T cell receptor repertoire during thymic development. *eLife*, 12:e81622.

Carey, A. J., Hope, J. L., Mueller, Y. M., Fike, A. J., Kumova, O. K., van Zessen, D. B., Steegers, E. A., van der Burg, M., and Katsikis, P. D. (2017). Public clonotypes and convergent recombination characterize the naive CD8+ T-cell receptor repertoire of extremely preterm neonates. *Frontiers in Immunology*, 8:1–13.

Carter, J. A., Preall, J. B., Grigaityte, K., Goldfless, S. J., Jeffery, E., Briggs, A. W., Vigneault, F., and Atwal, G. S. (2019). Single T cell sequencing demonstrates the functional role of $\alpha\beta$ TCR pairing in cell lineage and antigen specificity. *Frontiers in Immunology*, 10.

Casanova, J. L., Romero, P., Widmann, C., Kourilsky, P., and Maryanski, J. L. (1991). T cell receptor genes in a series of class I Major Histocompatibility Complex-restricted cytotoxic T lymphocyte clones specific for a Plasmodium berghei nonapeptide: implications for T cell allelic exclusion and antigen-specific repertoire. *Journal of Experimental Medicine*, 174:1371–1383.

Castillo-González, R., Cibrian, D., and Sánchez-Madrid, F. (2021). Dissecting the complexity of $\gamma\delta$ T-cell subsets in skin homeostasis, inflammation, and malignancy. *Journal of Allergy and Clinical Immunology*, 147:2030–2042.

Chao, A., Colwell, R. K., Lin, C. W., and Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, 90:1125–1133.

Chao, A., *et al.* (2020). Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*, 35:292–314.

Chidrawar, S., Khan, N., Wei, W., McLarnon, A., Smith, N., Nayak, L., and Moss, P. (2009). Cytomegalovirus-seropositivity has a profound influence on the magnitude of major lymphoid subsets within healthy individuals. *Clinical and Experimental Immunology*, 155:423–432.

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.

Cose, S., Brammer, C., Khanna, K. M., Masopust, D., and Lefrançois, L. (2006). Evidence that a significant number of naive T cells enter non-lymphoid organs as part of a normal migratory pathway. *European Journal of Immunology*, 36:1423–1433.

Cukalac, T., Kan, W. T., Dash, P., Guan, J., Quinn, K. M., Gras, S., Thomas, P. G., and Gruta, N. L. L. (2015). Paired TCR$\alpha\beta$ analysis of virus-specific CD8+ T cells exposes diversity in a previously defined ‘narrow’ repertoire. *Immunology & Cell Biology*, 93:804.

Dalai, S. C. and Baldo, L. (2021). Letter to the editor regarding ‘perspective: diagnostic laboratories should urgently develop T cell assays for SARS-CoV-2 infection’. *Expert Review of Clinical Immunology*, 17(11):1155–1157.

Dash, P., *et al.* (2017). Quantifiable predictive features define epitope specific T cell receptor repertoires. *Nature*, 547:89.

Davis, M. M. and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395.

Dellabona, P., Padovan, E., Casorati, G., Brockhaus, M., and Lanzavecchia, A. (1994). An invariant V alpha 24-J alpha Q/V beta 11 T cell receptor is expressed in all individuals by clonally expanded CD4-8-T cells. *Journal of Experimental Medicine*, 180:1171–1176.

Desponds, J., Mayer, A., Mora, T., and Walczak, A. M. (2021). Population dynamics of immune repertoires. *Mathematical, computational and experimental T cell immunology*, pages 203–221.

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*, 113:274–279.

DeWitt, W. S., *et al.* (2015). Dynamics of the cytotoxic T cell response to a model of acute viral infection. *Journal of Virology*, 89:4517–4526.

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

Dupic, T., Koraichi, M. B., Minervina, A. A., Pogorelyy, M. V., Mora, T., and Walczak, A. M. (2021). Immune fingerprinting through repertoire similarity. *PLoS Genetics*, 17.

Dupic, T., Marcou, Q., Walczak, A. M., and Mora, T. (2019). Genesis of the αβ T-cell receptor. *PLoS Computational Biology*, 15:e1006874.

Dykema, A. G., *et al.* (2022). SARS-CoV-2 vaccination diversifies the CD4+ spike-reactive T cell repertoire in patients with prior SARS-CoV-2 infection. *eBioMedicine*, 80.

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*, 111:9875–9880.

Elhanati, Y., Sethna, Z., Jr, C. G. C., Mora, T., and Walczak, A. M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, 284:167–179.

Emerson, R. O., *et al.* (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49:659.

Ferrarini, M., Molina-París, C., and Lythe, G. (2017). Sampling from T cell receptor repertoires. *Modeling Cellular Systems*, pages 67–79.

Fink, K. (2019). Can we improve vaccine efficacy by targeting T and B cell repertoire convergence? *Frontiers in Immunology*, 10:110.

Gaimann, M. U., Nguyen, M., Desponds, J., and Mayer, A. (2020). Early life imprints the hierarchy of T cell clone sizes. *eLife*, 9:1–36.

Gallatin, W. M., Weissman, I. L., and Butcher, E. C. (1983). A cell-surface molecule involved in organ-specific homing of lymphocytes. *Nature*, 304:30–34.

Ganusov, V. V. and Boer, R. J. D. (2007). Do most lymphocytes in humans really reside in the gut? *Trends in Immunology*, 28:514–518.

Garcia, K. C. and Adams, E. J. (2005). How the T cell receptor sees antigen — a structural view. *Cell*, 122:333–336.

