Quantifying the Development, Size, and Repertoire Diversity of T Cell Populations

Edward S. Lee

Submitted in fulfilment of the requirements
for the Degree of Doctor of Philosophy

COLLEGE OF MEDICAL, VETERINARY AND LIFE SCIENCES
UNIVERSITY OF GLASGOW

COLOPHON

The serif text typeface is Sabon by Jan Tschichold. The sans serif text typeface is Neue Frutiger by Adrian Frutiger and Akira Kobayashi. The monospaced face is Fira Mono by Ralph du Carrois and Erik Spiekermann. The typefaces used in the figures are Trade Gothic Next by Akira Kobayashi and Jackson Burke, Neue Haas Grotesk by Christian Schwartz, and Myriad Pro by Robert Slimbach and Carol Twombly.

Contents

Acknowledgments

Abstract

1 Introduction

1.1 OVERVIEW 3
1.2 T CELL RECEPTORS 4
1.3 T CELL DEVELOPMENT 9
1.4 RECIRCULATION OF T CELLS 14
1.5 SEQUENCING 16
1.6 MATHEMATICAL MODELING IN IMMUNOLOGY 20
1.7 AIMS 21

2 Materials and Methods

2.1 MICE 25
2.2 CELL ISOLATION AND MATERIALS 25
2.3 FLOW CYTOMETRY PROTOCOLS 26
2.4 ONTOGENY EXPERIMENTS 27
2.5 SINGLE-CELL SEQUENCING 28
2.6 MATHEMATICAL AND STATISTICAL ANALYSIS 29
2.7 BAYESIAN DATA ANALYSIS 31

3 Ontogeny of peripheral T cells

3.1 INTRODUCTION 45
3.2 EXPERIMENTAL DATA 45
3.3 $CD4^+$ T CELL DEVELOPMENT IN THE THYMUS 52
3.4 $CD8^+$ T CELL DEVELOPMENT IN THE THYMUS 59
3.5 NAIVE $CD4^+$ T CELLS 66
3.6 SUMMARY 72
3.7 MATHEMATICAL MODELS 73

4 Counting Lymphocytes Using Thoracic Duct Cannulations

4.1 INTRODUCTION 93
4.2 RESULTS 97
4.3 EXTENDING THE MODEL 108

4.4 DISCUSSION 121

5 Obtaining paired αβ TCR sequences
5.1 INTRODUCTION 127
5.2 DESCRIPTION OF ALPHABETR AND HOW IT WAS TESTED 135
5.3 RESULTS 145
5.4 DISCUSSION 165
5.5 DETAILED DESCRIPTION OF THE ALGORITHM 170

6 Identifying paired TCR sequences with the stable matching problem
6.1 INTRODUCTION TO MATCHING PROBLEMS 189
6.2 FROM HOSPITAL-RESIDENTS TO T CELL RECEPTORS 195
6.3 RESULTS 196
6.4 DISCUSSION 204

7 Discussion and Conclusions

List of Figures

1.1 Structure of the αβ T cell receptor 5
1.2 Germline configuration of the TCRα and TCRβ loci. 6
1.3 V(D)J recombination 7
1.4 The journey of thymocyte development 10
1.5 Key events in the different stages of thymocyte development 12
1.6 An illustration of the lymphatics of the thorax and neck of the mouse 15

2.1 A simple model for peripheral T cell dynamics 29

3.2 Schematic of the developmental trajectory of CD4 and CD8 T cell lineages 48
3.3 Gating of lymphocytes/thymocytes, singlet events, and live cells 48
3.4 Gating strategy for thymocyte subsets 49
3.5 Gating strategy for peripheral T cells 50
3.6 Three models for SP4 thymocyte development 55
3.7 Posterior distribution of the fit for model $\mathbf{M}_3$ 56
3.8 Posterior distributions of the fits for models $\mathbf{M}_1$ and $\mathbf{M}_2$ 57
3.9 Posterior distributions of the fits for more simple SP4 thymocyte models 60
3.10 Three models for SP8 thymocyte development 62
3.11 Posterior distributions of the fits for SP8 thymocyte models 64
3.12 Posterior distribution of the fits for the best-fitting SP8 thymocyte model 65
3.13 Model for naive $CD4^+$ T cell counts and $Ki67^{hi}$ proportions 67
3.14 Model fit to naive $CD4^+$ T cell counts and $Ki67^{hi}$ proportions 68
3.15 Sensitivity of interdivision and residence times to the ratio of influxes 69
3.16 Exploring Ki67 expression levels in mature SP4 thymocytes and naive $CD4^+$ T cells 71
3.17 Empirical descriptor functions fitted to DP2 cell counts and $Ki67^{hi}$ proportions 74

4.1 Overview of how lymphocyte numbers were estimated using thoracic duct cannulation 96
4.2 Schematic of calculate total lymphocyte numbers 97
4.3 Gating strategy for $CD4^+$ and $CD8^+$ T cells 98
4.4 Gating strategy for B cells 99
4.5 Loss of T cells due to surgical stress 100
4.6 Loss of B cells due to surgical stress 101
4.7 The decrease in T cell counts due to thoracic duct cannulations 102
4.8 The effect of thoracic duct cannulations on B cell counts 103
4.9 Proportions of CD8 naive, central memory, and effector memory T cells in the cannulated, sham surgery, and control mice 105
4.10 Proportions of $CD4^+$ naive, central memory, and effector memory T cells in the cannulated, sham surgery, and control mice. 106
4.11 Proportions of transitional and mature B cells in the cannulated, sham surgery, and control mice 107
4.12 Schematic of the lymphatic drainage of different lymph nodes 109
4.13 Extending the total number of lymphocytes calculation to account for differential effects from the cannulation 113
4.14 Estimated number of total CD4 T cell 115
4.15 Estimated number of total CD8 T cell 116
4.16 Estimated numbers of $CD4^+$ T cell subsets for different values of f_{spl} 117
4.17 Estimated numbers of $CD8^+$ T cell subsets for different values of f_{spl} 118
4.18 Estimated number of transitional B cell 119
4.19 Estimated number of mature B cells 120

5.1 Analysis of TCRα usage in human, YFV-specific peripheral-blood $CD8^+$ T cells 133
5.2 An overview of the implementation of ALPHABETR 139
5.3 Depth and accuracy of αβ pairings generated by ALPHABETR 147
5.4 Assessment of the precision of clonal frequency estimation 150
5.5 Discriminating between dual-TCRα and β-sharing clones 153
5.6 Discriminating between β-sharing and dual-TCRα clones 154
5.7 Simulations of populations with a high level of sharing 157

5.8 Simulations of populations with a low level of sharing 158
5.9 Simulations of populations consisting of 500 unique clones 159
5.10 Simulations of populations consisting of 3000 unique clones 160
5.11 Simulations with no dual-TCRβ clones 161
5.12 Comparison of well occupancy patterns of the clones identified by ALPHABETR and by pairSEQ 163
5.13 Comparison of single-cell approaches and ALPHABETR 167
5.14 Sample space for calculating likelihoods of two- and three-way co-occurrences of chains under the hypotheses of CDR3β-sharing 184

6.1 Depth of the common clones representing the top 50% of the population in frequency 197
6.2 Depth of the rare clones representing the bottom 50% of the population in frequency 198
6.3 Depth of all clones in the simulated populations 199
6.4 False pairing rates in the simulated populations 200
6.5 Results from simulations of populations exhibiting medium levels of sharing 201
6.6 Expected number of unique β chains recovered from the naive repertoire 203

List of Tables

3.1 Markers used to define thymocyte and peripheral T cell subsets 51
3.2 Estimated parameter values for SP4 thymocytes, SP8 thymocytes, and naive CD^+ T cells 61
3.3 Parameter values for empirical descriptor functions for DP2, DP3, and mature SP4 thymocytes 73

5.1 A summary of the degrees of sharing of CDR3α and CDR3β at the amino acid level across clones within epitope-specific T cell populations 134
5.2 The three different levels of sharing used in the simulations 141
5.3 The mixed sampling strategies used in the simulations 142
5.4 Assessing the impact of underestimation of sequencing error on clonal frequency estimation 151
5.5 Recovery of tumor-infiltrating lymphocyte TCR pairs using ALPHABETR and data from Reference [1] 163
5.6 Simulations without the resampling procedure 173

Declaration

I declare that, except where reference is made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Edward S. Lee
April 2018

Acknowledgments

I often hear that it takes a village to raise a child, but nobody ever says how it takes a village to earn a PhD. The work represented in this thesis is the culmination of the support and encouragement from mentors, friends, and family who chose to invest their time and energy in me. It's fitting that thanking them is the first thing I do in this thesis.

Andy, it's hard to express how thankful I am for your guidance and mentorship throughout these years. Thank you for always seeking to help me grow as a scientist by allowing me to explore new questions and techniques and by always challenging me to do better science. You have continually been a fierce advocate for me and my education, and I am truly grateful for all of your help and support. I promise that I will master subject-verb agreement one day.

Simon, thank you for having the courage to adopt me into your lab, especially knowing that I hadn't touched a pipette for 5 years before working in the lab. Thanks to Ben for being a second unofficial supervisor of sorts and answering all of my random questions about T cells and experiments; I hope that one day my banter will be half as witty as yours is.

To my fellow labmates—Graeme, Sanket, Daniel, and Luise—thanks for all the chats about immunology, statistics. Special thanks to Sanket for helping me with a lot of experiments and co-creating Paperball™. Thanks to Thea and Melissa for helping me plan out details about my experiments and answering all my questions about T cells. Thank you Mowlings for helping me find my way in lab and answering all of my dumb questions—particular thanks to Verena for teaching me how to cannulate and Alberto for teaching me pretty much everything else.

I also want to thank several people who helped me with my experiments. The staff at the CRF deserve a long round of applause, particularly Tony and Joanna for assisting me with surgeries. Special thanks to Tony, who stood by me when I was utterly useless in the operating room and helped become the cannulator that I am now. Many thanks to Diane for teaching me pretty much everything I know about flow cytometry.

None of the work in this thesis would have possible without the amazing work and support from the open-source software community, especially those involved in the R project and the Stan project. I'm indebted to the RStudio team, particularly Hadley Wickham for the tidyverse and his clear teaching materials and Jenny Bryan for teaching me how to use Git at useR!2017—both of you have made my life better.

To my church families at Maranatha Grace Church and the Tron Church, thank you for all of your support and your prayers throughout the past four years. Glasgow has become a second home for us because of the fellowship we had at the Tron, and we are so thankful that you continue to be a bold witness for the gospel of Jesus.

And finally, to my family who have been so supportive over the years. Ernest and Eric, thank you for both for your friendship and companionship. To Mama and Papa Hsieh, thank you for your constant care for and feeding of me and Tiff. Dad, thanks for always being so proud of my academic endeavors and supporting me in everything that I do. Mom, thank you for tirelessly and sacrificially loving us and our family. Your faith in Jesus and the hope you have in the gospel always encourages me to remember that our God is indeed a living God. The achievement in this thesis is just as much yours as it is mine.

To Tiff, thank you for being so supportive and walking with me through these past four years. Moving away from home to Glasgow was kind of crazy, but I'm glad I got to do it with you.

And to the only Triune God—the one who upholds the universe by the word of his power and yet gave his Son up for us all—be all glory forever.

Soli Deo gloria.

Abstract

The adaptive immune system must be able to respond to virtually any pathogen that the body encounters. T cell immunity is able to do so by developing a diverse repertoire of T cell receptors and maintaining large numbers of T cells. These two quantitative properties are fundamental for the ability of T cell-mediated immunity to clear infections and generate memory cells for future protection. The aims of this thesis are to quantify the sizes of T cell populations, to develop tools to measure the diversity of T cell repertoires, and to describe how T cell populations develop in neonatal mice.

We studied the development of T cell populations in neonatal mice by measuring cell counts and Ki67 expression in thymocyte and peripheral T cell subsets from mice soon after birth to late adulthood. The presumed lymphopenic environment of the neonatal mouse is thought to cause T cells to undergo lymphopenia-induced proliferation, and we wanted to quantify the balance between thymic output and peripheral expansion in the naive T cell compartment during development with mathematical modeling. We also used modeling to find the most parsimonious description of differentiation within the thymus that explains the dynamically growing thymus.

We then sought to quantify the sizes of the peripheral T cell compartments in the adult mouse. Understanding the characteristics of healthy T cell immunity requires knowing the precise numbers of the different T cell subsets found in the body. We performed thoracic duct cannulations in adult mice to collect recirculating T cells and reduce cell numbers in the lymph nodes and spleens; by counting the number of collected T cells and its effect on cell numbers on the secondary lymphoid organs, we sought to back-calculate the total number of T cells in the mouse. Finally, we developed tools that provide high-throughput and cost-effective methods for identifying paired TCR sequences. By using computational techniques, we were able to adapt standard sequencing protocols to identify many paired TCR sequences without resorting to large and expensive single-cell sequencing techniques. By leveraging experimental design with mathematical methods, we were able to quantify and characterize many properties of effective T cell immunity.

CHAPTER 1

Introduction

1.1 OVERVIEW

The adaptive immune system has the extraordinary task of responding to virtually any pathogen without reacting to self molecules. This feat is achieved by creating a diverse repertoire of receptors—each with its own antigen specificity—that is vetted for self-reactive receptors. In the case of T lymphocytes, each T cell clonotype is the set of all T cells that expresses the same unique T cell receptor (TCR), which is created by a random gene rearrangement process that can potentially form 10^{15} different TCRs in mice and 10^{18} in humans. T cells use their TCRs to recognize antigens as short peptides presented on major histocompatibility complex (MHC) molecules.

When a pathogen enters peripheral tissues, local antigen-presenting cells (APCs) such as dendritic cells (DCs) take up the pathogen and other pathogenic antigens and migrate to the local secondary lymphoid organ (SLO). In the SLO, the DCs process their antigens and present them to T cells. T cells whose TCRs can react with these peptide-MHC (pMHC) epitopes become activated, differentiate, and proliferate, producing up to 10^7 progeny cells to clear the pathogens [2, 3]. When the infection is cleared, many of the responding T cells die and leave behind memory cells. These include two recirculating subsets called central memory T cells (T_{CM}) and effector memory T cells (T_{EM}) and one subset called tissue-resident memory T cells (T_{RM}) that permanently remain in peripheral tissues. T_{CM} cells are defined by expression of CCR7 and CD62L in humans and CD44 and CD62L in mice whereas T_{EM} cells do not express CD62L. T_{CM} primarily recirculating through SLOs while T_{EM} recirculate through peripheral tissues. These memory T cells recirculate in order to provide immediate responses to subsequent exposure to the same pathogens.

In order for the immune system to able to mount robust T cell responses, the T cell repertoire must be able to respond to any given pathogenic antigen and respond vigorously enough to clear the pathogen. The naive T cell repertoire achieves this property by maintaining a high number of unique TCRs that must be distributed across a large but limited number of T cells. Detailed quantification of the T cell repertoire is emerging with the develop-

ment of advanced experimental and quantitative tools, but many questions about T cell repertoires are still open. How can we identify the numerous unique T cell clonotypes that respond to a given epitope? How many naive T cells does the body maintain, and how are the TCRs distributed across the naive pool? How is this diverse naive T cell compartment formed during development in neonates? In this thesis, we explore these questions by combining careful experimental design with mathematical and computational methods to quantitatively analyze and interpret immunological data.

1.2 T CELL RECEPTORS

1.2.1 *Structure of the aß T cell receptor*

The interaction between the αβ T cell receptor (henceforth referred to simply as TCR) and a peptide-MHC (pMHC) molecule is fundamental to the function of the adaptive immune system. The TCR is composed of an α chain (TCRα) and an β chain (TCRβ), each which have three hypervariable complementarity-determining regions (CDRs) that determine the set of pMHC antigens recognized by the TCR (Figure 1.1). In a TCR-pMHC interaction, the CDR1 and CDR2 loops primarily make contacts with the MHC portion of the pMHC molecule, and the CDR3 loops make contacts with the peptide portion of the pMHC [4, 5]. Since the CDR3 loops of the TCR have the most contact with the peptide, the CDR3 regions make the major contributions to determining the epitope specificity of the TCR [6]. When the TCR interacts with a cognate pMHC molecule, the TCR transmits signals to the CD3 complex and triggers cascades of signaling that lead to robust T cell responses [7].

Studies of many TCR-pMHC crystal structures have demonstrated that both the CDR3α and CDR3β loops interact with the peptide portion of pMHC molecules. In 34 different TCR-pMHC-I structures, both chains contribute to interactions between the TCR and the pMHC, as measured by buried surface area (BSA) (as percentages of BSA: TCRα, 33–78%; CDR3α, 4.6–34.7%; TCRβ, 22–67%; CDR3β, 8.3–42%) [9]. In 22 different TCR-pMHC-II structures, both chains also contribute to interactions between the TCR and the pMHC (as percentages of BSA: TCRα, 26.4–61.3%; CDR3α, 3–37%; TCRβ, 38.7–73.6%; CDR3β, 16.2–49.4%) [9]. These data demonstrate that both the TCRα chain and TCRβ chain (in particular, both the

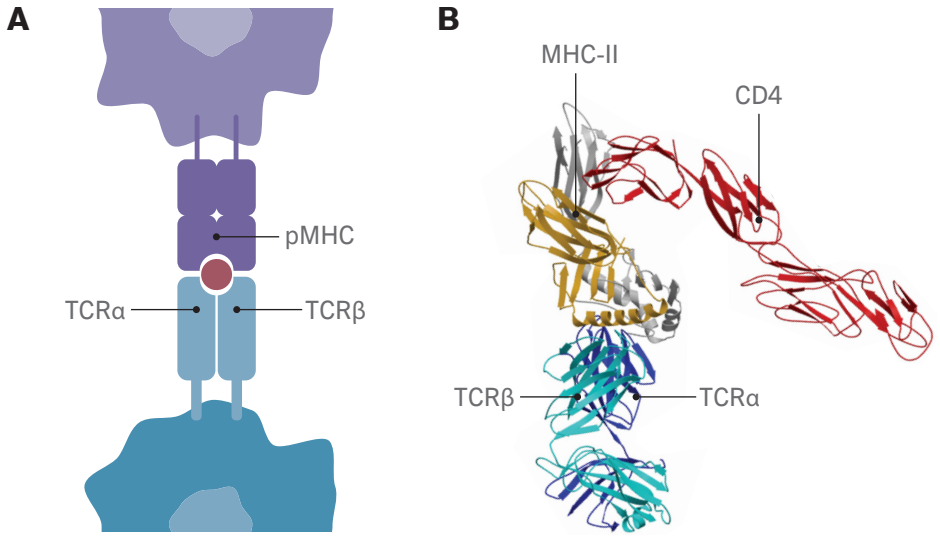


Figure 1.1: Structure of the αβ T cell receptor. (A) The αβ T cell receptor is a heterodimer of a TCRα chain and a TCRβ chain. The TCR interacts with peptides presented on MHC molecules (depicted here is a TCR from $CD4^+$ T cell interacting with pMHC-II). The CDR3 regions of both chains contribute to the interactions with the peptide portion of the pMHC epitope. **(B)** Crystal structure of a TCR–pMHC–CD4 complex where the TCRα chain is shown in dark blue and the TCRβ chain is shown in cyan (image modified from Li *et al*. [8], released under CC BY 4.0).

CDR3α and CDR3β regions) determine the antigen specificity of the TCR.

1.2.2 *Rearrangement of T cell receptor genes*

The wide diversity of receptors found in the TCR repertoire is created by a gene rearrangement process called V(D)J recombination that chooses random gene segments and combines them to create unique receptors (Figure 1.3). This process combines one random variable (V) gene segment to either one random J segment in the TCRα chain or one random diversity (D) and one joining (J) segment in the TCRβ chain [10] and is mediated by the RAG-1/RAG-2 enzymes [11]. In the germline configuration, the human TCRβ locus contains 76 different TRBV segments, 2 TRBD segments, and 14 TRBJ segments, and the murine TCRβ locus contains 35 different TRBV segments, 2 TRBD segments, and 14 TRBJ segments (Figure 1.2) [12]. The human TCRα locus contains 54 TRAV segments and 61 TRAJ segments, and the murine TCRα locus contains 98 TRAV genes and 60 TRAJ genes [12]. The TCRα locus also has one constant TRAC gene, and the TCRβ locus has

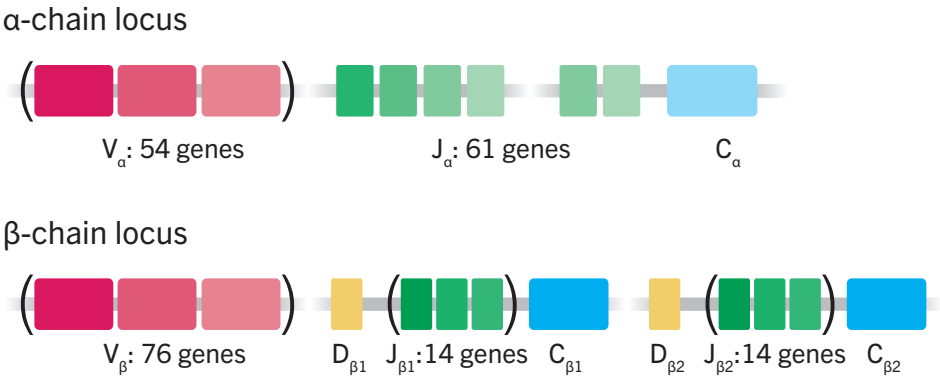


Figure 1.2: Germline configuration of the TCRα and TCRβ loci. The TCRα and TCRβ loci contain different V segments, D segments (in the TCRβ locus only), and J segments that are randomly chosen to form TCRs. In humans, the TCRα locus contains 54 V_α gene segments, 61 J_α segments, and one C_α segment. It is unknown how many of the V_α segments are functional. The TCRβ locus contains 76 V_β segments, and two clusters containing one D_β segments in front of 6–7 J_β segments and a single C_β segment. The TCRδ locus is found in between the V and J segments of the TCRα (not shown here).

two TRBC genes that are homologues with no functional differences; the C-region codes for the transmembrane polypeptides of the TCR.

Additional diversity is introduced at the junctions between the different gene segments by random addition and subtraction of nucleotides. One type of nucleotide, called N-nucleotides, are added to the junctions by the enzyme Tdt [13, 14]. N-nucleotides are non-template-encoded nucleotides that are randomly added to the junctions during the joining process of the gene segments. The *combinatorial* diversity produced by joining random V(D)J segments and the *junctional* diversity provided by these added/subtracted nucleotides at the junctions can create a theoretical diversity of ~10^{15} TCRs [15].

The CDR1 and CDR2 regions of the TCR chains are encoded entirely by the TRAV and TRBV genes [15]. The CDR3 region is encoded by the junction of the V(D)J segments [5]. Hence, the site of the most diversity—diversity that derives from the combinatorial diversity of the different V(D)J segments and junctional diversity from the addition and deletion of nucleotides—determines the peptide-recognizing loops of TCRs whereas the germline V segments predominantly determine MHC-specificity. Since the theoretical maximum diversity of ~10^{15} is much greater than the number of T cells in the mouse (~10^8), the T cell repertoire contains a small fraction of all of the possible CDR3 sequences.

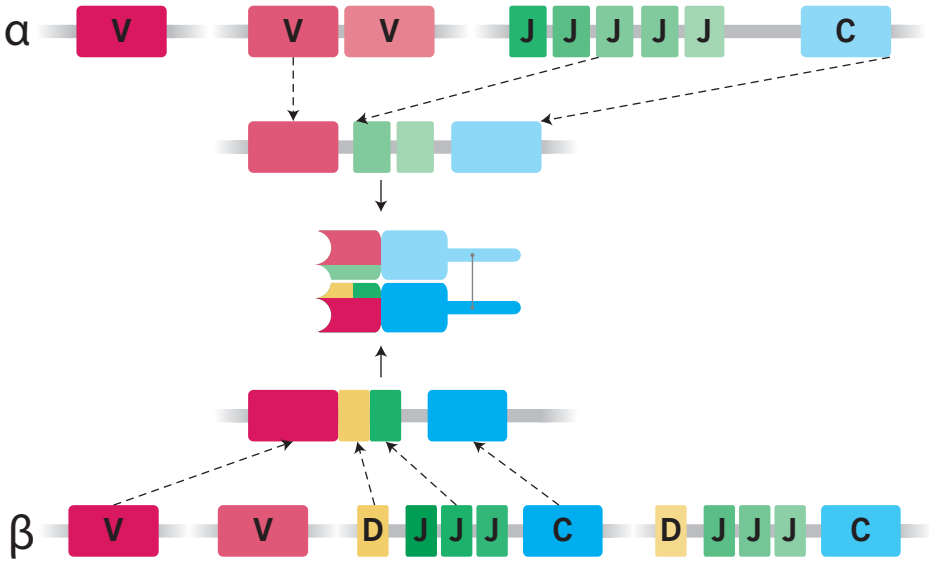


Figure 1.3: V(D)J recombination. TCRα and TCRβ are created by combining different segments together, and the resulting TCRα and TCRβ pair to form a complete TCR. In the upper half of the figure, the TCRα chain is formed by combining a V_α segment with a J_α segment and a C_α segment. In the lower half of the figure, the TCRβ chain is formed by combining a V_β segment with a D_β segment, J_β segment, and a C_β segment. The CDR3 of TCRα chain is encoded by the V segment, the J segment, and the junction between the two segments, and the CDR3 of the TCRβ chain is encoded by the V, D, and J segments and the VD and DJ junctions.

1.2.3 *Diversity of the TCR repertoire*

Direct measurement of the total diversity of the TCR repertoire has eluded the field due to two features of the naive T cell repertoire. The large number of unique TCRs and the skewed frequency distribution of these TCRs prevent a blood sample from representing the full diversity of the repertoire [16]. Estimates of the diversity of the human T cell repertoire have been made using various approaches, including spectratyping (described in Section 1.5.1) and TCRβ sequencing (described in Section 1.5.2). One of the first estimates was made by Arstila *et al.* [17], who sequenced a few hundred CDR3β sequences and then used spectratyping to extrapolate these numbers to the entire repertoire. Their approach yielded a lower-bound estimate of approximately 10^6 different CDR3β sequences. Robins *et al.* [18] used high-throughput sequencing (HTS) of the CDR3β and estimated the diversity in the peripheral blood to be $3–4\times10^6$ T cell clonotypes. Warren *et al.* [19]

found ~10^6 TCRβ sequences in the blood of one patient, corroborating this lower bound. Finally, Qi *et al.* [20] found a higher estimate of 100×10^6 unique CDR3β sequences in the naive T cell repertoire. The differences in these estimates reflect the difficulty of making inferences about the whole repertoire based on blood samples that capture only a small fraction of its diversity. The different methods used to extrapolate estimates of diversity in small samples to that of the full repertoire is out of the scope of this thesis (an overview is presented in Reference [16]).

Simply quantifying the number of unique T cell receptors in the T cell repertoire does not fully describe the diversity of the "peptide" repertoire to which the TCR repertoire is capable of responding [21]. The adaptive immune system does not have to create an unique TCR for each pathogenic peptide encountered by the body. Instead, a given TCR can respond to a range of different peptides through cross-reactivity. Cross-reactivity is mediated by mechanisms such as induced fit by pMHC ligands [22] and flexibility in the pMHC ligand to conform to the confirmation of a TCR [23]. Thus, cross-reactivities of the TCRs would need to considered to quantify the range of peptides that the TCR repertoire can respond to and to fully characterize the functional capacity of the TCR repertoire.

In addition, measurements of diversity of TCR repertoires will be affected by factors that are intrinsic and extrinsic to the individual that is studied. First, diversity differs between mice and humans in absolute numbers of unique TCRs and in clonal sizes. The murine T cell repertoire is estimated to have approximately 2×10^6 unique TCRs [24] that are distributed across approximately 10^8 T cells, and the human T cell repertoire has 10^8 unique TCRs [20] distributed across 10^{11} T cells. By dividing the number of T cells by the number of unique TCRs, clone sizes on average are expected to be roughly orders of magnitude of 10 cells in mice and 1000 cells in humans. Second, TCR diversity decreases with age [25], presumably due to decreases in thymic output from thymic involution and the expansion of existing memory T cells [26, 27]. Finally, the history of infections and antigen exposure of an individual is reflected in the TCR repertoire [28].

With the ability to quantify TCR diversity with all of the considerations mentioned in this section, we would be able to answer many fundamental questions about T cell mediated immunity. We would obtain quantitative characterizations of healthy T cell immunity and how changes diversity af-

fect the function immune system. For example, tracking changes in diversity with age could provide immense insights into how the function of the whole T cell population changes with aging. These types of insight will not only expand our knowledge of T cell physiology but also can motivate clinical applications in designing immunotherapies and other T cell dependent approaches.

1.3 T CELL DEVELOPMENT

The diverse set of TCRs produced by V(D)J recombination must be vetted so that the chosen TCRs can interact with the MHC molecules found in the body yet not react strongly against self peptides. This filtering process occurs in the thymus, where developing T cell precursors called thymocytes rearrange their TCRs and test their TCRs on self-pMHC molecules presented by thymic epithelial cells.

T cell development begins with in the bone marrow (BM), from which BM-derived precursors migrate to the thymus [29]. In the thymus, T cell progenitors rearrange their TCR genes to commit to either the γ:δ or the α:β lineage. They then sample self-peptide:self-MHC molecules to ensure MHC restriction through a process called positive selection and self-tolerance through a process called negative selection. Positive selection selects for thymocytes whose TCRs have specificity for the self-MHC molecules, and negative selection deletes thymocytes with TCRs that react strongly to self-peptides. These two selection processes create a repertoire of TCRs that can react to peptides presented by MHC molecules in the body but does not react strongly to self-antigens.

1.3.1 *Thymocyte development*

Development of the αβ T cell lineage in the thymus can be split up into three broad stages defined by the expression of CD4 and CD8: double-negative ($CD4^-CD8^-$), double-positive ($CD4^+CD8^+$), and single-positive ($CD4^+CD8^-$ or $CD4^-CD8^+$).

Progenitor cells enter the thymus at the cortico-medullary junction (CMJ) and begin as double-negative (DN) thymocytes [31]. In mice, the double-negative stage is split into four substages based on CD44 and CD25 expression. DN1 cells are $CD44^+$ and $CD25^-$, and DN2 cells are $CD44^+CD25^+$.

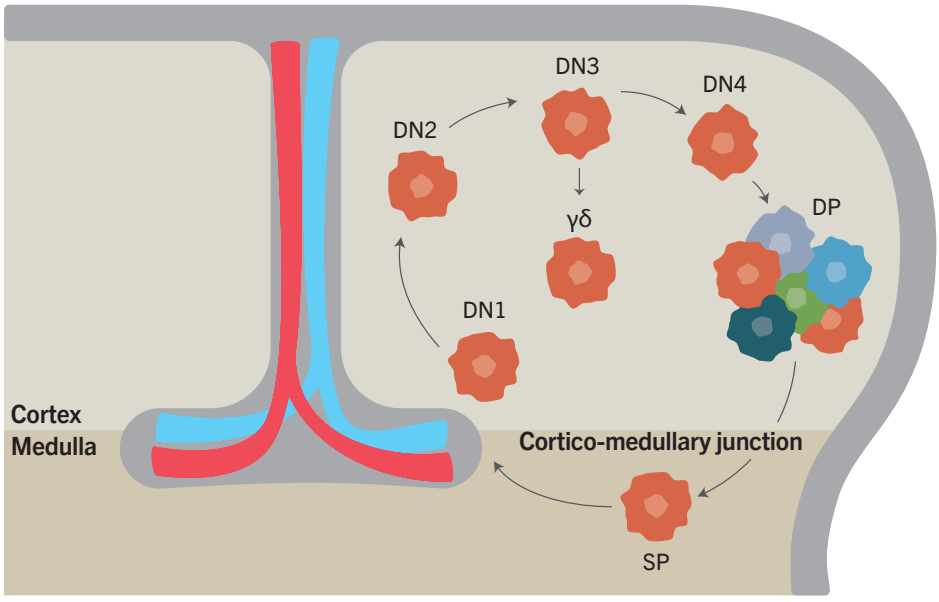


Figure 1.4: The journey of thymocyte development. BM-derived precursors enter the CMJ and start as DN thymocytes. DN thymocytes travel through the cortex, rearrange the TCRβ chain. Successful TCRβ rearrangement results in maturation into the DP stage, in which the coexpression of CD4 and CD8 and TCRα rearrangement occur. DP thymocytes undergo positive and negative selection in the cortex. Once a DP thymocyte receives a positive selection signal, it moves to the medulla and enters the SP stage, where it matures and undergoes negative selection. SP thymocytes that pass selection leave the thymus through the CMJ to enter the periphery. (Figure adapted from [30]).

These cells move from the CMJ to the cortex, where the thymocytes enter DN3 ($CD44^+CD25^+$) at the subcapsular epithelium. DN2 thymocytes begin rearranging their β, γ, and δ loci, and it is thought that these cells commit to the αβ or γδ lineages depending on whether a β chain or a functional γδ chain is rearranged first (γδ T cell development will not be discussed further) [32].

For the DN3 thymocytes that successfully rearrange their β chain first, the expressed β chain pairs with the invariant pre-T-cell α chain (pTα) to form a pre-T-cell receptor (preTCR), which marks the entrance into the DN4 stage ($CD44^-CD25^+$) [33]. The preTCR complex signals the downregulation of RAG1/2 to stop TCRβ chain rearrangement, signals the expression of CD4 and CD8, and causes the thymocytes to proliferate 6–9 times in mice [34, 35, 36, 37]. Thus, each DN3 thymocyte will yield many daughter DN4 cells with an identical TCRβ chain. The co-expression of CD4 and

CD8 define the beginning of the double positive (DP) stage.

Once proliferation from the DN4 stage has ceased and CD4/CD8 co-expression occurs, DP thymocytes will begin rearranging the TCRα loci on both chromosomes. When a thymocyte forms a complete TCR after successfully rearranging a TCRα locus, it begins to sample self-pMHC presented by thymic epithelial cells. TCRs must recognize self-peptide:self-MHC to pass positive selection, ensuring that the TCR can interact with the MHC molecules used by the body. A DP thymocyte makes multiple attempts at rearranging the TCRα genes until it passes positive selection or until it receives a signal to die. Since thymocytes attempt to rearrange the TCRα locus on both chromosomes until they receive a positive selection signal, T cells can have two in-frame TCRα rearrangements, which we will discuss in more detail in Section 5.1.3. DP thymocytes that pass positive selection then down-regulate either CD4 or CD8, move from the thymic cortex to the medulla, and become CD4 single-positive (SP4) or CD8 single-positive (SP8) thymocytes. In the medulla, SP thymocytes undergo negative selection, in which thymocytes with TCRs that bind too strongly to self-pMHC are signaled to die. Studies have estimated that 75–80% of DP thymocytes fail positive selection [38, 39] and that 20–50% of the positively-selected thymocytes survive negative selection [38, 39, 40], yielding approximately 5% of DP thymocytes that survive selection. Finally, thymocytes that pass both positive and negative selection undergo 2–4 divisions before leaving the thymus [38]. Key events in murine thymocyte development are shown in Figure 1.5.

The DP stage can divided into three stages based on expression levels of TCR and CD5 [41]. DP1 thymocytes are $TCR^{lo}CD5^{lo}$ and consist of thymocytes that have not undergone selection. DP2 thymocytes are $TCR^{int}CD5^{hi}$ and consist of a mix of thymocytes that give rise to CD4 and CD8 lineages. DP3 thymocytes are $TCR^{hi}CD5^{int}$ and consist of just cells of the CD8 lineage. Mathematical modeling by Sinclair *et al.* has shown that DP1 cells give rise to DP2 cells, which give rise to DP3 cells; SP4 thymocytes differentiate from DP2 thymocytes only, and SP8 thymocytes differentiate from DP3 thymocytes only [38]. They estimated residence times of 3.5, 1.4, and 7 days for DP1, DP2, and DP3 thymocytes respectively.

SP thymocytes can be divided into two stages based on CD62L and heat-stable antigen (HSA) expression: immature SPs ($CD62L^{lo}HSA^{hi}$) and mature SPs ($CD62L^{hi}HSA^{lo}$) [42, 43]. SP cells that are ready to emigrate from the

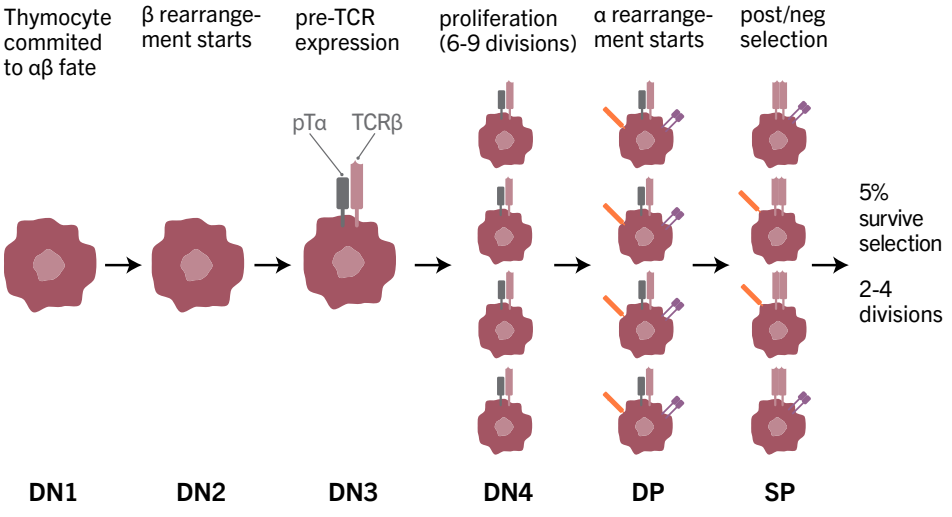


Figure 1.5: Key events in the different stages of thymocyte development. The 6–9 divisions in the DN4 stage results in multiple daughter cells with the same TCRβ chain, each of which then undergoes independent TCRα rearrangement. Most TCRs produced by V(D)J rearrangement will not pass selection. The ~5% of DP thymocytes that survive selection and become SP thymocytes, which can undergo 2–4 divisions before emigrating from the thymus.

thymus express Kruppel-like factor 2 (KLF2), which is a transcription factor that drives the expression of sphingosine-1-phosphate receptor-1 (S1P1). The mature SP thymocytes are associated with a phenotype that is less susceptible to apoptosis and poised to proliferate; only mature SP thymocytes emigrate from the thymus [43]. Mature SP thymocytes, by expressing S1P1, follow a S1P gradient created by neural-crest-derived pericytes and leave the thymus through the CMJ [44]. SP4 and SP8 thymocytes have been estimated to have residence times of 4.4 and 4.6 days respectively [43, 41] whereas Sinclair *et al.* [38] estimated SP4 and SP8 residence times of 5 and 3.7 days respectively.

Human thymocyte development also follow the progression from DN to DP to SP, but subdivisions within these stages are not as well-characterized compared to murine thymocyte development. The double negative stage is subdivided into three stages based on CD45, CD38, and CD1a expression. DN1 thymocytes are $CD34^{+}CD38^{-}CD1a^{-}$ and are equivalent to murine DN1 thymocytes. DN2 thymocytes are $CD34^{+}CD38^{+}CD1a^{-}$, and DN3 thy-

mocytes are $CD34^+CD38^+CD1a^+$; these latter two stages correspond to the murine DN2, DN3, and DN4 stages. The double positive stage is subdivided into two stages based on CD3 expression. $DP3^-$ thymocytes ($CD4^+CD8^+CD3^-$) differentiate into $DP3^+$ ($CD4^+CD8^+CD3^+$) thymocytes, which become single positive thymocytes that are $CD3^+$ and either $CD4^+$ or $CD8^+$.

1.3.2 *Development of peripheral T cell compartments in neonatal mice*

The development of peripheral T cell compartments depends on the export of T cells from the thymus and possibly the expansion of peripheral T cells [45, 46]. Peripheral T cell populations in mice grow in numbers from $\sim 0.02 \times 10^6$ cells at birth to $\sim 50 \times 10^6$ cells at 2 months of age [45]. It has been suggested that thymic export is more important than peripheral expansion in establishing peripheral T cell compartments [45], but the presumed lymphopenic environment of neonatal mice may also support peripheral expansion through a mechanism called lymphopenia-induced proliferation (LIP) [46].

LIP occurs when naive T cells are placed in a lymphopenic environment—whether naturally occurring following an insult to the immune system such an HIV infection or experimentally introduced through drugs or irradiation. The "open space" in the lymphopenic environment causes naive T cells to proliferate and take on a memory-like phenotype. This process is driven by signals such as IL-7 and self-pMHC complexes, and it is thought that naive T cells have greater access to these signals in a lymphopenic environment [47].

These observations led to the question of whether LIP occurs in neonates as a physiological mechanism for rapidly growing T cells compartments. Since the peripheral T cell compartments in neonates contain very few T cells, Min *et al.* [46] asked whether the neonatal environment supports LIP. They adoptively transferred CFSE-labeled $CD4^+$ T cells from adult mice into 1 day old pups and observed 17–19 days after the transfer that a significant proportion of transferred cells divided many times and were $CD44^{bright}$. Similar results occurred when adult $CD44^{dull}$ $CD4^+$ T cells and $CD4^+$ SP thymocytes from 2-week old mice were transferred into 1-day old neonates. These observations indicate that neonates have lymphopenic environments that allow for LIP of $CD4^+$ T cells. However, most of the data from this study looked at transferred T cells from older animals and thus do not nec-

essarily indicate that peripheral expansion is a physiological mechanism for the development of neonatal peripheral T cell compartments. As argued by others, too much peripheral expansion in relation to thymic output during T cell ontogeny would result in a T cell repertoire with relatively few over-represented clones that is not diverse enough for adequate protection [45]. In Chapter 3, we attempt to quantify the balance between thymic output and peripheral expansion in naive T cells during ontogeny.

1.4 RECIRCULATION OF T CELLS

1.4.1 *Anatomy of the lymphatic system*

The lymphatic system is responsible for returning interstitial fluid from tissues into systemic circulation, absorbing digested fats from the intestines, and playing a major role in the immune system. The major secondary lymphoid organs (SLOs) of the mouse are the lymph nodes, spleen, and Peyer's patches. In lymph nodes, afferent lymphatic vessels drain fluid from surrounding tissues and bring in pathogens and antigen-bearing DCs. Lymphocytes enter the lymph node from the blood through walls of blood vessels called high endothelial venules. Peyer's patches are specialized lymph-node-like structures in the small intestine where antigens from the gut are sampled in order to develop protective mucosal immunity. Lymphocytes enter Peyer's patches through high endothelial venules, and efferent lymphatics connect naive and activated lymphocytes to the mesenteric lymph nodes. The spleen does not have any direct connections to the lymphatic system; lymphocytes enter and leave the spleen through the blood.

Interstitial fluid from tissue is drained by afferent lymphatics to lymph nodes, which in turn are drained by efferent lymphatics into larger lymphatic vessels that return lymph to the blood [48]. In both humans and mice, the thoracic duct is the biggest vessel of the lymphatic system and collects lymph from the body (approximately 3/4 of all interstitial fluid) other than the right thorax, arm, head, and neck. The latter regions are drained by the right lymphatic duct (the other 1/4 of all interstitial fluid). The thoracic duct travels from the abdomen and ascends to the thorax, where it drains into the venous angle of the left subclavian vein and the internal jugular veins. The mesenteric lymph nodes drain into the thoracic duct. In mice and rats, other lymph nodes, such as the inguinal, axillary, and cervical lymph nodes,

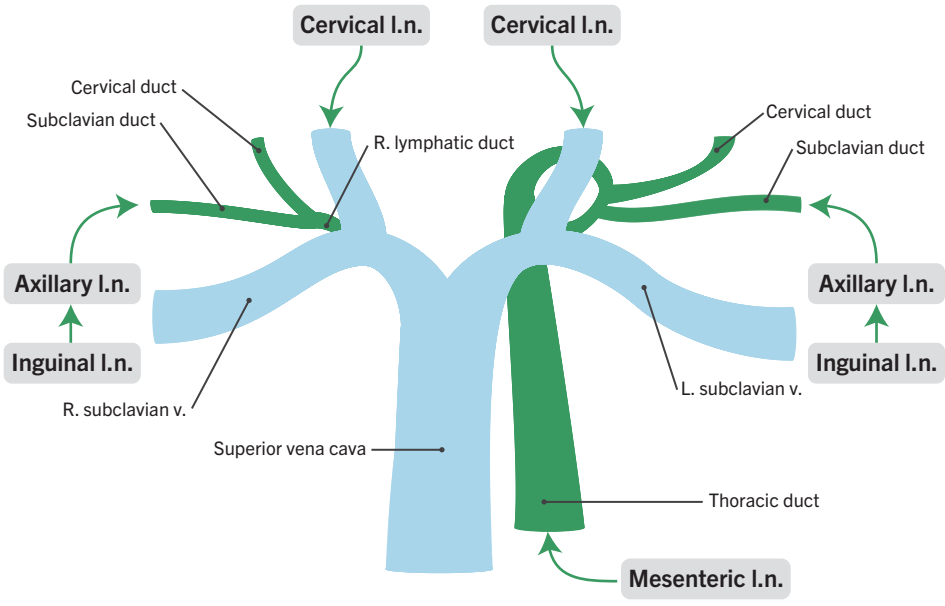


Figure 1.6: An illustration of the lymphatics of the thorax and neck of the mouse. This figure illustrates how inguinal, axillary, cervical, and mesenteric lymph nodes are drained by the lymphatic system and how the draining major lymphatic vessels enter back into venous blood [50, 48, 49].

are drained by other large ducts that return lymph back to the system circulation without connecting to the thoracic duct [48]. In the rat, the right inguinal lymph nodes drain into the axillary lymph nodes, which then drain through the subclavian duct into the right lymphatic duct before returning back into venous blood [48].[1] The left inguinal lymph nodes drain into the axillary lymph node, which then drain into the subclavian duct to return to venous blood in the left subclavian vein [48]. The cervical lymph nodes have their own cervical ducts on either side that eventually drain into the systemic circulation via the subclavian veins. We assume that the anatomy of the lymphatics of rat and mice are similar (Figure 1.6) [50].

[1]This contrasts with humans, where the inguinal lymph nodes are drained by the external and common iliac lymph nodes, which eventually drain into the cisterna chyli and the thoracic duct [49].

1.4.2 *Recirculation of T cells*

The diverse naive T cell repertoire can recognize virtually any pathogen, but the naive T cells must come in contact with their cognate antigen to mount a response to foreign antigens. Naive T cells must be constantly surveying different foreign antigens in order for this to happen. Since these interactions happen in the SLOs—particularly in lymph nodes—T cell recirculate from lymph node to lymph node in order to survey many antigens. This is especially true since estimates for the number of naive T cells that can respond to a given foreign antigen is 200–1400 in mice [2, 51, 52], and these antigen-specific precursors will be distributed across ~30 lymph nodes [38, 50], the spleen, and other SLOs. Thus, naive T cells must recirculate between blood and SLOs in order for all of these antigen-specific precursors to be recruited and respond to a pathogenic insult [53, 54] Central memory T cells and some effector memory T cells also recirculate between blood and SLOs [55]. T_{CM} cells, which express CD62L, enter lymph nodes through HEVs and spend 6–16 hours sampling different antigens before entering the efferent lymphatics to return to the systemic circulation [56]. T_{EM} cells migrate between the spleen, blood, and peripheral tissues. These memory subsets contrasts with T_{RM} cells, which permanently reside in peripheral tissues of the site of infection [55]. By recirculating from site to site, T cells are able to sample different antigens, facilitating the recognition of foreign antigens by T cells to mount an efficient response. We attempt to quantify the number of circulating T cells in the SLOs of the mouse in Chapter 4.

1.5 SEQUENCING

Numerous techniques exist for identifying TCR sequences and probing the diversity of the repertoire of T cell populations. These methods can be divided into four broad categories: anti-TRBV antibodies and spectratyping, TCRα- or TCRβ-only sequencing, single-cell paired TCRα/TCRβ sequencing, and statistical multi-cell TCRα/TCRβ sequencing.

1.5.1 *Anti-V-segment antibodies and spectratyping*

Flow cytometry with monoclonal antibodies against TRBV segments and spectratyping were commonly used to probe the TCR repertoire before the

emergence of high-throughput sequencing (HTS) techniques. One of the original techniques employs a panel of antibodies against TRBV segments of interest and uses flow cytometry to quantify the proportion of T cells expressing each of those TRBV segments. Although relatively simple to implement, this approach suffers from two main drawbacks. First, identifying TRBV segments used in a TCR does not identify the specificity-conferring CDR3 sequences. Second, a panel of anti-TRBV antibodies cannot identify all TRBV subfamilies since only a limited number of TRBV antibodies is commercially available for mice and for humans [57].

To partially overcome these problems, CDR3 spectratyping (also known as Immunoscope) took repertoire analysis further by identifying the distribution of CDR3 lengths found in a sample of T cells [58]. Spectratyping can be performed on entire T cell populations or in conjunction with anti-TRBV antibodies to identify CDR3 length distributions with TRBV subfamilies. The many variations of this technique all involve using PCR to amplify the CDR3β region of the mRNA coding for the TCRβ chain with primers specific for each TRBV segment and a primer specific for the TRBC region. Since β-rearrangement creates variability in CDR3 length due to the addition of N-nucleotides during V(D)J rearrangement, CDR3β lengths can vary by up to nine amino acids, and thus a collection of many different T cell clonotypes should yield a distribution of CDR3β lengths. The resulting PCR products of different lengths are then separated and quantified by using electrophoresis or by labeling products with fluorochromes and analyzing with gel readers. The data are a series of 8–9 peaks of the frequency of PCR products of a given length (which have been incorrectly assumed to be normally distributed) [58], and the area of each peak gives the frequency of PCR products associated with a given CDR3 length.

These two techniques have been used to answer questions about TCR repertoire diversity and probing changes in T cell clonality in disease states. Studies have estimated the diversity of the human αβ TCR repertoire by using antibodies against specific V segments [17, 24]. Many studies have used spectratyping to associate perturbations in the distribution of CDR3 lengths with disease states (reviewed in Reference [59]). Although these two techniques probe information about TCRs indirectly, they were instrumental in the quantification of clonal diversity before HTS became available.

1.5.2 *TCRa- and TCRß-sequencing*

Obtaining the nucleotide and amino acid sequences of the CDR3 regions of either only the TCRα chains or only the TCRβ chains used in T cell populations has become common with the advent of efficient and economical HTS techniques [18, 60, 61, 62, 63]. Standard sequencing techniques cannot obtain paired CDRα and CDRβ sequences from a sample of T cells. Since the TCRα and TCRβ loci are found on different chromosomes, the DNA or mRNA encoding these genes are not anchored together and thus bulk sequencing of a T cell population loses pairing information. This technical obstacle has limited most studies to sequencing only one of the chains, typically the TCRβ chain (and from this point on, single-chain TCR sequencing will be referred as TCRβ sequencing).

Many of the technical challenges in TCRβ sequencing, such as design of primers, reduction in amplification bias, choices in HTS technologies, and data processing and aligning of raw sequence reads, have been solved (reviewed in [64]). TCRβ sequencing has been used to answer many questions about TCR repertoires, such as estimating the total number of αβ TCRs found in the naive repertoire [18]. Another study utilized TCRβ sequencing to study the dynamics of the TCRβ repertoire in T cells responding to YF-17D yellow fever vaccine in human patients [62]. TCRβ sequencing was a significant progression from antibody and spectratyping techniques because it directly identifies the peptide-recognizing region of TCRs rather than yielding indirect measures of TCR diversity.

However, single-chain sequencing loses information about TCRs since both chains of the TCR contribute to its antigen specificity. As we will discuss in Section 5.1.3, different clones can have TCRs that share the same CDR3 sequence in one of the chains, and this information is lost by sequencing only one of the TCR chains.

1.5.3 *Single-cell sequencing*

Many studies have developed methods to identify paired TCR sequences by sequencing single cells. One of the first single-cell sequencing studies was described by Dash *et al.* [65], where they performed single-cell sorting on murine K^bPB1_{703}-specific $CD8^+$ T cells using tetramer staining. cDNA libraries were made using single-cell multiplex reverse transcription poly-

merase chain reaction (RT-PCR) with TRAV-, TRBV-, TRAC-, and TRBC-specific primers, and high-throughput sequencing was used to obtain CDR3α and CDR3β sequences. Similar approaches were used in both human and murine T cell populations [66, 67], and analogous approaches have been used for identifying paired heavy-chain and light-chain CDR3 sequences found in B cell populations [68]. Han *et al.* [69] extended single-cell sequencing by including the sequencing of other functional genes such as those coding for cytokines and transcription factors, allowing for the identification of paired CDR3α/CDR3β sequences and the phenotypes of the sampled T cells. Software has also been developed to identify TCR sequences (called TraCeR [70]) and BCR sequences (called BraCeR [71]) from single-cell RNA-seq (RNA sequencing) data.

DeKosky *et al.* [72] avoided using single-cell sorting by using of customized polydimethylsiloxane slides with 125 pL wells to isolate B cells with a 95% probability of wells containing exactly one cell. Turchaninova *et al.* [73] created single-cell emulsions and performed linkage RT-PCR in the emulsion droplets, where TCRα and TCRβ transcripts were fused before being sequenced. DeKosky *et al.* [74] extended their approach by using a flow-focusing apparatus to create a single-cell emulsification, performing linkage RT-PCR to link light-chain and heavy-chain transcripts, and then sequencing these linked transcripts. McDaniel *et al.* [75] designed and carefully described another flow-focusing device to emulsify many single cells for single-cell sequencing.

The drawbacks of these techniques include (*i*) limited scalability, which risks the undersampling of rare clones and thus underestimating diversity, (*ii*) imprecise and missing information regarding clonal frequencies, and (*iii*) the need to use customized equipment [72, 74, 75].

1.5.4 *Statistical multi-cell sequencing approaches*

To overcome the sample size limitations of single-cell sequencing approaches, an alternative strategy is to use a statistical method to associate CDR3α and CDR3β pairs from sequences obtained from multiple subsamples of T cells sampled from a target T cell population. Methods that employ this strategy capitalize on the fact that chains from the same TCR will tend to appear together in samples and use the frequencies of these co-occurrences to as-

sociate them together. Three approaches using "frequency-based pairing" currently exist: a preliminary attempt used to pair B cell receptors [76], a commercial methodology called pairSEQ developed by Adaptive Biotechnologies [1], and our open source methodology, presented in Chapter 5 [77].

The study by Reddy *et al.* [76] attempted a rudimentary version of frequency-based pairing in B cells by pairing the most abundant V_L and V_H CDR3 sequences by matching their relative frequencies in the sampled repertoire. Howie *et al.* [1] developed a more sophisticated method called pairSEQ that samples T cells into different subsets and then uses combinatorics and statistics to determine paired TCRα/TCRβ sequences. Finally, our approach called ALPHABETR [77] uses an experimental design of sequencing multiple subsamples of an antigen-specific T cell population and uses an algorithm to determined paired TCRβ/TCRβ sequences. ALPHABETR has been designed to capture characteristics specific to antigen-specific T cell repertoires that have not been explicitly considered in other studies. These frequency-based pairing methods are discussed in more detail in Chapter 5.

1.6 MATHEMATICAL MODELING IN IMMUNOLOGY

1.6.1 *Why do we need mathematical modeling?*

As immunological experiments produce more detailed and complex data, analyzing and interpreting the results of these experiments requires careful and reproducible quantitative approaches. There are least two uses for mathematical models: (*i*) estimation of parameters representing biological processes that cannot be directly measured experimentally and (*ii*) testing of hypotheses about underlying biological mechanisms that cannot be achieved experimentally. These two purposes are not necessarily mutually exclusive and often occur in the same modeling effort. In mathematical modeling efforts in T cell biology, parameters for T cell population dynamics such as lifespans and proliferation rates of naive and memory $CD4^+$ and $CD8^+$ T cells have been gleaned from bone-marrow chimeras in mice [78, 79], deuterium labeling experiments in mice [80], deuterium-labeling experiments in humans [81, 82, 83, 84], deuterated-glucose-labeling experiments in humans [85, 86, 87, 88, 89], and experiments using adoptive-transfers of CSFE-labelled T cells experiment [90, 91]. Mathematical models have also partially elucidated the structure of the $CD4^+$ and $CD8^+$ memory compartments

in mice, including testing different hypotheses regarding pathways of differentiation from naive to memory to effector T cells [92, 79] and clarifying the nature of heterogeneity in division and death rates found in T cell compartments [93, 89, 94, 95].

Interpretation of data using model-free intuition and simple statistical tests can often lead to erroneous conclusions due to implicit assumptions made in these interpretations [96, 97]. Mathematical modeling attempts to overcome these problems by making the assumptions explicit when the models are constructed. By finding models that explain the data of interest, we create mathematical descriptions and interpretations of the data that are easy to scrutinize. In addition, the process of mathematical modeling often find many mathematical models that cannot explain the data, which also can provide insight into interpretations about the data [98]. By combining careful mathematical modeling with good experimental design, more information can be drawn from data to make richer inferences and more informed predictions than possible with standard statistical analyses of data.

1.7 AIMS

The aims of this thesis are to provide quantitative descriptions the development, size, and diversity of T cell populations. We first develop mathematical models that describe and quantify the ontogeny of naive T cell populations. We then describe an attempt to quantify the sizes of various lymphocyte compartments using a thoracic duct cannulation technique in mice to collect circulating lymphocytes. Finally, we develop tools that are efficient at identifying paired TCRα and TCRβ sequences. Our attempts to quantify the sizes of T cell populations, to describe the mechanisms for the development of these populations, and to provide tools for identifying the richness of repertoires of these populations all serve the overarching goal of quantitatively characterizing effective T cell immunity.

CHAPTER 2

Materials and Methods

2.1 MICE

Male C57BL/6 mice were purchased from Harlan/Envigo and maintained in individually ventilated cages prior to surgeries. All protocols were approved by the local ethical committee and conducted under licenses issued by the UK Home Office, project license number PPL60-3822.

2.2 CELL ISOLATION AND MATERIALS

2.2.1 *Media*

Cells were prepared and washed in FACS buffer, which is Dulbecco's Phosphate-Buffered Saline (PBS) (Gibco) supplemented with 2% inactivated fetal bovine serum (Invitrogen). Fetal bovine serum was added to reduce non-specific binding by the antibodies used for staining for flow cytometry.

2.2.2 *Thoracic Duct Cannulation*

Male 12–18 week old mice were anesthetized with continual inhalation of isoflurane (Abbot Animal Health). The mice were gavaged with 200 μL of olive oil in order to visualize the lymphatics. After the induction of anesthesia, a lateral incision was made through the skin and muscle of the left side of the mouse just under the ribcage. Blunt dissection was used to visualize the thoracic duct. A small cut in the thoracic duct was made with microscissors, and a polyurethane cannula (2 Fr, Linton Instrumentation) was inserted into the duct. The cannula was secured to the site of insertion with super glue (Loctite), and then the muscle layer was sutured. The cannula was fed through a wire tube on the back of the mouse attached to a harness worn by the mouse in order to avoid biting and tearing of the cannula. For 18–24 hours after the surgery, lymph was collected on ice in 1 mL sterile PBS (Gibco) with 20 U/mL heparin sodium (Wockhardt).

2.2.3 *Lymph nodes and spleens*

Spleens and peripheral lymph nodes (cervical, axillary, inguinal, mesenteric) were dissected from mice euthanized with carbon dioxide inhalation and were teased between a pair of frosted slides to obtain single-cell suspensions. Cells were washed with FACS buffer and centrifuged at 340×g, 4°C. Cells were resuspended in FACS buffer and then filtered through a 40μm cell sieve (Greiner) in order to remove tissue debris. Aliquots of the samples were used for cell counting with the MACSQuant Analyzer 10 flow cytometer (Miltenyi Biotec). Cells were counted with a Neubauer Chamber hemocytometer before being stained with antibodies for flow cytometry analysis.

2.2.4 *Lymph*

Collected lymph was filtered through a 40 μm cell sieve (Greiner), and a small aliquot was used for cell counting with the MACSQuant Analyzer. The remaining lymph was washed with FACS buffer and centrifuged at 340×g, 4°C. The cells were then resuspended in FACS buffer and then used for antibody staining and analysis. Cells were counted with a Neubauer Chamber hemocytometer before being stained with antibodies for flow cytometry analysis.

2.3 FLOW CYTOMETRY PROTOCOLS

2.3.1 *Cell counting*

Cell counting was performed with the volumetric MACSQuant Analyzer 10 flow cytometer. 100–200 μL aliquots of each sample were diluted with FACS buffer, and 7-AAD viability dye (BioLegend) was used for viability staining. A fixed volume was analyzed by the MACSquant flow cytometer, and the number of events in the lymphocyte forward-scatter/side-scatter gate and live gate were used to count the number of cells.

2.3.2 *Antibodies*

Cells were stained with an antibody panel for T cell markers or stained with a panel for B cell markers. The following fluorochrome-conjugated antibodies

were used for T cell markers: BV421-CD4 (RM4-5), BV510-CD8 (53-6.7), FITC-CD44 (IM7), APC-CD62L (MEL-14), PE-CD122 (5H4), PerCP/Cy5.5-TCRβ (H57-597), PeCy7-CD25 (PC61) (BioLegend). The following fluorochrome-conjugated antibodies were used for B cell markers: BV421-IgD (11-26c.2a), BV510-CD23 (B3B4), APC-CD93 (AA4.1), PE/Cy7-IgM (RMM-1), FITC-B220 (RA3-6B2) (BioLegend). Viability staining was performed with Fixable Viability Dye eFluor 780 (eBioscience), which was used in a concentration of 1 μ/mL and mixed with the antibody panel in PBS.

2.3.3 *Staining for cell surface markers*

1×10^6 cells were incubated with 50 μL of the panel of the flurochrome-conjugated antibodies and the viability dye on ice from 20 min in the dark in a 96-well round-bottom polystrene plate. The cells were washed twice with FACS buffer. Data were acquired on a LSR II or a LSR-Fortessa flow cytometer (BD Biosciences) for analysis and analyzed using FlowJo 10.4.2 (Treestar).

2.4 ONTOGENY EXPERIMENTS

The experiments described in Chapter 3 were performed by Thea V. Hogan. Wildtype male and female mice (C57BL/6 background) were bred and maintained in conventional pathogen-free conditions at the animal facility at the Royal Free Campus of University College London. All experiments were performed in accordance with UK Home Office regulations, project license number PPL70-8310. Mice younger than 14 days old were euthanized by decapitation, and older mice were euthanized by either cervical dislocation or by carbon dioxide inhalation. Lymph nodes and spleens were collected from these mice. Cells were counted with a CASY Counter (Omni Life Sciences) and stained with the following antibody-panel: BV510-CD5, PerCP-Cy5.5-TCRβ, BV711-CD4, BV785-CD44 (BioLegend), FITC-Ki67, eFluor450-HSA, APC-FoxP3, PE-CD25, PE-Cy7-NK1.1, biotin-CD122 (eBioscience), BUV395-CD8, BUV737-CD62L (BD Biosciences), PE-Texas-Red-Streptavidin, and LIVE/DEAD nearIR (Invitrogen). Data were acquired on a LSR-Fortessa flow cytometer (BD Biosciences) and analyzed using FlowJo 9.9.6 (Treestar).

2.5 SINGLE-CELL SEQUENCING

The single-cell sequencing experiment described in Chapter 5, Section 5.1.4 was performed by Jeff E. Mold. A human volunteer was identified as HLA-$A2^+$/HLA-$B7^+$ and received the live attenuated yellow fever vaccine (YFV-17D). On day 15 post-vaccination, peripheral blood samples were taken, and live $CD3^+CD8^+$ T cells were isolated by negative selection using magnetic columns (Miltenyi Biotec, $CD8^+$ T cell negative isolation kit). Cells were labeled with a panel of antibodies and the HLA-A02:01/LLWNGPMAV dextramer representing the immunodominant response [99]. Single dextramer-specific $CD3^+CD8^+$ T cells were sorted into individual wells in 96 well plates containing a lysis buffer (0.4% Triton, RNAse inhibitor, dNTP, OligodT) and immediately stored on dry ice. Single cell transcriptome libraries were subsequently generated from these cells using an adapted version of the SMRT-Seq2 protocol [100]. Libraries were prepared for sequencing by tagmentation and labeling individual single cell transcriptomes with a custom Tn5 enzyme [101] and Nextera XT dual indexes. Pooled libraries were then sequenced using an Illumina Hiseq2500 on high output mode (2×100bp or 2×125bp reads), and individual CDR3α and CDR3β sequences were identified using the MiTCR algorithm with default parameters [102]. The default settings for MiTCR were used to align the CDR3 sequences. These were then manually filtered to remove erroneous sequences (e.g. early stop codons and CDR3 sequences that were greater than 30 amino aids in length), and then BLAST was used on the remaining sequences to check for mapping to other parts of the genome, removing as appropriate. All clones used in the comparative analysis of CDR3α lengths in Section 5.1.4 were curated manually to exclude the possibility of contaminating TCR sequences.

These experimental procedures were approved by the Regional Ethical Review Board in Stockholm, Sweden: 2008/1881-31/4, 2013/216-32, and 2014/1890-32.

Figure 2.1: A simple model for peripheral T cell dynamics. This figure is a schematic of a simple model that describes thymic export of T cells into the periphery, division, and cell death.

2.6 MATHEMATICAL AND STATISTICAL ANALYSIS

A range of mathematical and statistical methods are used in this thesis to analyze and to make inferences from data. Mathematical modeling centered on creating ordinary differential equation (ODE) models that represent the important processes and mechanisms occurring in different cell subsets, and we will briefly discuss how these equations are formed in Section 2.6.1. For parameter estimation and model selection, we used a combination of frequentist and Bayesian statistical approaches. We will briefly describe null-hypothesis significance testing and bootstrapping in Sections 2.6.2–2.6.3 and spend more time describing Bayesian approaches in Section 2.7.

2.6.1 *Differential equations*

Creating a differential equation model first involves identifying the key processes that are involved in the biological system of interest. For example, suppose we want to create a simple model for the number of T cells in the periphery over time. We assume that the thymus exports T cells into the periphery at a fixed rate and that T cells die at some constant rate. Let's represent these rates with variables:

- Thymic export: let θ be the number of T cells per day that the thymus exports into the periphery.
- Cell death: let δ rate of death of T cells where $1/\delta$ is the mean lifetime; this is one of many ways to model death where cells are lost with first order kinetics and their lifetimes are exponentially distributed.

We can draw a schematic of these two processes as arrows entering or leaving a box that represents the number of N peripheral T cell (Figure 2.1).

We now can write down a differential equation that contains mathematical representations of these processes and describes the rate of change of the number of T cells N over time, namely $\frac{dN}{dt}$.

- Thymic export: the number of cells per day entering the T cell compartment from the thymus is simply θ
- Cell death: the number of cells per day that are dying is δ multiplied by the number of T cells, namely δN.

We put these terms together to get the differential equation

$$\frac{dN}{dt} = \theta - \delta N\,.$$

Note the negative sign in the second term since cell death results in a loss of cells. The solution of this equation either is determined analytically or estimated with numerical techniques with software.

2.6.2 *Null hypothesis significance testing*

Null hypothesis testing was performed using the one-sample Student's t test or the Mann-Whitney U test. These tests were performed with R 3.4.3 [103]. p values less than 0.01 were considered statistically significant.

2.6.3 *Bootstrapping*

We use a statistical tool called the bootstrap to quantify the uncertainty around statistical estimates of data. Let $X = (x_1, x_2, \ldots, x_n)$ be n data points, and suppose we want to quantify the uncertainty on some statistical estimate θ of the dataset X. The bootstrap procedure begins by creating new datasets of size n by sampling with replacement from X. Each one of these resampled datasets is called a bootstrap sample, and we create a large number of these bootstrap samples, say 1000. For each bootstrap sample i, we calculate the statistical estimate θ_i^* for that sample and obtain 1000 bootstrap replicates θ_1^*, θ_2^*, …, θ_{1000}^*. This set of bootstrap replicates allows us to make inferences about our data. For example, determining the 2.5 percentile and 97.5 percentile of the replicates defines a 95% confidence interval for θ, and taking the standard deviation of the bootstrap replicates gives us the standard error for θ.

2.7 BAYESIAN DATA ANALYSIS

2.7.1 *Advantages of Bayesian statistics*

Two broad approaches to statistics and interpretations of probability exist. The *frequentist* approach views probability as long-run frequencies over a large number of repetitions of an experiment; a unbiased coin that has a 50% chance of showing heads after a toss means that the proportion of tosses showing heads will approach 0.5 with a large numbers of trials. The *Bayesian* approach views probability as a degree of plausibility for the event that we are interested in; in its essence, Bayesian inference counts the number of ways an event can occur, and events that occur in more ways are more plausible. These descriptions can feel abstract with no practical relevance, but the two different approaches to statistical inference give rise to profound differences in how we quantify and understand interpretations made in data analysis.

The interpretation of frequentist statistics often relies on imagining many multiple samples of the data, repeating an experiment many times until we see a pattern. For example, suppose we compute a frequentist 95% confidence interval on a parameter for an experiment. The strict interpretation is that if we were to repeat the study many many times, then 95% of the computed confidence intervals from the repeated studies contain the true value of the parameter. Thus in the frequentist framework, uncertainty is dependent on making repeated measurements over and over again. Although this type of resampling of measurements is not done in practice, it provides a framework to understand uncertainty in frequentist inferences.[1]

The Bayesian approach counts the number of ways that the data can occur as a way to measure plausibilities, and these plausibilities are assigned probabilities. This approach has the advantage of providing very direct interpretations of inferences about data. For example, a Bayesian 95% confidence interval (often called credible intervals in order to be distinct from a frequentist confidence interval) is constructed such that the true parame-

[1] Bootstrapping is one example of the frequentist approach used in practice. The process of resampling the data with replacement is an attempt to simulate the repeated sampling of data. The bootstrapping procedure assumes that the collected data is representative of the population of interest and that resampling from the data simulates multiple data sets from the population. By having thousands of resampled data sets, we capture the thinking of the frequentist framework.

ter value has a 0.95 probability of being within the interval. The Bayesian framework allows for this kind of common-sense and straightforward interpretation of inferences.

In Bayesian data analysis, we aim to make inferences on unknown quantities from data using probability models of observed quantities. In other words, we want to calculate the relative plausibility of different parameter values conditional on the observed data. For example, suppose we want to estimate the rate of proliferation of T cells with data of T cell counts in the spleen in 1 hour intervals for 2 days. Based on the cell count data and a statistical model, we aim to calculate the plausibility for all possible rates of proliferation. A Bayesian analysis would then result in inferences like the following: there's a 20% probability for the rate to be greater than 1 division per 200 days, 50% probability for the rate to range from 1 division per 201 to 300 days, and 30% probability for the rate to be greater than 1 division per 400 days. This is quite different from a frequentist analysis, which would typically provides a best estimate and a confidence interval but does give degrees of plausibilities for different estimates.

Bayesian inference, which describes the relative plausibility of different parameter values based on the data, is encapsulated in the *posterior distribution*. The posterior distribution describes all of the possible parameter values and assigns a probability to each one, conditioned on the observed data. Suppose θ is a parameter or a set of parameters we wish to estimate and y is observed data. The posterior distribution is the quantity

$$p(\theta|y) \tag{2.1}$$

which is the probability of a value θ conditioned on the observed data y and $p(\cdot)$ is the marginal distribution or density of the probability model. Equation 2.1 is calculated with Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}\,. \tag{2.2}$$

The first term $p(y|\theta)$ is called the *likelihood*, and this is the probability of seeing the observed data given the parameter value θ and the probability model. Observed data or log-transformed data is often assumed to be normally distributed—particularly in biological data—and so the likelihood of-

ten is calculated from the normal distribution. The second term $p(\theta)$ is called the *prior*, and this is the initial set of plausible values for the parameter. This is where previous evidence or intuitions regarding the parameters are incorporated into the model. The final term $p(y)$ is called the marginal likelihood of y and is given by $p(y) = \sum_\theta p(\theta)p(y|\theta)$. This can be thought of as the average likelihood of the data over the prior. This term in practice does not need to be computed. Thus, in words,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Average Likelihood}}$$

Priors are chosen by the data analyst performing the data analysis and reflect the information we have about the parameters. This can be as simple as constraining the parameters to be positive (e.g. a negative proliferation rate makes no physical sense) and/or ruling out unrealistic values (e.g. naive T cells do not divide once every 10 minutes). Many frequentist approaches implicitly utilize flat priors by design (meaning all values are possible). In contrast, Bayesian analysis allows the scientist to use prior information to help the statistical model learn from the data more accurately and efficiently. Bayesian statistics is often criticized for being "subjective" since the priors are chosen by the scientist. Although much ink has been spilled about this topic, priors should be viewed as the available information about the parameters, and explicitly testing how different priors affect inferences provides much information about the data [97, 104, 105]. In the end, priors are just part of the statistical model and thus need to be evaluated and revised just like one would with any other type of models.

2.7.2 *Steps in Bayesian data analysis*

Bayesian data analysis has three generalized steps [106]:

- Setting up a *full probability model*: this involves describing a probability model that describes all observable data and unobservable parameters that we are interested in estimating.
- Calculating the *posterior distribution*: here we calculate the conditional probability of all unobserved parameters conditioned on the observed data.

- Evaluation of the model: we attempt to answer how well the model explains the observed data and how sensitive the model is to our assumptions made in the probability model.

References [97] and [106] describe these steps in detail. Instead, we will describe how these steps are performed in Chapter 3. For probability models used in Chapter 3, we assume that log-transformed counts and logit-transformed Ki67 proportions are normally distributed. The means of these values are predicted by ordinary differential equation (ODE) models, and the unknowns are the parameters used in the ODEs. The posterior distributions are approximated with the Stan language [107], and priors are specified based on information from previous modeling studies. Model fits are evaluated by how well the model explains the observed data while correcting for overfitting, which we discuss in the next three sections.

2.7.3 *Evaluating model fits with log predictive densities*

Suppose Eilidh wants to be the next haggis hurling champion by competing in the Haggis Hurling World Championship at the 2018 Bearsden & Milngavie Highland Games.[2] In her last training session, she threw the haggis 120 feet, 108 feet, and 105 feet in three practice throws. Based on these data, we would like to estimate what her average haggis-throwing distance is.

Let $N(\mu, \sigma)$ denote a normal distribution with mean μ and variance σ^2. Suppose we have two models for her haggis-throwing arm: $N(112, 7)$ and $N(90, 7)$. How do we evaluate how well these models fit the data?

The first approach is to use the *sum of the squared errors* (also known as the sum of the squared residuals (SSR)), namely

$$\text{SSR} = \sum_{i=1}^{n} \big(y_i - E(y_i|\theta)\big)^2 \tag{2.3}$$

where $E(y_i|\theta)$ is the expected value for y_i given by the model, and $y_1, y_2, \ldots, y_n$ are n data points. This equation sums up the squared differences between each data point and the value predicted by the model for the data point.

[2] http://www.bearsdenmilngaviehighlandgames.com/about-the-games/

For our haggis throwing example, the model $N(112, 7)$ predicts an average distance of 112, so

$$\text{SSR}_1 = (120 - 112)^2 + (108 - 112)^2 + (105 - 112)^2 = 129\,.$$

The model $N(90, 7)$ predicts an average distance of 90, so

$$\text{SSR}_2 = (120 - 90)^2 + (108 - 90)^2 + (105 - 90)^2 = 1449\,.$$

It's very clear that the first model $N(112, 7)$ explains the data much better than the second model. Measures like SSR are easy to compute and interpret but fail to incorporate uncertainty about parameter values. For example, in the calculations for model $N(112, 7)$, we are stating with full certainty that μ is 112 and σ is 7. Uncertainty quantified in posterior distributions cannot be incorporated in measures like the SSR.

As an alternative, we use *probabilistic prediction* to incorporate uncertainty. In particular, we use the log predictive density

$$\log p(y|\theta)\,.$$

y is the data point we are trying to predict with our model, and $p(y|\theta)$ is the likelihood given by the data model. A deeper explanation and rationale for this measure is beyond the scope of this section, but it has connections to the Kullback-Leibler information measure—models with the lowest Kullback-Leibler information have the highest expected log predictive density and thus have the highest posterior probability. In practice, we do not know what θ is, but we have estimates of θ and our uncertainty on θ in the posterior distribution $p_{\text{post}}(\theta) = p(\theta|y)$. So, we summarize the predictive density of the model by using the posterior distribution,

$$\log p_{\text{post}}(y|\theta) = \log \int p(y|\theta) p_{\text{post}}(\theta)\, d\theta\,. \tag{2.4}$$

And if y is a whole data set $\{y_1, y_2, \ldots, y_n\}$, then we calculate the log predictive density of the whole data set

$$\begin{aligned} \text{lppd} &= \text{log pointwise predictive density} \\ &= \log \prod_{i=1}^{n} p_{\text{post}}(y_i|\theta) \\ &= \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{\text{post}}(\theta)\, d\theta \end{aligned} \tag{2.5}$$

For our haggis throwing example, suppose we have a model $N(\mu, 7)$ and the posterior distribution for μ is 0.75 probability for $\mu = 112$ and 0.25 probability for $\mu = 108$. The lppd for this model is given by

$$\begin{aligned} \text{lppd} = &\log \int p(120|\mu) p_{\text{post}}(\mu)\, d\mu + \\ &\log \int p(108|\mu) p_{\text{post}}(\mu)\, d\mu + \\ &\log \int p(105|\mu) p_{\text{post}}(\mu)\, d\mu \end{aligned}$$

Let $N(y|\mu, \sigma)$ denote the probability density of y for a normal distribution with mean μ and standard deviation σ.[3] Then,

$$\begin{aligned} \text{lppd} = &\log\Big(N(120|112, 7) \cdot 0.75 + N(120|108, 7) \cdot 0.25\Big) + \\ &\log\Big(N(108|112, 7) \cdot 0.75 + N(108|108, 7) \cdot 0.25\Big) + \\ &\log\Big(N(105|112, 7) \cdot 0.75 + N(105|108, 7) \cdot 0.25\Big) \\ = &-9.9 \end{aligned}$$

To give a comparison, suppose we proposed a second (and worse) model $N(\mu, 7)$ with a posterior distribution for μ of 0.75 probability for $\mu = 100$ and 0.25 probability for $\mu = 50$. This model gives us a lower lppd of -14.4.

[3] The probability density of a normal distribution with mean μ and standard deviation σ is given by

$$N(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}$$

2.7.4 *Evaluating model fits using out-of-sample predictive accuracy*

The ideal measure for a model's fit would be a measure of how well it fits a new data point, also called its out-of-sample predictive performance. If f is the the true data-generating model, y is the observed data, and $\tilde{y}_i$ is a new data point from the true data-generating process, then we would like to know the log predictive density of $\tilde{y}_i$, namely,

$$\log p_{\text{post}}(\tilde{y}_i|\theta)\,.$$

Since the future data $\tilde{y}_i$ are unknown, we cannot calculate this and must instead take a step forward and calculate the *expected* log predictive density for a new point,

$$E_f\big(\log p_{\text{post}}(\tilde{y}_i|\theta)\big) = \int \big(\log p_{\text{post}}(\tilde{y}_i|\theta)\big) f(\tilde{y}_i)\, d\tilde{y}_i \tag{2.6}$$

If we have a whole dataset $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n$, then we can calculate the expected out-of-sample log predictive density for the whole dataset,

$$\begin{aligned} \text{elpd} &= \text{expected log pointwise predictive density for a new dataset} \\ &= \sum_{i=1}^{n} E_f\big(\log p_{\text{post}}(\tilde{y}_i|\theta)\big) \end{aligned} \tag{2.7}$$

where we simply added up the elpd for each data point. However, we cannot calculate these values as well since f, the true-data generating process, is also unknown to us.

Since future new data are unknown, we are left to use the fit of the model to the observed data (such as the lppd). All models are subject to a phenomenon called overfitting, which is when a model learns too much from the observations and goes beyond representing its general features, resulting in poor predictions for future data. Thus, in order to estimate the elpd of a model, we can use the lppd then subtract a correction to account for overfitting,

$$\text{elpd} \;=\; \text{lppd} - \text{correction}\,.$$

Here we describe alternative approaches to calculating the correction term. We will also describe the equivalent *information criterion* forms of these estimates of the elpd. The information criteria are simply the elpd multiplied by -2, and this form has connections to information theory. In mathematical modeling efforts in immunology, information criteria have been traditionally used instead of elpd values to evaluate models, and so we present both forms below.

Akaike information criterion (AIC) approach. Let k be the number of parameters used by the model. The simplest approach is to subtract k from the log predictive density given by the maximum likelihood estimate $\hat{\theta}_{\text{mle}}$. This approach assumes that the more parameters that a model has, the more prone it is to overfitting. Thus, we get

$$\widehat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\theta}_{\text{mle}}) - k \tag{2.8}$$

and AIC is calculated by multiplying this quantity by -2, giving us

$$\text{AIC} = -2\log p(y|\hat{\theta}_{\text{mle}}) + 2k \tag{2.9}$$

The AIC approach has two drawbacks: the log predictive density uses only the maximum likelihood estimate and not the full posterior distribution, and the correction $-k$ makes sense only when non-informative flat priors are used since even weakly informative priors reduce overfitting.

Watanabe-Akaike information criterion (WAIC) approach. The WAIC is a fully Bayesian approach that uses the full lppd and a correction that incorporates the posterior distribution. We estimate the elpd with the equation

$$\widehat{\text{elpd}}_{\text{WAIC}} = \text{lppd} - p_{\text{WAIC}} \tag{2.10}$$

where the correction p_{WAIC} is given by

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}}\big(\log p(y_i|\theta)\big) \tag{2.11}$$

This term measures the variance of the log predictive density across all values of θ in the posterior distribution for each data point and then sums them all up. Greater variance of the log predictive density indicates a more flexible

model that is thus more prone to overfitting. This term is often thought of as the *effective number of parameters* since the variance is a reflection of the flexibility of the model. WAIC is calculated by multiplying Equation 2.10 by -2, giving us

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}} \tag{2.12}$$

Leave-one-out cross-validation approach. This approach repeatedly partitions the data into a training set y_{training} and a holdout set y_{holdout}. The model is fit to the training set y_{training} to obtain a posterior p_{train}, and then the model fit is evaluated by the log predictive density of the holdout data

$$\log p_{\text{train}}(y_{\text{holdout}}|\theta)\,. \tag{2.13}$$

Typically, y_{holdout} is one data point and y_{training} is the other $n-1$ data points; this process is called leave-one-out cross-validation (LOO-CV). Let $p_{\text{post}(-i)}(y_i|\theta)$ denote the posterior distribution of the model fit when $y_{\text{holdout}} = y_i$. Then, the Bayesian LOO-CV estimate for the out-of-sample predictive fit is

$$\text{lppd}_{\text{LOO-CV}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i|\theta)\,. \tag{2.14}$$

Similarly to the other approaches, we can calculate a correction to subtract from $\text{lppd}_{\text{LOO-CV}}$, correction can done for this value. However, in practice, overfitting is limited since each prediction is conditioned on $n-1$ data points instead of the whole data set. Thus, the correction is often not necessary [106], giving us

$$\text{elpd}_{\text{LOO-CV}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i|\theta) \tag{2.15}$$

For consistency, we can compute the effective number of parameters for the Bayesian LOO-CV as

$$p_{\text{LOO-CV}} = \text{lppd} - \text{lppd}_{\text{LOO-CV}} \tag{2.16}$$

The leave-one-out information criterion (LOOIC) can be calculated by multiplying Equation 2.14 by -2.

In practice, LOO-CV as described is very computationally expensive since the model needs to be fitted n times. An approach called Pareto-smoothed importance sampling (PSIS) provides an approximation to the LOO-CV that is quick to compute [108]. In Chapter 4, all LOOIC values are calculated by using the PSIS approximation of the Bayesian LOO-CV using Stan, Rstan, and the `loo` R package [108]. We use LOOIC to evaluate models since it is considered to be the best measure of the out-of-sample performance of a model [108, 106]

2.7.5 *Using information criteria for model selection*

Recent mathematical modeling efforts in immunology have used information criteria, in particular the AIC, to evaluate models [92, 78, 79]. More specifically, AIC values of different candidate models have been used to perform *model selection*, where models with sufficiently lower AIC values (and hence higher elpd values) are chosen to be the models with the most statistical support. For example, Buchholz *et al.* [92] created 304 mathematical models to describe different differentiation patterns among $CD8^+$ naive T cell, T_{CM} cells, T_{EM} cells, and effector T cells and found that two models had AIC values at least 10 units lower than all other models. One model described a linear differentiation framework where naive $\rightarrow T_{CM} \rightarrow T_{EM} \rightarrow$ effector T cells, the second model added an additional partial flow from naive $\rightarrow T_{EM}$, and all other models were discarded due to their relatively high AIC values.

The Bayesian statistical literature often recommends not selecting individual models based on information criteria [109, 108]. Information is lost when models are discarded; instead, quantifying the relative accuracy of the models compared to each other gives us information about how confident we should be about individual models. As an alternative to model selection, we perform model averaging, in which information criterion values are used to assign relative weights to candidate models.

In Chapter 3, we use Akaike weights based on LOOIC values. Suppose we have m candidate models, each evaluated with a LOOIC value of LOOIC_i. We define the difference of model i's LOOIC from the minimum LOOIC value from set of candidate models,

$$\Delta\text{LOOIC}_i = \text{LOOIC}_i - \min_k \text{LOOIC}_k \,. \tag{2.17}$$

Then the Akaike weight for model i is given by

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta\text{LOOIC}_i\right)}{\sum_{k=1}^{m} \exp\left(-\frac{1}{2}\Delta\text{LOOIC}_k\right)} \,. \tag{2.18}$$

The $\exp\left(-\frac{1}{2}\Delta\text{LOOIC}_i\right)$ term converts the information criteria to a probability (reversing the multiplication of the elpd by -2 and then reversing the log transformation). The denominator then standardizes all values so that all weights add up to 1. We interpret Akaike weights as follows [110, 97]: "A model's weight is an estimate of the probability that the model will make the best predictions on new data, conditional on the set of models considered." In Chapter 3, we will find that one model out of a set of candidate models will have an overwhelming favorable Akaike weight, allowing us to effectively rule out out other models.

CHAPTER 3

Ontogeny of peripheral T cells

3.1 INTRODUCTION

During early life, peripheral T cells grow in numbers from tens of thousands at birth to tens of millions in several weeks, peaking at around 2 months in mice [80]. This magnitude of growth during early life is possible due to a growing thymus that exports naive T cells into the periphery at a rate that is assumed to be proportional to the size of the SP thymocyte pool. As we discussed in Section 1.3.2, there is evidence that the presumed lymphopenic environment in a neonatal mouse can support the expansion of adoptively transferred T cells through a mechanism called lymphopenia-induced proliferation (LIP). However, it is unknown whether LIP makes a significant contribution to the growth of peripheral T cell populations during early life. If significant LIP does occur, the associated expansion of T cell clones populating the periphery early in life could potentially skew the repertoire. Furthermore, while the stages of thymic development are well-characterized, the thymus itself first grows and then involutes with age, with associated changes in output. Little is known regarding the effect of these changes on the dynamics of passage through the different stages of development within the thymus.

In order to study the dynamics of T cell ontogeny, we studied cell counts and Ki67 expression in the thymus and in the lymph nodes of mice from days after birth into late adulthood. We confronted these data with mathematical models of thymocyte and naive T cell dynamics using Bayesian methods. Using this approach, we constructed and tested detailed quantitative descriptions of αβ T cell development from soon after birth into old age.