Gascoigne, N. R. and Alam, S. M. (1999). Allelic exclusion of the T cell receptor α-chain: developmental regulation of a post-translational event. *Seminars in Immunology*, 11:337–347.

Gattinoni, L., *et al.* (2011). A human memory T cell subset with stem cell-like properties. *Nature Medicine*, 17:1290.

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(June):btw339.

Gielis, S., Moris, P., Bittremieux, W., Neuter, N. D., Ogunjimi, B., Laukens, K., and Meysman, P. (2019). Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Frontiers in Immunology*, 10:2820.

Gittelman, R. M., *et al.* (2022). Longitudinal analysis of T-cell receptor repertoires reveals shared patterns of antigen-specific response to SARS-CoV-2 infection. *JCI insight*.

Glanville, J., *et al.* (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547:94.

Glenn, T. C., *et al.* (2019). Adapterama II: universal amplicon sequencing on Illumina platforms (TaggiMatrix). *PeerJ*, 7.

Gonçalves, P., Ferrarini, M., Molina-Paris, C., Lythe, G., Vasseur, F., Lim, A., Rocha, B., and Azogui, O. (2017). A new mechanism shapes the naive CD8+ T cell repertoire: the selection for full diversity. *Molecular Immunology*, 85:66–80.

de Greef, P. C. and de Boer, R. J. (2021). TCRβ rearrangements without a D segment are common, abundant, and public. *Proceedings of the National Academy of Sciences*, 118:e2104367118.

de Greef, P. C., Oakes, T., Gerritsen, B., Ismail, M., Heather, J. M., Hermsen, R., Chain, B., and de Boer, R. J. (2020). The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife*, 9:1–24.

Gregoire, K. E., Goldschneider, I., Barton, R. W., and Bollum, F. J. (1979). Ontogeny of terminal deoxynucleotidyl transferase-positive cells in lymphohemopoietic tissues of rat and mouse. *Journal of immunology (Baltimore, Md. : 1950)*, 123:1347–52.

Guo, X., *et al.* (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine*, 24:978–985.

Hapuarachchi, T., Lewis, J., and Callard, R. (2013). A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. *Frontiers in Immunology*, 4:366.

Hayday, A. C., Diamond, D. J., Tanigawa, G., Heilig, J. S., Folsom, V., Saito, H., and Tonegawa, S. (1985). Unusual organization and diversity of T-cell receptor a-chain genes. *Nature*, 316:828–832.

Heather, J. M., Ismail, M., Oakes, T., and Chain, B. (2018). High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*, 19:554–565.

Heine, W. O. P. (1998). *Environmental management in laboratory animal units : basic technology and hygiene : methods and practice (Umweltmanagement in der Labortierhaltung : technisch-hygienische Grundlagen : Methoden und Praxis)*. Pabst Science Publishers.

Hogan, T., Gossel, G., Yates, A. J., and Seddon, B. (2015). Temporal fate mapping reveals age-linked heterogeneity in naive T lymphocytes in mice. *Proceedings of the National Academy of Sciences*, 112:E6917–E6926.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press.

Inaba, T., Koseki, H., Suzuki, M., and Taniguchi, M. (1991). Double-step and inverse polymerase chain reaction for sensitive detection and cloning of T cell receptor variable region sequences. *International Immunology*, 3:1053–1057.

Izraelson, M., *et al.* (2018). Comparative analysis of murine T-cell receptor repertoires. *Immunology*, 153:133–144.

Jenkins, M. K., Chu, H. H., McLachlan, J. B., and Moon, J. J. (2009). On the composition of the preimmune repertoire of T cells specific for peptide-Major Histocompatibility Complex ligands. *Annual Review of Immunology*, 28:275–294.

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.

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*, 109:21432–21437.

Jumper, J., *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

Kasatskaya, S. A., *et al.* (2020). Functionally specialized human CD4+ T-cell subsets express physicochemically distinct TCRs. *eLife*, 9:1–22.

Katzelnick, L. C., Montoya, M., Gresh, L., Balmaseda, A., and Harris, E. (2016). Neutralizing antibody titers against dengue virus correlate with protection from symptomatic infection in a longitudinal cohort. *Proceedings of the National Academy of Sciences*, 113:728–733.

Khoury, D. S., Cromer, D., Reynaldi, A., Schlub, T. E., Wheatley, A. K., Juno, J. A., Subbarao, K., Kent, S. J., Triccas, J. A., and Davenport, M. P. (2021). Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine*, 27:1205–1211.

Kieback, E., *et al.* (2016). Thymus-derived regulatory T cells are positively selected on natural self-antigen through cognate interactions of high functional avidity. *Immunity*, 44:1114–1126.

Kirberg, J., Berns, A., and Boehmer, H. V. (1997). Peripheral T cell survival requires continual ligation of the T cell receptor to Major Histocompatibility Complex–encoded molecules. *Journal of Experimental Medicine*, 186:1269–1275.

Koraichi, M. B., Touzel, M. P., Mazzolini, A., Mora, T., and Walczak, A. M. (2022). NoisET: Noise learning and expansion detection of T-cell receptors. *Journal of Physical Chemistry A*, 126:7407–7414.

Krishna, C., Chowell, D., Gönen, M., Elhanati, Y., and Chan, T. A. (2020). Genetic and environmental determinants of human TCR repertoire diversity. *Immunity and Ageing*, 17:1–7.

Kumar, B. V., Connors, T. J., and Farber, D. L. (2018). Human T cell development, localization, and function throughout life. *Immunity*, 48:202–213.