3.2 EXPERIMENTAL DATA

3.2.1 *The players: gating strategies and developmental pathways*

Cell counts and Ki67^{hi} fractions were measured in thymocyte and peripheral T cell subsets in mice of ages 5 days to 296 days old (total of 34 mice). Thymocyte subsets included DP1, DP2, DP3, immature SP4 (iSP4) and mature

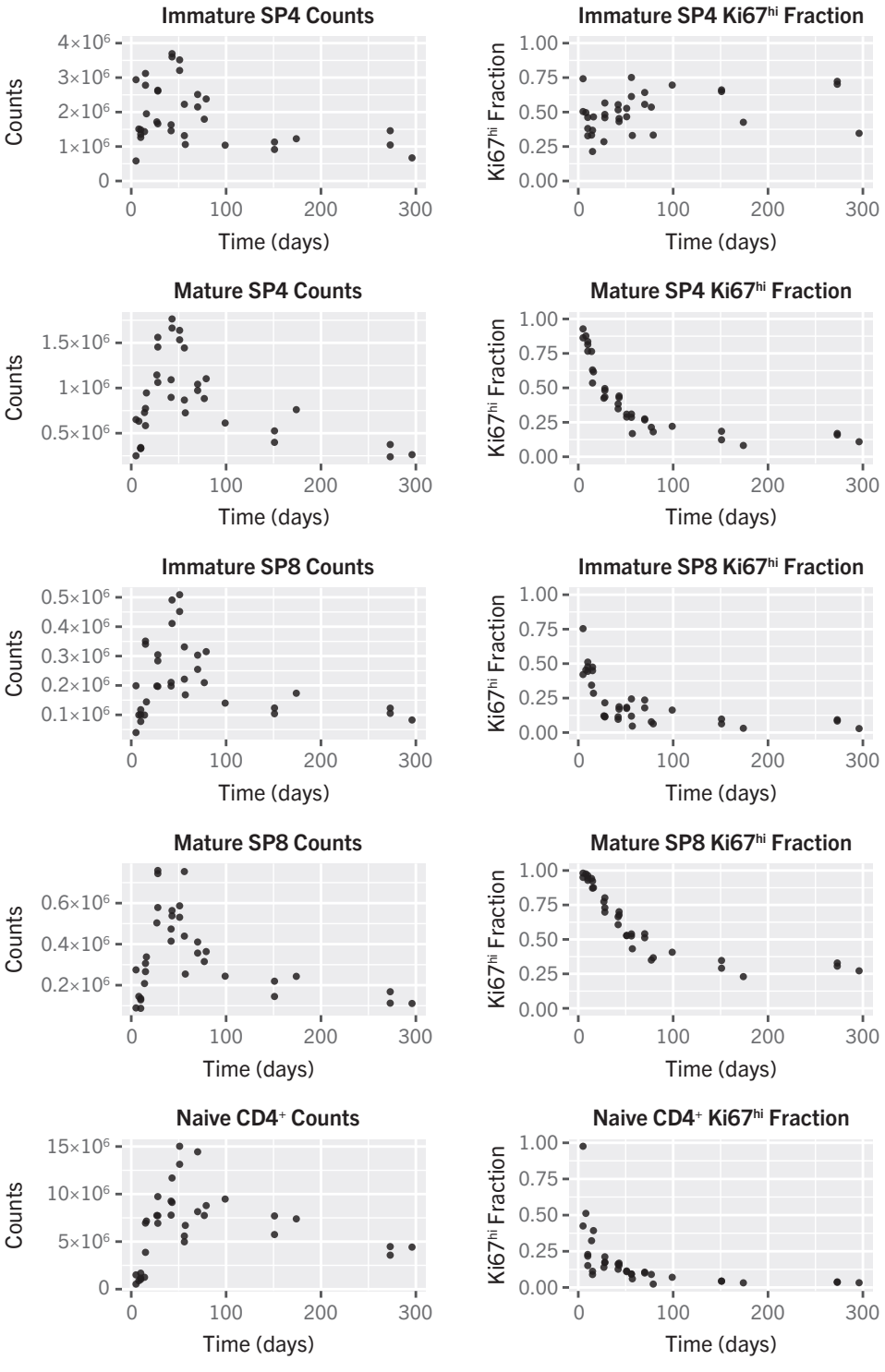

Immature SP4 Counts
Counts
4×10⁶
3×10⁶
2×10⁶
1×10⁶
0
0
100
200
300
Time (days)
Immature SP4 Ki67hi Fraction
Ki67hi Fraction
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (days)
Mature SP4 Counts
Counts
1.5×10⁶
1×10⁶
0.5×10⁶
0
100
200
300
Time (days)
Mature SP4 Ki67hi Fraction
Ki67hi Fraction
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (days)
Immature SP8 Counts
Counts
0.5×10⁶
0.4×10⁶
0.3×10⁶
0.2×10⁶
0.1×10⁶
0
100
200
300
Time (days)
Immature SP8 Ki67hi Fraction
Ki67hi Fraction
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (days)
Mature SP8 Counts
Counts
0.6×10⁶
0.4×10⁶
0.2×10⁶
0
100
200
300
Time (days)
Mature SP8 Ki67hi Fraction
Ki67hi Fraction
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (days)
Naive CD4+ Counts
Counts
15×10⁶
10×10⁶
5×10⁶
0
0
100
200
300
Time (days)
Naive CD4+ Ki67hi Fraction
Ki67hi Fraction
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Time (days)

Figure 3.1: (facing page) Cell counts and Ki67^{hi} fractions in the SP4, SP8, and naive CD4^{+} compartments. These data were obtained from thymii and lymph nodes of mice that are 5 days to 296 days old ($n = 34$, data from all compartments from the same mice). All cell counts data show a pattern of increasing in numbers and peaking at ~50 days. From that point, cell counts slowly decline with age. In every compartment except iSP4, Ki67^{hi} fractions start at a maximum shortly after birth and rapidly decrease with age. In iSP4 thymocytes, Ki67 expression starts lower and then increases into adulthood.

SP4 (mSP4), and immature SP8 (iSP8) and mature SP8 (mSP8) thymocytes (subsets defined in Table 3.1, gating strategy shown in Figure 3.4). Peripheral T cell subsets included naive CD4^{+}, CD4^{+} T_{CM}, CD4^{+} T_{EM}, naive CD8^{+}, CD8^{+} T_{CM}, and CD8^{+} T_{EM} cells (subsets defined in Table 3.1, gating strategy shown in Figure 3.5). Henceforth we drop the term "thymocyte" for DP and SP thymocytes unless it is needed for clarity.

The cells counts and Ki67^{hi} fractions for the iSP4/mSP4 compartments, the iSP8/mSP8 compartments, and the naive CD4^{+} T cell compartment are shown in Figure 3.1. In all compartments, the cell counts increase from birth until its peak at the age of ~50 days and then slowly declines. In every compartment except for iSP4, the Ki67^{hi} fractions start at its highest right after birth and decline very rapidly, reaching a low steady state at the age of ~125 days. The Ki67^{hi} fractions in the iSP4 compartment begins at a lower fraction and increases to a higher steady state Ki67^{hi} at age ~100 days. It should be noted that since Ki67 is expressed approximately for 3.5 days after a cell division event [79], Ki67 expression does not necessarily represent divisions that occurred within the compartment but instead could be inherited Ki67 expression from cell divisions in upstream compartments.

Figure 3.2 shows a schematic of the developmental trajectories within the thymus. We assumed that thymocytes committed to the CD4 lineage differentiate from DP2 directly to iSP4 and that thymocytes committed to the CD8 lineage pass through the DP3 stage before differentiating to iSP8 [38].

The structure of this chapter is as follows. We begin by modeling CD4^{+} T cell development from DP2 to mSP4 in Section 3.3 and study CD8 development from DP3 to mSP8 in Section 3.4. We then connect the dynamics of mSP4 cells with those of peripheral naive CD4 T cells in Section 3.5.

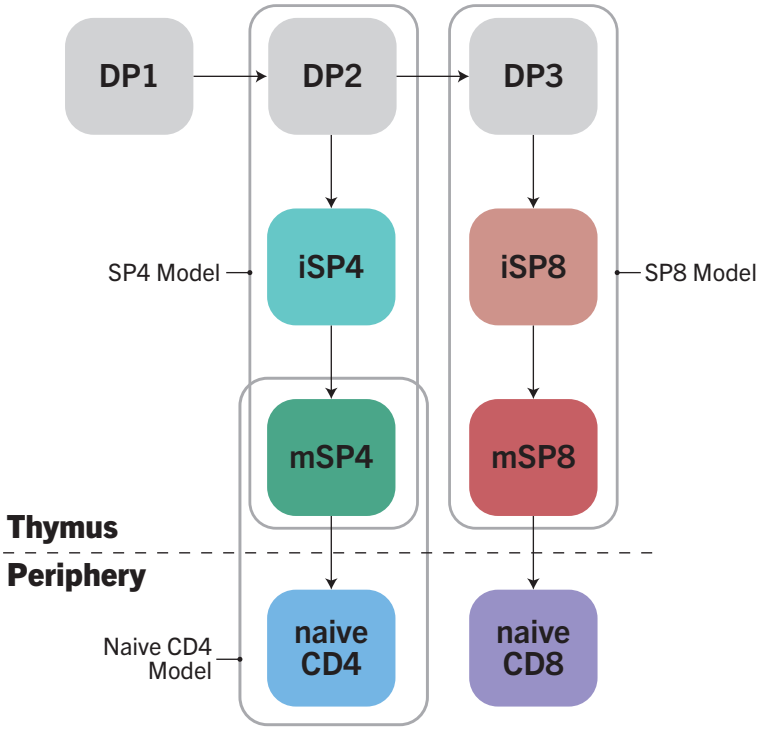


Figure 3.2: Schematic of the developmental trajectory of CD4 and CD8 T cell lineages. We assume a linear differentiation pattern of DP1 $\rightarrow$ DP2 $\rightarrow$ DP3. Thymocytes committed to the CD4 lineage differentiate from DP2 $\rightarrow$ iSP4 $\rightarrow$ mSP4, and thymocytes committed to the CD8 lineage differentiate from DP3 $\rightarrow$ iSP8 $\rightarrow$ mSP8. Naive CD4$^+$ T cells are formed from the export of mSP4 $\rightarrow$ naive CD4. We also highlighted how these compartments are split up for the modeling effort: the SP4 model considered DP2, iSP4, and mSP4 at the same time, the SP8 model considered DP3, iSP8, and mSP8, and the naive CD4$^+$ model considered mSP4 and naive CD4$^+$ T cells.

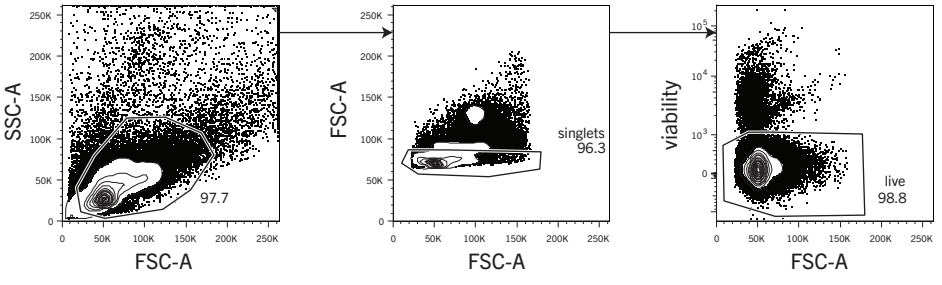


Figure 3.3: Gating of lymphocytes/thymocytes, singlet events, and live cells.

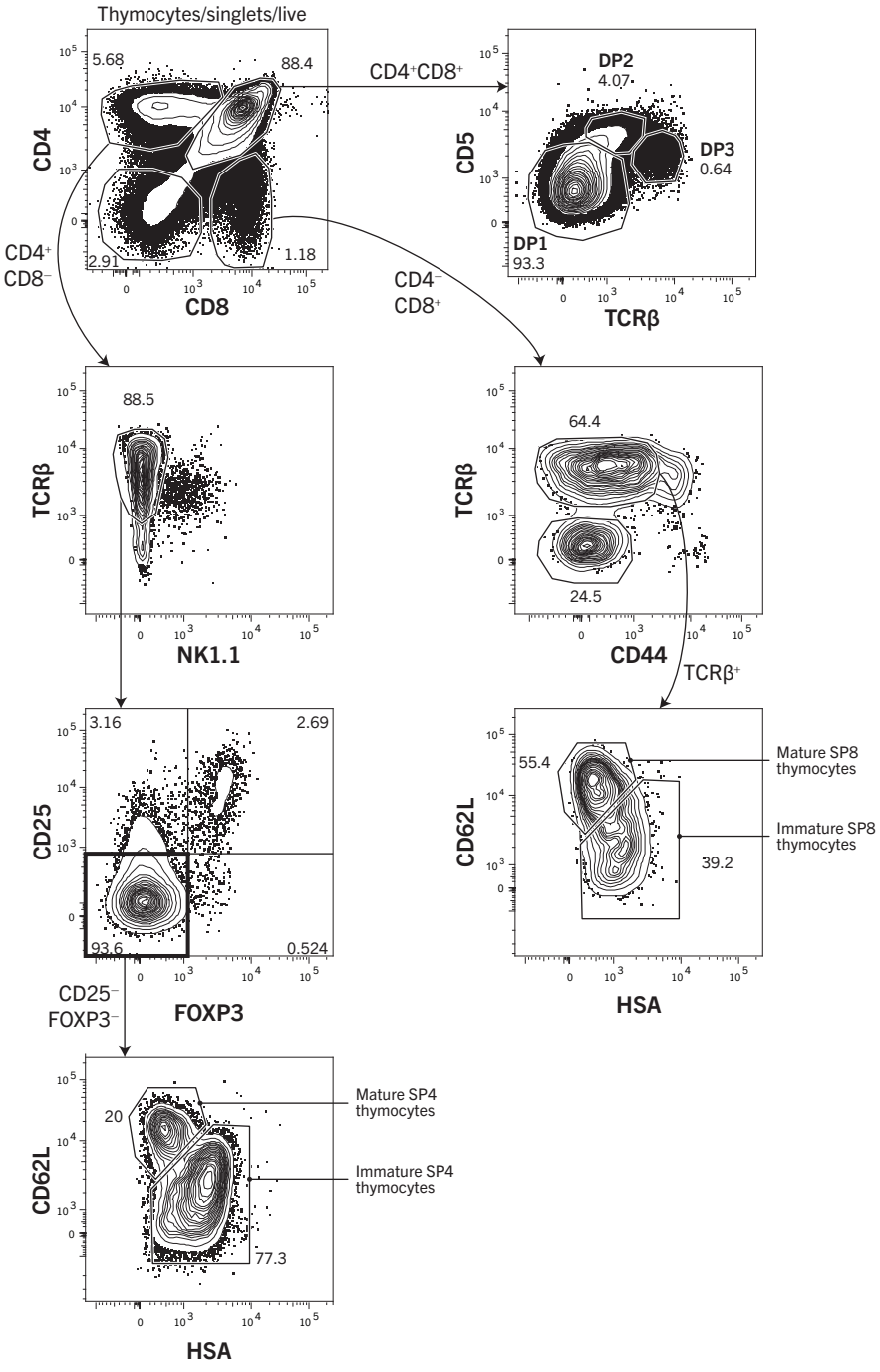


Figure 3.4: Gating strategy for thymocyte subsets.

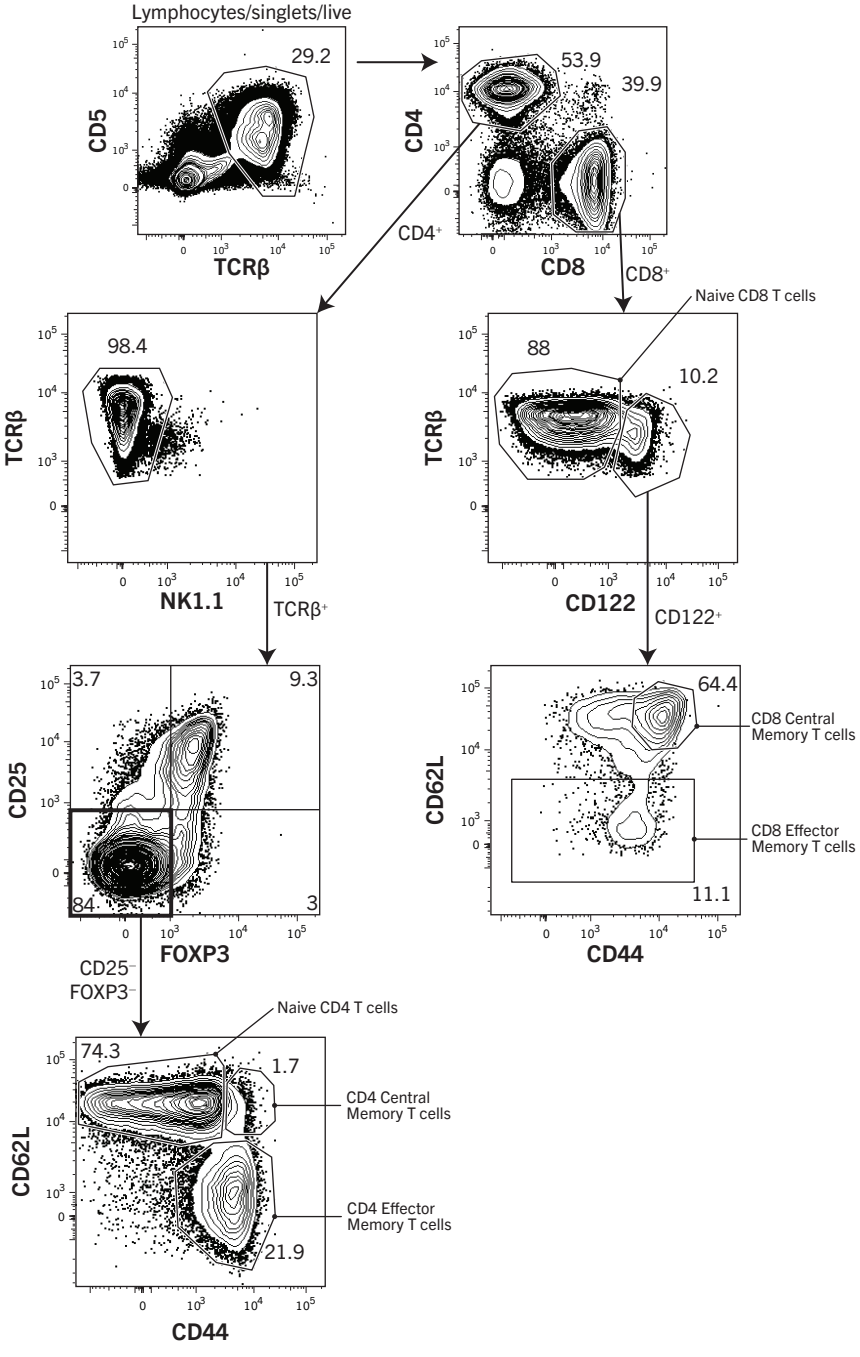


Figure 3.5: Gating strategy for peripheral T cells.

Subset	Markers
DP1	$CD4^+CD8^+TCR\beta^{lo}CD5^{lo}$
DP2	$CD4^+CD8^+TCR\beta^{int}CD5^{hi}$
DP3	$CD4^+CD8^+TCR\beta^{hi}CD5^{int}$
iSP4	$CD4^+CD8^-TCR\beta^{hi}NK1.1^-CD25^-Foxp3^-CD62L^{lo}HSA^{hi}$
mSP4	$CD4^+CD8^-TCR\beta^{hi}NK1.1^-CD25^-Foxp3^-CD62L^{hi}HSA^{lo}$
iSP8	$CD4^-CD8^+TCR\beta^{hi}CD44^-CD62L^{lo}HSA^{hi}$
mSP8	$CD4^-CD8^+TCR\beta^{hi}CD44^-CD62L^{hi}HSA^{lo}$
naive $CD4^+$	$TCR\beta^+CD4^+CD8^-NK1.1^-CD25^-Foxp3^-CD44^-CD62L^+$
$CD4^+$ T_{CM}	$TCR\beta^+CD4^+CD8^-NK1.1^-CD25^-Foxp3^-CD44^+CD62L^+$
$CD4^+$ T_{EM}	$TCR\beta^+CD4^+CD8^-NK1.1^-CD25^-Foxp3^-CD44^+CD62L^-$
naive $CD8^+$	$TCR\beta^+CD4^-CD8^+CD44^-$
$CD8^+$ T_{CM}	$TCR\beta^+CD4^-CD8^+CD122^+CD44^+CD62L^+$
$CD8^+$ T_{EM}	$TCR\beta^+CD4^-CD8^+CD122^+CD44^+CD62L^-$

Table 3.1: Markers used to define thymocyte and peripheral T cell subsets.

3.3 CD4$^+$ T CELL DEVELOPMENT IN THE THYMUS

3.3.1 *Modeling CD4$^+$ T cell development in the thymus*

We explored an array of mathematical models to explain the division, loss and (time-varying) flows between the DP $\rightarrow$ iSP4 $\rightarrow$ mSP4 compartments, and the subsequent loss or export of mSP4 thymocytes into the periphery. To do this, we took advantage of the breakdown of each stage into Ki67$^{hi/lo}$ cells and used ordinary differential equation (ODE) models to describe the flows between them.

We assumed that DP2 Ki67$^{hi/lo}$ thymocytes differentiate to iSP4$^{hi/lo}$ respectively. In addition, data from TetZap70 mice in which single cohorts of thymocytes were tracked through development in the thymus clearly show that iSP4 become Ki67^{lo} before mSP4 thymocytes begin to appear (unpublished data from Louise Webb and Benedict Seddon).[1] Thus, we assumed that there is no direct flow from Ki67^{hi} iSP4 $\rightarrow$ mSP4. However, the following two points are unclear:

1. Do iSP4 thymocytes lose Ki67 expression before differentiating into mSP4? Or can they differentiate into the mSP4 compartment while losing Ki67 expression?
2. Must iSP4 thymocytes divide before differentiating into mSP4?

We proposed three mathematical models to test these possible differentiation patterns (shown in Figure 3.6).

3.3.2 *Modeling the DP2 thymocyte compartment*

In all three models, we modeled the DP2 compartment by fitting empirical descriptor functions to timecourses of cell counts and Ki67 profiles (Equations 3.1–3.2, Section 3.7.1). Every model assumes that DP2 $\rightarrow$ iSP4 occurs at a constant per capita rate γ_{DP2}.

[1] TetZap70 mice are *Zap70*$^{-/-}$ mice with a tetracycline-inducible Zap70 transgene [41, 38]. In these mice, thymocytes are arrested at the early DP1 stage since they cannot receive positive selection signals from TCR-pMHC interactions without Zap70 expression. Doxycycline can be fed to these mice to induce Zap70 and allow thymocyte development to progress.

3.3.3 *Model #1: Immature SP4 thymocytes must lose Ki67 expression before differentiating to mature SP4 cells*

The first model ($\mathbf{M}_1$ in Figure 3.6*A*) assumes differentiation from only Ki67$^{\text{lo}}$ iSP4 $\rightarrow$ Ki67$^{\text{lo}}$ mSP4 at a constant rate. Ki67$^{\text{hi}}$ iSP4/mSP4 become Ki67$^{\text{lo}}$ at a per capita rate β_{SP4}, meaning that the mean time spent as Ki67$^{\text{lo}}$ is $1/\beta_{\text{SP4}}$. iSP4 and mSP4 are assumed to have constant turnover rates, but we allowed the turnover rates of Ki67$^{\text{hi}}$/Ki67$^{\text{lo}}$ cells within these compartments to differ (discussed in Section 3.3.7). No proliferation occurs in iSP4 thymocytes, and mSP4 thymocytes proliferate at a declining rate; a constant proliferation rate does not explain the data (discussed in Section 3.3.7). Equations 3.4–3.7 are the ODEs for model $\mathbf{M}_1$.

3.3.4 *Model #2: Immature SP4 thymocytes can differentiate to mature SP4 thymocytes as they lose Ki67 expression*

The second model ($\mathbf{M}_2$ in Figure 3.6*B*) extends model $\mathbf{M}_1$ by letting Ki67$^{\text{hi}}$ iSP4 $\rightarrow$ Ki67$^{\text{lo}}$ mSP4, which allows for the possibility of iSP4 maturing while they lose Ki67 expression. This model uses the same form of Ki67 expression lifetimes, turnover rates, and proliferation used by model $\mathbf{M}_1$. Equations 3.8–3.11 are the ODEs for model $\mathbf{M}_2$.

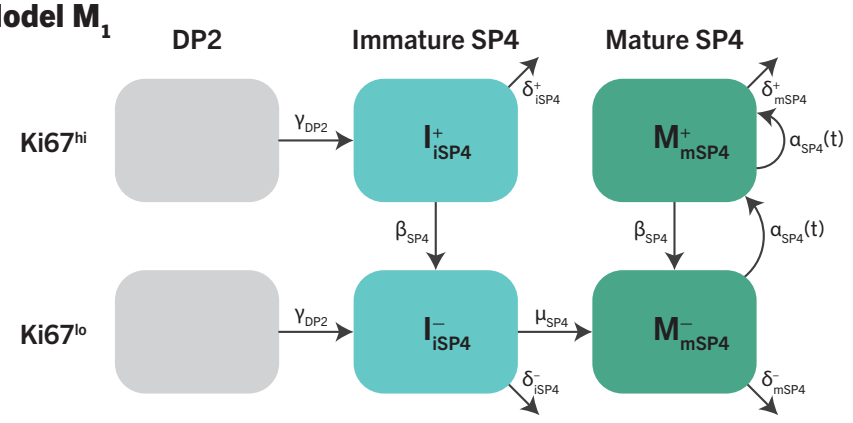
3.3.5 *Model #3: Differentiation from immature SP4 to mature SP4 requires division after loss of Ki67 expression*

The last model ($\mathbf{M}_3$ in Figure 3.6*C*) requires division in the transition iSP4 to mSP4. Thus, the cells dividing from Ki67$^{\text{lo}}$ iSP4 $\rightarrow$ Ki67$^{\text{hi}}$ mSP4 is the only way iSP4 mature to mSP4. This model again uses the same form of Ki67 expression lifetimes, turnover rates, and proliferation used in models $\mathbf{M}_1$ and $\mathbf{M}_2$. Equations 3.12–3.15 are the ODEs for model $\mathbf{M}_3$.

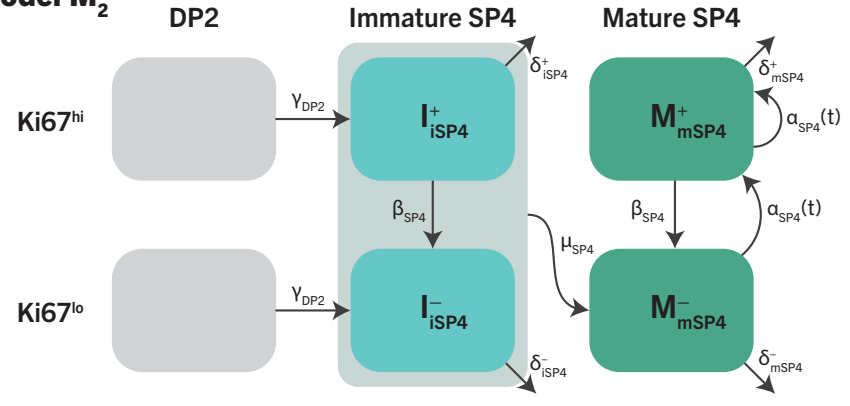
3.3.6 *Differentiation from immature SP4 to mature SP4 requires division*

We assessed the abilities of these models to describe the kinetics of the sizes and the Ki67$^{\text{hi}}$ fractions of the iSP4 and mSP4 compartments. We used a Bayesian approach to fit these models to the data and determine posterior distributions for the parameters of each model using the Stan programming language. The statistical models are described in Section 3.7.2. We took a

Model M_1



Model M_2



Model M_3

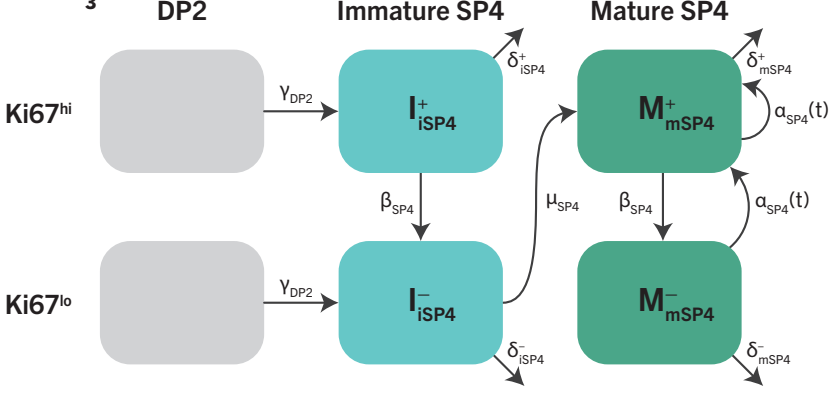


Figure 3.6: (facing page) Three models for SP4 thymocyte development. This figure depicts three models that describe in different ways how immature SP4 thymocytes differentiate to mature SP4 thymocytes. Model $\mathbf{M}_1$ requires immature SP4 thymocytes to lose Ki67 expression before they differentiate to mature SP4 thymocytes. Model $\mathbf{M}_2$ extends the previous model by allowing $Ki67^{hi}$ immature SP4 thymocytes to differentiate to $Ki67^{lo}$ mature SP4 thymocytes as well. Model $\mathbf{M}_3$ requires immature SP4 thymocytes to lose Ki67 expression and then divide in order to differentiate to mature SP4 thymocytes.

Bayesian approach in part because we wanted to describe the flows between $Ki67^{hi/lo}$ populations within DP2 $\rightarrow$ iSP4 $\rightarrow$ mSP4 simultaneously, and the large number of parameters involved complicated the use of standard frequentist techniques that employ parameter-space search algorithms; but primarily, a Bayesian approach allowed us to more carefully characterize the uncertainty in parameters and support for different models.

We show the fits of the models to the data as the mean of the posterior distribution of predicted values at every time point. These "posterior mean fits" for model $\mathbf{M}_3$ are shown in Figure 3.7 with 97% credible intervals for the model fit and 97% prediction intervals. The posterior mean fits for models $\mathbf{M}_1$ and $\mathbf{M}_2$ are shown in Figure 3.8 with 97% credible intervals for the model fit. These fits are shown with two uncertainty envelopes. The smaller bands ('credible intervals') indicate the uncertainty in how well the model fits the observed data (i.e. they reflect the posterior distributions of the dynamic model parameters). The wider envelopes indicate the degree of our uncertainty in predicting new data with the model; this uncertainty reflects both the posterior distributions of the dynamic parameters and posterior distributions of the scatter in the observations in each dataset due to biological variability and measurement error. The latter are also parameters that are estimated in the fits.

Model $\mathbf{M}_3$ had the strongest statistical support (ΔLOOIC = 10.2, Akaike weight = 0.99; see Section 2.7.4 for a description of these metrics) and explains the early kinetics of Ki67 expression in iSP4 thymocytes and the cell counts in mSP4 thymocytes more accurately than the other two models. The fits of models M_1 and M_2 to cell counts and $Ki67^{hi}$ fractions were visually and statistically similar (ΔLOOIC = 0.07, Figure 3.8).

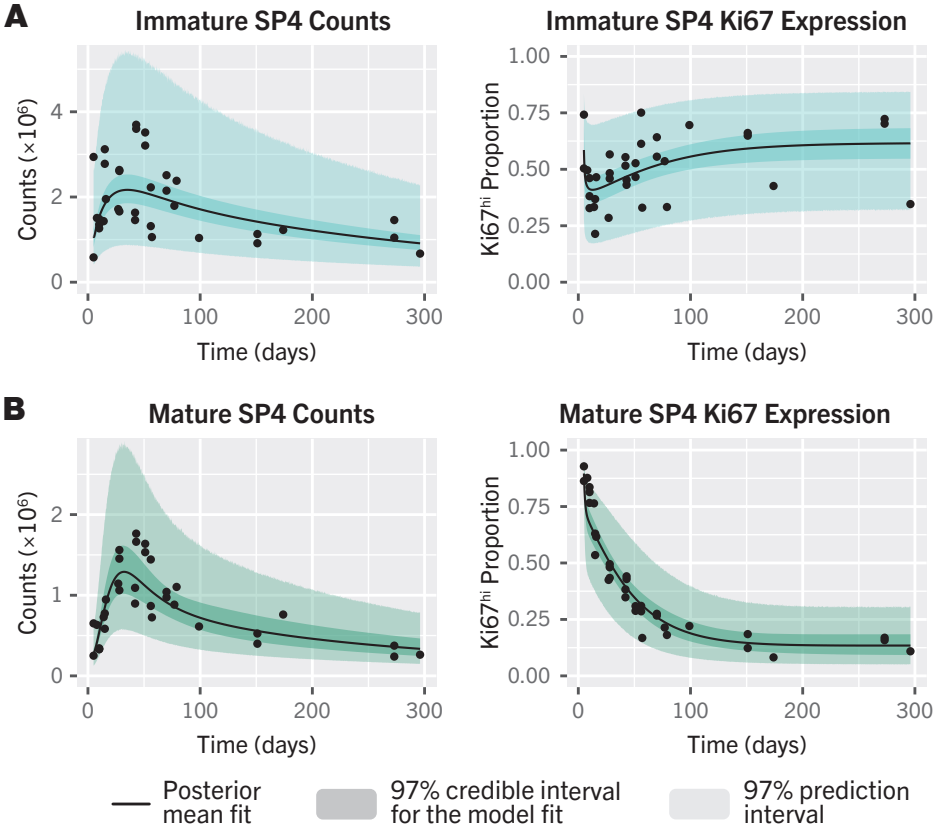


Figure 3.7: Posterior distribution of the fit for model M_3. In this model, immature SP4 thymocytes differentiate into the mature SP4 compartment only by losing Ki67 expression first and then dividing. The black curve shown is the posterior mean fit to the data, the darker envelope indicates the 97% credible interval for the model fit, and the lighter envelope indicates the 97% prediction intervals.

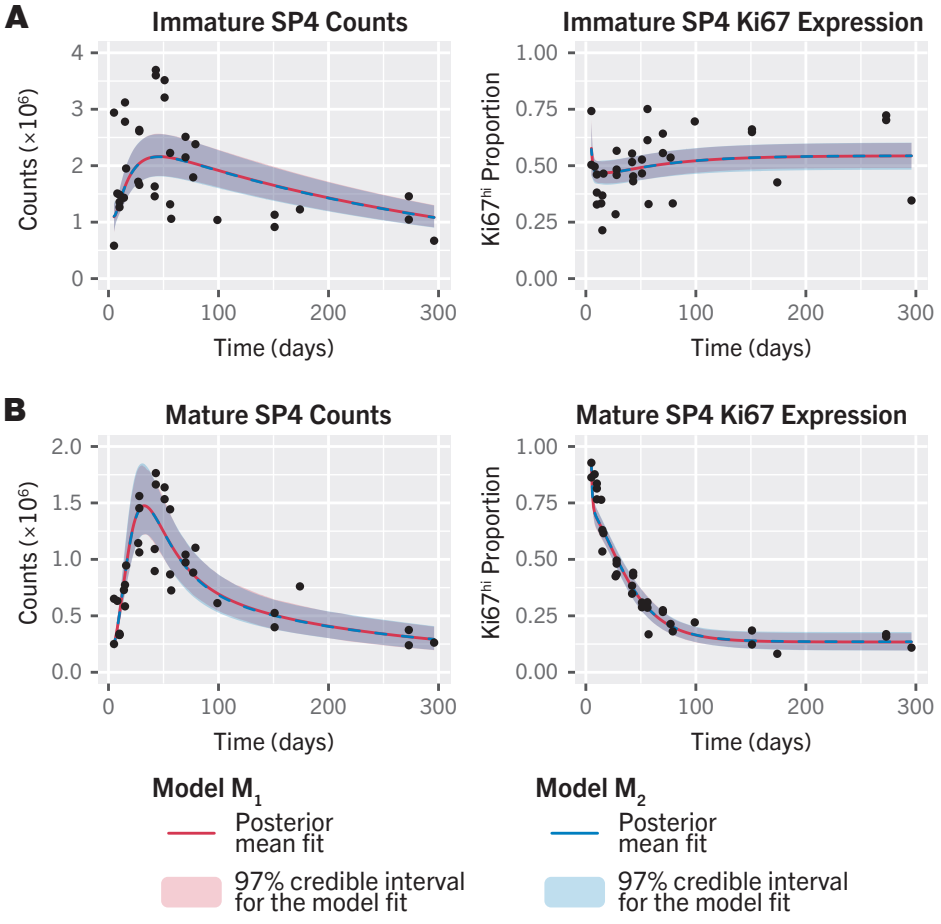


Figure 3.8: Posterior distributions of the fits for models M_1 and M_2. The posterior mean fits and the 97% credible intervals for model M_1 (in blue) and model M_2 (in red) are shown. Both models have similar fits, and the posterior distributions overlap, resulting in the light purple envelopes around the mean fits seen here. Both models fail to capture the dynamics of the Ki67 profile in the immature SP4 thymocytes in the early time points and overestimate the number of mature SP4 thymocytes in the early time points.

We found that model $\mathbf{M}_3$ of SP4 thymocyte maturation was strongly favored, as indicated by the 99% Akaike weight for the model. The estimated parameter values are shown in Table 3.2. Turnover rates are the rates at which cells are lost to death in the immature compartments or to death and thymic export in the mature compartments, and the mean interdivision time is the average time time it takes for a cell in the compartment to divide and is given by inverse of the division rate. The mean residence time is the average time a cell spends in an compartment, which is calculated by taking the inverse of all of the efflux rates from the compartment, and the Ki67 lifetime is calculated by taking the inverse of the Ki67 loss rate and represents the mean number of days a recently divide cell will express Ki67. This model suggests that $Ki67^{hi}$ iSP4 thymocytes are required to lose Ki67 expression and then divide again in order to differentiate to mSP4 thymocytes. In addition, the model requires the division of mSP4 thymocytes to occur more rapidly in early life. The fastest mean interdivision time is 2.5 days at the age of 5 days (which is the earliest time point in the data), the slowest is 105 days in adulthood, and the interdivision time decreases with a half-life of 22 days. This increased division rate early in life may increase the size of naive αβ TCR clones exported early in life. The model also indicates that $Ki67^{hi}$ mSP4 thymocytes are lost 14 times faster than $Ki67^{lo}$ mSP4, suggesting that cell division predisposes mature thymocytes for egress, and/or division is associated only with the final stages of egress.

Model $\mathbf{M}_3$ has the clearest statistical support and seems to provide the clearest explanation of SP4 thymocyte dynamics. Both models $\mathbf{M}_1$ and $\mathbf{M}_2$ do not have capture the Ki67 expression profile of the immature SP4 thymocytes and the mature SP4 cell counts. The failure of model $\mathbf{M}_1$ points to the conclusion that immature SP4 cells need to divide in order to enter the mature SP4 compartment, and the failure of model $\mathbf{M}_2$ indicates that immature SP4 thymocytes must lost Ki67 expression before they can make these divisions to transition into the mature SP4 compartment. Since mature SP4 $Ki67^{hi}$ fractions are high in the first 50 days in life, this division requirement for the transition from immature to mature SP4 is consistent with the observed data.

3.3.7 *Exploration of alternative models supports the conclusion that SP4 thymocytes show different turnover rates based on Ki67 expression, and proliferation rates decrease with time*

We also explored simpler versions of model $\mathbf{M}_3$ in which (*i*) $Ki67^{hi}/Ki67^{lo}$ thymocytes have the same turnover rates within the iSP4 and mSP4 compartments and (*ii*) the turnover rates are different for the $Ki67^{hi}/Ki67^{lo}$ iSP4 thymocytes but the same for $Ki67^{hi}/Ki67^{lo}$ mSP4 thymocytes. Both models yielded poorer fits to the data (ΔLOOIC > 35 in favor of model $\mathbf{M}_3$, with an Akaike weight of 1). The first model failed to capture the dynamics of Ki67 expression in iSP4 by overestimating the $Ki67^{hi}$ fractions early in life and in mSP4 thymocytes by overestimating $Ki67^{hi}$ fractions later in life (Figure 3.9, green curve). The second model failed to explain the dynamics of $Ki67^{hi}$ mSP4 thymocytes at late times (results not shown). We tested analogous versions of models $\mathbf{M}_1$ and $\mathbf{M}_2$, which also yielded very poor fits (results not shown).

We examined whether mSP4 kinetics could be explained with a constant division rate. This model also yielded a poorer fit (ΔLOOIC = 52.3 in favor of model $\mathbf{M}_3$ with an Akaike weight of 1) by failing to explain the $Ki67^{hi}$ fractions in mSP4 thymocytes after day 100 (Figure 3.9, orange curve).

3.4 CD8$^+$ T CELL DEVELOPMENT IN THE THYMUS

3.4.1 *Modeling CD8$^+$ T cell development in the thymus*

We aimed to describe the kinetics of the SP8 compartment by exploring different mathematical models to describe the flow from DP3 $\rightarrow$ iSP8 $\rightarrow$ mSP8. We assume that $Ki67^{hi/lo}$ DP3 $\rightarrow$ $Ki67^{hi/lo}$ iSP8 respectively. In contrast with the SP4 compartments, data from TetZap70 mice do not clearly rule out the differentiation of iSP8 to mSP8 before the loss of Ki67 expression (unpublished data from Louise Webb and Benedict Seddon). Thus, the models allow $Ki67^{hi}/Ki67^{lo}$ iSP8 $\rightarrow$ $Ki67^{hi/lo}$ mSP8 respectively. Ki67 lifetimes and turnover rates were modeled similarly to the SP4 models, and proliferation occurs in both iSP8 and mSP8.

We assessed the abilities of the models to explain the kinetics of cell numbers and $Ki67^{hi}$ fractions of the iSP8/mSP8 compartments. We first tested a model where all turnover and division rates are constant (Figure 3.10*A*,

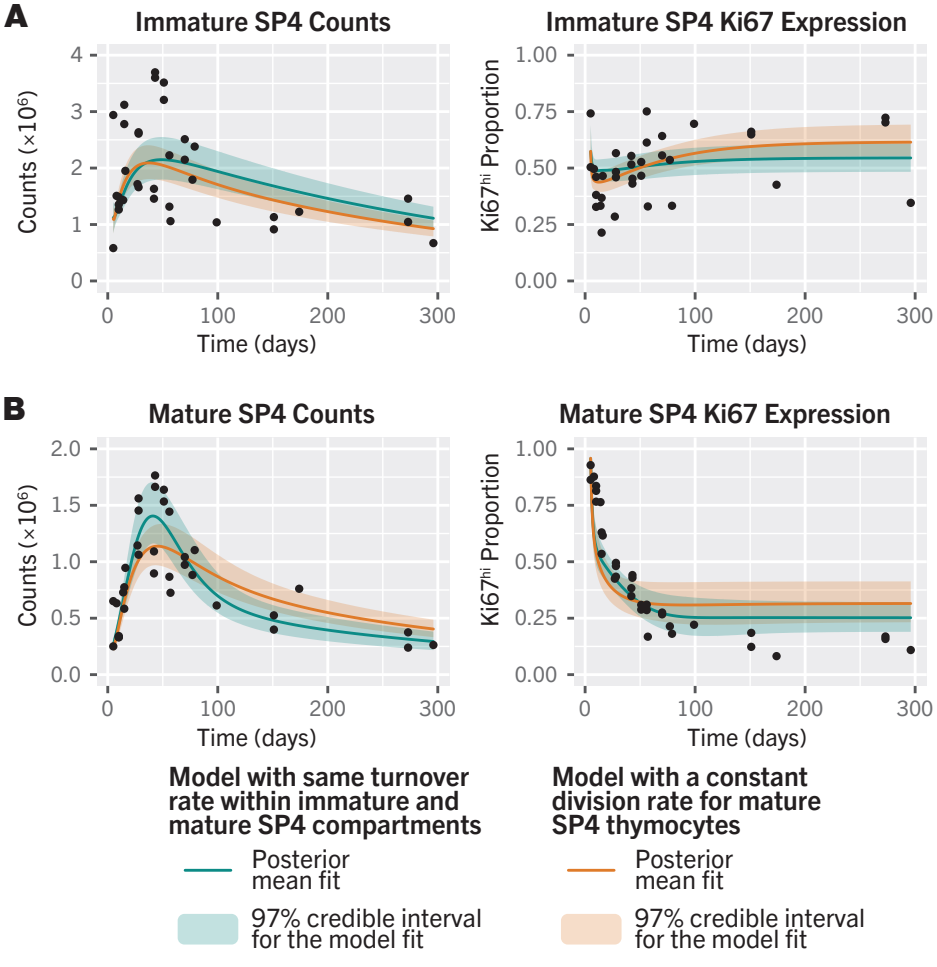


Figure 3.9: Posterior distributions of the fits for more simple SP4 thymocyte models. The posterior mean fits and the 97% credible intervals for a model with same turnover rates for Ki67^{hi} and Ki67^{lo} thymocytes within the immature and mature SP4 thymocyte compartments (in green) and for a model where proliferation of mature SP4 thymocytes occurs with a constant rate (in orange).

Compartment		Turnover Rates (day^{-1})		Mean interdivision time (days)		Ki67 Lifetime (days)	
		Mean	97% CI	Mean	97% CI	Mean	97% CI
Immature SP4	$Ki67^{hi}$	1.8	(0.94, 2.8)	NA	NA	4.0	(3.0, 5.6)
	$Ki67^{lo}$	0.32	(0.20, 0.45)				
Mature SP4	$Ki67^{hi}$	0.50	(0.34, 0.68)	**Fastest:** 2.2 **Slowest:** 101 $t_{1/2}$: 22	**Fastest:** (1.6, 4.1) **Slowest:** (49, 1321) $t_{1/2}$: (16, 31)		
	$Ki67^{lo}$	0.032	(0.015, 0.053)				
Immature SP8	$Ki67^{hi}$	4.2	(2.3, 6.1)	**Fastest:** 0.3 **Slowest:** 101 $t_{1/2}$: 9.2	**Fastest:** (0.2, 0.6) **Slowest:** (62, 317) $t_{1/2}$: (5.7, 17.6)	6.1	(3.7,13.4)
	$Ki67^{lo}$	0.54	(0.26, 1.0)				
Mature SP8	$Ki67^{hi}$	**Fastest:** 2.0 **Slowest:** .50 $t_{1/2}$: 6.4	**Fastest:** (0.97, 3.5) **Slowest:** (0.19, 1.1) $t_{1/2}$: (4.8, 9.2)	**Fastest:** 1.2 **Slowest:** 53 $t_{1/2}$: 19	**Fastest:** (0.8, 2.0) **Slowest:** (14, 4818) $t_{1/2}$: (12, 30)		
	$Ki67^{lo}$	**Fastest:** 4.2 **Slowest:** .21 $t_{1/2}$: 6.4	**Fastest:** (0.17, 6.0) **Slowest:** (.055, .53) $t_{1/2}$: (4.8, 9.2)				
Naive CD4		0.019	NC	236	NC	3.2	NC

Table 3.2: Estimated parameter values for SP4 thymocytes, SP8 thymocytes, and naive CD^+ T cells. Abbreviations: NA, not applicable; NC, not calculated; CI, credible interval.

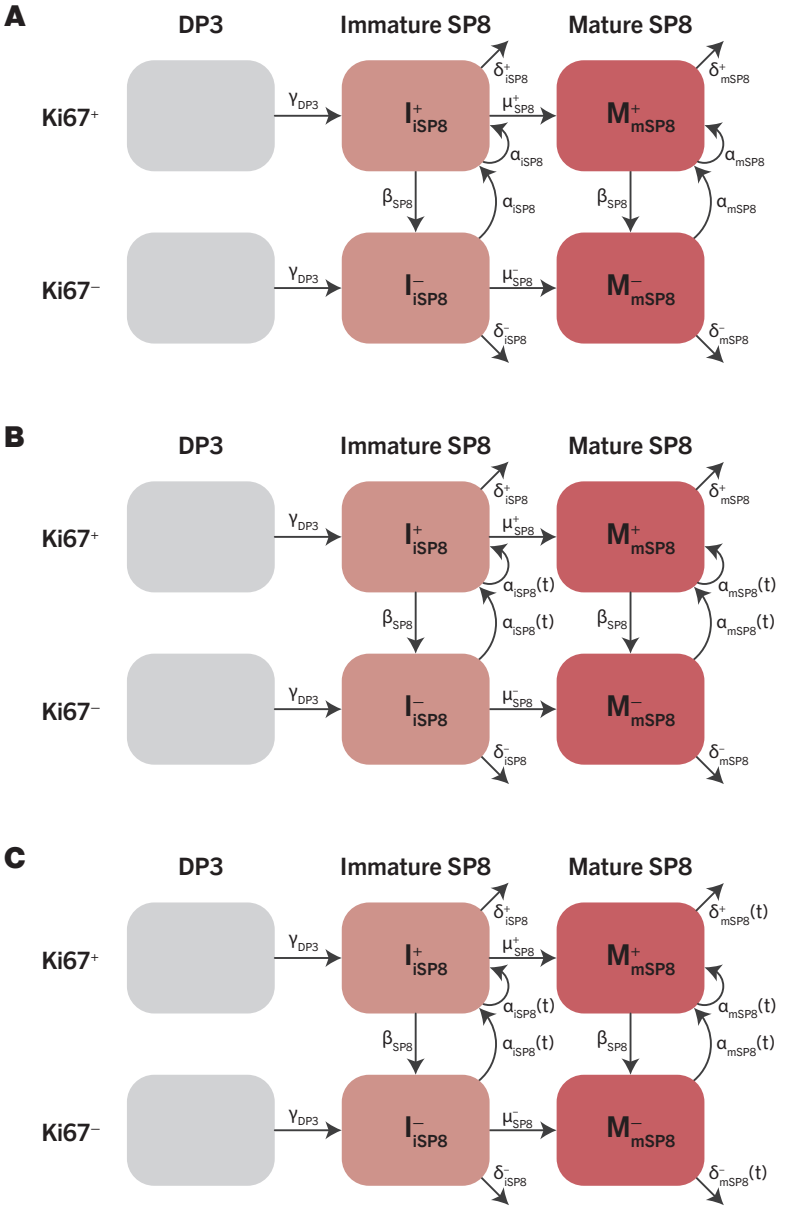


Figure 3.10: (facing page) Three models for SP8 thymocyte development. This figure depicts three different models for the SP8 thymocyte compartments. **(A)** This model depicts constant turnover, differentiation, and proliferation rates. **(B)** This model allows the proliferation rates of the immature and mature SP8 compartments to change with time. **(C)** This model extends the previous models by also allowing the turnover rates of the mature SP8 compartments to change with time.

Equations 3.18–3.21). The model fits explained the cells counts in both iSP8 and mSP8 compartments but completely failed to capture the dynamics of Ki67 expression levels (Figure 3.11, red curves). We extended this model by allowing division rates to decline with age for iSP8 and mSP8 thymocytes (Figure 3.10*B*, Equations 3.24–3.27) These division rates were modeled similarly to the mSP4 proliferation rate of model $\mathbf{M}_3$ (Equations 3.22–3.23). This model was able to explain the $Ki67^{hi}$ profile of iSP8 thymocytes, but it still did not completely explain the $Ki67^{hi}$ profile of mSP8 thymocytes and had worse fits for cell counts (Figure 3.11, blue curves).

We extended the model further by allowing turnover rates in the mature SP8 compartments to decline with age (Figure 3.10*C*, Equations 3.30–3.33, turnover rates in Equations 3.28–3.29). This model was able to explain the cell counts and Ki67 expression data from both immature and mature SP8 thymocytes (Figure 3.12) and had the most statistical support out of all three models ($\Delta LOOIC > 70$, Akaike weight = 1).

The estimated parameters values for this model are shown in Table 3.2. Immature SP8 thymocytes start with a mean interdivision time of 0.3 days in early life that exponentially increases to a interdivision time of 101 days in adulthood with a half-life of 9.2 days. Mature SP8 thymocytes start with a mean interdivision time of 1.2 days in early life that exponentially increases to a interdivision time of 53 days in adulthood with a half-life of 19 days. The model suggests that both immature and mature SP8 thymocytes are dividing much more rapidly in early life than in adulthood. This conclusion is similar to how division differs in early life for mature SP4 thymocytes. In both SP4 and SP8 thymocyte populations, fast division rates are needed to explain the high $Ki67^{hi}$ profiles found in early life. These patterns are not surprising since the thymus is rapidly growing and expanding early in life; the initial abundant proliferation in thymocytes reflects rapid thymic growth.

The model invokes changing turnover rates for mature SP8 thymocytes as well. Thus, these thymocytes have very short residence times early in life (0.49 days for $Ki67^{hi}$ and 0.24 days for $Ki67^{lo}$ SP8 thymocytes) that increase to longer residence times in adulthood (2.0 days for $Ki67^{hi}$ and 4.9 days for $Ki67^{lo}$ SP8 thymocytes). These turnover rates are a combination of cell death and thymic egress, so we cannot quantify how thymic output differs between neonates and adults. However, the very short residence times in

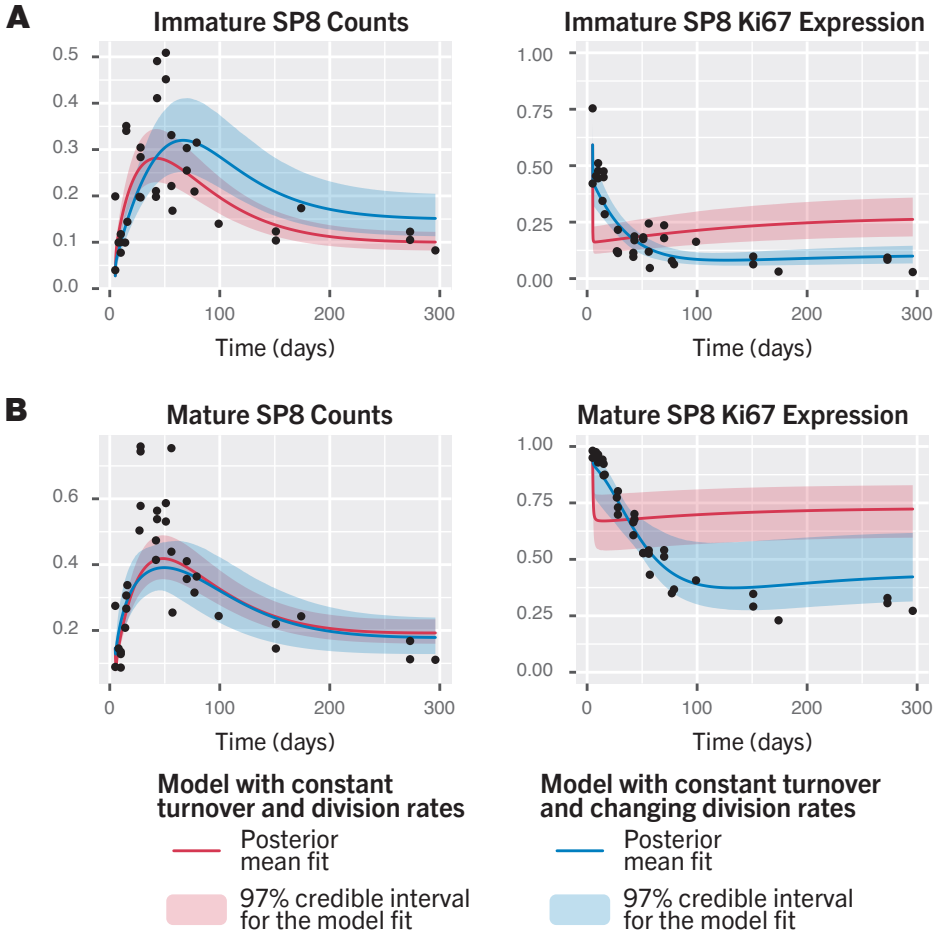


Figure 3.11: Posterior distributions of the fits for SP8 thymocyte models. The posterior mean fits and the 97% credible intervals for the SP8 thymocyte model with constant turnover and proliferation rates (in red) and for the model with constant turnover rates and changing proliferation rates (in blue).

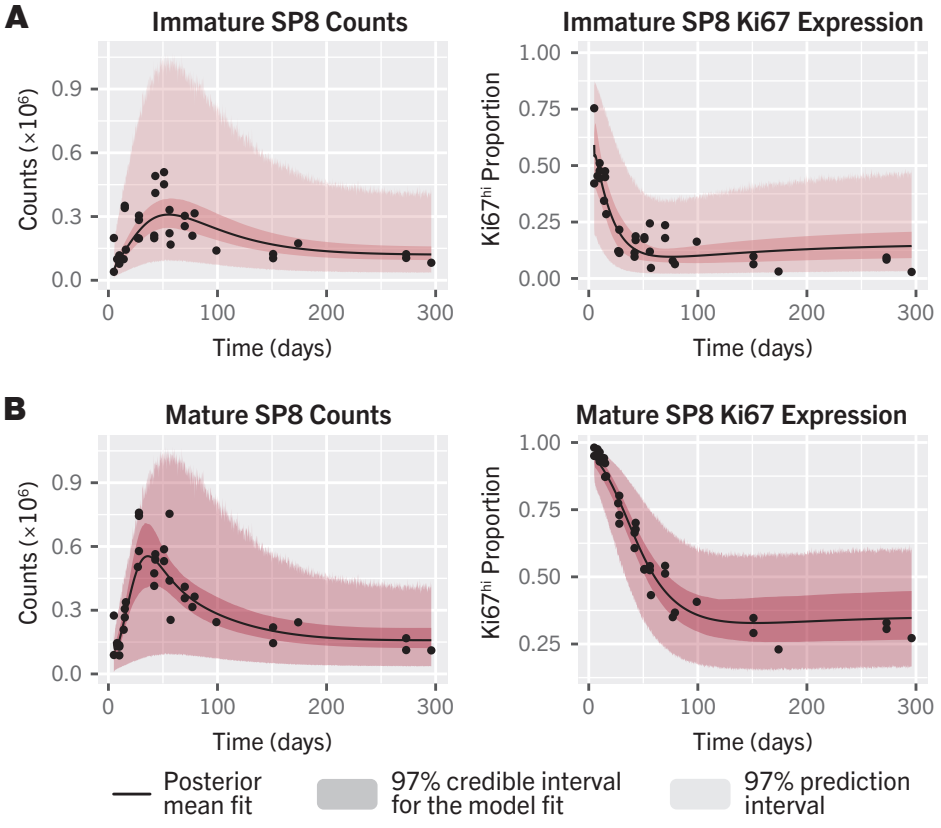


Figure 3.12: Posterior distribution of the fits for the best-fitting SP8 thymocyte model. The posterior mean fits, the 97% credible intervals for the fits, and 97% prediction intervals are shown for the SP8 thymocyte model with changing proliferation rates and changing turnover rates for mature SP8 thymocytes.

early life do suggest that both cell death is faster in early life (e.g. due to more thymocytes dying from negative selection) and thymic output is greater in early life. This contrasts with the mSP4 thymocyte compartment, in which we saw no evidence of changing turnover rates.

SP8 thymocytes have a qualitatively different maturation pattern than SP4 thymocytes since both iSP8 and mSP8 thymocytes divide with rates that decrease into adulthood and since mSP8 thymocytes have turnover rates that decrease with age as well. The models predict fundamentally different differentiation patterns for SP4 and SP8 thymocytes such that SP8 thymocytes do not have to lose Ki67 expression before transitioning from the immature to mature stages whereas SP4 thymocytes do. This conclusion is reflective of the idea that DP2 thymocytes directly mature into the SP4 compartment whereas DP2 thymocytes must pass through the DP3 compartment before differentiating into the SP8 compartment. SP8 thymocytes have a longer history in the DP compartment, and thus DP2 thymocytes entering the SP4 compartment and DP3 thymocytes entering the SP8 compartment are in fundamentally different states of differentiation. The different model structures supported by the data demonstrate the different developmental histories of the two SP compartments.

3.5 NAIVE CD4$^+$ T CELLS

3.5.1 *No evidence for lymphopenia-induced proliferation of naive CD4$^+$ T cells in neonates*

With a working model for the dynamics in the SP4 compartment, our next aim was to find the most parsimonious description of the dynamics of naive CD4$^+$ T cell numbers in lymph nodes. We aimed to estimate the proliferation rate of naive CD4$^+$ T cells in order to assess if LIP occurs within the naive CD4$^+$ compartment.

We considered the simplest model in which the Ki67$^{hi/lo}$ naive CD4$^+$ T cells are fed by Ki67$^{hi/lo}$ mSP4 cells respectively (Figure 3.13). This assumes that the timecourses of mSP4 cells were proportional to their rates of export from the thymus The subcompartments have the same constant turnover rate δ_n, which encompasses both cell death and differentiation from the naive CD4$^+$ compartment. Ki67^{hi} naive CD4$^+$ T cells lose their Ki67 expres-

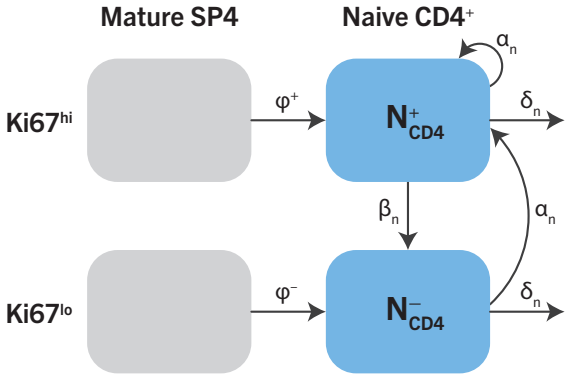


Figure 3.13: Model for naive CD4⁺ T cell counts and Ki67^{hi} proportions. In this model, naive CD4 T cells have different influx rates from Ki67^{hi} and Ki67^{lo} mature SP4 thymocytes. Ki67^{hi} and Ki67^{lo} naive CD4 T cells have the same turnover rates and divide with a constant proliferation rate.

sion with a mean lifetime of $1/\beta_n$. Naive CD4⁺ T cells divide with a constant rate α_n. The respective flows from Ki67$^{hi/lo}$ mSP4 $\rightarrow$ Ki67$^{hi/lo}$ naive CD4⁺ compartments have different per capita rates φ^+ and φ^-. The ODEs for this model are Equations 3.36–3.37.

We determined the maximum-likelihood estimates of these parameters by fitting the model to the naive CD4⁺ T cell counts and Ki67 expression profiles in the lymph nodes (Table 3.2) (fitting procedure described in Reference [78]. The model fit to the naive CD4⁺ T cell data is shown in Figure 3.14. We estimate a per capita rates of influx to the naive pool from the Ki67$^{hi/lo}$ mature SP4 thymocyte compartments of $\varphi^+ = 0.55$ and $\varphi^- = 0.036$ respectively. These values are in general agreement with the turnover rates of the Ki67^{hi}/Ki67^{lo} mature SP4 compartments, which we estimated to be 0.49 (97% credible interval: 0.35–0.65) and 0.03 (97% credible interval: 0.01–0.05). The mean residence time for naive CD4⁺ T cells is 53 days, which is in agreement with estimates of ~10-60 days from previous studies [78, 80].

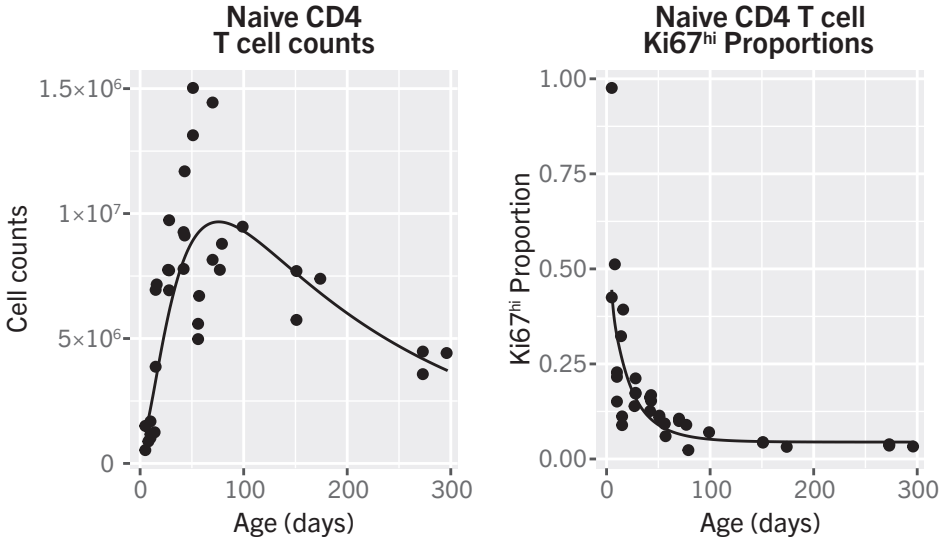


Figure 3.14: Model fit to naive CD4⁺ T cell counts and Ki67^{hi} proportions.

The estimated mean interdivision time of naive CD4⁺ T cells is 236 days ($\alpha_n = 0.00423$ days^{-1}). We next evaluated how sensitive this estimate is to different influx rates from the thymus by scanning through different ratios of φ^+ to φ^-. The influx into the naive CD4⁺ compartment from the Ki67^{hi}/Ki67^{lo} mSP4 compartments are different since the turnover rates in these compartments are different. Because of the uncertainty of the ratio of the turnover rates of the Ki67$^{hi/lo}$ mSP4 compartments (posterior mean 16.6; 97% credible interval 9.2–31.2), there is uncertainty on the ratio of the influxes into the naive CD4⁺ compartment. We scanned through different values of φ^+/φ^- and fitted the model to the naive CD4⁺ T cell data with these fixed ratios (Figure 3.15). We found that within a reasonable range of ratios (shown in dark gray in Figure 3.15*A*), the mean interdivision time for naive CD4⁺ T cells remained greater than 200 days, which is too slow to be explained by LIP. Thus, we conclude that our inferences regarding the interdivision time of naive CD4 T cells in lymph nodes is robust, despite any uncertainty on thymic output rates.

Our estimate of 236 days is within the range of estimated interdivision times of 127–530 days for naive CD4⁺ T cells in adult mice from Hogan *et al*. [78]. In addition, since Ki67 expression has a mean lifetime of ~3.5 days [79], we would expect $0.00423 \times 3.5 = 1.5\%$ of naive CD4⁺ T cells

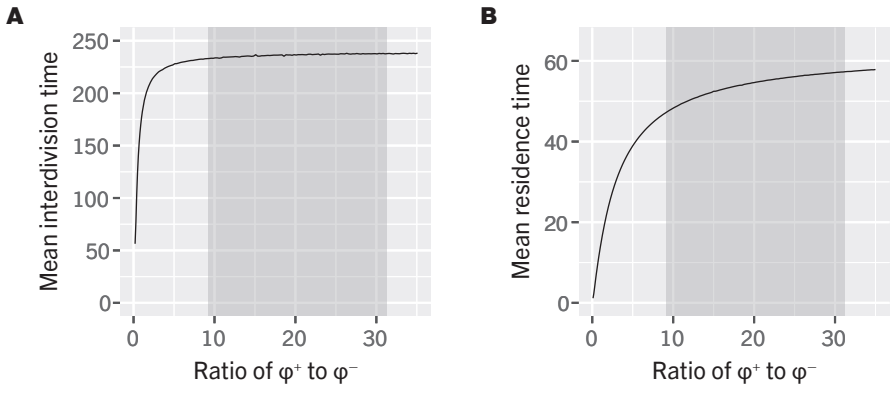


Figure 3.15: Sensitivity of interdivision and residence times to the ratio of influxes. In order to test the sensitivity of the estimated **(A)** mean interdivision times and **(B)** mean residence times of naive CD4 T cells to the ratio of influx of Ki67^{hi} and Ki67^{lo} mature SP4 thymocytes, we scanned through different values of the ratio *f* and fit the model to the data. The darker gray intervals highlight the possible values of *f* based on the 97% credible intervals of the ratio of the turnover rates of Ki67^{hi} and Ki67^{lo} mature SP4 thymocytes. We find robust estimates that are not sensitive to different values of *f* until *f* becomes very small. However, within the most likely ratio values (shown in the darker gray intervals), we find that the estimated interdivision time for naive CD4^{+} T cells remains above 200 days.

to be Ki67^{hi} in adult mice, which agrees with the Ki67^{hi} proportions found in the latest time points in the oldest mice. These agreements, and the visual quality of the fit, lend support to the conclusion that naive CD4^{+} T cells in neonatal mice do not undergo substantial levels of LIP, at least at the high levels of one division every 3–5 days observed in highly lymphopenic mice [111] and that the elevated level of Ki67 in the periphery during the first ~75 days of life (Figure 3.14, right panel) is inherited, short-lived expression on cells recently exported from the highly proliferative young thymus. A caveat is that any putative lymphopenia would likely span only a small portion of the early part of the timecourse we are fitting to, and so any LIP might have a relatively small effect on the peripheral division rate averaged across the first 9 months of life.

In order to further explore the source of elevated Ki67 expression of naive CD4^{+} T cells seen in early life, we compared the Ki67^{hi} proportions in mature SP4 thymocytes to Ki67^{hi} proportions in naive CD4^{+} T cells (Figure 3.16). Every data point except for one is below the diagonal, indicating that Ki67^{hi} proportions are greater in mSP4 thymocytes than in naive CD4^{+} T cells up to age 296 days. This observation suggests that recent thymic

emigrants in mice express higher levels of Ki67 than mature cells, which is largely residual expression from recent proliferation in the thymus. To illustrate this point further, we performed linear regression on the ratio of the $Ki67^{hi}$ proportion in mSP4 to that in naive $CD4^+$ T cells against time, and found no evidence for a non-zero slope (for days 5–296, slope = 0.003, $p = 0.25$ for all data, Figure 3.16*B*, left panel; for days 5–100, slope = 0.006, $p = 0.56$, Figure 3.16*B*, right panel). This 'lockstep' of the two proportions is consistent with either (*i*) LIP occurring in equal measure in the thymus and periphery; or (*ii*) our model of low levels of peripheral division and Ki67 being largely inherited from the thymus early in life.

The model describes only the dynamics within the naive $CD4^+$ compartment and makes no predictions about the dynamics of the $CD4^+$ memory compartments. We conclude from the model that in early life there is no significant degree of divisions by naive $CD4^+$ T cells that form more naive $CD4^+$ T cells. However, the model does not provide any insight if LIP occurs in which naive T cells take on a memory phenotype. Future modeling work will attempt to elucidate how naive $CD4^+$ T cells differentiate into the effector and central memory pools, which may include significant degree of LIP. High levels of LIP early in life could possibly explain the presence of "virtual memory" T cells found in mice, which are antigen-inexperienced T cells that express the same markers as typical memory cells (such as CD44) but display a functionally naive phenotype [112]. These cells have been thought to arise from T cell responses against gut bacteria or low-level exposure to pathogens, but many of these virtual memory T cells have not been stimulated by antigen but arise from cytokine signaling [113]. T cells with memory-like phenotypes that are specific for non-encountered antigens in mice have been shown to have the same phenotype as T cells that have undergone LIP [113, 114]. Although naive $CD4^+$ T cells do not exhibit increased proliferation within the naive pool, some maybe undergoing LIP events that result in the development of these virtual memory T cells.

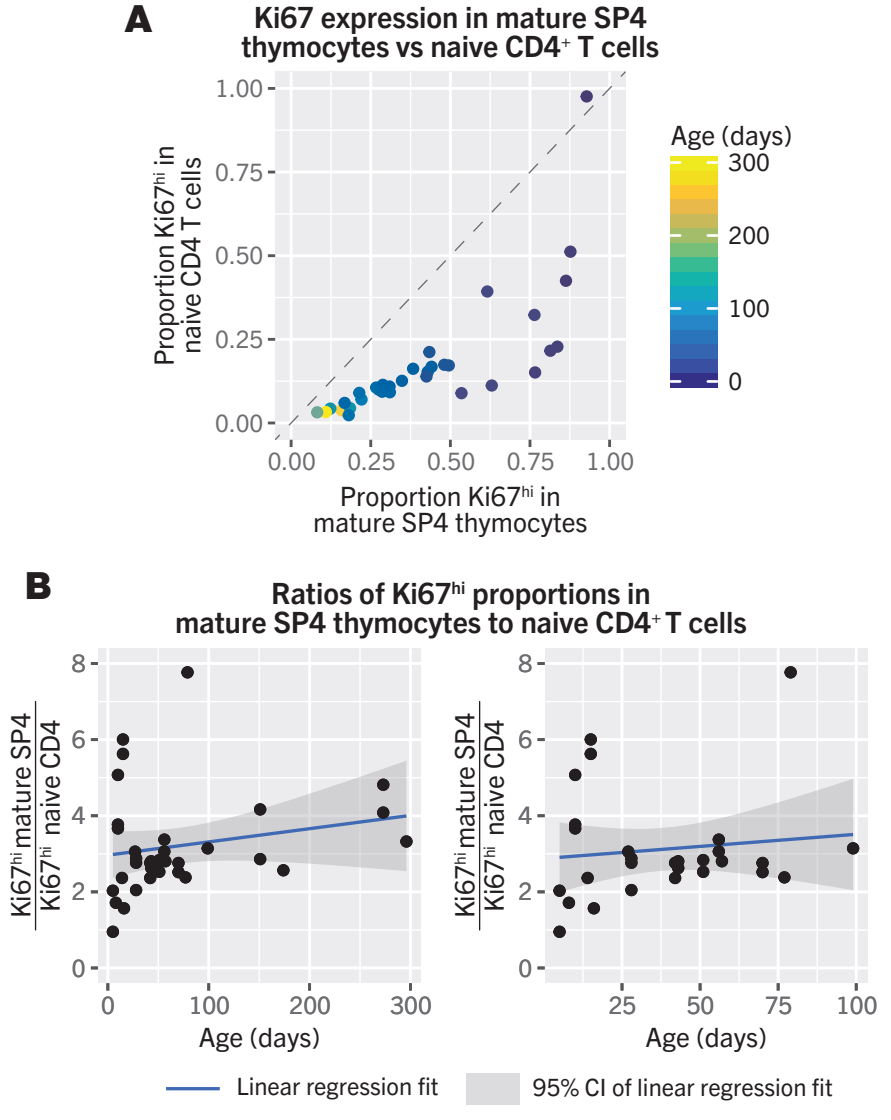


Figure 3.16: Exploring Ki67 expression levels in mature SP4 thymocytes and naive CD4⁺ T cells. (A) The proportions of Ki67^{hi} cells in naive CD4⁺ T cells are plotted against the Ki67^{hi} proportions in mature SP4 thymocytes. **(B)** The ratios of Ki67^{hi} proportions in naive CD4⁺ T cells to those in mature SP4 T cells are plotted against the age of the mouse. Linear regression was performed on all data points (left plot) and on data from mouse younger than 101 days (right plot).

3.6 SUMMARY

Understanding the ontogeny of T cells provides a framework for studying the characteristics of a robust and healthy T cell immunity. Using mathematical models, we dissected the developmental dynamics of murine single-positive CD4 and CD8 thymocytes and naive $CD4^+$ T cells from day 5 to day 296 of life. Our key conclusions were:

- Our models are consistent with previously reported differentiation patterns of DP2 $\rightarrow$ SP4 and DP2 $\rightarrow$ DP3 $\rightarrow$ SP8, and constant per capita differentiation rates explain the influx from the DP compartments to the SP compartments.
- Increased proliferation rates in mSP4s and both iSP8 and mSP8 early in life are needed to explain the high Ki67 levels found in neonatal mice; models that invoked constant division rates failed to explain the Ki67 data (Figure 3.8; Figure 3.11, red curve).
- Our estimates of the lifetime of Ki67 expression both in the thymus and periphery are consistent with previous estimates of a few days [79].
- The ability of the neonatal environment to support LIP and the high Ki67 expression levels found in T cells of neonatal mice have led to the conclusion that significant peripheral expansion of T cells occurs in neonatal mice [46, 115]. However, we find (*i*) Ki67 levels in naive CD4 T cells track (and are consistently lower than) Ki67 levels in mSP4 thymocytes; (*ii*) a model with constant and low levels of peripheral homeostatic division gives good fits to the data; and (*iii*) the estimated interdivision time is consistent with a previous study. Our results therefore support the conclusion that early in life, the elevated Ki67 expression in naive CD4 T cells is not due to LIP but is residual expression from a highly proliferative young thymus.

More questions must be answered to obtain a fuller picture of the ontogeny of peripheral T cells. First, the dynamics of naive $CD8^+$ T cells will be studied with mathematical models to determine if LIP occurs in the $CD8^+$ compartment. We saw that SP4 and SP8 thymocytes have different differentiation and development patterns; in particular, mature SP8 thymocytes exhibit turnover rates that change with time (Figure 3.10*C*). Second, using counts and Ki67 profiles we will develop mathematical models to describe

Function	Parameter	Value	Units
$n_{DP2}(t)$	a_1	0.068	$Days^{-1}$
	a_2	0.0029	$Days^{-1}$
	b_1	-1730000	Number of cells
	b_2	2280000	Number of cells
$k_{DP2}(t)$	c_1	0.018	$Days^{-1}$
	d_1	0.20	Number of cells
$n_{DP3}(t)$	a_3	0.025	$Days^{-1}$
	a_4	0.0041	$Days^{-1}$
	b_3	18000	Number of cells/day
	b_4	180000	Number of cells
	g_1	6.8	Days
	g_2	1.4	Days
	g_3	256	Days
$k_{DP3}(t)$	c_2	0.0094	Days
	d_2	0.33	Number of cells
	d_3	0.63	Number of cells
$n_{mSP4}(t)$	a_5	0.05	$Days^{-1}$
	a_6	0.005	$Days^{-1}$
	b_5	1440000	Number of cells
	b_6	1500000	Number of cells
$k_{mSP4}(t)$	c_4	2.93	$Days^{-1}$
	d_4	0.033	Number of cells
	d_5	0.14	Number of cells

Table 3.3: Parameter values for empirical descriptor functions for DP2, DP3, and mature SP4 thymocytes.

the developmental pathways and dynamics of $CD4^+$ and $CD8^+$ effector and central memory compartments in neonatal mice. Finally, we would like to quantify the size of exported clones during development. Since mid- and post-selection thymocyte proliferation rates are falling with time, we would expect exported clone sizes to fall with age. We could quantify this by implementing a stochastic version of our ODE models (using for example the Gillespie algorithm) to generate the distribution of exported clone sizes.

3.7 MATHEMATICAL MODELS

3.7.1 *DP2 thymocytes*

In all three models, we describe the DP2 compartment by fitting DP2 cell counts and Ki67 profiles with empirical descriptor functions. These forms of these functions were chosen to describe the patterns observed in the data and does not explicitly model the dynamics in these compartments. For the

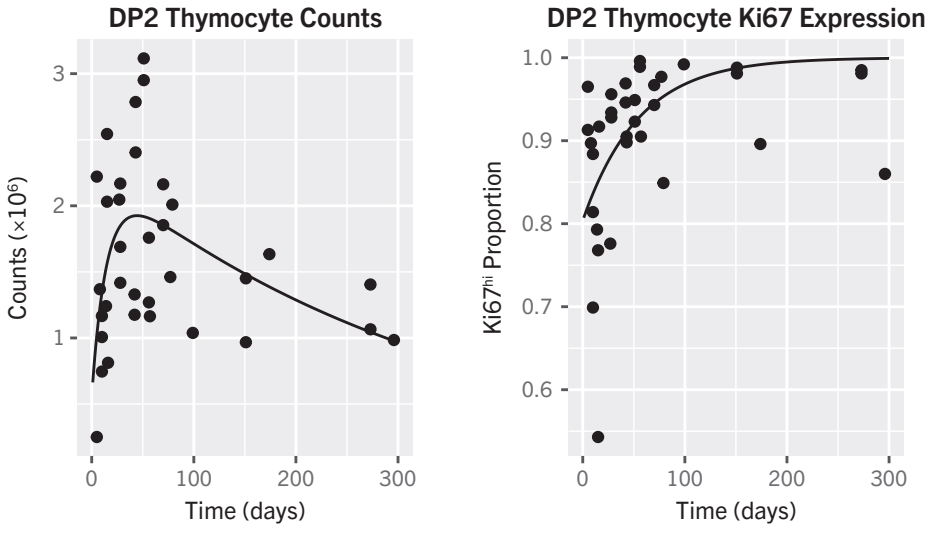


Figure 3.17: Empirical descriptor functions fitted to DP2 cell counts and Ki67^{hi} proportions.

DP2 counts, we used a function $n_{DP2}(t)$ in the form

$$n_{DP2}(t) = b_1 e^{-a_1 t} + b_2 e^{-a_2 t} \tag{3.1}$$

where t is in days, $a_1 > 0$, and $a_2 > 0$. For the DP2 Ki67^{hi} fractions, we used a function $k_{DP2}(t)$ in the form

$$k_{DP2}(t) = -d_1 e^{-c_1 t} + 1 \tag{3.2}$$

where t is in days, $c_1 > 0$, and $d_1 > 0$. Values for the parameters are shown in Table 3.3. Every model assumes that DP2 thymocytes differentiate to immature SP4 thymocytes with a constant rate γ_{DP2}, which estimated to be 0.424 days^{-} (0.237, 0.635; 95% credible interval).

3.7.2 *SP4 thymocytes*

Model #1: Immature SP4 thymocytes differentiate to mature SP4 cells at a constant rate after losing Ki67 expression

The first model ($\mathbf{M}_1$ in Figure 3.6*A*) is described by the following processes:

- **Differentiation**: only Ki67$^{\text{lo}}$ iSP4 $\rightarrow$ mSP4 with a constant rate μ_{SP4}.
- **Ki67 lifetimes**: Ki67$^{\text{hi}}$ iSP4/mSP4 thymocytes lose Ki67 expression at a constant rate β_{SP4}.
- **Turnover of iSP4**: Ki67$^{\text{hi}}$/Ki67$^{\text{lo}}$ iSP4 thymocytes are lost with turnover rates δ^{+}_{iSP4} and δ^{-}_{iSP4} respectively.
- **Turnover of mSP4**: Ki67$^{\text{hi}}$/Ki67$^{\text{lo}}$ mSP4 thymocytes are lost by either thymic egress or death with one-parameter turnover rates δ^{+}_{mSP4} and δ^{-}_{mSP4} respectively.
- **Proliferation**: iSP4 thymocytes do not undergo cell proliferation, and mSP4 thymocytes proliferate with a time-dependent rate $\alpha_{\text{mSP4}}(t)$.

The proliferation rate $\alpha_{\text{mSP4}}(t)$ begins at a maximum rate $\alpha_{\text{mSP4}_{\text{max}}}$, decays exponentially to a minimum rate $\alpha_{\text{mSP4}_{\text{min}}}$, and is given by

$$\alpha_{\text{mSP4}}(t) = \left(\alpha_{\text{mSP4}_{\text{max}}} - \alpha_{\text{mSP4}_{\text{min}}}\right) e^{-\lambda_{\text{mSP4}} t} + \alpha_{\text{mSP4}_{\text{min}}} \tag{3.3}$$

The ODEs for model $\mathbf{M}_1$ are

$$\frac{dI^{+}_{\text{SP4}}}{dt} = \gamma_{\text{DP2}} k_{\text{DP2}}(t) d_{\text{DP2}}(t) - \left(\beta_{\text{SP4}} + \delta^{+}_{\text{iSP4}}\right) I^{+}_{\text{SP4}} \tag{3.4}$$

$$\frac{dI^{-}_{\text{SP4}}}{dt} = \gamma_{\text{DP2}} \left(1 - k_{\text{DP2}}(t)\right) d_{\text{DP2}}(t) + \beta_{\text{SP4}} I^{+}_{\text{SP4}} - \left(\delta^{-}_{\text{iSP4}} + \mu_{\text{SP4}}\right) I^{-}_{\text{SP4}} \tag{3.5}$$

$$\frac{dM^{+}_{\text{SP4}}}{dt} = 2\alpha_{\text{SP4}}(t) M^{-}_{\text{SP4}} + \left(\alpha_{\text{SP4}}(t) - \delta^{+}_{\text{mSP4}} - \beta_{\text{SP4}}\right) M^{+}_{\text{SP4}} \tag{3.6}$$

$$\frac{dM^{-}_{\text{SP4}}}{dt} = \mu_{\text{SP4}} I^{-}_{\text{SP4}} + \beta_{\text{SP4}} M^{+}_{\text{SP4}} - \left(\alpha_{\text{SP4}}(t) + \delta^{-}_{\text{mSP4}}\right) M^{-}_{\text{SP4}} \tag{3.7}$$

with initial conditions $(I^{+^o}_{\text{SP4}}, I^{-^o}_{\text{SP4}}, M^{+^o}_{\text{SP4}}, M^{-^o}_{\text{SP4}})$.

For the Bayesian model, let $D_i^{\text{SP4}} = \left(t_i, x_i^{\text{iSP4}}, k_i^{\text{iSP4}}, x_i^{\text{mSP4}}, k_i^{\text{mSP4}}\right)$ denote data from mouse i at age t_i days, with x_i^{iSP4} iSP4 cells in the thymus that have a proportion k_i^{iSP4} that are Ki67$^{\text{hi}}$ and x_i^{mSP4} mSP4 cells in the thymus that have a proportion k_i^{mSP4} that are Ki67$^{\text{hi}}$. We then define the following variables from the solutions to Equations 3.4–3.7:

$$
\begin{aligned}
X_{\text{iSP4}}(t) &= I^{+}_{\text{SP4}}(t) + I^{-}_{\text{SP4}}(t) \\
K_{\text{iSP4}}(t) &= \frac{I^{+}_{\text{SP4}}(t)}{I^{+}_{\text{SP4}}(t) + I^{-}_{\text{SP4}}(t)} \\
X_{\text{mSP4}}(t) &= M^{+}_{\text{SP4}}(t) + M^{-}_{\text{SP4}}(t) \\
K_{\text{mSP4}}(t) &= \frac{M^{+}_{\text{SP4}}(t)}{M^{+}_{\text{SP4}}(t) + M^{-}_{\text{SP4}}(t)}
\end{aligned}
$$

Model $\mathbf{M}_1$ is then:

$$\begin{bmatrix} \log\left(x_i^{\text{iSP4}}\right) \\ \text{logit}\left(k_i^{\text{iSP4}}\right) \\ \log\left(x_i^{\text{mSP4}}\right) \\ \text{logit}\left(k_i^{\text{mSP4}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP4}}(t_i)\right) \\ \log\left(X_{\text{mSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP4}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP4}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP4}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP4}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP4}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP4}} &\sim \text{PNormal}(0.2857, 0.02) \\
\gamma_{\text{DP2}} &\sim \text{PNormal}(0.8, 0.4) \\
\delta_{\text{iSP4}}^{+} &\sim \text{PNormal}(0.1, 0.2) \\
\delta_{\text{iSP4}}^{-} &\sim \text{PNormal}(0.1, 0.2) \\
\mu_{\text{SP4}} &\sim \text{PNormal}(0.1, 0.1) \\
\delta_{\text{mSP4}}^{+} &\sim \text{PNormal}(0, 0.09) \\
\delta_{\text{mSP4}}^{-} &\sim \text{PNormal}(0, 0.4) \\
\alpha_{\text{mSP4}_{\text{max}}} &\sim \text{PNormal}(1, 0.1) \\
\alpha_{\text{mSP4}_{\text{min}}} &\sim \text{PNormal}(0.01, 0.05) \\
\lambda_{\text{mSP4}} &\sim \text{PNormal}(0.01, 0.01) \\
I_{\text{SP4}}^{-o} &\sim \text{PNormal}(473482, 100000) \\
I_{\text{SP4}}^{+o} &\sim \text{PNormal}(580857, 100000) \\
M_{\text{SP4}}^{+o} &\sim \text{PNormal}(20000, 2000) \\
M_{\text{SP4}}^{-o} &\sim \text{PNormal}(250000, 30000) \\
\sigma_{X_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

Here PNormal(μ, σ) indicates a normal distribution with mean μ and standard deviation σ that is greater than or equal to 0.[2] MVNormal$(\vec{\mu}, S)$ denotes a multivariate normal distribution with mean vector $\vec{\mu}$ and convariance matrix S. Note that the covariances in the matrix $\mathbf{S}$ have been set to 0 since we are assuming that the counts and Ki67 proportions in the mature and immature SP4 compartments are independent from each other.

Model #2: Immature SP4 thymocytes differentiate to mature SP4 cells at a constant rate during or after losing Ki67 expression

The second model ($\mathbf{M}_2$ in Figure 3.6*B*) extends model $\mathbf{M}_1$ also allowing Ki67$^{\text{hi}}$ iSP4 thymocytes to differentiate into the Ki67$^{\text{lo}}$ mSP4 compartment This is modeled by pooling all iSP4 thymocytes together and letting them differentiate into the Ki67$^{\text{lo}}$ mature SP4 compartment with a constant per capita rate μ_{SP4}. The loss of Ki67 expression, turnover rates, and proliferation is modeled identically to model $\mathbf{M}_1$. In particular, the proliferation rate $\alpha_{\text{mSP4}}(t)$ is the same as Equation 3.3.

The ODEs for model $\mathbf{M}_2$ are

$$\frac{dI^+_{\text{SP4}}}{dt} = \gamma_{\text{DP2}} k_{\text{DP2}}(t) d_{\text{DP2}}(t) - \left(\beta_{\text{SP4}} + \delta^+_{\text{iSP4}}\right) I^+_{\text{SP4}} \tag{3.8}$$

$$\frac{dI^-_{\text{SP4}}}{dt} = \gamma_{\text{DP2}} \left(1 - k_{\text{DP2}}(t)\right) d_{\text{DP2}}(t) + \beta_{\text{SP4}} I^+_{\text{SP4}} - \delta^-_{\text{iSP4}} I^-_{\text{SP4}} \tag{3.9}$$

$$\frac{dM^+_{\text{SP4}}}{dt} = 2\alpha_{\text{SP4}}(t) M^-_{\text{SP4}} + \left(\alpha_{\text{SP4}}(t) - \delta^+_{\text{mSP4}} - \beta_{\text{SP4}}\right) M^+_{\text{SP4}} \tag{3.10}$$

$$\frac{dM^-_{\text{SP4}}}{dt} = \mu_{\text{SP4}} \left(I^-_{\text{SP4}} + I^+_{\text{SP4}}\right) + \beta_{\text{SP4}} M^+_{\text{SP4}} - \left(\alpha_{\text{SP4}}(t) + \delta^-_{\text{mSP4}}\right) M^-_{\text{SP4}} \tag{3.11}$$

with initial conditions $(I^{+^o}_{\text{SP4}}, I^{-^o}_{\text{SP4}}, M^{+^o}_{\text{SP4}}, M^{-^o}_{\text{SP4}})$.

[2] To be precise, the prior

$$x \sim \text{PNormal}(\mu, \sigma)$$

has a distribution of

$$\frac{\text{Normal}(x|\mu, \sigma)}{1 - \text{NormalCDF}(0|\mu, \sigma)}$$

where $x \geq 0$ and NormalCDF$(0|\mu, \sigma)$ is the normal cumulative distribution function at 0 to normalize the density to 1.