Lagattuta, K. A., Kang, J. B., Nathan, A., Pauken, K. E., Jonsson, A. H., Rao, D. A., Sharpe, A. H., Ishigaki, K., and Raychaudhuri, S. (2022). Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate. *Nature Immunology*, 23:446–457.

Lefranc, M.-P. (2001). IMGT, the international ImMunoGeneTics database. *Nucleic acids research*, 29(1):207–209.

Lefranc, M. P., *et al.* (2009). IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 37.

Lepault, F., Gagnerault, M., Faveeuw, C., and Boitard, C. (1994). Recirculation, phenotype and functions of lymphocytes in mice treated with monoclonal antibody MEL-14. *European Journal of Immunology*, 24:3106–3112.

Lepore, M., *et al.* (2014). Parallel T-cell cloning and deep sequencing of human MAIT cells reveal stable oligoclonal TCRβ repertoire. *Nature Communications*, 5:3866.

Lu, T., *et al.* (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence*, 3:864–875.

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.

Lugli, E., *et al.* (2013b). Superior T memory stem cell persistence supports long-lived T cell memory. *The Journal of Clinical Investigation*, 123.

Lythe, G., Callard, R. E., Hoare, R. L., and Molina-Paris, C. (2016). How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224.

Ma, K. Y., He, C., Wendel, B. S., Williams, C. M., Xiao, J., Yang, H., and Jiang, N. (2018). Immune repertoire sequencing using molecular identifiers enables accurate clonality discovery and clone size quantification. *Frontiers in Immunology*, 9.

Ma, L., Yang, L., Shi, B., He, X., Peng, A., Li, Y., Zhang, T., Sun, S., Ma, R., and Yao, X. (2016). Analyzing the CDR3 repertoire with respect to TCR - beta chain V-D-J and V-J rearrangements in peripheral T cells using HTS. *Scientific Reports*, 6:1–10.

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.

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.

Marcou, Q., Mora, T., and Walczak, A. M. (2018). High-throughput immune repertoire analysis with IGoR. *Nature communications*, 9:561.

Mark, M., Reich-Zeliger, S., Greenstein, E., Reshef, D., Madi, A., Chain, B., and Friedman, N. (2022). A hierarchy of selection pressures determines the organization of the T cell receptor repertoire. *Frontiers in Immunology*, 13.

Marraco, S. A. F., *et al.* (2015). Long-lasting stem cell-like memory CD8+ T cells with a naive-like profile upon yellow fever vaccination. *Science Translational Medicine*, 7:282ra48–282ra48.

Mason, D. (1998). A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology today*, 19(9):395–404.

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.

Merkenschlager, M., Graf, D., Lovatt, M., Bommhardt, U., Zamoyska, R., and Fisher, A. G. (1997). How many thymocytes audition for selection? *Journal of Experimental Medicine*, 186:1149–1158.

Messaoudi, I., Patiño, J. A. G., Dyall, R., Lemaoult, J., and Nikolich-Zugich, J. (2002). Direct link between MHC polymorphism, T cell avidity, and diversity in immune defense. *Science*, 298:1797–1800.

Miyasaka, A., Yoshida, Y., Wang, T., and Takikawa, Y. (2019). Next-generation sequencing analysis of the human T-cell and B-cell receptor repertoire diversity before and after hepatitis B vaccination. *Human Vaccines and Immunotherapeutics*, 15:2738–2753.

Mombaerts, P., Clarke, A. R., Hooper, M. L., and Tonegawa, S. (1991). Creation of a large genomic deletion at the T-cell antigen receptor beta-subunit locus in mouse embryonic stem cells by gene targeting. *Proceedings of the National Academy of Sciences*, 88:3084–3087.

Mombaerts, P., *et al.* (1992). Mutations in T-cell antigen receptor genes α and β block thymocyte development at different stages. *Nature*, 360:225–231.

Mora, T. and Walczak, A. M. (2018). Quantifying lymphocyte receptor diversity. *Systems Immunology*, pages 183–198.

Mudd, P. A., *et al.* (2022). SARS-CoV-2 mRNA vaccination elicits a robust and persistent T follicular helper cell response in humans. *Cell*, 185:603–613.e15.

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*, 109:16161–16166.

Niederberger, N., Holmberg, K., Alam, S. M., Sakati, W., Naramura, M., Gu, H., and Gascoigne, N. R. J. (2003). Allelic exclusion of the TCR alpha-chain is an active process requiring TCR-mediated signaling and c-Cbl. *Journal of immunology (Baltimore, Md. : 1950)*, 170:4557–4563.

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

Oakes, T., *et al.* (2017). Quantitative characterization of the T cell receptor repertoire of naive and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Frontiers in Immunology*, 8:1–17.

Pizzolla, A., Nguyen, T. H., Smith, J. M., Brooks, A. G., Kedzierska, K., Heath, W. R., Reading, P. C., and Wakim, L. M. (2017). Resident memory CD8+ T cells in the upper respiratory tract prevent pulmonary influenza virus infection. *Science Immunology*, 2:6970.

Pogorelyy, M. V., Rosati, E., Minervina, A. A., Mettelman, R. C., Scheffold, A., Franke, A., Bacher, P., and Thomas, P. G. (2022). Resolving SARS-CoV-2 CD4+ T cell specificity via reverse epitope discovery. *Cell Reports Medicine*, 3:100697.

Pogorelyy, M. V., *et al.* (2017). Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Computational Biology*, 13:e1005572.

Pogorelyy, M. V., *et al.* (2018). Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences*, 115(50):12704–12709.