For the Bayesian model, let $D_i^{\text{SP4}} = \left(t_i, x_i^{\text{iSP4}}, k_i^{\text{iSP4}}, x_i^{\text{mSP4}}, k_i^{\text{mSP4}}\right)$ denote data from mouse i at age t_i days, with x_i^{iSP4} iSP4 cells in the thymus that have a proportion k_i^{iSP4} that are Ki67$^{\text{hi}}$ and x_i^{mSP4} mSP4 cells in the thymus that have a proportion k_i^{mSP4} that are Ki67$^{\text{hi}}$. We then define the following variables from the solutions to Equations 3.8–3.11:

$$
\begin{aligned}
X_{\text{iSP4}}(t) &= I_{\text{SP4}}^{+}(t) + I_{\text{SP4}}^{-}(t) \\
K_{\text{iSP4}}(t) &= \frac{I_{\text{SP4}}^{+}(t)}{I_{\text{SP4}}^{+}(t) + I_{\text{SP4}}^{-}(t)} \\
X_{\text{mSP4}}(t) &= M_{\text{SP4}}^{+}(t) + M_{\text{SP4}}^{-}(t) \\
K_{\text{mSP4}}(t) &= \frac{M_{\text{SP4}}^{+}(t)}{M_{\text{SP4}}^{+}(t) + M_{\text{SP4}}^{-}(t)}
\end{aligned}
$$

Model $\mathbf{M}_2$ is then:

$$\begin{bmatrix} \log\left(x_i^{\text{iSP4}}\right) \\ \text{logit}\left(k_i^{\text{iSP4}}\right) \\ \log\left(x_i^{\text{mSP4}}\right) \\ \text{logit}\left(k_i^{\text{mSP4}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP4}}(t_i)\right) \\ \log\left(X_{\text{mSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP4}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP4}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP4}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP4}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP4}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP4}} &\sim \text{PNormal}(0.2857, 0.02) \\
\gamma_{\text{DP2}} &\sim \text{PNormal}(0.8, 0.4) \\
\delta_{\text{iSP4}}^{+} &\sim \text{PNormal}(0.1, 0.2) \\
\delta_{\text{iSP4}}^{-} &\sim \text{PNormal}(0.1, 0.2) \\
\mu_{\text{SP4}} &\sim \text{PNormal}(0.1, 0.1) \\
\delta_{\text{mSP4}}^{+} &\sim \text{PNormal}(0, 0.09) \\
\delta_{\text{mSP4}}^{-} &\sim \text{PNormal}(0, 0.4) \\
\alpha_{\text{mSP4}_{\text{max}}} &\sim \text{PNormal}(1, 0.1) \\
\alpha_{\text{mSP4}_{\text{min}}} &\sim \text{PNormal}(0.01, 0.05) \\
\lambda_{\text{mSP4}} &\sim \text{PNormal}(0.01, 0.01) \\
I_{\text{SP4}}^{-o} &\sim \text{PNormal}(473482, 100000) \\
I_{\text{SP4}}^{+o} &\sim \text{PNormal}(580857, 100000) \\
M_{\text{SP4}}^{+o} &\sim \text{PNormal}(20000, 2000) \\
M_{\text{SP4}}^{-o} &\sim \text{PNormal}(250000, 30000) \\
\sigma_{X_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

The covariances in the matrix $\mathbf{S}$ again have been set to 0.

Model #3: Differentiation from immature SP4 to mature SP4 requires division to occur

The third and best fitting model ($\mathbf{M}_3$ in Figure 3.6C) describe differentiation by allowing immature SP4 thymocytes to differentiate to mature SP4 only by dividing at a constant rate μ_{SP4} after losing Ki67 expression. Thus, differentiation occurs with a flow from Ki67$^{\text{lo}}$ iSP4 $\rightarrow$ Ki67$^{\text{hi}}$ mSP4 only by dividing at a constant rate μ_{SP4}. The loss of Ki67 expression, turnover rates, and proliferation is modeled identically to model $\mathbf{M}_1$ and $\mathbf{M}_2$.

The ODE equations for model $\mathbf{M}_3$ is

$$\frac{dI^{+}_{\text{SP4}}}{dt} = \gamma_{\text{DP2}} k_{\text{DP2}}(t) d_{\text{DP2}}(t) - \left(\beta_{\text{SP4}} + \delta^{+}_{\text{iSP4}}\right) I^{+}_{\text{SP4}} \tag{3.12}$$

$$\frac{dI^{-}_{\text{SP4}}}{dt} = \gamma_{\text{DP2}}\left(1 - k_{\text{DP2}}(t)\right) d_{\text{DP2}}(t) + \beta_{\text{SP4}} I^{+}_{\text{SP4}} - \left(\delta^{-}_{\text{iSP4}} + \mu_{\text{SP4}}\right) I^{-}_{\text{SP4}} \tag{3.13}$$

$$\frac{dM^{+}_{\text{SP4}}}{dt} = 2\mu_{\text{SP4}} I^{-}_{\text{SP4}} + 2\alpha_{\text{SP4}}(t) M^{-}_{\text{SP4}} + \left(\alpha_{\text{SP4}}(t) - \delta^{+}_{\text{mSP4}} - \beta_{\text{SP4}}\right) M^{+}_{\text{SP4}} \tag{3.14}$$

$$\frac{dM^{-}_{\text{SP4}}}{dt} = \beta_{\text{SP4}} M^{+}_{\text{SP4}} - \left(\alpha_{\text{SP4}}(t) + \delta^{-}_{\text{mSP4}}\right) M^{-}_{\text{SP4}} \tag{3.15}$$

with initial conditions $(I^{+^o}_{\text{SP4}}, I^{-^o}_{\text{SP4}}, M^{+^o}_{\text{SP4}}, M^{-^o}_{\text{SP4}})$.

For the Bayesian model, let $D_i^{\text{SP4}} = \left(t_i, x_i^{\text{iSP4}}, k_i^{\text{iSP4}}, x_i^{\text{mSP4}}, k_i^{\text{mSP4}}\right)$ denote data from mouse i at age t_i days, with x_i^{iSP4} iSP4 cells in the thymus that have a proportion k_i^{iSP4} that are Ki67$^{\text{hi}}$ and x_i^{mSP4} mSP4 cells in the thymus that have a proportion k_i^{mSP4} that are Ki67$^{\text{hi}}$. We then define the following variables from the solutions to Equations 3.12–3.15:

$$\begin{aligned}
X_{\text{iSP4}}(t) &= I^{+}_{\text{SP4}}(t) + I^{-}_{\text{SP4}}(t) \\
K_{\text{iSP4}}(t) &= \frac{I^{+}_{\text{SP4}}(t)}{I^{+}_{\text{SP4}}(t) + I^{-}_{\text{SP4}}(t)} \\
X_{\text{mSP4}}(t) &= M^{+}_{\text{SP4}}(t) + M^{-}_{\text{SP4}}(t) \\
K_{\text{mSP4}}(t) &= \frac{M^{+}_{\text{SP4}}(t)}{M^{+}_{\text{SP4}}(t) + M^{-}_{\text{SP4}}(t)}
\end{aligned}$$

Model $\mathbf{M}_3$ is then:

$$\begin{bmatrix} \log\left(x_i^{\text{iSP4}}\right) \\ \text{logit}\left(k_i^{\text{iSP4}}\right) \\ \log\left(x_i^{\text{mSP4}}\right) \\ \text{logit}\left(k_i^{\text{mSP4}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP4}}(t_i)\right) \\ \log\left(X_{\text{mSP4}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP4}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP4}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP4}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP4}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP4}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP4}} &\sim \text{PNormal}(0.2857, 0.05) \\
\gamma_{\text{DP2}} &\sim \text{PNormal}(0.8, 0.4) \\
\delta_{\text{iSP4}}^{+} &\sim \text{PNormal}(0.6, 1) \\
\delta_{\text{iSP4}}^{-} &\sim \text{PNormal}(0.26, 0.2) \\
\mu_{\text{SP4}} &\sim \text{PNormal}(0.1, 0.5) \\
\delta_{\text{mSP4}}^{+} &\sim \text{PNormal}(0.3, 0.3) \\
\delta_{\text{mSP4}}^{-} &\sim \text{PNormal}(0.3, 0.2) \\
\alpha_{\text{mSP4}_{\text{max}}} &\sim \text{PNormal}(6, 3) \\
\alpha_{\text{mSP4}_{\text{min}}} &\sim \text{PNormal}(0.01, 0.005) \\
\lambda_{\text{mSP4}} &\sim \text{PNormal}(0.01, 0.01) \\
I_{\text{SP4}}^{-o} &\sim \text{PNormal}(473482, 100000) \\
I_{\text{SP4}}^{+o} &\sim \text{PNormal}(580857, 100000) \\
M_{\text{SP4}}^{+o} &\sim \text{PNormal}(30000, 5000) \\
M_{\text{SP4}}^{-o} &\sim \text{PNormal}(250000, 30000) \\
\sigma_{X_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP4}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

The covariances in the matrix $\mathbf{S}$ again have been set to 0.

3.7.3 *DP3 thymocytes*

We describe the DP3 compartment by fitting DP3 cell counts and Ki67 profiles with empirical descriptor functions. For the DP3 counts, we used a function $d_{\mathrm{DP3}}(t)$ in the form

$$d_{\mathrm{DP3}}(t) = b_3(t - g_2)e^{a_3(t+g_3)} + b_4(1 - e^{-a_4(t+g_4)}) \tag{3.16}$$

and for $\mathrm{Ki67^{hi}}$ fractions,

$$k_{\mathrm{DP3}}(t) = -d_2 e^{-c_2 t} + d_3 \tag{3.17}$$

where t is in days, $a_3, a_4, b_3, b_4, g_2, g_3, g_4, c_2, d_2, d_3 > 0$. Values for the parameters are shown in Table 3.3. Every model assumes that $\mathrm{Ki67^{hi}}$/$\mathrm{Ki67^{lo}}$ DP3 thymocytes differentiate to $\mathrm{Ki67^{hi}}$/$\mathrm{Ki67^{lo}}$ iSP8 thymocytes respectively with a constant rate γ_{DP3}.

3.7.4 *SP8 thymocytes*

The first model shown in Figure 3.10*A* is the simplest model considered and describes the following processes:

- **Differentiation**: $\mathrm{Ki67^{hi/lo}}$ iSP8 $\rightarrow$ $\mathrm{Ki67^{hi/lo}}$ mSP4 with a constant rate μ_{SP4} with constant per capita rates μ^{+}_{SP8} and μ^{-}_{SP8} respectively.
- **Ki67 lifetimes**: $\mathrm{Ki67^{hi}}$ iSP8/mSP8 thymocytes lose Ki67 expression at a constant rate β_{SP8}.
- **Turnover of iSP8**: $\mathrm{Ki67^{hi}}$/$\mathrm{Ki67^{lo}}$ iSP8 thymocytes are lost with turnover rates $\delta^{+}_{\mathrm{iSP8}}$ and $\delta^{-}_{\mathrm{iSP8}}$ respectively.
- **Turnover of mSP8**: $\mathrm{Ki67^{hi}}$/$\mathrm{Ki67^{lo}}$ mSP8 thymocytes are lost by either thymic egress or death with one-parameter turnover rates $\delta^{+}_{\mathrm{mSP8}}$ and $\delta^{-}_{\mathrm{mSP8}}$ respectively.
- **Proliferation**: iSP8 thymocytes proliferation with a constant rate $\alpha^{c}_{\mathrm{iSP8}}$, and mSP8 thymocytes proliferate with a time-dependent rate $\alpha^{c}_{\mathrm{mSP8}}$.

The ODEs for this model are

$$\frac{dI^{-}_{\mathrm{SP8}}}{dt} = \gamma_{\mathrm{DP3}}\left(1 - k_{\mathrm{DP3}}(t)\right) d_{\mathrm{DP3}}(t) + \beta_{\mathrm{SP8}} I^{+}_{\mathrm{SP8}} - \left(\alpha^{c}_{\mathrm{iSP8}} + \mu^{-}_{\mathrm{iSP8}} + \delta^{-}_{\mathrm{iSP8}}\right) I^{-}_{\mathrm{SP8}} \tag{3.18}$$

$$\frac{dI^{+}_{\mathrm{SP8}}}{dt} = \gamma_{\mathrm{DP3}} k_{\mathrm{DP3}}(t) d_{\mathrm{DP3}}(t) + 2\alpha^{c}_{\mathrm{iSP8}} I^{-}_{\mathrm{SP8}} + \left(\alpha^{c}_{\mathrm{iSP8}} - \mu^{+}_{\mathrm{SP8}} - \beta_{\mathrm{SP8}} - \delta^{+}_{\mathrm{iSP8}}\right) I^{+}_{\mathrm{SP8}} \tag{3.19}$$

$$\frac{dM^{-}_{\mathrm{SP8}}}{dt} = \mu^{-}_{\mathrm{SP8}} I^{-}_{\mathrm{SP8}} + \beta_{\mathrm{SP8}} M^{+}_{\mathrm{SP8}} - \left(\alpha^{c}_{\mathrm{mSP8}} + \delta^{-}_{\mathrm{mSP8}}\right) M^{-}_{\mathrm{SP8}} \tag{3.20}$$

$$\frac{dM^{+}_{\mathrm{SP8}}}{dt} = \mu^{+}_{\mathrm{SP8}} I^{+}_{\mathrm{SP8}} + 2\alpha^{c}_{\mathrm{mSP8}} M^{-}_{\mathrm{SP8}} + \left(\alpha^{c}_{\mathrm{mSP8}} - \beta_{\mathrm{SP8}} - \delta^{+}_{\mathrm{mSP8}}\right) M^{+}_{\mathrm{SP8}} \tag{3.21}$$

with initial conditions $\left(I^{-^{o}}_{\mathrm{SP8}}, I^{+^{o}}_{\mathrm{SP8}}, M^{+^{o}}_{\mathrm{SP8}}, M^{-^{o}}_{\mathrm{SP8}}\right)$.

For the Bayesian model, let $D^{\mathrm{SP8}}_{i} = \left(t_i, x^{\mathrm{iSP8}}_{i}, k^{\mathrm{iSP8}}_{i}, x^{\mathrm{mSP8}}_{i}, k^{\mathrm{mSP8}}_{i}\right)$ denote data from mouse i at age t_i days, with x^{iSP8}_{i} iSP8 cells in the thymus that have a proportion k^{iSP8}_{i} that are Ki67$^{\mathrm{hi}}$ and x^{mSP8}_{i} mSP8 cells in the thymus that have a proportion k^{iSP8}_{i} that are Ki67$^{\mathrm{hi}}$. We then define the following variables from the solutions to Equations 3.18–3.21:

$$\begin{aligned}
X_{\mathrm{iSP8}}(t) &= I^{+}_{\mathrm{SP8}}(t) + I^{-}_{\mathrm{SP8}}(t) \\
K_{\mathrm{iSP8}}(t) &= \frac{I^{+}_{\mathrm{SP8}}(t)}{I^{+}_{\mathrm{SP8}}(t) + I^{-}_{\mathrm{SP8}}(t)} \\
X_{\mathrm{mSP8}}(t) &= M^{+}_{\mathrm{SP8}}(t) + M^{-}_{\mathrm{SP8}}(t) \\
K_{\mathrm{mSP8}}(t) &= \frac{M^{+}_{\mathrm{SP8}}(t)}{M^{+}_{\mathrm{SP8}}(t) + M^{-}_{\mathrm{SP8}}(t)}
\end{aligned}$$

$$\begin{bmatrix} \log\left(x_i^{\text{iSP8}}\right) \\ \text{logit}\left(k_i^{\text{iSP8}}\right) \\ \log\left(x_i^{\text{mSP8}}\right) \\ \text{logit}\left(k_i^{\text{mSP8}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP8}}(t_i)\right) \\ \log\left(X_{\text{mSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP8}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP8}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP8}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP8}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP8}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP8}} &\sim \text{PNormal}(0.2857, 0.05) \\
\gamma_{\text{DP3}} &\sim \text{PNormal}(2.6, 0.7) \\
\delta_{\text{iSP8}}^{+} &\sim \text{PNormal}(3, 1) \\
\delta_{\text{iSP8}}^{-} &\sim \text{PNormal}(3.5, 1) \\
\mu_{\text{SP8}}^{+} &\sim \text{PNormal}(5, 5) \\
\mu_{\text{SP8}}^{-} &\sim \text{PNormal}(0.5, 0.2) \\
\delta_{\text{mSP8}}^{+} &\sim \text{PNormal}(3, 1) \\
\delta_{\text{mSP8}}^{-} &\sim \text{PNormal}(3, 1) \\
\alpha_{\text{iSP8}}^{c} &\sim \text{PNormal}(1, 1) \\
\alpha_{\text{mSP8}}^{c} &\sim \text{PNormal}(0.5, 0.2) \\
I_{\text{SP8}}^{-o} &\sim \text{PNormal}(10000, 2000) \\
I_{\text{SP8}}^{+o} &\sim \text{PNormal}(15000, 2000) \\
M_{\text{SP8}}^{+o} &\sim \text{PNormal}(10000, 3000) \\
M_{\text{SP8}}^{-o} &\sim \text{PNormal}(120000, 30000) \\
\sigma_{X_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

The covariances in the matrix $\mathbf{S}$ have been set to 0.

The second model in Figure 3.10*B* extends the first model by allowing for proliferation rates that change with time in both iSP8 and mSP8 compartments. These proliferation rates start at a maximum rate and exponentially decay to a minimum rate. They are given by the following equations:

$$\alpha_{\text{iSP8}}(t) = \left(\alpha_{\text{iSP8}_{\max}} - \alpha_{\text{iSP8}_{\min}}\right)e^{-\eta_{\text{iSP8}}t} + \alpha_{\text{iSP8}_{\min}} \tag{3.22}$$

$$\alpha_{\text{mSP8}}(t) = \left(\alpha_{\text{mSP8}_{\max}} - \alpha_{\text{mSP8}_{\min}}\right)e^{-\eta_{\text{mSP8}}t} + \alpha_{\text{mSP8}_{\min}} \tag{3.23}$$

The ODEs for this model are

$$\begin{aligned}\frac{dI^{-}_{\text{SP8}}}{dt} = {} & \gamma_{\text{DP3}}\left(1 - k_{\text{DP3}}(t)\right)d_{\text{DP3}}(t) + \beta_{\text{iSP8}}I^{+}_{\text{SP8}} \\ & - \left(\alpha_{\text{iSP8}}(t) + \mu^{-}_{\text{iSP8}} + \delta^{-}_{\text{SP8}}\right)I^{-}_{\text{SP8}}\end{aligned} \tag{3.24}$$

$$\begin{aligned}\frac{dI^{+}_{\text{SP8}}}{dt} = {} & \gamma_{\text{DP3}}k_{\text{DP3}}(t)d_{\text{DP3}}(t) + 2\alpha_{\text{iSP8}}(t)I^{-}_{\text{SP8}} \\ & + \left(\alpha_{\text{iSP8}}(t) - \mu^{+}_{\text{SP8}} - \beta_{\text{SP8}} - \delta^{+}_{\text{iSP8}}\right)I^{+}_{\text{SP8}}\end{aligned} \tag{3.25}$$

$$\frac{dM^{-}_{\text{SP8}}}{dt} = \mu^{-}_{\text{SP8}}I^{-}_{\text{SP8}} + \beta_{\text{SP8}}M^{+}_{\text{SP8}} - \left(\alpha_{\text{mSP8}}(t) + \delta^{-}_{\text{mSP8}}\right)M^{-}_{\text{SP8}} \tag{3.26}$$

$$\begin{aligned}\frac{dM^{+}_{\text{SP8}}}{dt} = {} & \mu^{+}_{\text{SP8}}I^{+}_{\text{SP8}} + 2\alpha_{\text{mSP8}}(t)M^{-}_{\text{SP8}} \\ & + \left(\alpha_{\text{mSP8}}(t) - \beta_{\text{SP8}} - \delta^{+}_{\text{mSP8}}\right)M^{+}_{\text{SP8}}\end{aligned} \tag{3.27}$$

For the Bayesian model, let $D_i^{\text{SP8}} = \left(t_i, x_i^{\text{iSP8}}, k_i^{\text{iSP8}}, x_i^{\text{mSP8}}, k_i^{\text{mSP8}}\right)$ denote data from mouse i at age t_i days, with x_i^{iSP8} iSP8 cells in the thymus that have a proportion k_i^{iSP8} that are Ki67$^{\text{hi}}$ and x_i^{mSP8} mSP8 cells in the thymus that have a proportion k_i^{iSP8} that are Ki67$^{\text{hi}}$. We then define the following variables from the solutions to Equations 3.24–3.27:

$$\begin{aligned}X_{\text{iSP8}}(t) &= I^{+}_{\text{SP8}}(t) + I^{-}_{\text{SP8}}(t) \\ K_{\text{iSP8}}(t) &= \frac{I^{+}_{\text{SP8}}(t)}{I^{+}_{\text{SP8}}(t) + I^{-}_{\text{SP8}}(t)} \\ X_{\text{mSP8}}(t) &= M^{+}_{\text{SP8}}(t) + M^{-}_{\text{SP8}}(t) \\ K_{\text{mSP8}}(t) &= \frac{M^{+}_{\text{SP8}}(t)}{M^{+}_{\text{SP8}}(t) + M^{-}_{\text{SP8}}(t)}\end{aligned}$$

$$\begin{bmatrix} \log\left(x_i^{\text{iSP8}}\right) \\ \text{logit}\left(k_i^{\text{iSP8}}\right) \\ \log\left(x_i^{\text{mSP8}}\right) \\ \text{logit}\left(k_i^{\text{mSP8}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP8}}(t_i)\right) \\ \log\left(X_{\text{mSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP8}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP8}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP8}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP8}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP8}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP8}} &\sim \text{PNormal}(0.2857, 0.05) \\
\gamma_{\text{DP3}} &\sim \text{PNormal}(2.6, 0.7) \\
\delta_{\text{iSP8}}^{+} &\sim \text{PNormal}(3, 1) \\
\delta_{\text{iSP8}}^{-} &\sim \text{PNormal}(3.5, 1) \\
\mu_{\text{SP8}}^{+} &\sim \text{PNormal}(5, 5) \\
\mu_{\text{SP8}}^{-} &\sim \text{PNormal}(0.5, 0.2) \\
\delta_{\text{mSP8}}^{+} &\sim \text{PNormal}(3, 1) \\
\delta_{\text{mSP8}}^{-} &\sim \text{PNormal}(3, 1) \\
\alpha_{\text{iSP8}_{\text{max}}} &\sim \text{PNormal}(3.1, 1) \\
\alpha_{\text{iSP8}_{\text{min}}} &\sim \text{PNormal}(0.01, 0.003) \\
\eta_{\text{iSP8}} &\sim \text{PNormal}(0.1, 0.03) \\
\alpha_{\text{mSP8}_{\text{max}}} &\sim \text{PNormal}(0.5, 0.2) \\
\alpha_{\text{mSP8}_{\text{min}}} &\sim \text{PNormal}(0.1, 0.3) \\
\eta_{\text{mSP8}} &\sim \text{PNormal}(0.1, 0.03) \\
I_{\text{SP8}}^{-^o} &\sim \text{PNormal}(10000, 2000) \\
I_{\text{SP8}}^{+^o} &\sim \text{PNormal}(15000, 2000) \\
M_{\text{SP8}}^{+^o} &\sim \text{PNormal}(10000, 3000) \\
M_{\text{SP8}}^{-^o} &\sim \text{PNormal}(120000, 30000) \\
\sigma_{X_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

The covariances in the matrix $\mathbf{S}$ have been set to 0.

The third model in Figure 3.10*C* extends the previous model by allowing turnover rates of mSP8 thymocytes to change with time. The turnover rates start at a maximum and exponentially decay to a minimum rate. They are given by the following equations:

$$\delta^{+}_{\mathrm{mSP8}}(t) = \left(\delta^{+}_{\mathrm{mSP8}_{\max}} - \delta^{+}_{\mathrm{mSP8}_{\min}}\right)e^{-\nu t} + \delta^{+}_{\mathrm{mSP8}_{\min}} \tag{3.28}$$

$$\delta^{-}_{\mathrm{mSP8}}(t) = \left(\delta^{-}_{\mathrm{mSP8}_{\max}} - \delta^{-}_{\mathrm{mSP8}_{\min}}\right)e^{-\nu t} + \delta^{-}_{\mathrm{mSP8}_{\min}} \tag{3.29}$$

The ODEs for this model are

$$\frac{dI^{-}_{\mathrm{SP8}}}{dt} = \gamma_{\mathrm{DP3}}\left(1 - k_{\mathrm{DP3}}(t)\right)d_{\mathrm{DP3}}(t) + \beta_{\mathrm{SP8}}I^{+}_{\mathrm{SP8}} - \left(\alpha_{\mathrm{iSP8}}(t) + \mu^{-}_{\mathrm{iSP8}} + \delta^{-}_{\mathrm{SP8}}\right)I^{-}_{\mathrm{SP8}} \tag{3.30}$$

$$\frac{dI^{+}_{\mathrm{SP8}}}{dt} = \gamma_{\mathrm{DP3}}k_{\mathrm{DP3}}(t)d_{\mathrm{DP3}}(t) + 2\alpha_{\mathrm{iSP8}}(t)I^{-}_{\mathrm{SP8}} + \left(\alpha_{\mathrm{iSP8}}(t) - \mu^{+}_{\mathrm{SP8}} - \beta_{\mathrm{iSP8}} - \delta^{+}_{\mathrm{iSP8}}\right)I^{+}_{\mathrm{SP8}} \tag{3.31}$$

$$\frac{dM^{-}_{\mathrm{SP8}}}{dt} = \mu^{-}_{\mathrm{SP8}}I^{-}_{\mathrm{SP8}} + \beta_{\mathrm{SP8}}M^{+}_{\mathrm{SP8}} - \left(\alpha_{\mathrm{mSP8}}(t) + \delta^{-}_{\mathrm{mSP8}}(t)\right)M^{-}_{\mathrm{SP8}} \tag{3.32}$$

$$\frac{dM^{+}_{\mathrm{SP8}}}{dt} = \mu^{+}_{\mathrm{SP8}}I^{+}_{\mathrm{SP8}} + 2\alpha_{\mathrm{mSP8}}(t)M^{-}_{\mathrm{SP8}} + \left(\alpha_{\mathrm{mSP8}}(t) - \beta_{\mathrm{SP8}} - \delta^{+}_{\mathrm{mSP8}}(t)\right)M^{+}_{\mathrm{SP8}} \tag{3.33}$$

For the Bayesian model, let $D^{\mathrm{SP8}}_i = \left(t_i, x^{\mathrm{iSP8}}_i, k^{\mathrm{iSP8}}_i, x^{\mathrm{mSP8}}_i, k^{\mathrm{mSP8}}_i\right)$ denote data from mouse i at age t_i days, with x^{iSP8}_i iSP8 cells in the thymus that have a proportion k^{iSP8}_i that are Ki67$^{\mathrm{hi}}$ and x^{mSP8}_i mSP8 cells in the thymus that have a proportion k^{iSP8}_i that are Ki67$^{\mathrm{hi}}$. We then define the following variables from the solutions to Equations 3.30–3.33:

$$\begin{aligned}
X_{\mathrm{iSP8}}(t) &= I^{+}_{\mathrm{SP8}}(t) + I^{-}_{\mathrm{SP8}}(t) \\
K_{\mathrm{iSP8}}(t) &= \frac{I^{+}_{\mathrm{SP8}}(t)}{I^{+}_{\mathrm{SP8}}(t) + I^{-}_{\mathrm{SP8}}(t)} \\
X_{\mathrm{mSP8}}(t) &= M^{+}_{\mathrm{SP8}}(t) + M^{-}_{\mathrm{SP8}}(t) \\
K_{\mathrm{mSP8}}(t) &= \frac{M^{+}_{\mathrm{SP8}}(t)}{M^{+}_{\mathrm{SP8}}(t) + M^{-}_{\mathrm{SP8}}(t)}
\end{aligned}$$

$$\begin{bmatrix} \log\left(x_i^{\text{iSP8}}\right) \\ \text{logit}\left(k_i^{\text{iSP8}}\right) \\ \log\left(x_i^{\text{mSP8}}\right) \\ \text{logit}\left(k_i^{\text{mSP8}}\right) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \log\left(X_{\text{iSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{iSP8}}(t_i)\right) \\ \log\left(X_{\text{mSP8}}(t_i)\right) \\ \text{logit}\left(K_{\text{mSP8}}(t_i)\right) \end{bmatrix}, \mathbf{S} \right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{X_{\text{iSP8}}} & 0 & 0 & 0 \\ 0 & \sigma_{K_{\text{iSP8}}} & 0 & 0 \\ 0 & 0 & \sigma_{X_{\text{mSP8}}} & 0 \\ 0 & 0 & 0 & \sigma_{K_{\text{mSP8}}} \end{pmatrix}$$

$$\begin{aligned}
\beta_{\text{SP8}} &\sim \text{PNormal}(0.2857, 0.05) \\
\gamma_{\text{DP3}} &\sim \text{PNormal}(2.6, 0.7) \\
\delta^{+}_{\text{iSP8}} &\sim \text{PNormal}(3, 1) \\
\delta^{-}_{\text{iSP8}} &\sim \text{PNormal}(3.5, 1) \\
\mu^{+}_{\text{SP8}} &\sim \text{PNormal}(3, 3) \\
\mu^{-}_{\text{SP8}} &\sim \text{PNormal}(0.5, 0.2) \\
\delta^{+}_{\text{mSP8}_{\text{max}}} &\sim \text{PNormal}(2, 1) \\
\delta^{+}_{\text{mSP8}_{\text{min}}} &\sim \text{PNormal}(0.1, 0.5) \\
\delta^{-}_{\text{mSP8}_{\text{max}}} &\sim \text{PNormal}(1, 1) \\
\delta^{-}_{\text{mSP8}_{\text{min}}} &\sim \text{PNormal}(0.1, 0.5) \\
\nu &\sim \text{PNormal}(0.1, 0.025) \\
\alpha_{\text{iSP8}_{\text{max}}} &\sim \text{PNormal}(3.1, 1) \\
\alpha_{\text{iSP8}_{\text{min}}} &\sim \text{PNormal}(0.01, 0.003) \\
\eta_{\text{iSP8}} &\sim \text{PNormal}(0.1, 0.03) \\
\alpha_{\text{mSP8}_{\text{max}}} &\sim \text{PNormal}(0.5, 0.2) \\
\alpha_{\text{mSP8}_{\text{min}}} &\sim \text{PNormal}(0.1, 0.3) \\
\eta_{\text{mSP8}} &\sim \text{PNormal}(0.1, 0.03) \\
I^{-^o}_{\text{SP8}} &\sim \text{PNormal}(10000, 2000) \\
I^{+^o}_{\text{SP8}} &\sim \text{PNormal}(15000, 2000) \\
M^{+^o}_{\text{SP8}} &\sim \text{PNormal}(10000, 3000) \\
M^{-^o}_{\text{SP8}} &\sim \text{PNormal}(120000, 30000) \\
\sigma_{X_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{iSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{X_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2) \\
\sigma_{K_{\text{mSP8}}} &\sim \text{HalfCauchy}(0, 2)
\end{aligned}$$

The covariances in the matrix $\mathbf{S}$ have been set to 0.

3.7.5 *Naive CD4⁺ T cells*

The model for naive CD4⁺ T cells is shown in Figure 3.13. The counts and Ki67^{hi} proportions were described with empirical descriptor functions $d_{mSP4}(t)$ and $k_{mSP4}(t)$ respectively:

$$d_{mSP4}(t) = -b_5 e^{-a_5 t} + b_6 e^{-a_6 t} \tag{3.34}$$

$$k_{mSP4}(t) = d_4 e^{-c_4 t} + d_5 \tag{3.35}$$

Ki67^{hi}/Ki67^{lo} mSP4 thymocytes differentiate to the Ki67^{hi}/Ki67^{lo} naive CD4⁺ compartment with constant per capita rates φ^+/φ^- respectively. Naive CD4⁺ lose Ki67 expression with a rate β_n and divide with rate α_n. Both compartments have a common turnover rate δ_n. The ODEs for this model are

$$\frac{dN^-_{CD4}}{dt} = \varphi^-\big(1 - k_{mSP4}(t)\big)d_{mSP4}(t) + \beta_n N^+_{CD4} - \big(\alpha + \delta_n\big)N^-_{CD4} \tag{3.36}$$

$$\frac{dN^+_{CD4}}{dt} = \varphi^+ k_{mSP4}(t) d_{mSP4}(t) + 2\alpha N^-_{CD4} + \big(\alpha - \beta_n - \delta_n\big)N^+_{CD4} \tag{3.37}$$

CHAPTER 4

Counting Lymphocytes Using Thoracic Duct Cannulations

4.1 INTRODUCTION

4.1.1 *Aim of the chapter*

Accurately quantifying the number of T cells and B cells in the secondary lymphoid organs (SLOs) of a mouse has eluded the field since obtaining all SLOs from a mouse is not possible. Counting lymphocytes by dissecting the spleen and every visible lymph node provides only a lower bound on these numbers because some SLOs might be missed. In particular, we would be counting recirculating T cells such as T_{CM} and T_{EM} cells, in contrast with T_{RM} cells which reside in peripheral sites like the gut. Knowing the number of T cells and B cells and their subsets would provide a basis for understanding the characteristics of a mature adaptive immune system. For example, $CD4^+$ T cells were once thought to primarily reside in the gut rather than the SLOs, leading to inaccurate conclusions about T cell biology; quantification of the lymphocyte population sizes in individual organs is needed for a precise understanding of the immune system [116]. In particular, these numbers would help us quantify the diversity of the T cell receptor and B cell receptor repertoires; for example, we could estimate clonal sizes by knowing the number of unique receptors in the repertoire and the number of lymphocytes. In addition, quantifying exact number of T cell subsets would characterize how many naive and memory T cells are needed for physiologically healthy immune system. Carefully quantifying the numbers of T cells and their subsets in the SLOs will help us elucidate the properties needed for healthy immune responses.

We can do a back-of-the-envelope calculation to roughly estimate the number of T cells and B cells in the mouse. Consider that a mouse spleen has about 100×10^6 white blood cells. Approximately 30% of these are αβ T cells, of which 70% are $CD4^+$ and 30% are $CD8^+$ T cells [117]. Thus,

- **$CD4^+$:** 21×10^6 cells in the spleen
- **$CD8^+$:** 9×10^6 cells in the spleen

For B cells, they approximately comprise 50% of splenic white blood cells, giving us an estimate of

- **B cells:** 50×10^6 cells in the spleen

For lymph nodes, we can roughly estimate another 100×10^6 white cells by estimating about 30 lymph nodes in the mouse [118], each with 3×10^6 cells. Assuming 70% of lymph nodes are αβ T cells (with the same 70%/30% split for $CD4^+$ and $CD8^+$ T cells) and 30% are B cells, we get estimates of

- **$CD4^+$:** 49×10^6 cells in the LNs
- **$CD8^+$:** 21×10^6 cells in the LNs
- **B cells:** 30×10^6 cells in the LNs

By combining spleen and LN estimates together, we get rough estimates of

- **$CD4^+$:** 60×10^6 cells
- **$CD8^+$:** 30×10^6 cells
- **B cells:** 80×10^6 cells

In this chapter, we aim to determine accurate estimates of the number of T cells and B cells found in the lymph nodes and spleens of the mouse. We attempted to do so by utilizing a thoracic duct (TD) cannulation technique.

4.1.2 *Overview of the experimental approach*

We performed TD cannulation on male wild-type C57Bl6 mice of ages 10–12 weeks to collect circulating T cells and B cells and then used these counts to estimate the total number of T cells and B cells in the mouse. Figure 4.1 shows a schematic of how we aimed to estimate lymphocyte numbers with this approach. We cannulated the thoracic ducts of wild-type mice in order to collect circulating T cells and B cells (Figure 4.1*A*). A proportion of these circulating lymphocytes are directed to the collection vessel instead of recirculating back into the lymph nodes and spleens, and this results in a decrease in lymphocyte numbers in these organs (Figure 4.1*B*–*C*). Thus, the number of lymphocytes collected during the cannulation should be proportional to the drop in lymphocyte numbers in the spleen and lymph nodes. We collected counts from spleens and lymph nodes from non-cannulated mice and from mice that undergone sham surgeries as well. These data served as the

baseline for lymphocyte numbers in order measure the effect on cell numbers in the cannulated mice.

We also performed sham surgeries in which every step was performed except the actual cannulation of the thoracic duct so that we could observe if the surgery procedure itself had an effect on cell counts. In these sham surgery experiments, we found that the surgery itself causes a drop in lymphocyte numbers in the lymph nodes and spleens of mice. As we will discuss in Section 4.3.2, the loss of cells in the lymph nodes and spleens could be a response to many effects due to the surgery, including the induction of glucocorticoids from the stress of surgery and the use of general anesthesia. Thus, the drop in cell numbers in the cannulated mice is a combination of lymphocytes collected from the cannulation, a loss of cells due to the surgical procedure itself, and potentially the redistribution of cells to tissues.

In order to minimize the loss of cells due to sample preparation, the number of times that samples were filtered and centrifuged were kept at a minimum. Lymph samples were filtered once before being counted, and lymph node and spleen samples were filtered and centrifuged once before being counted. Each filtering step and centrifugation would cause some experimental loss of cells, but these were kept to a minimum and should cancel each other out in the calculations done in this chapter. The sample sizes are 6 mice for cannulations, 7 mice for sham surgeries, and 7 mice as controls.

We illustrate how we calculate total cell numbers in Figure 4.2. The cannulation causes some fractional drop in the cell counts in the SLOs (Figure 4.2*A*–B), with the cells being collected by the cannula in the collecting vessels (Figure 4.2*C*). Since the fractional drop in cell numbers in the SLOs is proportional to the number collected, we can back-calculate the total number of lymphocytes in the mouse. If f is the fractional drop in the cell counts in the lymphoid tissue due to the cannulation and X is the number of cells collected from the cannulation, then the total number of cells N is

$$\begin{aligned} fN &= X \\ N &= \frac{X}{f} \end{aligned} \tag{4.1}$$

where f is estimated by dividing the spleen/LN counts in cannulated mice by the spleen/LN counts in non-cannulated mice (Figure 4.2*C*).

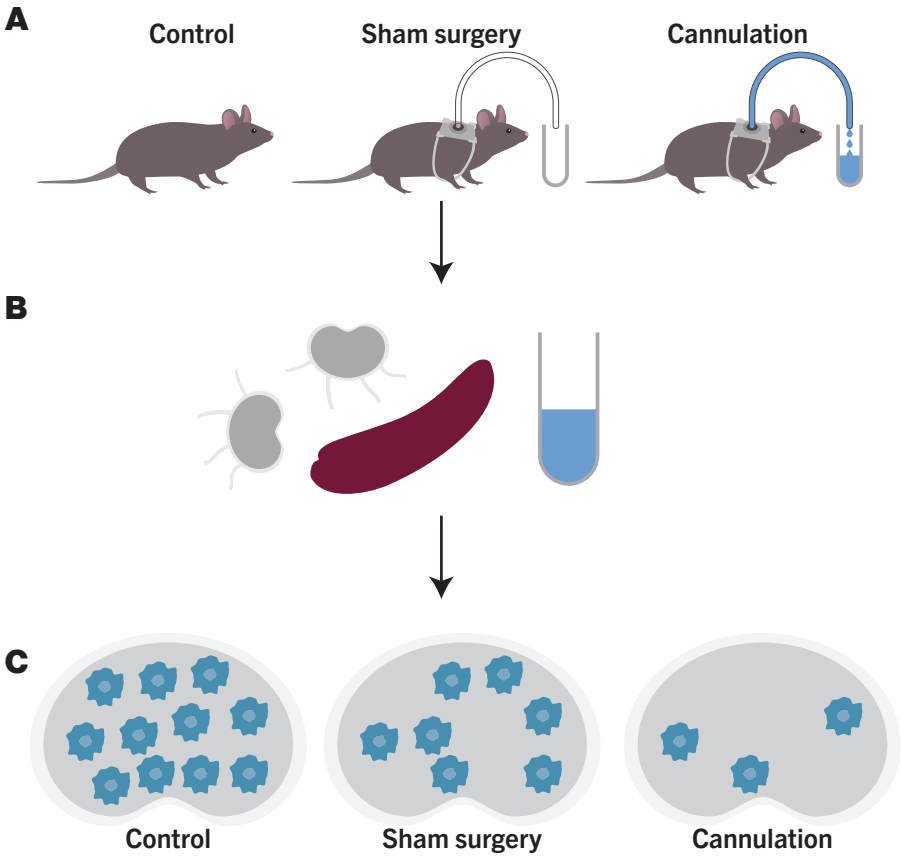


Figure 4.1: Overview of how lymphocyte numbers were estimated using thoracic duct cannulation. (A) We studied three groups of mice: control mice, mice that underwent sham surgeries, and mice that had their thoracic ducts cannulated. **(B)** We collected the lymph nodes, spleens, and lymph (for the cannulated mice) from these mice. **(C)** We obtained cell counts and quantified the proportions of different T cell and B cell subsets. We determined the drop in counts due to cells being lost to surgical stress by comparing the counts in control mice to the counts in the sham surgery mice. We also determined the drop in counts due to the cannulation by comparing the counts in sham surgery mice to cannulated mice. In the cannulated mice, the drop in the SLOs due to the cannulation after correcting for cell loss from the surgery should be represented by the number of cells collected.

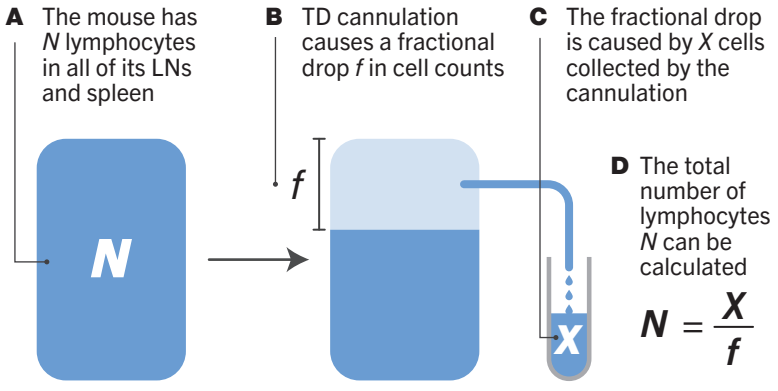


Figure 4.2: Schematic of calculate total lymphocyte numbers (A) Suppose that a mouse has *N* lymphocytes in all of its lymph nodes and spleen. **(B)** The thoracic duct cannulation drains recirculating lymphocytes away from the body, causing a fractional decrease *f* in the cell counts in the secondary lymphoid organs. **(C)** The fractional drop is caused by the cells collected during the cannulation. The number of cells collected is denoted as *X*. **(D)** The total number *N* can be determined with *X* and *f* using Equation 4.1.

4.2 RESULTS

4.2.1 *Stress from the surgical procedure has differential effects on cell numbers in secondary lymphoid organs*

We wanted to see how cell counts in SLOs change due to the stress from surgical procedures. We performed sham surgeries in which every step of the TD cannulation procedure except for the insertion of the cannula into the thoracic duct. The lymph nodes (mesenteric, cervical, inguinal, and axillary) and spleens were collected from control mice and sham surgery mice. The mesenteric lymph nodes (MLNs) were analyzed by themselves, and the cervical, inguinal, and axillary lymph nodes were pooled together in these experiments since these lymph nodes are not directly drained by the thoracic duct (henceforth referred to as OLNs for the *other* lymph nodes). Lymphocytes were counted and stained for T cell and B cell subsets (gating strategies shown in Figures 4.3–4.4).

The following T cell subsets were identified: naive CD4 (live TCRβ$^+$CD4$^+$ CD25$^-$ CD44$^-$ CD62L$^+$), CD4 effector memory (live TCRβ$^+$ CD4$^+$ CD25$^-$ CD44$^+$ CD62L$^-$), CD4 central memory (live TCRβ$^+$ CD4$^+$ CD25$^-$ CD44$^+$ CD62L$^+$), total CD4 (live TCRβ$^+$ CD4$^+$), naive CD8 (live TCRβ$^+$ CD8$^+$

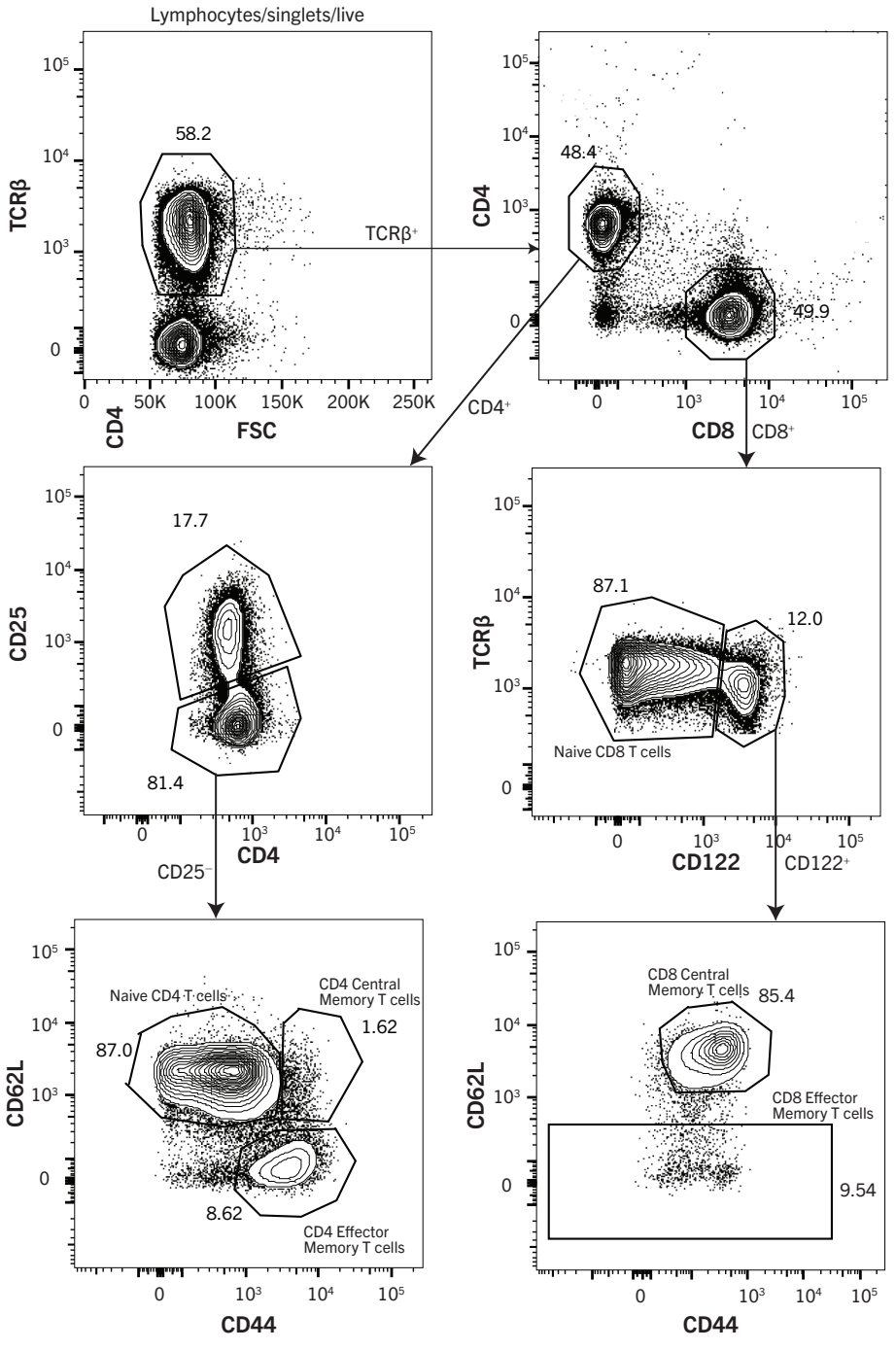


Figure 4.3: Gating strategy for CD4 and CD8 T cells. Gating strategy shown already have lymphocytes, singlets and live cells gated (Figure 3.3).

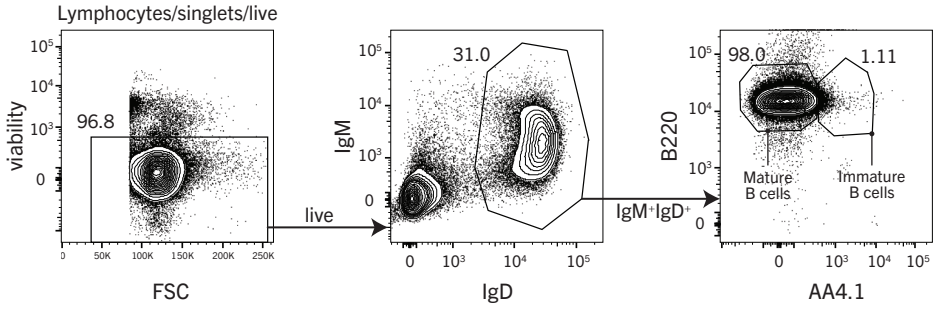


Figure 4.4: Gating strategy for B cells. Gating strategy shown already have lymphocytes, singlets and live cells gated (Figure 3.3)

CD122$^-$), CD8 effector memory (live TCRβ$^+$ CD8$^+$ CD122$^+$ CD62L$^-$), CD8 central memory (live TCRβ$^+$ CD8$^+$ CD122$^+$ CD62L$^+$ CD44$^+$), and total CD8 (live TCRβ$^+$ CD8$^+$). The B cell subsets identified were mature B cells (live IgM$^+$ IgD$^+$ AA4.1$^-$) and transitional B cells (live IgM$^+$ IgD$^+$ AA4.1$^+$).

In Figure 4.5, we calculated the ratios of cell counts of T cell subsets from control mice to cell counts from mice that underwent sham surgeries. We calculated 99% confidence intervals by performing bootstrapping with 2000 replicates. We see consistently that the OLNs show the largest decrease in cells counts in the sham surgery mice. In all T cell subsets except for CD4$^+$ effector memory and CD8$^+$ effector memory T cells, we observed decreases in counts in the MLNs in the sham surgery group. The spleen showed no discernible effects on cell counts from the surgery, which is clearly shown by the fact that all confidence intervals overlap with the ratio of 1.

We repeated the same calculations in transitional and mature B cells in Figure 4.6 by performing bootstrapping with 2000 replicates. We again see that OLNs exhibit the largest decrease in cell numbers in both transitional and mature B cells. MLNs show decreases in cell counts as well, except the confidence interval for transitional B cells in the MLNs contain the ratio value of 1. Unlike the T cells subsets, both transitional and mature B cell numbers decrease in the spleens due to the sham surgery.

The data clearly show that T cells and B cells in the different SLOs are reduced in numbers to different extents by surgery-induced stress. Since MLNs, OLNs, and spleens show differential changes from the stress of the surgical process itself, we could not pool all SLOs together and treat them as one homogenous group. This prevented us from using the approach used to

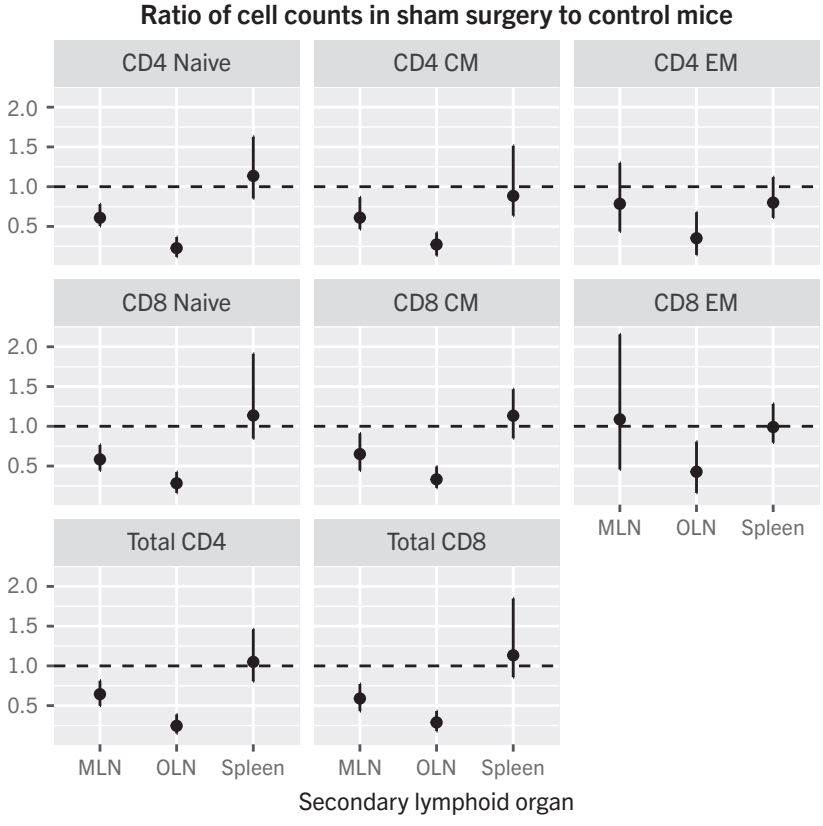


Figure 4.5: Loss of T cells due to surgical stress. The ratio of the mean number of cells in mice that underwent sham surgery to that of control mice are shown for all T cell subsets. The mean ratios are plotted as circles, and the bars represent bootstrap 99% confidence intervals. OLNs consistently showed the largest decrease due to surgical stress whereas MLNs showed decreases in all compartments except the CD4 effector memory and CD8 effector memory compartments. Spleens appeared to not be affected by the stress from surgery. (7 sham surgery mice, 7 control mice).

derive Equation 4.1, and so we will discuss in Section 4.3.3 how the model was extended to include these differential effects.

4.2.2 *Cannulating the thoracic duct causes different secondary lymphoid organs to lose proportionally different numbers of cells*

We cannulated the thoracic ducts of mice and collected lymph for 18–24 hours. After collecting the lymph, we collected the lymph nodes (mesenteric, cervical, inguinal, and axillary) and spleens of the mice. We counted the number of cells in the lymph, lymph nodes, and spleens and stained for T cell

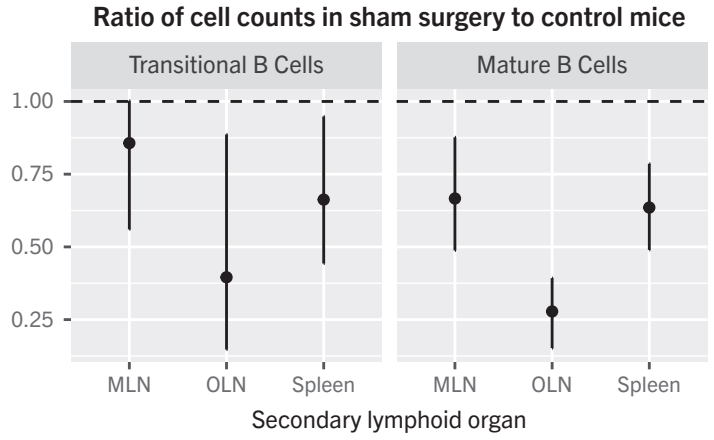


Figure 4.6: Loss of B cells due to surgical stress. The ratio of the mean number of cells in mice that underwent sham surgery to that of control mice are shown for transitional B cells and mature B cells. The mean ratios are plotted as circles, and the bars represent bootstrap 99% confidence intervals. As observed in the T cell subsets, OLNs consistently showed the largest decrease in cell counts due to surgical stress. MLNs showed decreases in cell counts as well, but the 99% confidence intervals includes a ratio of 1 for transitional B cells. Unlike T cell subsets, spleens appeared to have decreases in B cell counts due to the stress from surgery. (7 sham surgery mice, 7 control mice).

and B cell subsets described in Section 4.2.1. As before, the cervical, inguinal, and axillary LNs (OLNs) were pooled together in these experiments.

In Figure 4.7, we show the ratios of cell counts of T cell subsets from cannulated mice to the mean of the counts of mice that underwent sham surgery mice. Bootstrapping with 2000 replicates was performed to create 99% confidence intervals. The ratios for each cannulated mouse are shown as gray circles, and mean ratios are shown in black circles. The mean ratios of the MLNs were less than 1 in all T cell compartments, with statistical significance found in the $CD4^+$ naive, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, total $CD4^+$, and total $CD8^+$ compartments (one-tailed 1-sample t-test, $p < 0.01$). The mean ratios in the spleens showed less consistent decreases in the cannulated mice, with only the $CD4^+$ naive, $CD8^+$ naive, total $CD4^+$, and total $CD8^+$ compartments showing statistically significance means less than 1 (one-tailed 1-sample t-test, $p < 0.01$). The OLNs had no compartment where the mean ratio was less than 1, and all compartments showed no statistical difference from mean ratios of 1 (two-tailed 1-sample t-test, $p < 0.01$).

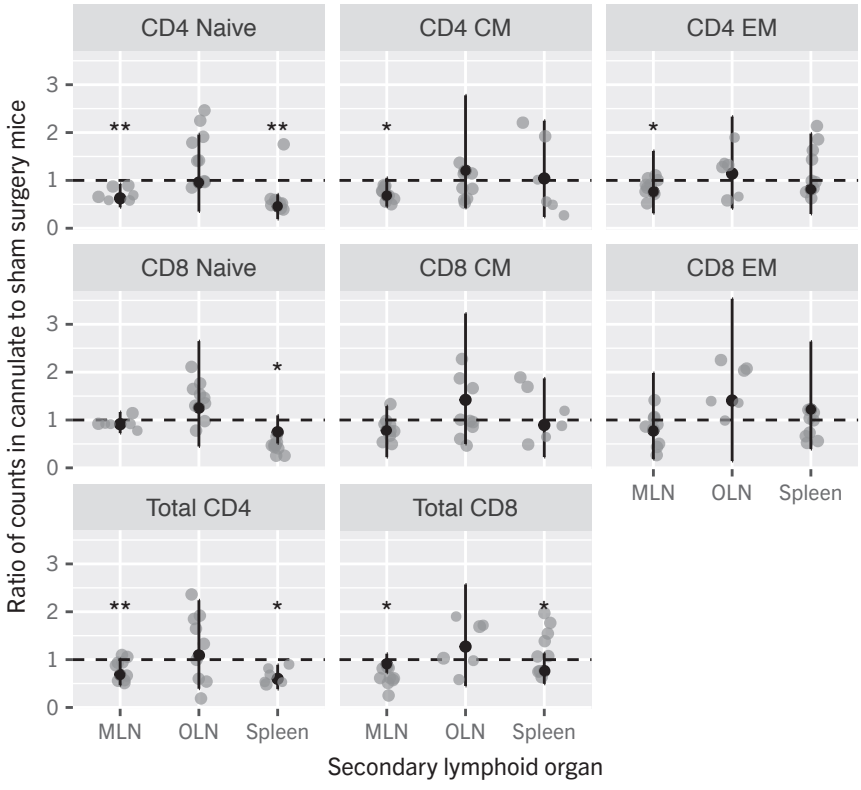


Figure 4.7: The decrease in T cell counts due to thoracic duct cannulations. The ratio of the number of cells in cannulated mice to the mean number of cells in mice that underwent sham surgery are shown for all T cell subsets. The mean ratios are plotted as black circles, and the bars represent bootstrap 99% confidence intervals. Ratios from the counts of individual cannulated mice are shown in lighter gray circles. (7 sham surgery mice, 6 cannulated mice).

In Figure 4.8, we performed similar calculations for transitional and mature B cells. The 99% confidence intervals were calculated with bootstrapping with 2000 replicates, ratios for each cannulated mouse are shown as gray circles, and mean ratios are shown in black circles. Unlike the T cell subsets, neither the transitional B cells or mature B cells show statistically significant decreases in counts in the cannulated mice compared to the mice that underwent sham surgeries (one-tailed 1-sample t-test, $p < 0.01$).

The data exhibit heterogeneity in the changes caused by TD cannulation across the lymphocyte subsets and the different SLOs. Because of these differential changes, these data suggest again that SLOs cannot be pooled.

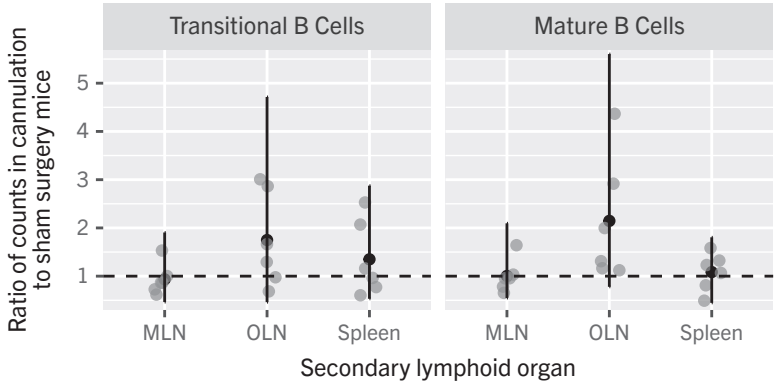


Figure 4.8: The effect of thoracic duct cannulations on B cell counts. The ratio of cell numbers in cannulated mice to the mean number of cells in surgery mice for B cells. None of the SLOs in either transitional B cells and mature B cells show any statistical differences between cannulated and sham surgery mice. (7 sham surgery mice, 6 cannulated mice).

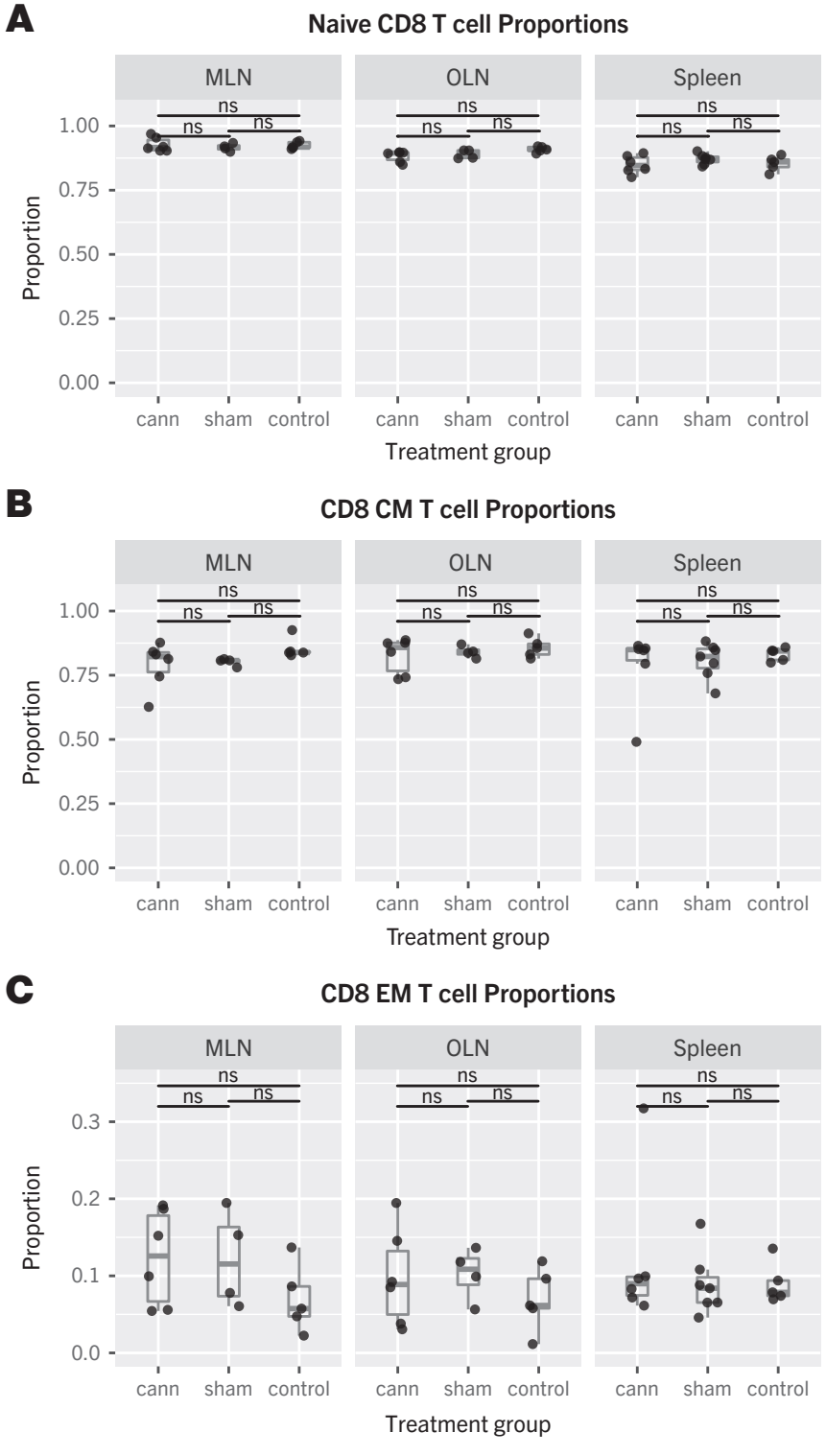

A
Naive CD8 T cell Proportions
MLN
OLN
Spleen
ns
Proportion
1.00
0.75
0.50
0.25
0.00
cann
sham
control
Treatment group
B
CD8 CM T cell Proportions
C
CD8 EM T cell Proportions
0.3
0.2
0.1
0.0

Figure 4.9: (facing page) Proportions of CD8 naive, central memory, and effector memory T cells in the cannulated, sham surgery, and control mice. (A) Proportions of naive CD8 T cells out of $CD8^+$ cells in all three treatment groups. **(B)** Proportions of CD8 central memory T cells out of $CD8^+$ cells in all three treatment groups. **(C)** Proportions of CD8 effector memory T cells out of $CD8^+$ cells in all three treatment groups. No statistically significant differences were found in any groups. (7 sham surgery mice, 6 cannulated mice)

4.2.3 *Surgical stress and thoracic duct cannulations do not change the relative frequencies of major T and B cell subsets within SLOs*

We were curious if certain T cell and B cell subsets were more susceptible to the effect of surgical stress. Figures 4.9–4.10 show the proportions of the $CD4^+$ and $CD8^+$ T cell subsets within all three treatment groups. In the $CD8^+$ T cell compartments, no statistically significant differences can be seen among the control, sham surgery, and cannulated mice in $CD8^+$ naive, central memory, and effector memory T cells in all SLOs (Mann-Whitney U Test, $p < 0.01$). In the $CD4^+$ T cell compartments, no statistically significant differences can be seen among the control, sham surgery, and cannulated mice in the CD4 central memory and effector memory T cells (Mann-Whitney U Test, $p < 0.01$). There was a statistically significant difference in the naive $CD4^+$ T cells in the OLNs between cannulated and control mice (Mann-Whitney U Test, $p < 0.01$), but the effect size was small as the difference between the mean proportions of the two groups is 0.069. Similarly, Figure 4.11 show that proportions of transitional and mature B cells were not statistically different among the three treatment groups in any of the SLOs (Mann-Whitney U Test, $p < 0.01$).

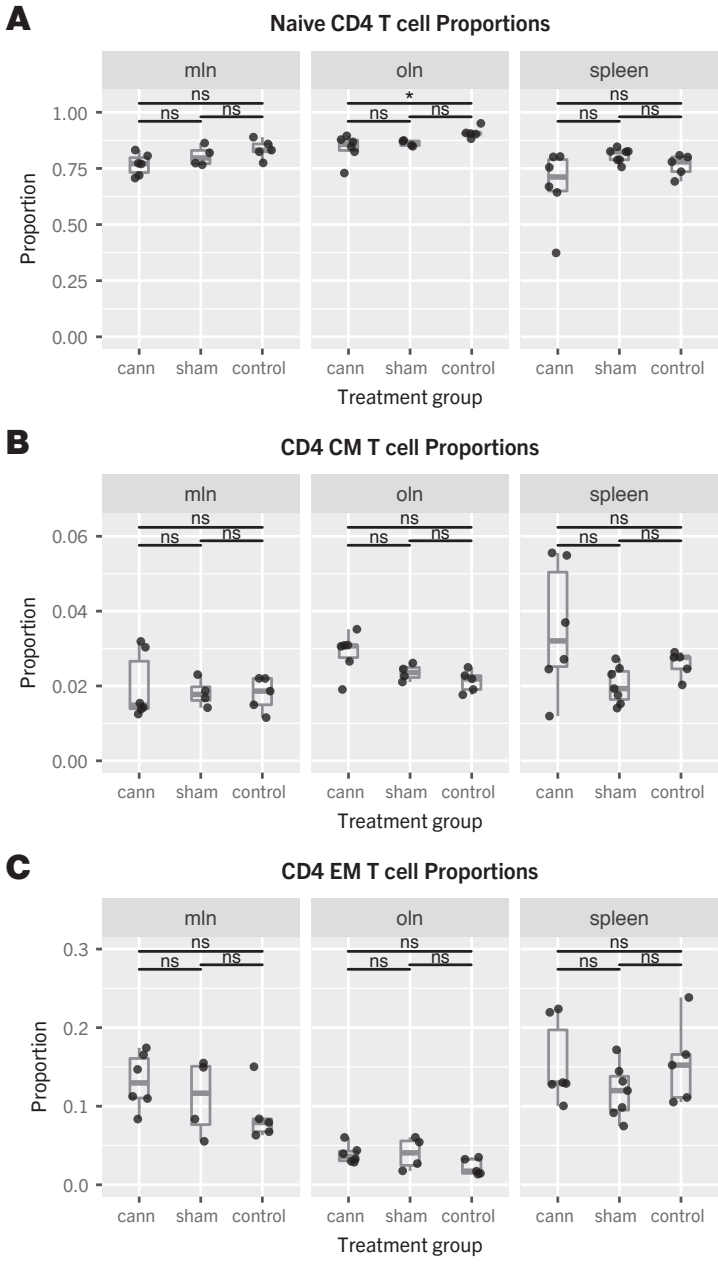


Figure 4.10: Proportions of CD4⁺ naive, central memory, and effector memory T cells in the cannulated, sham surgery, and control mice. (A) Proportions of naive CD4⁺ T cells out of total CD4⁺ cells in all three treatment groups. **(B)** Proportions of CD4⁺ central memory T cells out of total CD4⁺ cells in all three treatment groups. **(C)** Proportions of CD4⁺ effector memory T cells out of total CD4⁺ cells in all three treatment groups. No statistically significant differences were found in any groups except for naive CD4⁺ T cells in the OLNs between cannulated and control mice. (7 control mice, 7 sham surgery mice, 6 cannulated mice).

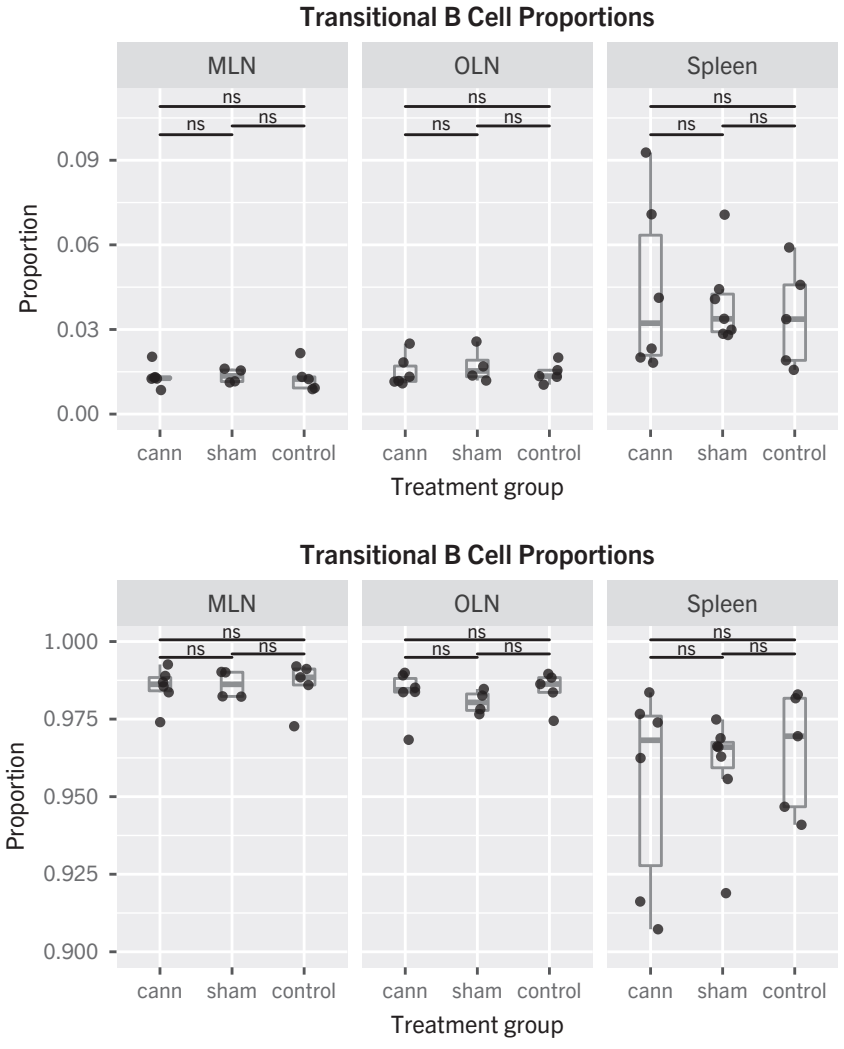


Figure 4.11: Proportions of transitional and mature B cells in the cannulated, sham surgery, and control mice. (7 control mice, 7 sham surgery mice, 6 cannulated mice).

4.3 EXTENDING THE MODEL

4.3.1 *Extending The Model To Account For Differential Effects On Different Secondary Lymphoid Organs*

The different effects on cell numbers by the sham surgeries and the thoracic duct cannulations challenged the assumption that all SLOs are affected identically by these procedures. This heterogeneity prevented us from using Equation 4.1 and required a more complex model that captures these effects. The different changes in cell numbers due to the thoracic duct cannulation may reflect the anatomy of lymphatic drainage, which we will discuss in the next section.

4.3.2 *The anatomy of the lymphatics in the mouse influences how thoracic duct cannulation effects cell numbers in the spleen and in different lymph nodes*

Carefully understanding the anatomy of the lymphatic drainage of the lymph nodes studied here helps us predict the effect of draining circulating lymphocytes with the cannulation of the thoracic duct (Figure 4.12). Since the cannulation drains the thoracic duct, the outflow from the mesenteric lymph nodes is being drained directly. The spleen has no direct lymphatic inputs or outputs; instead, recirculating lymphocytes enter and leave the spleen through the blood. The TD cannulation reduces the number of lymphocytes that recirculate back into the blood, which in turn reduces the number of cells that to return the spleen. The cervical, inguinal, and axillary lymph nodes are indirectly affected by the cannulation since the afferent lymphatics will bring fewer cells into these lymph nodes due to recirculating lymphocytes being shunted away by the cannulation. Because these SLOs are affected by the TD cannulation differently, we expect different fractional changes in counts in TD cannulation experiments in short time scales, as we observed in Section 4.2.2.

Although we might intuitively expect counts in the OLNs to decrease as well, the data show no difference in counts between cannulated mice and mice that underwent sham surgeries for T cells. The relatively short time scales of the cannulation might be not long enough to cause decreases in cell counts in these lymph nodes. Since the lymphatic ducts that ultimately

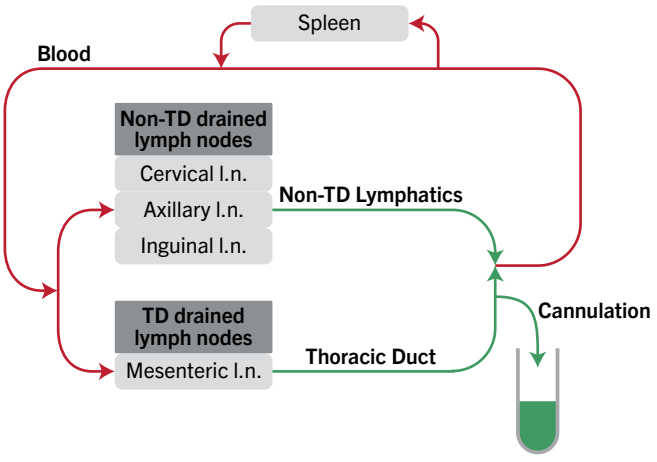


Figure 4.12: Schematic of the lymphatic drainage of different lymph nodes. The mesenteric lymph nodes are drained by the thoracic duct, from which the TD cannulation directly shunts lymph away. The other lymph nodes are drained by other lymphatic vessels, and thus the cannulation does not directly shunt lymph away from these lymph nodes. Lymphocytes recirculate through the spleen via blood.

return lymph from these lymph nodes to the systemic circulation are not cannulated, their drainage rates are mostly likely not increased in the short term. So despite the fact that the inflow of lymphocytes is decreased due to the TD cannulation, the outflow of these lymph nodes likely are not affected, and the short period of time that the cannulation collects in these experiments lymph might not show a discernible effect in counts.

4.3.3 *Estimating total lymphocyte counts from thoracic duct cannulation data*

The data clearly showed that the TD-drained lymph nodes, the non-TD-drained lymph nodes, and the spleen must be considered separately. Equation 4.1 needs to be expanded to account for cell loss from the surgical procedure while considering the differential effects on cell counts from the TD cannulation. We sketch out the details of the extended model in Figure 4.13.

Figure 4.13*A–C* shows how we account for the fact that TD cannulation causes different fractional drops in cell counts in the lymph nodes directly drained by the TD, the non-TD drained lymph nodes, and the spleen. The lymphocyte subset of interest have a total count of N cells in all of the

SLOs. A proportion f_m of these cells is in the TD-drained lymph nodes (represented by the **m**esenteric lymph nodes), a proportion f_o of these cells is in the non-TD-drained lymph nodes (represented the **o**ther lymph nodes), and the remaining $f_{spl} = (1 - f_m - f_o)$ proportion is in the spleen (Figure 4.13*A*). After cannulating the mouse and collecting cells for a period of time, drained recirculating cells do not return to the SLOs, resulting in a drop in counts. We denote the proportional decrease in the counts of each compartment as p, q, and r for the TD-drained LNs, the non-TD-drained LNs, and the spleen respectively (Figure 4.13*B*). Multiplying the proportion of the counts represented by the compartment, the proportional drop due to cannulation, and the total number of lymphocytes N gives us the number of cells of that compartment that were collected by the cannulation (Figure 4.13*C*). For example, the number of cells collected from the spleen is given by

$$p(1 - f_m - f_o)N\,.$$

This does not account for the cells that die or disappear due to the stress of the surgical procedure. To account for these cells, we denote p^s as the proportion of the compartment that is lost due to the stress of the surgery itself (Figure 4.13, lightest colored portion). Then, after the mouse is cannulated, we denote p^c as the proportion of the remaining $(1 - p^s)$ proportion of the compartment that is drained by the cannulation (Figure 4.13, middle colored portion). p^c is defined relative to the remaining compartment after cells lost to the surgical stress and not to the full compartment. The middle colored portion in Figure 4.13 represents the number of cells that should be collected by the cannulation.

Combining the approaches in Figures 4.13*C* and 4.13*D* together, we get the full model in Figure 4.13*E*. Thus, the total number of lymphocytes collected from the cannulation from all three compartments after accounting for cell loss due to the surgical stress is

$$(1 - p^s)p^c(1 - f_m - f_o)N + (1 - q^s)q^c f_m N + (1 - r^s)r^c f_o N \qquad (4.2)$$

where p^s, q^s, and r^s are the fractional losses due to surgical stress, p^c, q^c, and r^c are the fractional losses due to the cannulation after correcting for the loss due to surgical stress, and N is the total number of lymphocytes in all SLOs. If X is the number of cells collected by the cannulation, then we

can set Equation 4.2 equal to X and solve for N, yielding

$$N = \frac{X}{(1-p^s)p^c(1-f_m-f_o)+(1-q^s)q^c f_m+(1-r^s)r^c f_o}. \qquad (4.3)$$

We estimated p^s, q^s and r^s by comparing cell counts in mice that received sham-surgeries to control mice. We estimated p^c, q^c and r^c by comparing cell counts in cannulated mice to mice that received sham-surgeries.

4.3.4 *The differential changes in the secondary lymphoid organs by the cannulations and surgical stress prevent the calculation of total lymphocyte numbers*

Since we cannot pool cell counts from the different, Equation 4.2 was used to estimate total numbers of the T cell and B cell subsets. This equation has three unknown parameters that we cannot estimate, namely f_m, f_o, and f_{spl} = $(1 - f_m - f_o)$. These fractions represent what fraction of the total lymphocytes in the SLOs is represented by the TD-drained lymph nodes, the non-TD-drained lymph nodes, and the spleen respectively. Since these fractions are unknown, we scanned across all possible combination of values for f_m and f_o and used these in Equation 4.2 to calculate total numbers. We constrained the fractions for each compartment to be between 0.1 and 0.9 since values outside of this range seem to be physiologically unrealistic values.

We estimated p^s, q^s, and r^s by dividing the mean sham surgery counts by the mean control count of the spleens, MLNs, and OLNs respectively. For every cannulated mouse, we then estimated p^c, q^c, and r^c by dividing the cell count of the lymphocyte subset by the mean sham surgery counts of the spleens, MLNs, and OLNs respectively. We also bootstrapped the cannulation and sham surgery data to create 95% confidence intervals for all values of f_m and f_o.

In Figures 4.14–4.15, the calculated total cell counts for the total CD4 and total CD8 T cell compartments. In Figure 4.14*A* and Figure 4.15*A*, we show heatmaps that represent the estimated numbers of total CD4 and CD8 T cells respectively by assuming different values for f_m, f_o, and f_s. The estimated CD4 T cell numbers ranged from 11×10^6 to 46×10^6, and the estimated CD8 T cell numbers ranged from to 35×10^6 to 100×10^6, which do not agree with our rough estimates derived in Section 4.1.1. In

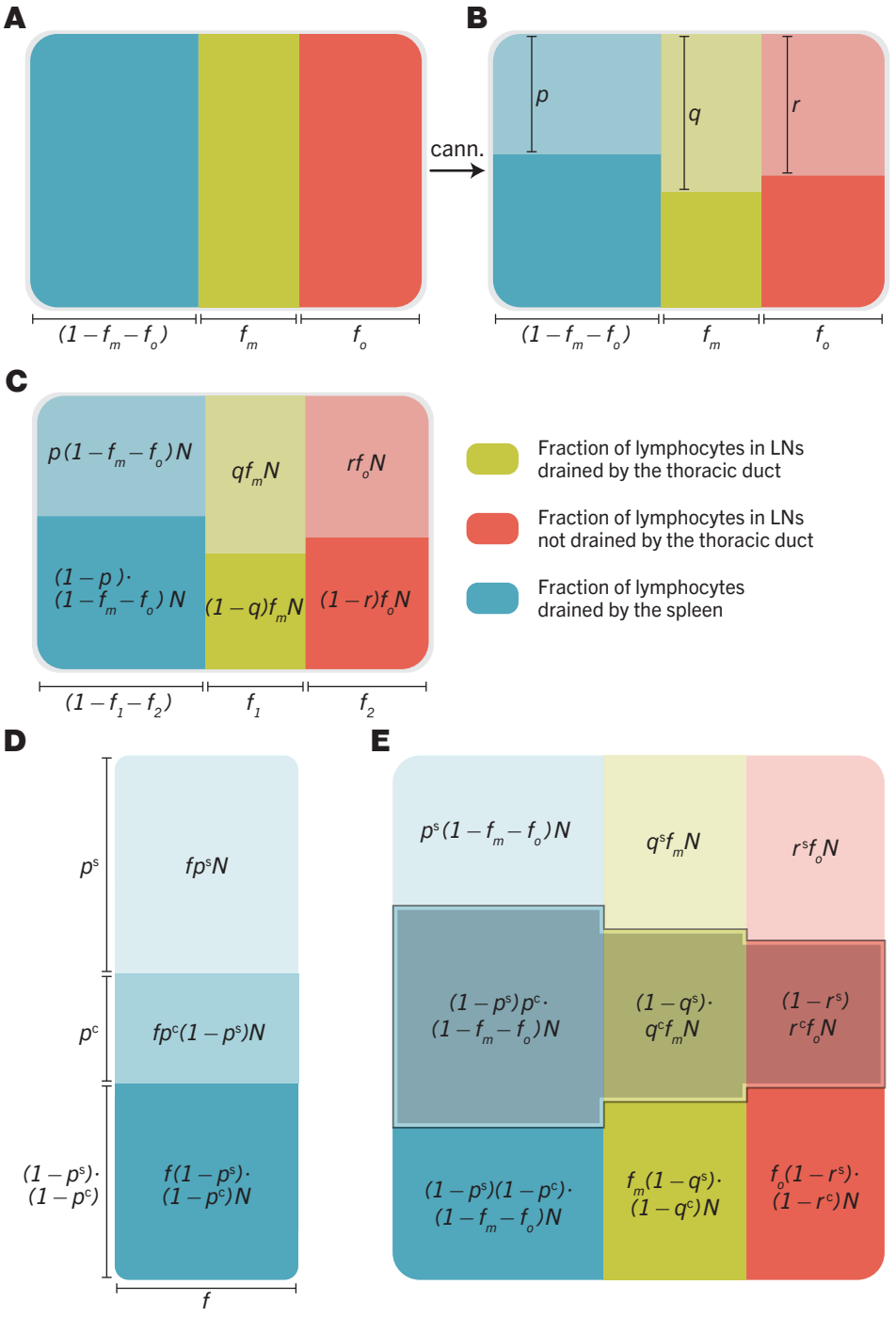

A
B
cann.
p
q
r
$(1-f_m-f_o)$
f_m
f_o
C
$p(1-f_m-f_o)N$
qf_mN
rf_oN
$(1-p)\cdot(1-f_m-f_o)N$
$(1-q)f_mN$
$(1-r)f_oN$
$(1-f_1-f_2)$
f_1
f_2
Fraction of lymphocytes in LNs drained by the thoracic duct
Fraction of lymphocytes in LNs not drained by the thoracic duct
Fraction of lymphocytes drained by the spleen
D
p^s
p^c
$(1-p^s)\cdot(1-p^c)$
fp^sN
$fp^c(1-p^s)N$
$f(1-p^s)\cdot(1-p^c)N$
f
E
$p^s(1-f_m-f_o)N$
q^sf_mN
r^sf_oN
$(1-p^s)p^c\cdot(1-f_m-f_o)N$
$(1-q^s)\cdot q^cf_mN$
$(1-r^s)r^cf_oN$
$(1-p^s)(1-p^c)\cdot(1-f_m-f_o)N$
$f_m(1-q^s)\cdot(1-q^c)N$
$f_o(1-r^s)\cdot(1-r^c)N$

Figure 4.13: (Facing page) Extending the total number of lymphocytes calculation to account for differential effects from the cannulation. The spleen and different lymph nodes are affected by the TD cannulation differently, and thus we need to extend Equation 4.1 to account for these differences. **(A)** Let N be the number of lymphocytes in all spleens and lymph nodes. We split these lymphocytes into three groups, indicated by the three colors: a proportion f_m of them are found in lymph nodes drained by the thoracic duct (red), a proportion f_o of them are found in lymph nodes drained by other lymphatic ducts (green), and the remaining $(1 - f_m - f_o)$ are found in the spleen (blue). **(B)** After the mouse is cannulated, a fractional drop occurs in each compartment due to recirculating lymphocytes being drained by the cannulation and not returning to the SLOs. We denote p, q, and r as the fractional drops for the TD-drained LNs, the non-TD-drained LNs, and the spleen respectively. The portion of these compartments that are drained by the cannulation are shown in the light colors. **(C)** The figure is split into the compartments and the drained and non-drained parts of each compartment. The equations that relate the total number of lymphocytes N to the number of cells in each part is shown. The sum of the three light-colored parts should equal the number of cells collected in the cannulation (assuming no loss of cells due to surgical stress). **(D)** Cells die or are lost due to the stress from the surgery itself, and we must account for this loss. We use p^s to denote the proportion of lymphocytes that are lost due to the surgery. We then use p^c to denote denote the proportion of the remaining lymphocytes that are drained by the cannulation after accounting for loss due to surgical stress. **(E)** This is the generalization of (C) by accounting for the loss of cells due to the surgery itself in all compartments. We show how to calculate the number of cells in each compartment in relation to the total number of lymphocytes N. The boxes labeled with a gray outline represent the number of cells collected by the cannulation.

Figure 4.14*B–D* and Figure 4.15*B–D*, we show respectively the estimated CD4 and CD8 T cell numbers with 95% confidence intervals for fixed values of f_{spl}. The estimated total T cell numbers are very sensitive to the values of f_m, f_o, and f_{spl}. In Figures 4.16–4.17, we show the estimated total numbers for the CD4 and CD8 T cell subsets and 95% confidence intervals for fixed values of f_{spl}. These numbers are also very sensitive to the chosen values of f_m, f_{spl}, and f_o. We performed the same calculations for transitional and mature B cells (Figures 4.18–4.19).

Making inferences about total numbers for the T cell and B cell subsets is difficult since we do not know how the three different types of SLOs contribute to the lymphocyte pool. Because we cannot pool all the SLOs and use the simple approach employed for Equation 4.1, we need to know the values of f_m, f_{spl}, and f_o. Since these data show that the estimated total numbers using Equation 4.2 is dependent on accurate values of f_m, f_{spl}, and f_o, we are unable to make accurate estimates for the sizes of any of these T cell and B cell subsets.

In addition, the large confidence intervals indicate uncertainty about these estimates; there is uncertainty both in the effect on cell counts by the sham surgeries and in the effect on cell counts by the cannulations. These calculations incorporate uncertainty on both processes when calculating these estimates, so these large confidence intervals are not surprising.