Pulko, V., *et al.* (2016). Human memory T cells with a naive phenotype accumulate with aging and respond to persistent viruses. *Nature Immunology*, 17:966.

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*, 111:13139–13144.

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*, 107:19414–19419.

Rane, S., Hogan, T., Seddon, B., and Yates, A. J. (2018). Age is not just a number: Naive T cells increase their ability to persist in the circulation over time. *PLoS Biology*, 16:e2003949.

Reantragoon, R., *et al.* (2013). Antigen-loaded MR1 tetramers define T cell receptor heterogeneity in mucosal-associated invariant T cells. *Journal of Experimental Medicine*, 210:2305–2320.

Reynaldi, A., Smith, N. L., Schlub, T. E., Tabilas, C., Venturi, V., Rudd, B. D., and Davenport, M. P. (2019). Fate mapping reveals the age structure of the peripheral T cell compartment. *Proceedings of the National Academy of Sciences*, 116:3974–3981.

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 αβ T cells. *Blood*, 114:4099–4107.

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.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139.

Roden, A. C., Morice, W. G., and Hanson, C. A. (2008). Immunophenotypic attributes of benign peripheral blood gammadelta T cells and conditions associated with their increase. *Archives of Pathology & Laboratory Medicine*, 132:1774–1780.

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.

Roy, S., Scherer, M. T., Briner, T. J., Smith, J. A., and Gefter, M. L. (1989). Murine MHC polymorphism and T cell specificities. *Science*, 244:572–575.

Rudd, B. D., Venturi, V., Li, G., Samadder, P., Ertelt, J. M., Way, S. S., Davenport, M. P., and Nikolich-Zugich, J. (2011). Nonrandom attrition of the naive CD8+ T-cell pool with aging governed by T-cell receptor:pmhc interactions. *Proceedings of the National Academy of Sciences*, 108:13694–13699.

Rupp, L. J., Chen, L., Krangel, M. S., and Bassing, C. H. (2016). Molecular analysis of mouse T cell receptor α and β gene rearrangements. *Methods in Molecular Biology*, 1323:179–202.

Samson, L. D., *et al.* (2020). Limited effect of duration of CMV infection on adaptive immunity and frailty: insights from a 27-year-long longitudinal study. *Clinical & Translational Immunology*, 9:e1193.

Sant, S., *et al.* (2018). Single-cell approach to influenza-specific CD8+ T cell receptor repertoires across different age groups, tissues, and following influenza virus infection. *Frontiers in Immunology*, 9:1453.

Schattgen, S. A., Guion, K., Crawford, J. C., Souquette, A., Barrio, A. M., Stubbington, M. J., Thomas, P. G., and Bradley, P. (2021). Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nature Biotechnology*, 40:54–63.

Sethna, Z., Elhanati, Y., Callan Jr, C. G., Walczak, A. M., and Mora, T. (2019). OLGA: fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981.

Sethna, Z., Isacchini, G., Dupic, T., Mora, T., Walczak, A. M., and Elhanati, Y. (2020). Population variability in the generation and selection of T-cell repertoires. *PLoS Computational Biology*, 16(12):e1008394.

Sewell, A. K. (2012). Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12:669.

Shugay, M., *et al.* (2017). Vdjdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic acids research*, 46:D419–D427.

Sidhom, J. W., Larman, H. B., Pardoll, D. M., and Baras, A. S. (2021). DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, 12.

Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. *Frontiers in Immunology*, 12.

Steinmann, G. G., Klaus, B., and Müller-Hermelink, H.-K. (1985). The involution of the ageing human thymic epithelium is independent of puberty: a morphometric study. *Scandinavian Journal of Immunology*, 22:563–575.

Sterrett, S., *et al.* (2020). Peripheral CD4 T follicular cells induced by a conjugated pneumococcal vaccine correlate with enhanced opsonophagocytic antibody responses in younger individuals. *Vaccine*, 38:1778–1786.

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

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.

Sycheva, A. L., Pogorelyy, M. V., Komech, E. A., Minervina, A. A., Zvyagin, I. V., Staroverov, D. B., Chudakov, D. M., Lebedev, Y. B., and Mamedov, I. Z. (2018). Quantitative profiling reveals minor changes of T cell receptor repertoire in response to subunit inactivated influenza vaccine. *Vaccine*, 36:1599–1605.

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.

Tanchot, C., Lemonnier, F. A., Pérarnau, B., Freitas, A. A., and Rocha, B. (1997). Differential requirements for survival and proliferation of CD8 naive or memory T cells. *Science*, 276:2057–2062.

Tanno, H., *et al.* (2020). Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proceedings of the National Academy of Sciences*, 117:532–540.

Tcherniaeva, I., den Hartog, G., Berbers, G., and van der Klis, F. (2018). The development of a bead-based multiplex immunoassay for the detection of IgG antibodies to CMV and EBV. *Journal of Immunological Methods*, 462:1–8.

Textor, J., Henrickson, S. E., Mandl, J. N., von Andrian, U. H., Westermann, J., de Boer, R. J., and Beltman, J. B. (2014). Random migration and signal integration promote rapid and robust T cell recruitment. *PLoS Computational Biology*, 10.

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*, 29:542–550.

Thomas, P. G. and Crawford, J. C. (2019). Selected before selection: A case for inherent antigen bias in the T cell receptor repertoire. *Current Opinion in Systems Biology*.

Thome, J. J. C., Grinshpun, B., Kumar, B. V., Kubota, M., Ohmura, Y., Lerner, H., Sempowski, G. D., Shen, Y., and Farber, D. L. (2016). Long-term maintenance of human naive T cells through in situ homeostasis in lymphoid tissue sites. *Science Immunology*, 1:eaah6506–eaah6506. DATA available via https://clients.adaptivebiotech.com/pub/2e838d8c-99cb-4103-b79e-7ee32768536d.

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*, 33:2924–2929.

Tillman, R. E., Wooley, A. L., Khor, B., Wehrly, T. D., Little, C. A., and Sleckman, B. P. (2003). Cutting edge: targeting of Vβ to Dβ rearrangement by RSSs can be mediated by the V (D) J recombinase in the absence of additional lymphoid-specific factors. *The Journal of Immunology*, 170(1):5–9.

Touzel, M. P., Walczak, A. M., and Mora, T. (2020). Inferring the immune response from repertoire sequencing. *PLoS Computational Biology*, 16:e1007873.

Trepel, F. (1974). Number and distribution of lymphocytes in man. A critical analysis. *Klinische Wochenschrift 1974 52:11*, 52:511–515.

Trowsdale, J. and Knight, J. C. (2013). Major Histocompatibility Complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, 14:301.

Tsang, T. K., Cauchemez, S., Perera, R. A., Freeman, G., Fang, V. J., Ip, D. K., Leung, G. M., Peiris, J. S. M., and Cowling, B. J. (2014). Association between antibody titers and protection against influenza virus infection within households. *The Journal of Infectious Diseases*, 210:684–692.

Uddin, I., Joshi, K., Oakes, T., Heather, J. M., Swanton, C., Chain, B., *et al.* (2019). An economical, quantitative, and robust protocol for high-throughput T cell receptor sequencing from tumor or blood. *Cancer Immunosurveillance*, pages 15–42.

Uematsu, Y., Ryser, S., Dembić, Z., Borgulya, P., Krimpenfort, P., Berns, A., von Boehmer, H., and Steinmetz, M. (1988). In transgenic mice the introduced functional T cell receptor β gene prevents expression of endogenous β genes. *Cell*, 52:831–841.

Vazquez-Lombardi, R., *et al.* (2022). High-throughput T cell receptor engineering by functional screening identifies candidates with enhanced potency and specificity. *Immunity*, 55:1953–1966.e10.

Venturi, V., *et al.* (2011). A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *The Journal of Immunology*, 186:4285–4294.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47:D339–D343.

Wang, Y., Prosen, D. E., Mei, L., Sullivan, J. C., Finney, M., and Horn, P. B. V. (2004). A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic acids research*, 32:1197–1207.

Wang, Y., *et al.* (2008). Th2 lymphoproliferative disorder of LatY136F mutant mice unfolds independently ofTCR-MHC engagement and is insensitive to the action of Foxp3+ regulatory T cells. *The Journal of Immunology*, 180:1565–1575.

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.

Wertheimer, A. M., Bennett, M. S., Park, B., Uhrlaub, J. L., Martinez, C., Pulko, V., Currier, N. L., Nikolich-Zugich, D., Kaye, J., and Nikolich-Zugich, J. (2014). Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans. *The Journal of Immunology*, 192:2143–2155.

Westera, L., van Hoeven, V., Drylewicz, J., Spierenburg, G., van Velzen, J. F., de Boer, R. J., Tesselaar, K., and Borghans, J. A. M. (2015). Lymphocyte maintenance during healthy aging requires no substantial alterations in cellular turnover. *Aging Cell*, 14:219–227.

Wilkinson, T. M., *et al.* (2012). Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans. *Nature Medicine*, 18(2):274–280.

Willyard, C. (2018). Expanded human gene tally reignites debate. *Nature*, 558:354–355.

Xu, Y., *et al.* (2020). Age-related immune profile of the T cell receptor repertoire, thymic recent output function, and miRNAs. *BioMed Research International*, 2020.

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.

Yoshida, K., Cologne, J. B., Cordova, K., Misumi, M., Yamaoka, M., Kyoizumi, S., Hayashi, T., Robins, H., and Kusunoki, Y. (2017). Aging-related changes in human T-cell repertoire over 20 years delineated by deep sequencing of peripheral T-cell receptors. *Experimental Gerontology*, 96:29–37.

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.

Zemmour, D., Zilionis, R., Kiner, E., Klein, A. M., Mathis, D., and Benoist, C. (2018). Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by theTCR. *Nature Immunology*, 19:291.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30:614–620.

Zhang, J., Wang, Y., Yu, H., Chen, G., Wang, L., Liu, F., Yuan, J., Ni, Q., Xia, X., and Wan, Y. (2021). Mapping the spatial distribution of T cells in repertoire dimension. *Molecular Immunology*, 138:161–171.

Zheng, C., *et al.* (2017). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell*, 169:1342–1356.e16.

Samenvatting

Het afweersysteem in mensen en gewervelde dieren bestaat uit zowel de aangeboren als de verworven afweer. Deze verschillende verdedigingsmechanismen voorkomen samen veel infecties die veroorzaakt worden door bijvoorbeeld virussen en bacteriën. Verworven afweer werkt door middel van grote populaties van zeer specifieke B-cellen en T-cellen, die samen gevaarlijke binnendringers en beschadigde cellen onderscheiden van onschuldige eiwitten en gezonde cellen. De specificiteit van de T-cellen, die de focus van dit proefschrift vormen, wordt bepaald door hun T-celreceptor (TCR). Een T-cel kan activeren wanneer zijn TCR bindt aan een lichaamsvreemd eiwitfragment, waarna de cel zich gaat vermenigvuldigen. Dit resulteert in een grote groep T-cellen met dezelfde TCR, die de geïnfecteerde cellen doden of ondersteuning bieden aan andere afweercellen. Een deel van de groep specifieke T-cellen blijft leven na deze respons en ontwikkelt zich tot geheugencellen. Door deze verhoogde aanwezigheid en paraatheid verloopt een volgende respons tegen hetzelfde pathogeen vaak sneller, waardoor ziekte in veel gevallen wordt voorkomen. Dit principe van verworven immuniteit vormt de basis van het succes van bescherming door vaccinatie.