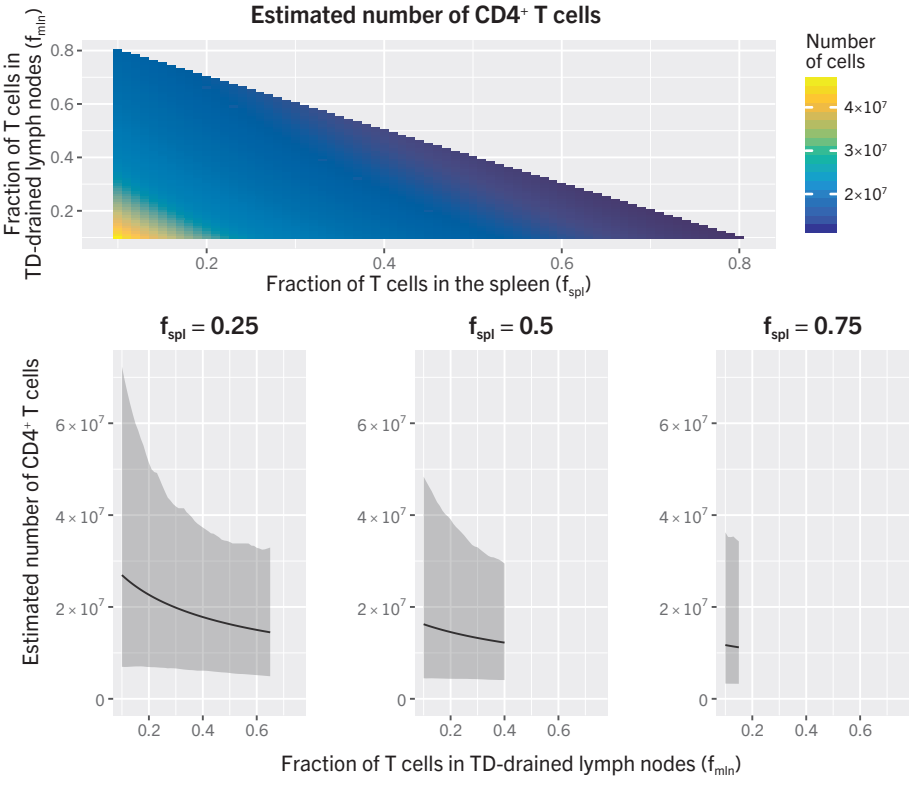


Figure 4.14: Estimated number of total CD4 T cell. (A) The estimates of the total number of CD4 T cells for different combinations of f_m, f_o, and f_{spl}. **(B)–(D)** Estimates of the total number of CD4 T cells and bootstrap 95% confidence intervals for fixed values of f_{spl} (0.25, 0.5, and 0.75)

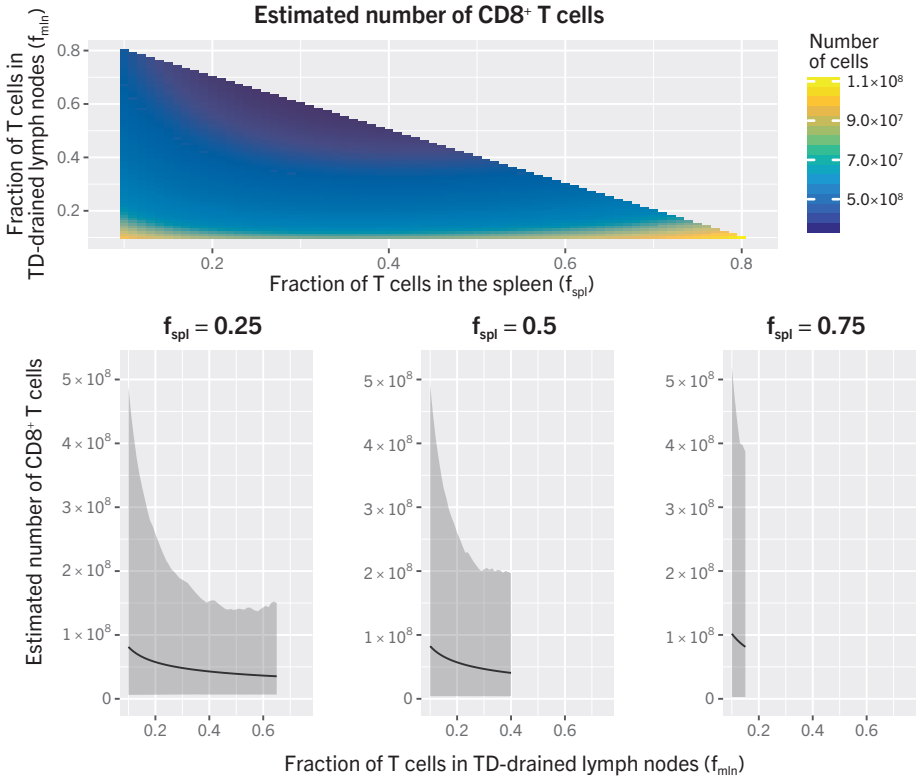


Figure 4.15: Estimated number of total CD8 T cell. (A) The estimates of the total number of CD8 T cells for different combinations of f_m, f_o, and f_{spl}. **(B)–(D)** Estimates of the total number of CD4 T cells and bootstrap 95% confidence intervals for fixed values of f_{spl} (0.25, 0.5, and 0.75)

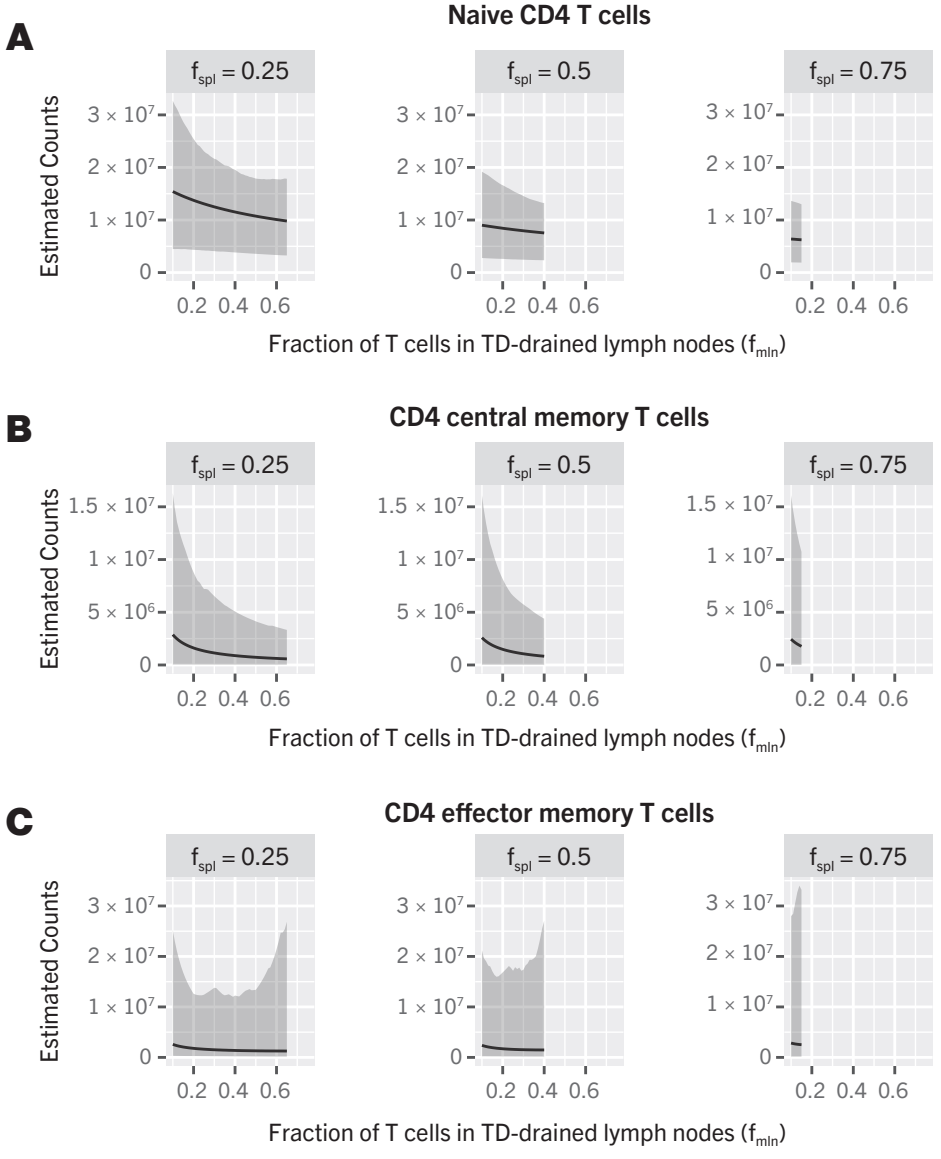


Figure 4.16: Estimated numbers of CD4⁺ T cell subsets for different values of f_{spl}.

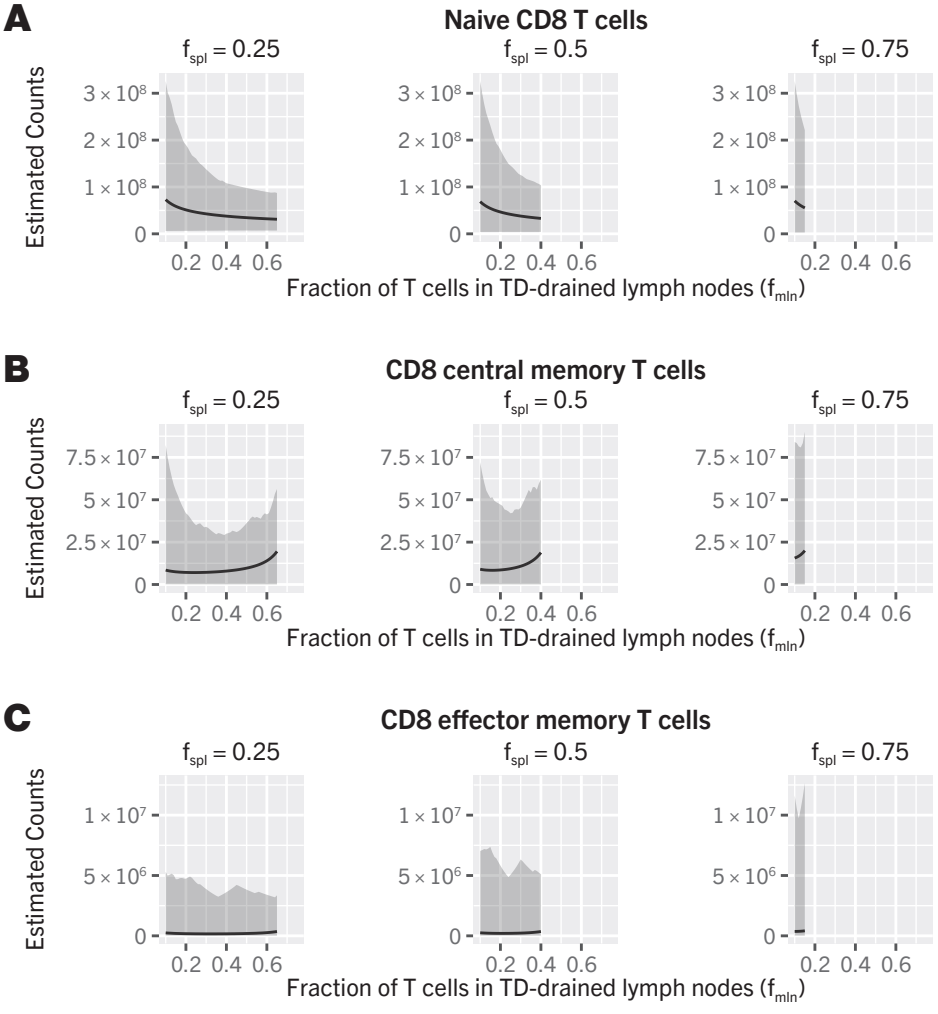


Figure 4.17: Estimated numbers of CD8⁺ T cell subsets for different values of f_{spl}.

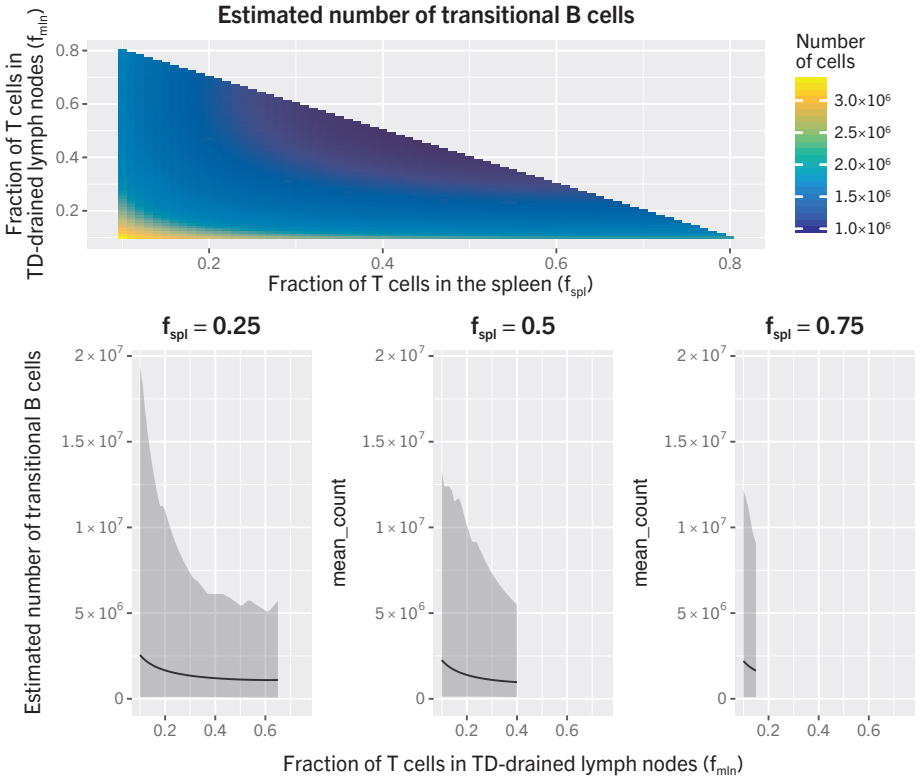


Figure 4.18: Estimated number of transitional B cells.
(A) The estimates of the total number of transitional B cells for different combinations of f_m, f_o, and f_{spl}. **(B)–(D)** Estimates of the total number of transitional B cells and bootstrap 95% confidence intervals for fixed values of f_{spl} (0.25, 0.5, and 0.75)

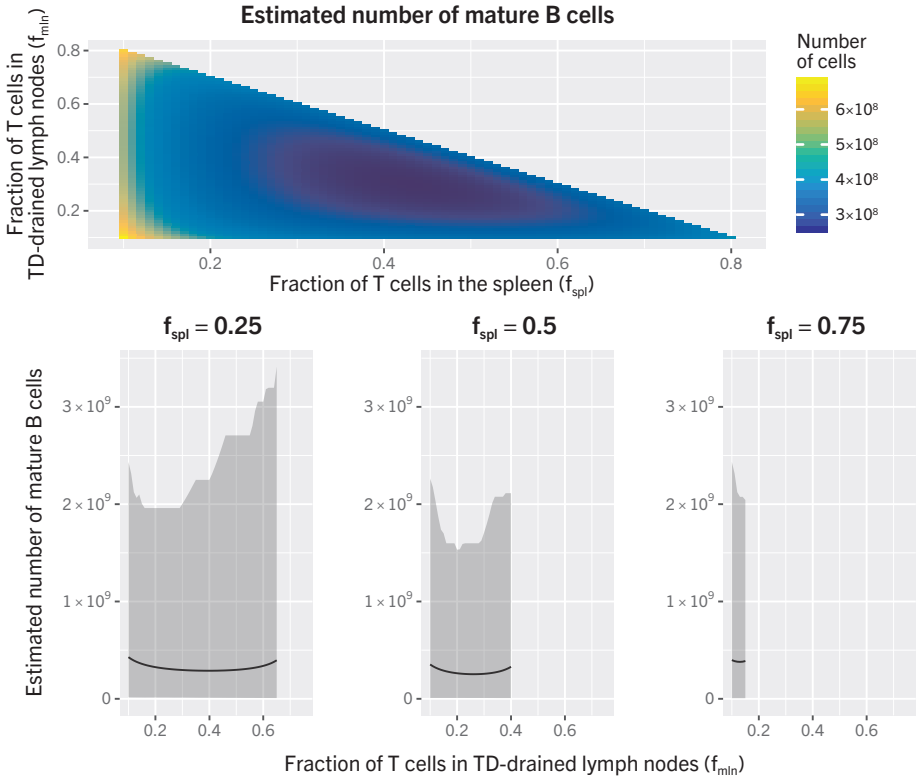


Figure 4.19: Estimated number of mature B cells. (A) The estimates of the total number of mature B cells for different combinations of f_m, f_o, and f_s. **(B)–(D)** Estimates of the total number of mature B cells and bootstrap 95% confidence intervals for fixed values of f_s (0.25, 0.5, and 0.75)

4.4 DISCUSSION

We sought to accurately estimate the number of T cells and B cells and their subsets in the SLOs of the mouse using thoracic duct cannulations. Our idea centered on using TD cannulations to perturb the SLOs by draining away recirculating lymphocytes. We expected the fold reduction of cell numbers in all SLOs to be directly proportional to the number of lymphocytes collected by the cannulation. However, heterogenous effects on the different lymphocyte subsets and on the different SLOs complicated the mathematical model for estimating the size of the lymphocyte compartments. T cells and B cells in the SLOs were lost due to effects of the surgeries that were independent of the cannulation itself. These observations prevented us from reliably estimating total T cell and B cell numbers but illustrated important consequences of surgical procedures on the immune system.

Surgical procedures have been shown to have significant effects on the immune system in humans and in mice. These effects include decreases in T cell counts [119], shifts in T_H1/T_H2 ratios [120], and increases in serum cytokine levels [121]. These changes in the immune system are mediated by at least two factors: the anesthetics used in surgeries and the release of steroids and hormones in response to surgical stress. Isoflurane (the anesthetic used in our experiments) has been shown to decrease T_H1/T_H2 ratios and changes in $CD3^+$, $CD4^+$, and $CD8^+$ counts [120, 122]. The stress on the body induced by surgical procedures has been shown to cause an increase in levels of glucocorticoids, catecholamines, and ACTH in humans [119]. One study showed that human patients undergoing distal gastrectomies or colectomies had decreases in $CD3^+$ cell counts in blood samples with a shift in more $CD8^+$ cells and less $CD4^+$ cells [119]. Increases in glucocorticoid levels in mice have also been shown to cause cell death in most T cell thymocyte subsets as well [123]. Although the study by Majumdar *et al*. studied the effects of glucocorticoids in the context of *Salmonella typhimurium* infections, they showed that blocking the glucocorticoid receptor rescues the thymocyte subsets, indicating that glucocorticoids have a important role in the death of these cells [123].

The procedure performed for the sham surgeries included every step of the cannulation procedure except for the insertion of the cannula into the thoracic duct. This included the incisions through the skin and muscle layers, the manipulation of the spleen and other tissues in order to visualize the thoracic duct, the placement and securing of the cannula near the thoracic duct, and the placement of the mouse in a harness for recovery. The sham surgery placed an exceptional amount of stress on the body that is comparable to the stress from any other major surgical procedure.

The large changes in cell numbers in the SLOs observed in Section 4.2.1 due to the sham surgeries represent a real effect of surgeries that must considered in the design of experiments. We showed that cell numbers of many T cell subsets decrease in the lymph nodes—with no apparent changes in the spleen—and that B cell numbers decrease in lymph nodes and spleens. Results from experiments that involve surgeries can potentially be affected by these quantitative changes in the number of T cells in lymph nodes and B cells in lymph nodes and the spleen. For example, surgical experiments studying T cell responses may find responses that are quantitatively different due to the shrinkage in T cell numbers. In addition, we did not study the changes in cell numbers in the thymus. It is plausible that thymic population sizes might shrink and have changes in phenotype due to the increased corticosteroid levels induced by the surgery [123]. Further experiments measuring cell counts of thymocyte subsets in the thymii of mice that have undergone sham surgeries would characterize the degree of shrinkage in response to surgeries.

Our study does not untangle the roles of anesthesia and the surgery itself in the loss of lymphocytes. Previous studies have shown contrasting effects on the immune system by different anesthetics, which indicates that anesthetics have some role in these changes [121, 120, 122]. This could be studied by giving mice general anesthesia with isoflurane and counting lymphocytes in their SLOs. This experiment would help us to characterize how much of the decrease in cell numbers is due to the stress of the anesthetic and to the stress of the surgery. Detailed understanding of how anesthetics affect the immune system can have clinical applications in trying to minimize these effects; for example, general anesthesia has been associated with increased rates of surgical site infections, which corroborates with anesthesia affecting the immune system [124, 125].

In summary, knowing the total number of T cells and B cells would help us quantify the characteristics of a mature adaptive immune system, particularly in quantifying the diversity of the TCR and BCR repertoire. We were unable to determine accurate estimates of these numbers, but our results point towards heterogeneous effects from surgical procedures on lymphocytes in SLOs that may have implications in experimental design and have clinical consequences in regards to surgeries affecting the immune systems of patients.

CHAPTER 5

Obtaining paired αβ TCR sequences

5.1 INTRODUCTION

5.1.1 *Identifying TCR sequences can answer many important questions in immunology*

The set of antigens that a T cell clonotype can respond to is defined by its T cell receptor. Recombination of the V(D)J segments and the addition of nucleotides in the junctions of these segments results in an enormous diversity of receptors in the TCR repertoire.

Many fundamental questions about T cell biology can be answered by sequencing TCRs. How many T cell clonotypes does the body maintain? How does the TCR repertoire differ among a group of people, and how does it change with age? Are there specific TCRs that consistently result in large, expanded clonotypes against a specific antigen or is this process inherently stochastic? TCR sequencing is an important tool in attempting to characterize these fundamental physiologically properties of T cell biology.

In clinical immunology, identifying clonotypes that are important to the pathogenesis of diseases and disorders could have immense value in our understanding of disease mechanisms and in developing treatments. For example, identifying T cell clonotypes that respond well to a patient's tumor could have profound influence in the design of personalized immunotherapy for the patient. Similarly, sequencing the dominant clones involved in autoimmune conditions can help the design of therapies for controlling or eliminating autoreactive T cells.

The number of questions that can be answered and the potential for clinical applications with TCR sequencing demands for efficient, high-throughput, and economical methods for determining TCR sequences of T cells. The field of lymphocyte-receptor sequencing is rapidly changing with continual advances that allow us to probe TCR repertoires in greater detail and with greater efficiency. In this chapter, we will discuss a method that we developed called ALPHABETR that efficiently and accurately determines paired TCR sequences of epitope-specific T cell populations.

Approaches like ALPHABETR belong to a class of TCR sequencing methods called frequency-based pairing. The other two approaches that have been described in the literature are a preliminary attempt used to pair B cell receptors [76] and a commercial product called pairSEQ developed by Adaptive Biotechnologies [1].

Reddy *et al.* [76] attempted a preliminary version of frequency-based pairing in B cells by matching the relative frequencies of the most abundant V_L and V_H CDR3 sequences. In this study, the V_L and V_H CDR3s were sequenced in bone marrow plasma cells from mice inoculated with purified C1, with chicken egg ovalbumin, or with recombinant bacterially expressed human B-cell regulator of IgH transcription. V_L and V_H CDR3 sequences represented with approximately equal frequencies and has frequencies greater than 0.5% of the repertoire were assumed to derive from the same B cell.

A methodology called pairSEQ by Howie *et al.* [1] attempts to identify TCRα/TCRβ pairs by sampling T cells (typically thousands) from the PBMCs of a subject into the wells of 96-wells plates and sequencing the CDR3α and CDR3β chains sampled in these wells. RNA is extracted from the cells and reverse transcribed into cDNA. The CDR3α and CDR3β sequences are amplified using primers specific for the V regions and C regions, and DNA barcodes were added as well-identifiers. These amplified products are pooled together and then sequenced, and the resulting data are CDR3α and CDR3β sequences and their barcodes that correspond to the well from which they originate. Due to various sources of experimental error, a sampled T cell will not necessarily have its CDR3α and/or its CDR3β sequenced, so pairSEQ accounts for this "dropping" of chains. The algorithm then produces a list CDR3α/CDR3β pairs and an estimated false discovery rate (FDR) for each pair for a user-specified rejection threshold.

The pairSEQ approach assumes that the CDR3α and CDR3β sequences are practically unique for every clone in a T cell population—an assumption not necessarily incorrect when taking samples from the naive repertoire pool (see Section 6.3.3) but highly inaccurate for antigen-specific T cell population (discussed below in Section 5.1.4). By sampling a T cell population into multiple subsets as done in a pairSEQ experiment, the chains of a clone should be found in an unique pattern of subsets since any given clone is

found with a relatively low frequency in the general T cell repertoire. With a procedure to optimize the number subsets taken and the number of cells sampled in each subset, pairSEQ uses the pattern of co-appearances of CDR3α and CDR3β sequences to determine TCR pairs. The statistical procedure used by pairSEQ calculates the p-value of the number of co-appearances of every pair of CDR3α and CDR3β sequences under the null hypothesis that these sequences are not from the same clone and then estimates an cutoff p-value for rejecting the null hypothesis.

pairSEQ begins by computing the probability (i.e. the p-value) of seeing the pattern of shared wells from chain α_i and chain β_j given the numbers of wells in the experiment by chance, wells containing chain α_i, and wells containing chain β_j for all pairs i and j.[1] Let the pair $\{r, s\}$ denote an occupancy pattern in the data where an α chain occupies r wells and a β chain occupies s wells. Then, for every possible occupancy pattern, a cutoff p-value δ_{rs} is calculated by performing multiple simulations of how two independent α and β chains would be sampled in the pairSEQ experiment and finding the smallest p-value calculated from these simulations. Thus, if the p-value for α_i and β_j given its co-appearance pattern in the data $\{r, s\}$ is less than δ_{rs}, then α_i and β_j is a TCR pair. An approach based on one developed by Bandcroft *et al.* [126] is then used to estimate the FDR for these pairs.

Howie *et al.* [1] attempted to demonstrate the accuracy and throughput of their algorithm by applying pairSEQ on two categories of samples: mixing T cells from whole PBMCs from two patients into the same 96-well plate and combining tumor-infiltrating lymphocytes (TILs) from nine tumor samples into one 96-well plate. The former was designed to test the accuracy of pairSEQ by showing that TCR pairs derived from both patients—namely, pairs such that one chain is from the 1st patient and the other chain is from the 2nd patient—were identified at a rate estimated by the FDR. In "Experiment 1" of their paper, the percentage of cross-subject TCR pairs was 0.98% and was in agreement with the estimated 1% FDR. In addition, Jurkat T cells were added to the samples, and the Jurkat CDR3α/CDR3β pair was successfully identified as well. It should be noted that these two points of analysis provide circumstantial evidence for the accuracy of pairSEQ. The cross-subject experiment indirectly shows agreement in the predicated FDR

[1] We denote the ith unique CDR3α sequence as α_i and the jth CDR3β sequence as β_j. If a clone has a TCR made up of the pair α_i and β_j, we simply denote the clone as $\alpha_i\beta_j$.

and the cross-subject pairing rate, but their analysis does not test if the TCR pairs themselves identified in each subjects are correct (except for the one TCR from the Jurkat cells). The TIL experiment is discussed in Section 5.2.2, and these data were used to test ALPHABETR in Section 5.3.2.

5.1.3 *Features of TCR repertoires*

T cell repertoires have a significant number of dual-TCRα clones and a small number of dual-TCRß clones

Studies have demonstrated that 10–30% of T cells possess two productive TCRα chains, defined as a T cell that contains mRNA transcripts of two different CDR3α sequences that do not contain stop codons or shortened sequences [127, 65, 70]. A rough calculation predicts a dual-TCRα clone prevalence of 20% in the naive pool.[2] The presence of dual-TCRα clones is a consequence of how thymocyte development occurs. In the DP thymocyte stage, thymocytes make repeated attempts at α-chain rearrangement on both chromosomes. Rearrangement continues until a thymocyte receives a positive selection signal from a self-pMHC molecule; simply expressing a full TCR molecule is not sufficient for stopping α-rearrangement. As a result, a thymocyte can potentially rearrange two in-frame TCRα chains before α-rearrangement is shut off. It is unclear if the presence of two different TCRα chains at the mRNA level always results in the expression of two TCRα proteins, but expression of two TCRα chains has been shown by labeling V_α segments with antibodies [127]. The functional significance is not fully

[2] We can make a very rough estimate for the proportion of developing DN4 thymocytes that will rearrange two productive α chains. The probability of a T cell having two productively rearranged TCRα chains is dependent on the T cell forming in-frame joins between the V and J segments, which we assume to be simply 1/3. Thus,

$$P(\text{no productive chains}) = \left(1 - \frac{1}{3}\right)^2 = \frac{4}{9}$$

$$P(\text{exactly one productive chain}) = P(\text{first chain productive}) + P(\text{second chain productive})$$
$$= \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$$

$$P(\text{two productive chains}) = \left(\frac{1}{3}\right)^2 = \frac{1}{9}$$

Since T cells that have no productive chains would fail positive selection, we would expect $\frac{1/9}{4/9} = 25\%$ of thymocytes to be dual-TCRα clones.

understood, but dual-TCRα T cells have been postulated to mediate autoimmunity by allowing the second 'hitchhiker' TCRα to enter the peripheral repertoire without necessarily undergoing negative selection [127, 128]. Other studies have proposed physiological purposes of these dual TCRs, including a mechanism to facilitate thymocyte commitment to the regulatory T cell lineage [129] and the "rescue" of non-selected TCRs through successful selection of the one of the two TCRs, effectively expanding the TCR repertoire [130]. The significant prevalence of dual-TCRα clones necessitates special considerations for frequency-based pairing techniques.

The presence of dual-TCRβ clones has been shown in a limited number of the most recent single-cell sequencing studies [131, 70]. These studies have found the expression of two CDRβ sequences of 6%–7% of their sampled T cells that contain at least one productive CDR3β. The low prevalence of dual-TCRβ clones can be ignored with little consequence, which we will discuss in the end of Section 5.3.1.

Epitope-specific T cell populations can exhibit a significant degree of sharing of CDR3 sequences

Epitope-specific T cell populations contain many clones that share CDR3 sequences, which we define as two or more clones that have the same CDR3α sequence but have different CDR3β sequences (or vice versa). This is a significant feature of epitope-specific T cell repertoires, which we discuss in detail here. Frequency-based pairing approaches for epitope-specific populations must be specifically tailored to deal with the sharing of CDR3 sequences.

5.1.4 *Epitope-specific T cell populations display a significant degree of CDR3a and CDR3ß sharing*

Previous single-cell TCR sequencing studies of epitope-specific T cell populations and TIL populations have not fully appreciated the degree of CDR3α and CDR3β sharing that can be found in populations of limited clonality. We analyzed the data from past single-cell analyses of epitope-specific T cell populations in mice and humans and discovered notable levels of sharing of both CDR3α and CDR3β sequences at the amino acid level across clones within individuals (Table 5.1). In the study by Cukalac *et al.*, the antigen-experienced epitope-specific T cell populations exhibit more sharing than

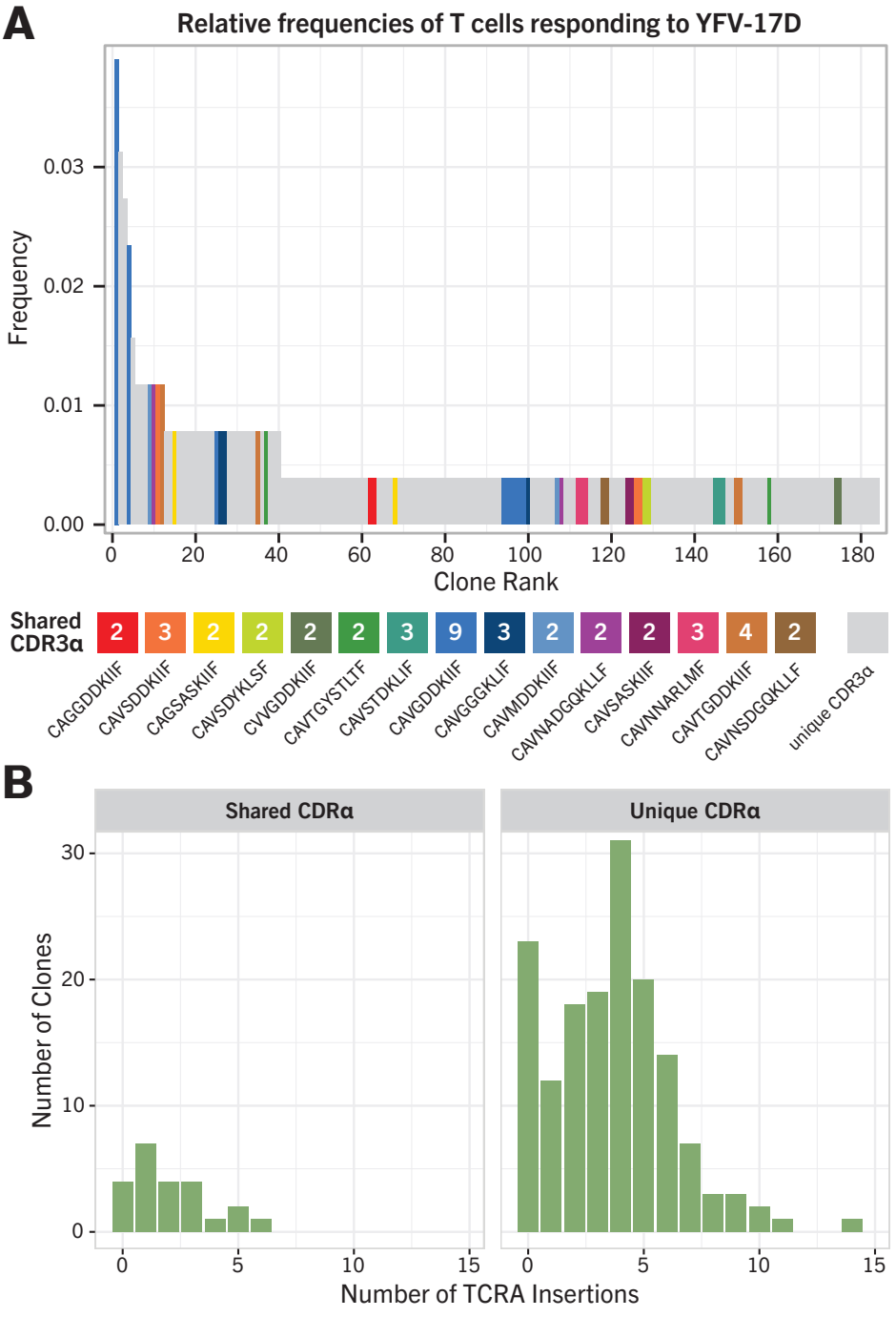

A
Relative frequencies of T cells responding to YFV-17D
Frequency
0.03
0.02
0.01
0.00
0
20
40
60
80
100
120
140
160
180
Clone Rank
Shared CDR3α
2
3
2
2
2
2
3
9
3
2
2
2
3
4
2
CAGGDDKIIF
CAVSDDKIIF
CAGSASKIIF
CAVSDYKLSF
CVVGDDKIIF
CAVTGYSTLTF
CAVSTDKLIF
CAVGDDKIIF
CAVGGGKLIF
CAVMDDKIIF
CAVNADGQKLLF
CAVSASKIIF
CAVNNARLMF
CAVTGDDKIIF
CAVNSDGQKLLF
unique CDR3α
B
Shared CDRα
Unique CDRα
Number of Clones
30
20
10
0
0
5
10
15
0
5
10
15
Number of TCRA Insertions

Figure 5.1: (Facing page) Analysis of TCRα usage in human, YFV-specific peripheral-blood CD8$^+$ T cells. **(A)** Observed distribution of relative clone sizes within the population specific for the HLA-A02:01/LLWNGPMAV epitope. Clones expressing a unique CDR3α are shown in grey; clones that share a CDR3α are colored, and the numbers in the coloured boxes represent the number of clones sharing each CDR3α. **(B)** The distributions of CDR3α nucleotide insertion lengths in clones with shared CDR3α (left hand panel) and unique CDR3α (right hand panel).

naive populations [67]. This potentially indicates that recruitment of T cells in an immune responses has a bias for TCRs with specific CDR3β and/or CDR3α sequences, which would result in increased sharing in epitope-specific populations. The small sizes of the sampled repertoires obscures any generalizations about the magnitude of sharing exhibited in antigen-specific populations, but the consistent findings across many studies demonstrate this common characteristic of these repertoires.

The phenomenon of CDR3β sharing can be demonstrated by tracking the sequence of events that are involved in generating TCRs in the thymus. In mice, thymocytes undergo 6–9 divisions following TCRβ rearrangement at the DN3 stage [34, 35, 36, 37], generating 64–512 cells with the same TCRβ chain that then undergo independent TCRα rearrangement. If we assume that 5% of these precursors survive selection [132, 133, 134, 38], then selection results in TCRβ clone sizes of 3–25 cells [24]. Thymocytes may undergo 1 or 2 divisions at the single-positive CD4 or CD8 stage before leaving the thymus [38]. Assuming 2-fold expansion here on average, each αβ T cell precursor at DN3 generates 6–50 new naive cells with identical TCRβ chains, comprising of 3–25 unique TCRαβ clones of typically 2 cells. Comparable estimates of TCRβ clone sizes have been obtained elsewhere [37, 24]. There is also evidence that TCRβ-clone sizes can be augmented by convergent recombination of the same TCRβ chain [18, 135]. If a particular CDR3β contributes strongly to the affinity of binding to a given peptide-MHC as described in the previous paragraph, our rough quantification of TCRαβ clonality in thymopoesis is consistent with the observation that TCRβ-sharing is commonly found within epitope-specific populations (Table 5.1).

Because TCRα rearrangement occurs after TCRβ rearrangement, any sharing of CDR3α sequences across clones presumably arises from convergent recombination, where the same CDR3α is formed from independent

Citation	Peptide/System	Status	# of clones	# of distinct α chains	# of distinct β chains	# of shared α chains	# of shared β chains	% of α chains that are shared	% of β chains that are shared
[65]	K^bPB1_{703}	Immune	35	16	24	3	2	18.8	8.3
[69]	Human $CD4^+$ TILs	Colon cancer	216	226	216	7	0	3.1	0.0
	$CD4^+$ T cells from adjacent colon		305	239	237	15	0	6.3	0.0
[67]	D^bNP_{366}	Naive 1	17	17	15	0	2	0.0	13.3
		Naive 2	11	11	11	0	0	0.0	0.0
		Naive 3	7	7	7	0	0	0.0	0.0
		Naive 4	10	7	9	3	1	42.9	11.1
		Naive 5	13	13	12	0	1	0.0	8.3
		Naive 6	9	9	9	0	0	0.0	0.0
		Immune 1	12	10	8	2	3	20.0	37.5
		Immune 2	15	9	8	4	3	44.4	37.5
		Immune 3	12	11	8	1	1	9.1	12.5
		Immune 4	10	10	8	0	1	0.0	12.5
	D^bPA_{244}	Naive 1	11	11	11	0	0	0.0	0.0
		Naive 2	10	10	10	0	0	0.0	0.0
		Naive 3	8	8	8	0	0	0.0	0.0
		Naive 4	25	25	25	0	0	0.0	0.0
		Naive 5	43	40	43	2	0	5.0	0.0
		Immune 1	17	15	15	2	1	13.3	6.7
		Immune 2	27	21	20	5	6	23.8	30.0
		Immune 3	14	14	12	0	2	0.0	16.7
		Immune 4	20	14	20	3	0	21.4	0.0
	$D^bPB1\text{-}F2_{62}$	Naive 1	13	13	13	0	0	0.0	0.0
		Naive 2	13	12	13	1	0	8.3	0.0
		Naive 3	9	9	9	0	0	0.0	0.0
		Naive 4	41	41	41	0	0	0.0	0.0
		Naive 5	21	21	21	0	0	0.0	0.0
		Naive 6	24	22	23	2	1	9.1	4.4
		Naive 7	16	16	16	0	0	0.0	0.0
		Immune 1	9	9	8	0	1	0.0	12.5
		Immune 3	11	11	11	0	0	0.0	0.0
		Immune 4	20	15	17	1	2	6.7	11.8
		Immune 5	16	15	16	1	0	6.7	0.0
This study	Human $CD8^+$ YFV	Immune	184	169	179	15	3	8.9	1.7

Table 5.1: A summary of the degrees of sharing of CDR3α and CDR3β at the amino acid level across clones within epitope-specific T cell populations, found in published single-cell TCR sequencing data and our own. Unless indicated otherwise, the samples were obtained from influenza-infected mice. The data clearly demonstrate that sharing of both α and β chains within an individual occurs in different infection/inoculation settings.

rearrangement events. In this hypothesis is true, then sharing is expected to arise most frequently for sequences that are similar to germline VJ sequences and thus contain relatively few random N-nucleotide insertions. To examine this possibility, we immunized an HLA-A2 human volunteer with the live attenuated yellow fever vaccine YFV-17D, took a peripheral blood sample 15 days post-vaccination, and used dextramer staining and single-cell RNAseq to recover paired TCRαβ sequences from $CD8^+$ T cells specific for the immunodominant epitope HLA-A02:01/LLWNGPMAV (described in Section 2.5). Out of 256 cells, we observed 169 unique CDR3α sequences, with 15 (8.9%) of them shared between two or more clones (Fig-

ure 5.1*A*, colored bars represent clones with shared CDR3α sequences, and the numbers in the boxes indicate the number of clones sharing each CDR3α sequence). We examined the numbers of nucleotide insertions at the V-J junction of the CDR3α and indeed saw significantly fewer in CDR3α sequences that were shared between two or more clones (mean of 2.04 insertions, $n = 23$) than in sequences that were unique to a single clone (mean of 3.62 insertions, $n = 154$; $p < 0.005$, Wilcoxon rank sum test; Figure 5.1*B*). It appears that convergent TCRα recombination may derive at least in part from the reduced junctional diversity of clones possessing CDR3 regions that are closer to germline.

Despite our incomplete understanding of the mechanisms and processes that cause CDR3α and CDR3β sharing, single-cell sequencing data clearly suggest that any method for sequencing epitope-specific populations and other populations of limited clonality will need to explicitly handle these features in the repertoire.

5.1.5 *Aim of the chapter*

The aim of this chapter is to describe a frequency-based pairing methodology called ALPHABETR to efficiently and accurately identify paired TCR sequences. ALPHABETR is designed to identify TCR pairs from epitope-specific T cell populations that contain dual-TCRα clones and the sharing of CDR3α and CDRβ sequences; it also estimates the frequencies of the identified clones within their parent populations. The chapter will also discuss the extensive testing performed on the algorithm through simulated and real sequencing data sets to demonstrate its accuracy and its advantages and limitations over single-cell sequencing technologies.

5.2 DESCRIPTION OF ALPHABETR AND HOW IT WAS TESTED

5.2.1 *An overview of the* ALPHABETR *algorithm*

We developed a procedure named ALPHABETR (**al**gorithm for **p**airing alp**ha** and **be**ta **T** cell **R**eceptors) that recovers TCRαβ pairs from high-throughput sequencing data, shown schematically in Figure 5.2. The experimental procedure is to sequence the CDR3α and CDR3β regions from multiple samples of T cells from the same parent population (Figure 5.2*A–B*). In contrast

with single-cell sequencing, multiple samples of many cells are sorted into wells of 96-well plates (Figure 5.2*A*). The number of cells in each well can be freely varied, and as described in Section 5.3.1, varying the sample size across the plate(s) helps to increase both the number and accuracy of pairings. The CDR3α and CDR3β sequences found in each well are sequenced (Figure 5.2*B*), and a list of the set of unpaired sequences from each sample are inputs for the ALPHABETR algorithm. Although Figure 5.2*B* depicts amino acid sequences as inputs, the inputted data can comprise of the CDR3 nucleotide sequences and/or the addition of V(D)J segment information.

With these data, ALPHABETR chooses a random subsample of wells and then calculates association scores between every α and every β chain (Figure 5.2*C*(*i*)-(*ii*)). The score is the sum of the number of co-occurrences of the chains weighted inversely by the number of unique CDR3α sequences in the well and the number of co-occurrences in each well weighted inversely by the number of unique CDR3α sequences in the well (Figure 5.2*C*(*ii*)). Weighting the number of co-occurrences in this way reflects the intuitive idea that our confidence that a co-occurring CDR3α and CDR3β pair derive from the same clone decreases as the number of unique CDR3 sequences recovered from that well increases. The algorithm then solves a linear sum assignment problem within each well based on these plate-wide association scores to generate a list of candidate pairs of CDR3α and CDR3β sequences within each well (Figure 5.2*C*(*iii*)). Each well yields a list of αβ pairs—where each CDRα is paired with only one CDR3β and vice versa—that maximizes the sum of the association scores. After repeating this assignment problem for every well in the random subset, the algorithm generates a list of the number times every CDR3α and CDR3β pair was associated with each other. This list then allows for sharing of chains across clones. Those αβ pairs that appear in a number of wells greater than a calculated filter level then form a refined list of candidate pairs; this filter is calculated as the mean of the number of associations of every αβ pair that had at least one association.

This pairing and filtering process is repeated on many different randomly chosen subsets of the data (Figure 5.2*C*), and a consensus list of putative paired CDR3 sequences comprises those appearing in more than a consensus threshold of these lists (Fig 5.2*D*). This resampling procedure of repeatedly using different subsets of the data acts to reduce the effect of very common clones pushing up the consensus threshold for inclusion in the filtered list

and increases the efficiency of pairing of rarer clones while minimizing the inclusion of incorrect αβ pairs (discussed more fully in Section 5.5.2). The consensus threshold is chosen by the user, and a higher threshold will decrease the number of incorrectly paired CDR3 sequences while decreasing the number of correctly identified pairs. Steps shown in Figure 5.2*A–C* are described in more detail in Section 5.5.2.

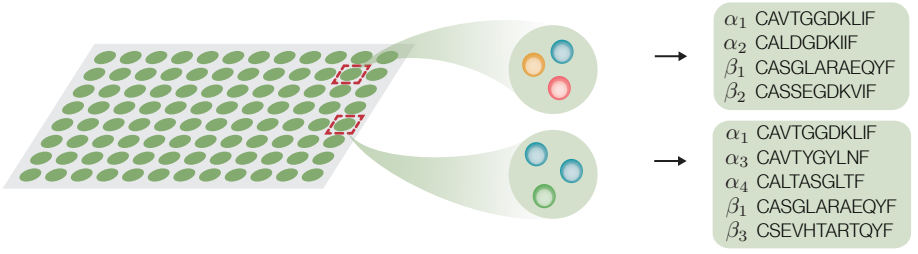
The algorithm then uses a maximum likelihood approach to estimate the relative frequencies of the clones associated with each candidate αβ pair (Figure 5.2*D*). The probability of all chains associated with a clone co-appearing in a given number of wells can be calculated from the binomial distribution, and thus the number of wells that contain the chains of a clone in the data provides information about the frequency of that clone. We calculate the maximum likelihood estimates for this distribution given the data to estimate clonal frequencies within the parent population (mathematical details in Section 5.5.3).