Doordat T-cellen elk een specifieke TCR hebben en in aantal kunnen veranderen door nieuwe productie, celdeling en sterfte, is het repertoire van TCRs erg dynamisch. We kunnen deze dynamica in kaart brengen door de volgorde van nucleotiden die coderen voor de receptor te achterhalen. Dit gebeurt met behulp van *sequencing*, waarmee het tegenwoordig mogelijk is om de TCR-identiteit van miljoenen T-cellen tegelijkertijd op te helderen. De informatie die volgt uit zulke experimenten vereist een gedegen analyse: naast een grote TCR-diversiteit bevat de data ook veel kleine foutjes. Een belangrijke kanttekening bij TCR-*sequencing* is dat het menselijk lichaam ongeveer een biljoen T-cellen bevat, waarvan dus slechts een klein deel in kaart kan worden gebracht. Hierdoor hebben kleine onzorgvuldigheden in het experiment en de analyse mogelijk een groot effect op de geschatte dynamica van de T-cellen in het lichaam. Dit probleem ondervangen we in dit proefschrift door de informatie uit meerdere monsters (*samples*) slim te combineren: zodoende kunnen we de onzekerheden van de analyse goed inschatten en de TCRs uit de monsters correct classificeren.

In **Hoofdstuk 2** bestuderen we het menselijk TCR-repertoire van de T-cellen die nog niet zijn geactiveerd. Deze zogeheten naïeve T-cellen hebben een grote verscheidenheid aan TCRs, die we met behulp van *sequencing* in kaart brengen. Opvallend is dat de TCRs in de monsters onderling sterk verschillen in hoeveel ze voorkomen. Deze verschillen worden deels verklaard door de verschillende productiekansen van elk van de TCRs. Daarnaast vergelijken we meerdere monsters van hetzelfde TCR-repertoire om uit te sluiten dat de veelvoorkomende TCRs enkel veroorzaakt worden door dezelfde T-cel meerdere keren te meten. Door de voorspellingen op basis van vele wiskundige modellen te vergelijken met de data hebben we de verschillen in TCR-aanwezigheid verder in kaart gebracht. Hieruit blijkt dat ook vóór T-cel activatie sommige TCRs in veel meer T-cellen voorkomen dan de meeste andere TCRs. Dit betekent dat een klein aantal TCRs een aanzienlijk deel van het naïeve T-celrepertoire beslaat.

De analyse van **Hoofdstuk 3** gaat verder in op de verschillen tussen TCRs die veel en weinig voorkomen in het repertoire van naïeve T-cellen. We identificeren specifieke eigenschappen in de nucleotiden- en eiwitvolgorde van veelvoorkomende TCRs. De meest opvallende eigenschap heeft te maken met de opbouw van de TCR-nucleotidenvolgorde. Deze wordt gevormd door het aan elkaar plakken van verschillende gen-segmenten, waaronder een zogeheten D-segment. Onze analyse laat zien dat een aanzienlijk deel van de TCRs het D-segment mist, en dat dit bovendien gebruikelijker is bij TCRs die veel voorkomen. TCRs die gedeeld worden tussen vele individuen missen ook vaak een D-segment, wat mogelijk verklaard wordt doordat ze vroeg in de ontwikkeling van het afweersysteem geproduceerd worden.

Een belangrijke eigenschap van een functioneel T-celrepertoire is een grote verscheidenheid aan TCRs. Omdat met de leeftijd de aanmaak van nieuwe TCRs afneemt is een grote vraag of dit leidt tot afname van TCR-diversiteit en daarmee tot verminderde T-celimmuniteit. **Hoofdstuk 4** richt zich op de eerste van deze twee vragen, door de gemeten TCR-diversiteit te vergelijken tussen verschillende leeftijdsgroepen. Hoewel dit klinkt als een relatief eenvoudige analyse, identificeren we verschillende haken en ogen aan deze aanpak. Zo laten we zien dat het sterk uitmaakt in welke mate de verschillende types T-cellen voorkomen in het genomen monster. Naïeve T-cellen bevatten bijvoorbeeld meestal veel meer verschillende TCRs dan geheugencellen. Daarnaast blijkt het TCR-repertoire zo divers dat een enkel monster vaak onvoldoende is om de diversiteit goed te kunnen schatten. Zulke schattingen worden beter door de informatie uit meerdere monsters te combineren. Dit stelt ons in staat om verschillen in TCR-repertoire diversiteit tussen jonge en oudere mensen aan te tonen. Al met al laat dit zien dat het koppelen van TCR-diversiteit aan kwaliteit van het afweersysteem een gedegen analyse vereist.

Waar de voorgaande hoofdstukken zich met name richtten op een momentopname van het naïeve T-celrepertoire, gaat **Hoofdstuk** 5 over de veranderingen in het TCR-repertoire van geheugencellen na vaccinatie. Om dit te doen nemen we T-cellen af van mensen, voor en nadat ze een vaccinatie tegen pneumokokken hebben ontvangen. Door het TCR-repertoire in deze monsters in kaart te brengen proberen we te achterhalen welke T-cellen geactiveerd zijn door het vaccin. Ook dit blijkt makkelijker gezegd dan gedaan, omdat het voorkomen van TCRs zelfs aanzienlijk kan verschillen tussen twee monsters vanuit hetzelfde TCR-repertoire. Hieruit blijkt hoe belangrijk het is om meerdere monsters te analyseren, zodat een betrouwbare classificatie van geactiveerde T-cellen kan worden gemaakt. In de meerderheid van de gevaccineerde donoren vinden we bewijs voor T-celdeling, maar het geschatte aantal specifieke TCRs blijkt afhankelijk van veel factoren. Een bepalende factor voor deze schatting blijkt de grootte van elk van de monsters, wat het vergelijken van de responsen tussen donoren bemoeilijkt. Als een gevolg hiervan worden T-cel responsen die niet heel groot of juist erg divers zijn moeilijk herkend. Het gebruik van het TCR-repertoire als indicatie van de effectiviteit van een vaccin in een individu is in zulke situaties dan ook voorlopig niet haalbaar.

In **Hoofdstuk 6** veranderen we van organisme dat we bestuderen: van de mens naar een specifieke muizenstam met een beperkte TCR-diversiteit. Dit maakt het mogelijk om bijna het volledige TCR-repertoire in kaart te brengen met behulp van *sequencing*, en om dit te vergelijken tussen meerdere muizen. Opnieuw hebben we meerdere monsters met T-cellen geanalyseerd, ditmaal vanuit de verschillende lymfeklieren in een muis. Door de TCR-repertoires met elkaar te vergelijken laten we zien dat er een strakke organisatie van het repertoire is, die gedeeld wordt door elk van de muizen. Een gevolg van deze organisatie is dat bepaalde TCRs consistent meer voorkomen in specifieke lymfeklieren. We vinden sterke aanwijzingen dat deze voorkeur qua verblijfplaats bepaald wordt door de TCR-specificiteit. Daarnaast vinden we sterk bewijs dat de TCR bepaalt in welke T-celtype de T-cel zich ontwikkelt. Deze inzichten laten zien dat de verdeling van TCRs over T-celsubgroepen en organen niet willekeurig is, maar een organisatie volgt die in deze muizen makkelijker aan te tonen is dan in mensen.

Samenvattend gebruiken de studies in dit proefschrift de verdeling en verschuiving van het TCR-repertoire om de onderliggende T-celdynamica in kaart te brengen. Door vooruitgang in de experimentele aanpak en de data-analyse ontstaat een steeds completer beeld van de onderliggende processen. Door ook in de toekomst informatie uit meerdere bronnen slim te combineren werken we steeds verder aan het persoonlijk maken van vaccinaties en medicijnen.

Curriculum Vitae

Peter de Greef was born in Wageningen, the Netherlands, on the 10th of May 1994. He grew up in Ermelo and received his pre-university education (vwo) at Christelijk College Nassau Veluwe in Harderwijk. He attended a personal development programme (Basisjaar) at the Evangelische Hogeschool in Amersfoort, after which he started his studies in Biology at Utrecht University. He complemented this programme with a minor in Computational Science at the University of Amsterdam. After obtaining his Bachelor's degree in 2015, Peter continued his studies as a Master's student of the Molecular and Cellular Life Sciences programme at Utrecht University. He conducted a research project under the supervision of prof. Rob de Boer, which laid the foundation of Chapter 2 in this thesis. He performed another research internship at the University of Edinburgh, supervised by prof. Peter Doerner and prof. Kirsten ten Tusscher. He joined the Quantitative Biology honours programme, which allowed him to write a research proposal that was awarded a PhD position. After graduating *cum laude* in 2017, he started as a PhD candidate under the supervision of prof. Rob de Boer. The results of his PhD research are described in this thesis. Currently, Peter is working as a data analyst for the Dutch government.

List of Publications

de Greef, P. C., Oakes, T., Gerritsen, B., Ismail, M., Heather, J. M., Hermsen, R., Chain, B., & de Boer, R. J. (2020). The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife*, 9, e49900.

Lanfermeijer, J., **de Greef, P.C.**, Hendriks, M., Vos, M., van Beek, J., Borghans, J. A. M., & van Baarle, D. (2021). Age and CMV-infection jointly affect the EBV-specific CD8+ T-cell repertoire. *Frontiers in Aging*, 2, 11.

de Greef, P. C., & de Boer, R. J. (2021). TCRβ rearrangements without a D segment are common, abundant, and public. *Proceedings of the National Academy of Sciences*, 118(39), e2104367118.

de Greef, P. C., & de Boer, R. J. (2023). Towards a robust comparison of diversity between sampled TCR repertoires. *Preprint at bioRxiv*, 10.1101/2023.02.10.528010.

Acknowledgements

These last pages are dedicated to the people that contributed to the past five years that led to this thesis: a result that would have been less likely and less enjoyable without their support and guidance. This includes many contributions to scientific discoveries, to my personal growth, and to creating a supportive social environment.

First of all, I am grateful to my promotor **Rob** for having the confidence to start this PhD trajectory with me. When I started as a Master's student in your group, you gave the advice to 'own' the project from the beginning. That advice, and your ability to keep stimulating such an attitude, have been critical to make this PhD journey reach its goal. Thank you for keeping your door always open and for giving the freedom to discover where I wanted to go. I really appreciated that you adapted your supervision style to what I needed and always treated me as a colleague rather than a student. I enjoyed the (often food-related) social gatherings such as the pizza party in your garden and our breakfasts at the balcony in Rehovot. Your advice 'First do, then think' will be useful in many future situations.

The other immunology group members have played an important role in making these years a pleasant experience. **Can**, your genuine interest in people and their well-being, as well as their science, is inspiring. I enjoyed our casual chats and found your feedback and honest advice always helpful. Belonging to a close group of four PhD students working on immunology was a big advantage. **Juliane**, thank you for sometimes being my personal secretary and always being the social glue of our group. You took the initiative for many social events that helped us to be involved with each other. I have seen you grow in many ways and I am confident you will do well during your adventure in New York. **Arpit**, thanks for many nice discussions during the past years and for being a paranymph during my defence. We had a lot of fun during our trips to Paris and many planned or spontaneous dinners. I wish you all the best for what is coming next. **Erdem**, you were a great TCR sparring partner and I am very happy to also have you as a paranymph during my defence. Thank you for many coffee conversations and good luck with the final steps of your PhD!

Even when having a great supervisor and research group, I benefited from my supervisory committee that made sure that the project was developing in the right direction. I am thankful to **José** and **Aridaman** for their input during our yearly meetings and I always appreciated our discussions. Thank you for regularly checking on how things were going and your honest advice on how to lay out plans for the future. I am also grateful to all members of the **assessment committee** for taking the time to read this thesis and join the opposition during the defence.

What would doing research in a computational group mean without the expertise of people that really know how to handle T cells and sequence their receptors? Thank you **Benny** for taking the initiative to test the validity of our simple models against your data. It has been a long journey, but we have managed to learn a lot about T cells and their receptors, as well as our assumptions on what is going on. Thank you for having me in your group for a month, which allowed me to see all steps of TCR repertoire sequencing with my own eyes. Another important person in connecting my computational skills with immunological data is **Josien**. What started as sometimes solving pipeline errors or unexpected results, developed into a great collaboration. I always enjoyed our teamwork, in which our areas of expertise complemented each other. Thank you for all the hard work in the lab, your project management skills and your helpful advice. I learned a lot from the nice discussions with you, **Debbie**, and **José**. Biological questions that remain unanswered in the human context could be addressed by collaborating with **Jorg**. Thank you for including us in the project to analyse your beautiful mouse data and for all the discussions we had during these years. I learned a lot from those and look back on a fruitful collaboration with exciting biological findings as a result.

One of the aspects of science I enjoyed most is sharing knowledge and skills with others. Many students have helped me to grow in this role during various courses. Specifically, I would like to thank my students **Pauline** and **Daphne**. I enjoyed our nice conversations while supervising your research projects. During the IJclub and ImmuniTEA meetings I learned a lot from the interactions with **Nila**, **Rianne**, **Weiyang**, and **Abhinandan**. Thank you for sharing your knowledge and helping us to connect our models and analyses to the immunological reality.

Working in the Theoretical Biology and Bioinformatics group is a special experience because of the people that are part of it. Thank you **Paulien** for founding the group and always contributing so much to both the science and the social interactions. My first exposure to TBB was a research project supervised by **Kirsten**. Thank you for introducing me into the world of understanding biology better by integrating model results and experimental data. During my PhD, I was happy to share an office with **Thea**, **Margo**, **Ksenia**, **Joana**, **Milton**, **Leonie**, and **Laurens**. Thank you for many coffee breaks, for our random chats in the office, and for sharing your PhD experience. Special thanks to **Jan Kees**, who finds a solution to every practical or computational issue. The mutants and their users cannot wish for a better person to maintain them than you. **Rutger**, thank you for many helpful conversations about statistics, science in general, and doing good. I am grateful to all TBB members that contributed to this somewhat odd, but always kind and stimulating working environment.

On a more personal note, there are some long-lasting friendships that mean a lot to me. **Bastian**, it is over two decades ago that we got to know each other. You predicted that I would start a PhD, long before I considered that as the best step after my studies. Thank you for our friendship and our never-ending conversations. **Jesper**, I remember many hours of math classes that we actually spent playing card games. Luckily, it did not hurt our career too much, and our practice from then still helps during our klaverjas evenings with Bastian and **Tom**. Thank you guys for this weekly distraction from scientific challenges. **Emma**, thank you for becoming such a close friend over the years and for allowing me to win sometimes during our game nights. **Julian**, we shared a large chunk of our trajectory from high school to obtaining a PhD. Our conversations always take longer than planned, which is to me a clear sign of their value. Thank you for being a great friend and colleague at the same time. Coming home after a day of work was never boring thanks to my **WS88 roommates**. The pizza nights, lockdown lunches, and walking drinks contributed in unexpected ways to this thesis. I would also like to thank my **ILT colleagues** for the welcoming atmosphere and their flexibility while finishing the final steps of this PhD journey.

A special word of thanks to my family. **Papa & Mama**, thank you for always stimulating my curiosity and motivating us to go the extra mile. Your support over the years means a lot to me. **Carla & Jacob**, **Erik & Justien**, and **Marco & Joanne**, I am grateful for you as brothers and sisters. Looking back on many LEGO-related fights in our childhood, I now enjoy our open conversations and our shared interest in trying new board games. **Siebrand & Nynke**, **Wander & Amelia**, and **Tjarco**, thank you for treating me as part of the Wierda family. While having a different set of genes, I enjoy the fun we have together and appreciate your genuine interest and honest advice.

Dear **Janneke**, we celebrated my first PhD salary during our first date and never stopped dating afterwards. Thank you for adding colour to my socks and to the rest of my life. It is hard to imagine how I would have completed this PhD journey without your patience and support. Let’s keep complementing each other and celebrating together. I am thankful for you in my life and looking forward to our next adventures together.

Utrecht, February 2023

Peter