These estimated frequencies are then used with the patterns of co-occurrences of chains to distinguish between β-sharing and dual-TCRα clones. The algorithm decide whether each candidate pair of clones that share a β chain (e.g. $\alpha_i\beta$ and $\alpha_j\beta$) are indeed two clones or derive from one dual-TCRα clone (e.g. $\alpha_i\alpha_j\beta$). This is done by exploiting the fact that the pattern of co-occurrences of all three chains will be different under the two hypotheses, and two methods are used to determine with which hypothesis the observed data is consistent.

The first method—dubbed the *k*-means approach—finds rare dual-TCRα clones by using the estimated frequencies of a putative β-sharing clone pair $\alpha_i\beta$ and $\alpha_j\beta$ to calculate the expected number of wells in which all three chains should co-occur. The three chains will tend to co-occur more often if they derive from a dual-TCRα clone than if they derive from two CDR3β-sharing clones. For each candidate clone, we take the ratio of the number of co-occurrences of their clones in the data to the expected number of co-occurrences assuming that the chains derive from a CDR3β-sharing clone. We cluster these candidate clones using *k*-means clustering into two sets based on these ratios, and the set with higher ratios are chosen as dual-TCRα clones (a representative example is shown in Figure 5.6).

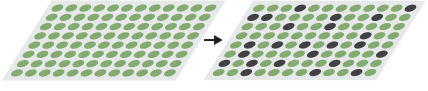
A T cells are sampled onto 96-well plates at 10-300 cells/well

B Unpaired CDR3α and CDR3β are sequenced in each well



C A resampling strategy is used to obtain a list of possible TCR pairs by repeatedly performing steps (i), (ii), (iii)

(i) A subset of wells is randomly selected for steps (ii) and (iii)



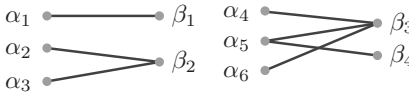
(ii) Association scores are calculated for every α and β chain across all wells in the subset

$$S_{ij} = \sum_{k=1}^{W} \left(\frac{\delta_{ij}^{k}}{c_{\alpha}^{k}} + \frac{\delta_{ij}^{k}}{c_{\beta}^{k}} \right)$$

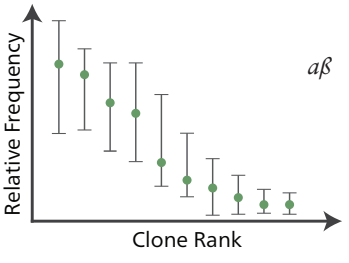
(iii) Scores are used to select likely αβ pairs within each well

	α_1	α_2	α_3	α_4	α_5
β_1	24.0	0.6	1.2	4.3	8.2
β_2	0.4	0.2	60.2	0.7	2.2
β_3	1.0	0.2	1.2	9.0	3.0
β_4	3.2	30.1	0.1	0.4	2.1

E Pairs from Step D present in more than a threshold proportion of replicates are candidate TCRs



F Clonal frequencies are estimated and used to distinguish β-sharing and dual TCRα clones



G The output is a list of single and dual TCR clones with their respective clonal frequencies

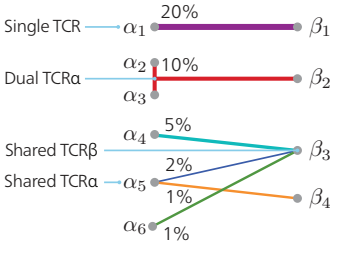


Figure 5.2: (Facing page) An overview of the implementation of alphabetr. (A) From the population of interest, multiple samples of 20-300 T cells are sorted into 96-well plates. This design allows for a given clone to be sampled in multiple wells. **(B)** High-throughput sequencing is used to recover the unpaired CDR3αand CDR3β sequences of the clones sampled in each well. No quantitative information regarding the number of times each clone was sampled in a given well; **(C)** (i) A random subset of the wells is chosen, (ii) association scores between every unique α and β found across the wells within this sample are calculated, and (iii) the set of one-to-one αβ pairs that maximizes the sum of association scores is identified using the Hungarian algorithm [136]. Step (iii) is illustrated for a particular set of CDR3α and CDR3β recovered from one well as a matrix of association scores calculated across all wells in the subsample. **(D)** Steps C(i)–(iii) are repeated to generate a consensus list of pairs, filtering out candidates that appear rarely across replicates. **(E)** The frequencies of each remaining candidate αβ pair within the parent population are estimated using a maximum-likelihood approach assuming only sharing and no dual-TCRα clones. Dual-TCRα clones (e.g. $\alpha_i\alpha_j\beta_k$) are then distinguished from clones apparently sharing a TCRβ chain (e.g. $\alpha_i\beta_k$ and $\alpha_j\beta_k$) with two methods—one using *k*-means and the other using full likelihood calculations based on the patterns of co-occurrences—and the frequencies of these clones are re-calculated. **(F)** The output of the algorithm is a list of single and dual-TCRα clones, each with their estimated frequency within the parent population.

The second method—dubbed the full-likelihood approach—finds more common dual-TCRα clones by using a more sophisticated approach that calculates the likelihoods of all three-way, two-way, and individual concurrences of α_i, α_j, and β under both hypotheses. Unlike the *k*-means approach, the full-likelihood approach incorporates additional information from the partial appearances of the chains of the candidate clone, which occurs due to the dropping of chains. These likelihoods are computationally expensive and in practice can be used for wells with ≤ 50 cells, which are conveniently the wells that contain maximal information regarding common clones. Thus, this approach designed to identify more abundant dual-TCRα clones. Candidate clones with larger likelihoods under the dual-TCRα hypothesis than the CDR3β-sharing hypothesis are then chosen as dual-TCRα clones—we empirically found that a difference of 10 in the log likelihoods identifies many correct dual-TCRα clones with a low rate of mistakes.

The output of the algorithm is a list of single or dual-TCRα clones together with estimates of their frequencies within the parent population (Figure 5.2*F*). ALPHABETR does not attempt to identify dual TCRβ expressing cells because dealing with this relatively infrequent phenomenon together with dual-TCRα chains and sharing of both TCRα and TCRβ chains across clones is extremely challenging algorithmically. However, we will demon-

strate that the presence of dual TCRβ clones at rates shown in previous studies has little effect on its performance (see the end of Section 5.3.1).

5.2.2 *Evaluating the algorithm with simulated and experimental data*

We evaluated the accuracy and precision of ALPHABETR by testing the algorithm with simulated data and with a tumor-infiltrating lymphocyte dataset from the Howie *et al*. study [1]. Simulating data allowed us to test how well the algorithm performed with data from T cell populations with different degrees of clonality, different degrees of CDR3α and CDR3β-sharing, and many different experimental conditions. In addition to testing a wide breadth of different immunological and experimental conditions, simulations have the advantage of knowing the "right answers," giving us the option to directly assess the accuracy and efficiency of the algorithm in a way that cannot be done with experimental data.[3] However, to demonstrate a real-world application of ALPHABETR, we successfully identified CDR3α/CDR3β pairs from T cells in the pairSEQ TIL sequencing dataset by comparing them to the pairs identified by pairSEQ [1].

Simulations of sequencing of epitope-specific T cell populations

To extensively test ALPHABETR, we created an comprehensive set of different synthetic datasets by simulating T cell populations with different numbers of clones, different immunodominance hierarchies, different degrees of CDR3α- and CDR3β-sharing, and different proportions of clones expressing dual-TCRα chains. In addition, we added experimental noise to simulate realistic errors in sequencing experiments, including sequencing errors and the failure to amplify and sequence a chain at all. We then simulated the sequencing experiments by sampling T cells from these populations into wells of virtual plates and then "sequenced" the CDR3s from sampled cells, allowing for the errors to occur.

[3] One could imagine an experiment where a T cell population is split into two samples, one sequenced with single-cell sequencing and one with the ALPHABETR approach. Although single-cell sequencing would provide us with some of the "right answers," current single-cell approaches cannot be scaled to sequence the thousands of clonotypes that would be identified by ALPHABETR. In addition, single-cell sequencing is itself prone to errors. It would be akin to trying to mark an exam with a partial answer sheet except the answer sheet was printed smudgy ink.

Percentage of CDR3α and CDR3β sequences shared	Level of sharing					
	Low		Medium		High	
	α	β	α	β	α	β
Not shared	90%	90%	81.6%	85.9%	60%	60%
2 clones	5%	5%	8.5%	7.6%	20%	20%
3 clones	5%	5%	2.1%	3.7%	10%	10%
4 clones	0%	0%	0.7%	1.9%	10%	10%
5 clones	0%	0%	3.3%	0.9%	0%	0%
6 clones	0%	0%	0.5%	0%	0%	0%
7 clones	0%	0%	3.3%	0%	0%	0%

Table 5.2: The three different levels of sharing used in the simulations. Three different levels of sharing were used to test the robustness of alphabetr to T cell populations with different degrees of sharing. The medium sharing level is the average of the rates of sharing found in all of the studies showing in Table 5.1. The low sharing level was chosen to have much fewer shared chains than the medium sharing level, and the high sharing level was chosen to have much more shared chains than the medium sharing level.

We simulated different degrees of CDR3α- and CDR3β-sharing at ranges of frequencies consistent with published single-cell TCR sequencing studies (Table 5.1) and our own data (Fig 5.1*A*). We also allowed between 10% and 30% of clones to express two productive TCRα chains and 6% of clones to express two productive TCRβ chains, both of which are at the upper limit of rates observed in T cell populations. The sequences in each 'well' were then generated by sampling between 10 and 300 T cells from the parent population with replacement. The choice in the pattern of sample sizes has huge implications in the accuracy and efficiency of the algorithm (Figure 5.3).

The breadth of the characteristics varied in our simulations reflect a range of plausible immunodominance hierarchies found in T cell responses, motivated by an analysis of epitope-specific cells recovered from human subjects immunized with live attenuated yellow fever virus vaccine based on analysis in Section 5.1.4 and from DeWitt et al. [62]). The following range of parameters were varied:

- **Number of distinct clones.** We tested populations with 500, 2100, and 3000 unique clonotypes. These numbers seem to cover a wide range of possible numbers of clonotypes found in antigen-specific populations that have been found based on β-only sequencing studies [62].
- **Different levels of skew in the immunodominance hierarchies.** We assumed skewed distributions of clone sizes, with between 5 and 50 clones comprising the most abundant 50% of the population and the

Sampling Strategy	Number of plates	Number of wells × **number of cells per well**					
High-Mixed	1	$26 \times \mathbf{20}$	$13 \times \mathbf{50}$	$19 \times \mathbf{100}$	$19 \times \mathbf{200}$	$19 \times \mathbf{300}$	
	5	$128 \times \mathbf{20}$	$64 \times \mathbf{50}$	$96 \times \mathbf{100}$	$96 \times \mathbf{200}$	$96 \times \mathbf{300}$	
Low-Mixed	1	$26 \times \mathbf{15}$	$6 \times \mathbf{20}$	$13 \times \mathbf{30}$	$19 \times \mathbf{50}$	$19 \times \mathbf{100}$	$19 \times \mathbf{150}$
	5	$96 \times \mathbf{15}$	$32 \times \mathbf{20}$	$64 \times \mathbf{30}$	$96 \times \mathbf{50}$	$96 \times \mathbf{100}$	$96 \times \mathbf{150}$

Table 5.3: The mixed sampling strategies used in the simulations. The bold text numbers indicate the number of cells sampled in the wells, and the associated normal text numbers indicate the number of wells with that sample size.

remainder forming a flat tail at low frequency (described mathematically in Section 5.5.5). These distributions test populations with different levels of skew in the clonal frequencies.

- **Different levels of sharing**. We simulated different degrees of CDR3α- and CDR3β-sharing at ranges of frequencies consistent with published single-cell TCR sequencing studies (Table 5.1) and our own data (Figure 5.1*A*). We tested populations that display low, medium, and high levels of sharing (defined in Table 5.2). The medium level of sharing is the average of the sharing levels of all studies shown in Table 5.2.
- **Proportion of clones expressing two productive TCRα chains**. We tested populations with 10% of its clones and 30% of its clones expressing two distinct CDR3α sequences. A level of 30% is used since it is an upper limit of estimates from the literature [127].
- **Different sampling strategies**. We explored different sampling strategies, i.e. the number of plates used and the number of cells sampled in each well of the plates. The different sampling strategies are described in Table 5.3.

To assess the robustness of ALPHABETR to experimental noise, we simulated the properties of two forms of sequencing error: dropping of chains and productive in-frame sequencing errors. Dropping of chains represents the failure to amplify or detect CDR3α and/or CDR3β sequences, a process which likely has both purely random and clone-specific elements [1]. In order to capture the dropping of chains in the simulations, every clone was randomly assigned a drop rate from a lognormal distribution with mean 0.15 and standard deviation of 0.01 with the rate capped at 0.9. Each instance of the CDR3α and CDR3β sequences from that clone (namely, the number of times that clone was sampled) was then removed from the well with

probability equal to the randomly assigned drop rate. In-frame sequencing errors occur when a sequencer mislabels a base or a series of bases in a read, potentially incorrectly identifying the CDR3 sequence with an erroneous nucleotide and/or amino acid sequence. To model these productive in-frame sequencing errors, every unique CDR3α and CDR3β sequence was assigned an error rate randomly drawn from a lognormal distribution with mean 0.02 and standard deviation 0.005. Each instance of a sequence at the per-cell level was replaced at random by one of three erroneous 'daughter' sequences—each set of three unique and specific to the parent sequence—with probability equal to the sequence-specific in-frame error rate. Thus, on average each CDR3α and CDR3β generated mutant daughter sequences at the rate of 2% per instance in each cell in the plate.

Testing with the pairSEQ TIL sequencing data

The TIL sequencing data described by Howie *et al*. provided an opportunity to test ALPHABETR with a real human TCR sequencing dataset of restricted clonal diversity [1]. The data were obtained by sampling T cells from nine different tumor samples from nine different patients into one 96-well plate and sequencing the CDR3α and CDR3β chains of the T cells sampled in each well using the pairSEQ platform. In addition, deep sequencing of the CDR3α and CDR3β sequences was performed on PMBCs from each patient using the immunoSEQ platform [18]. These sets of sequences from the patients served as libraries to resolve the sequences found in the mixed pairSEQ 96-well plate to their tumor sources.

Because the mixed plate contained 561452 unique CDR3α nucleotide sequences and 955987 unique CDR3β nucleotide sequences, these data are too diverse for direct input into ALPHABETR. [4] We therefore made tumor-sample-specific virtual plates by matching CDR3 regions in the pairSEQ wells to the immunoSEQ libraries of CDR3 sequences obtained from PMBCs sampled from each patient. This was done by exactly matching the first 76 bases of

[4]There are at least two reasons for this. First, is designed to handle the more restricted clonality found in epitope-specific populations and thus the hundreds of thousands of sequences represented in these data are too diverse for its design. Second, in the `alphabetr` R package, all of the possible αβ pairs are stored in an $n \times m$ matrix, where n is the number of unique CDR3α sequences and m is the number of unique CDR3β sequences. The 536740813124 possible αβ pairs are about 250 times greater than the largest positive integer that can be represented in R.

the nucleotide sequences of CDR3α pairSEQ sequences to the last 76 bases of the nucleotide sequences of the CDR3α libraries of each tumor sample. For the CDR3βsequences, matching regions were 81 bases in length. These choices reflect the different reads utilized by pairSEQ and immunoSEQ sequencing. Each of the nine plate of tumor-specific chains was then analyzed using ALPHABETR to obtain αβ pairs, and the pairs identified by ALPHABETR were compared to the pairs identified by pairSEQ. Since single-cell sequencing was not performed to validate the identified pairs, we can analyze only if the two methods agree; the possibility of both methods being wrong exists, which motivated the need to perform extensive simulations.

Metrics used to evaluated the algorithm

For the simulations, the αβ pairs identified by ALPHABETR were compared with the correct list of αβ pairs, and the following metrics were calculated:

1. *Overall depth*, the number of αβ pairs that were correctly identified, as a proportion of the total number in the parent population (in this calculation, a dual-TCRα clone $\alpha_j\alpha_k\beta$ is treated as two clones $\alpha_j\beta$ and $\alpha_k\beta$)
2. *Depth of top clones*, the proportion of the clones that comprise the top 50% of the population after ranking by frequency that were correctly identified
3. *False pairing rate*, the proportion of identified αβ pairs that were incorrect
4. *Adjusted dual depth*, a measure of how well dual-TCRα clones were be identified from candidate pairs, namely

$$\frac{\#\text{ correctly identified dual-TCR}\alpha\text{ clones}}{\#\text{ true dual-TCR}\alpha\text{ clones whose two }\alpha\text{ chains are in the list of candidate }\alpha\beta\text{ pairs}}$$

5. *False dual rate*, the proportion of candidate dual-TCRα clones that are incorrectly identified.

These five metrics evaluate different aspects of the accuracy and effectiveness of ALPHABETR. Overall depth measures how many correct CDR3α/CDR3β pairings that ALPHABETR can identify out of all possible pairs found

in the target populations, regardless of whether the CDR3α/CDR3β pair derive from a single TCR or a dual-α TCR clone. Depth of top clones indicates how well ALPHABETR identify the immunodominant clones of a T cell population. False pairing rate captures how often ALPHABETR identifies spurious CDRα/CDRβ pairs. Adjusted dual depth evaluates how well ALPHABETR distinguishes dual-TCRα clones from two clones sharing the same CDR3α sequence. Instead of using an absolute dual depth, we adjusted the dual depth for the fact ALPHABETR cannot identify a dual clone if both CDR3α sequences from a dual clone are not paired with the CDR3β of that clone. This measures when ALPHABETR makes a mistake of failing to identify a dual clone $\alpha_i\alpha_j\beta$ after it pairs $\alpha_i\beta$ and $\alpha_j\beta$. Finally, the false dual rate measures how many of the dual-TCRα clones identified by ALPHABETR are incorrect.

5.3 RESULTS

5.3.1 *Testing* ALPHABETR *with simulated datasets*

Mixed sampling strategies with stringent consensus threshold strikes a balance of depth and accuracy of pairing

We began testing ALPHABETR with simulations of populations with 2100 clones, medium levels of sharing (as defined in Table 5.2), 30% of clones expressing dual-TCRα chains, and experimental lognormal errors as described in Section 5.3.1. We evaluated how well the algorithm performs as measured by overall depth of the population, depth of top clones, depth of clones in the tail of the population, and the false pairing rate. The simulations also explored how different sampling strategies affect the performance measured by these metrics since the number of cells sampled in the wells of the sequencing plates can be varied. We also show results for different consensus thresholds uses for the pairings (described schematically in Section 5.2.1 and described mathematically in Section 5.5.1).

Figure 5.3 shows the results from these simulations. Figure 5.3*A* shows the depth of the clones in the top 50% of the population only, and Figure 5.3*B* shows the depth of the clones in the other 50% tail of the population only. Figure 5.3C shows the depth for the whole population, and Figure 5.3*D* shows the false pairing rates. The different number of top clones

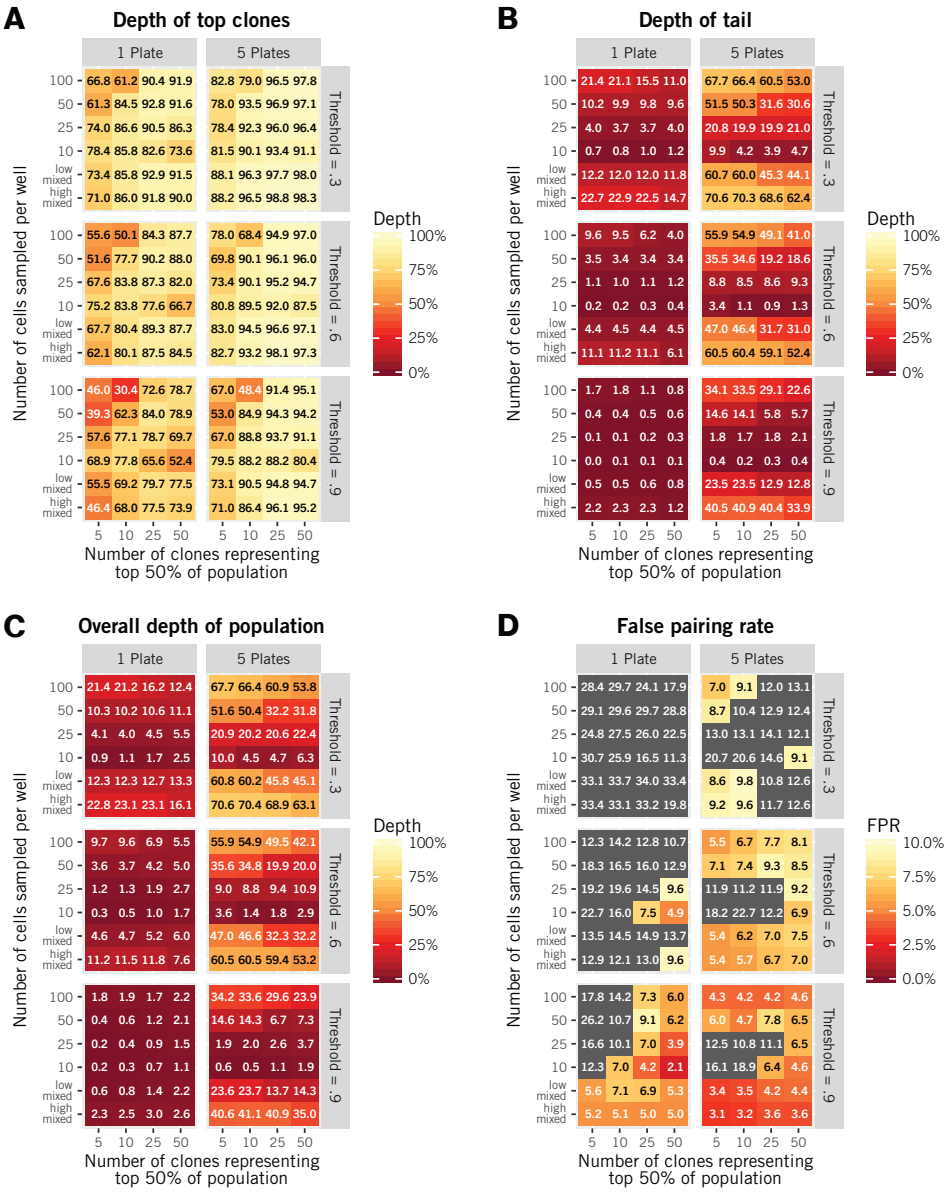

A
Depth of top clones
1 Plate
5 Plates
Threshold = .3
Threshold = .6
Threshold = .9
Number of cells sampled per well
Number of clones representing top 50% of population
Depth
100%
75%
50%
25%
0%
B
Depth of tail
C
Overall depth of population
D
False pairing rate
FPR
10.0%
7.5%
5.0%
2.5%
0.0%

Figure 5.3: (Facing page) Depth and accuracy of αβ pairings generated by alphabetr, for a range of overall sample sizes, sampling strategies and underlying distributions of clone sizes. Simulations were performed using simulated data sets of one or five plates using six different sampling strategies (Table 5.3) and different degrees of skew in clonal frequencies, as indicated by the number of clones comprising 50% of the population when ranked by frequency. 'Threshold' refers to the stringency of pair association, T (Section 5.5.1). **(A)** The proportion of the most abundant 50% of clones that were identified. **(B)** The proportion of the least abundant 50% of clones that were identified. **(C)** The overall depth is influenced strongly by the tail depth, indicating that data from one plate may be sufficient for recovering the most common clones. **(D)** The rate at which CDR3α and CDR3β sequences were incorrectly paired.

changes on the x-axis (populations become less skewed moving from left to right), and the different sampling strategies are represented on the y-axis. The rows of the facetted panels show the application of different consensus thresholds, and the columns of the facetted panels represent the use of either one 96-well plate or five 96-well plates. For each set of conditions, metrics were computed by averaging the results of 100 simulated experiments.

With only a single plate, the most abundant 50% of clones were recovered with depths between 62% and 89% with a moderate threshold of 0.6 and the mixed sampling strategies, improving with less skewed distributions (Figure 5.3*A*, left panels). Coverage of rare clones (Figure 5.3*B*, left panels) is much more limited, particularly for sparse sampling strategies, but improves with a more lenient consensus threshold of 0.3. The use of five plates considerably boosts the recovery of rare clones (Figure 5.3*B*, right panels), providing up to 61% depth with a threshold of 0.6 and 70% with a threshold of 0.3. As a result, for all sampling strategies, increasing the number of plates—and hence total sample size—increases overall depth (Figure 5.3*C*), almost entirely through greater recovery of rarer clones.

Increasing the number of plates also significantly reduces the false pairing rate (Figure 5.3*D*), which can be as low as 3.1% for 5 plates and a stringent threshold of 0.9 (Figure 5.3*D*, lower right panel). In general, increasing the stringency threshold reduces false pairing rates. The stringency of the threshold can be relaxed if there is no significant presence of dual-TCRβ clones in the T cell population, which will be discussed in the end of this section.

Increasing the consensus threshold of the resampling procedure—that is, requiring a high frequency of occurrence of candidate pairs across subsets of the data—results in a lower false pairing rate at the cost of lower depth,

largely for rarer clones (Figures 5.3*C–D*). This is because rarer clones will be excluded from the replicates more often than common ones; as the stringency of pair selection is increased, rare clones will tend to be filtered out.

In summary, mixed sampling strategies with moderate to high acceptance thresholds yield the lowest false pairing rates (Figure 5.3*D*) while maintaining good depth of recovery of rare clones (Figure 5.3*B*). The high-mixed strategy obtains a larger overall sample size than the other sampling strategies and thus achieves greater depths, particularly of rare clones.

In practice, the availability of cells may place constraints on the sampling strategy. For example, with five plates the high- and low-mixed strategies require a total of 64,000 and 33,000 cells respectively. A typical sample of four tubes (approximately a total of 30mL) of human blood yields roughly 3×10^7 PBMCs, of which roughly half are αβ T cells. With such a sample, numbers of T cells specific for immunodominant epitopes of highly immunogenic infections such as Epstein-Barr virus and cytomegalovirus [137, 138, 139, 140], the number of antigen-specific T cells is unlikely to be limiting. A conservative estimate is that to acquire ~100,000 cells needed for the high-mixed sampling strategy on five 96-well plates requires epitope-specific frequencies in excess of 1% of αβ T cells, or 0.5% of PBMC. Frequencies below this may dictate fewer plates and/or a sparser sampling strategy.

Precise estimation of frequencies of common clones benefits from sparse or mixed sampling strategies

We assessed the ability of ALPHABETR to estimate clonal frequencies over a range of clonal size distributions and sampling strategies (Figure 5.4). We show results only for the most abundant clones making up 50% of the population. The left and right panels of Figure 5.4*A* show typical sets of abundance estimates for populations with moderately and highly skewed clonal distributions, with 25 and 5 clones respectively making up the top 50% of clones by size. We tested the method of construction of point estimates and confidence intervals using simulated data and confirmed that close to 95% of such intervals contained the true frequency (results not shown).

Figure 5.4*B* summarizes the precision of the abundance estimation for a variety of sampling strategies and levels of skew. We also calculate the coefficient of variation $\hat{\sigma}/\hat{f}$, where $\hat{\sigma}$ is estimated by using a quadratic ap-

proximation to the 95% confidence interval $3.92\hat{\sigma}$, and $\hat{f}$ is the estimated frequency. The procedure yielded CVs in the range 0.13–0.41 for one plate and 0.07–0.20 for five plates (Figure 5.4*B*), indicating very precise estimates.

Intuitively, the influence of skew arises because the data provides the most information about the frequency of a given clone when sample sizes causes the clone to appear in an intermediate proportion of wells. For instance, as an extreme example, if a clone appears in 0 wells of the plates, then the data have no information about the frequency of the clone. At the other extreme, if a clone appears in all of the wells of the plate, then the data indicate only that the clone is very abundant but fails to be informative about the magnitude of the frequency (namely, a large range of frequencies can explain the data). Sampling low numbers of cells is therefore optimal for determining the abundance of highly immunodominant clones to ensure that they are not sampled in too many wells, and larger numbers are optimal for determining the abundance of rare clones so that they are sampled at all. For the clone distributions considered here, the sparsest sampling strategy (i.e. uniformly 10 cells/well in our simulations) gives the greatest precision for the common clones. In general, however a mixed sampling strategy strikes a balance between precision over a wide range of abundances (Figure 5.4*B*, bottom row in each panel), false pairing rates, and depth.

The clonal frequencies shown in Figure 5.4*A* depend on prior knowledge or estimation of the mean drop rate, or the mean probability that any CDR3α or CDR3β of a clone will fail to be sequenced. Although this drop rate is most likely specific to experimental protocol and equipment, our simulations used a mean of 15%, which appears to be an upper bound for these drop rates, and this rate can be roughly estimated using a protocol performed by Howie *et al*. [1].

Neglecting this error rate yields lower bounds on clonal abundances. We performed simulations to determine the sensitivity of clonal frequency estimation to inaccuracies in estimates of the average drop rate of CDR3α and CDR3β sequences. These simulations involved creating data sets of 5 plates, the high-mixed sampling strategy, and a constant mean drop rate of $\varepsilon = 0.15$. We then performed frequency estimation on these data sets using mean drop rates of $\varepsilon = 0.15$, $\varepsilon = 0.08$, and $\varepsilon = 0$. We show the mean bias

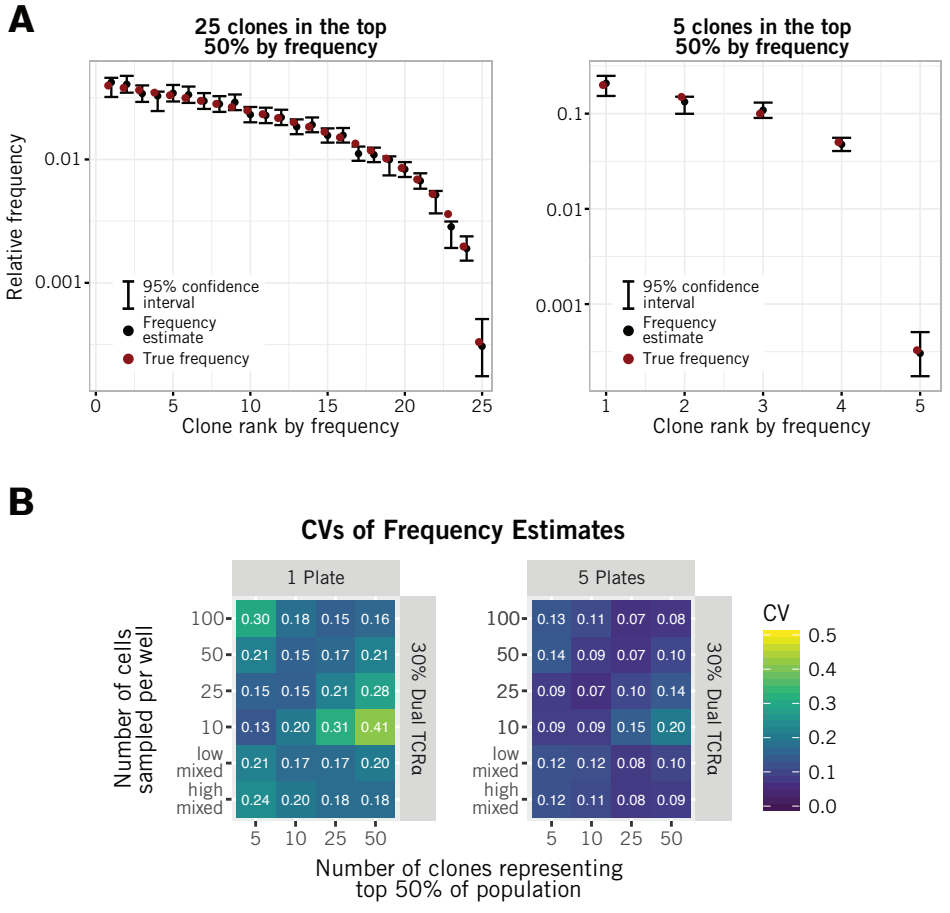


Figure 5.4: Assessment of the precision of clonal frequency estimation. **(A)** Point estimates of clonal frequencies calculated by alphabetr, derived from representative simulations using five plates and distributions with 25 and 5 clones in the top 50% (left and right panels respectively). **(B)** The CVs of frequency estimates for a range of skew of clone sizes and sampling strategies. Results here are averages over 100 simulations.

Dropping rate	Number of top clones	Percent of 95%-CI containing true frequency	Mean bias
$\varepsilon = 0.15$	5	91.5%	3.8%
	10	93.4%	2.4%
	25	92.7%	2.1%
	50	92.3%	3.0%
$\varepsilon = 0.08$	5	65.3%	-12.1%
	10	53.4%	-12.7%
	25	47.7%	-12.8%
	50	47.8%	-12.7%
$\varepsilon = 0$	5	31.2%	-25.2%
	10	12.8%	-26.4%
	25	6.7%	-26.3%
	50	7.8%	-26.2%

Table 5.4: Assessing the impact of underestimation of sequencing error on clonal frequency estimation. Using accurate estimates of the drop rate (here, $\varepsilon = 0.15$) results in accurate frequency estimates. Underestimating the drop rate leads to biased estimates that are lower bounds on the true frequencies.

in these frequency estimates, where

$$\text{Bias} = \frac{\text{Estimated frequency} - \text{True frequency}}{\text{True frequency}}$$

Table 5.4 shows the results of 200 simulations for each combination of ε and skewness of the clone size distribution indicated by the number of clones in the top 50% of the population when ranked by clonal frequency. Assuming no dropping of chains ($\varepsilon = 0$) leads to the underestimation of clonal frequencies and inaccuracies in the construction of the 95% confidence intervals.

Efficient discrimination of dual-TCRa and CDR3ß-sharing clones requires a mixed sampling strategy and distinct methods for common and rare clones

The final step in the algorithm is to decide whether each candidate pair of clones that share a β chain (e.g. $\alpha_i\beta$ and $\alpha_j\beta$) are indeed two clones or derive from one dual-TCRα clone (e.g. $\alpha_i\alpha_j\beta$). Two methods are used for this discrimination, one which identifies rare clones and another which identifies more common clones.

To do this, we exploit the fact that the pattern of co-occurrences of all three chains will be different under these two hypotheses. Initially, we use

the estimated frequencies of a putative β-sharing clone pair $\alpha_i\beta$ and $\alpha_j\beta$ to calculate the expected number of wells in which all three chains should co-occur. Essentially, the three chains will tend to co-occur more frequently if they derive from a dual-TCRα clone than if they derive from two β-sharing clones. We construct the ratio of the expected to the observed numbers of three-way co-occurrences for each β-sharing pair and perform *k*-means clustering on these ratios. The cluster of higher values forms the first list of candidate dual-TCRα clones. A visual example of the clustering of clones into two groups is shown in Figure 5.6.

However, performing *k*-means clustering on only the numbers of three-way occurrences is inefficient at discriminating β-sharing and dual-TCRα clones that are relatively abundant because the expected frequencies of co-occurrences become indistinguishable, particularly for rich sampling strategies in which the three chains co-occur in many wells. We therefore added a second step which utilizes more information from the plates, calculating the likelihoods of all three- and two-way concurrences of α_i, α_j, and β under both hypotheses. Exact computation of these likelihoods is only practical for the low-occupancy wells (less than 50 cells/well), which conveniently are also the wells that contain maximal information regarding common clones. As a result, this second approach can be performed only when using sparse sampling strategies or the low-occupancy wells used in the mixed sampling strategies. We determined empirically that differences in the log likelihoods of more than 10 distinguish the β-sharing and dual-TCRα hypotheses.

Figure 5.5 summarizes the ability of the algorithm to distinguish TCRβ-sharing and dual-TCRα clones. Common clones are identified through the three-way likelihood approach, and mixed sampling strategies give the best results in this case, with adjusted depths of up to 79% for less skewed distributions (Figure 5.5*A*). The likelihood approach still performs relatively poorly with very highly skewed populations, distinguishing dual-TCRα from β-sharers correctly at most 34% of the time for population with 5 clones making up the top 50% of the population (Figure 5.5*A*). Under these circumstances, the statistics of co-incidence of the three chains are very similar under the two hypotheses of dual-TCRα or TCRβ-sharing clones. In contrast, the *k*-means procedure achieves adjusted depths of 93–99% for rare clones when using 5 plates and the high-mixed strategy (Figure 5.5*B*). Averaging over all clones, this strategy gives false dual rates of between 10–13%

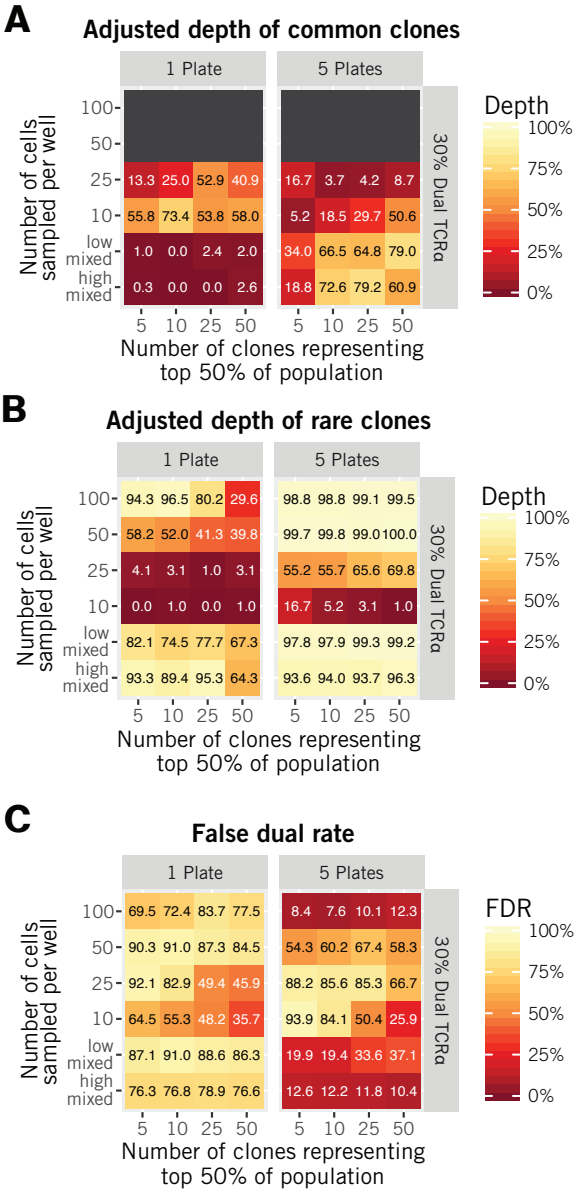


Figure 5.5: Discriminating between dual-TCRα and β-sharing clones. We assess the degree of recovery of dual-TCRα clones with the adjusted dual depth, which is the proportion of dual-TCRα clones correctly assigned out of the list of candidate dual-TCRα and TCRβ-sharing clones. **(A)** The adjusted dual depth of common clones. **(B)** The adjusted dual depth of rare clones. For common clones, we used likelihood-based discrimination; for rare clones we used a *k*-means clustering approach. **(C)** The false dual rate averaged over all clones—the proportion of identified dual-TCRα that are incorrect. All results are shown for a threshold of 0.3 with 30% prevalence of dual-TCRα and are averages over 100 simulations.

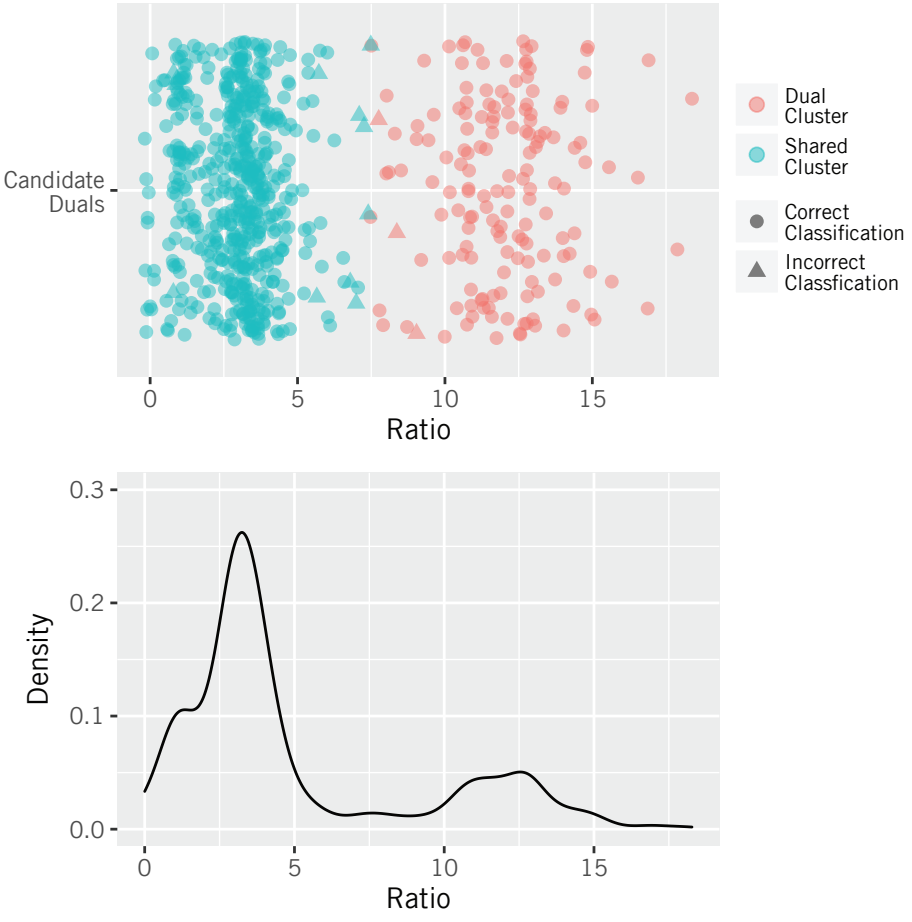


Figure 5.6: Discriminating between β-sharing and dual-TCRα clones. For each candidate β-sharing clone pair ($\alpha_i\beta$, $\alpha_j\beta$) we calculate the ratio of the observed number of co-occurrences of the three chains to the number expected under the hypothesis they are indeed distinct clones. The latter uses the estimated frequencies of the two clones. These ratios typically partition into a lower set of values that represent CDR3β sharers and a higher set of values that represent dual-TCRαclones. Note that the lower cluster is not centered on one; this is because calculating the expected number of co-occurrences of all three chains under the two hypotheses with a non-zero drop rate ε is computationally intractable due to large multinomial coefficients. However, assuming $\varepsilon = 0$ biases both estimates, and *k*-means still resolves the two clusters of ratios.

(Figure 5.5C).

Exploring different degrees of CDR3a and CDR3ß sharing, richness in clonal structure, and prevalence of dual TCRß

In order to evaluate the robustness of ALPHABETR, we performed simulations to measure its ability to determine TCR pairs from a wide range of antigen-specific T cell populations with different levels of sharing, with different numbers of distinct clones, and without the presence of dual-TCRβ clones.

Different levels of sharing. We performed simulations with the high level of sharing and low leves of sharing within populations comprised of 2100 clones (Table 5.1). The higher sharing level had a minimal effect on top and tail depths while causing an absolute increase in false pairing rates of only 1%-2% (Figure 5.7). The lower level of sharing decreased the false pairing rate by approximately 1% in absolute terms while having minimal effect on top and tail depths (Figure 5.8). In both cases, the depths of recovery were very similar to those presented in Figure 5.3 with very small differences in the false pairing rate.

Different levels of clonal diversity. We simulated populations with higher diversity by increasing the number of unique clones to 3000 clones and populations with lower diversity by decreasing the number of clones to 500 clones (with error models and levels of sharing identical to those used for simulations for Figure 5.3. For populations of 3000 clones, ALPHABETR maintained the same top depth while obtaining slightly lower tail depths, which is not surprising given that larger total sample sizes are needed to achieve coverage of the larger tail of more diverse populations (Figure 5.10). False pairing rates show no substantial difference. For populations of 500 clones, simulations show that ALPHABETR again have similar top depths and higher tail depths (Figure 5.9). The higher false pairing rates seen here are due to the fact that using both of the mixed sampling strategies described in the main text (Table 5.3) involve sufficiently large numbers of cells per well. For low-diversity populations, common clones will appear together in wells very often with these sample sizes, creating ambiguity in pairing and increasing the apparent degree of sharing of chains between clones. Populations comprising of fewer clones overall will by definition have higher rela-

tive abundances, and in such situations, frequency-based pairing approaches will benefit from sparser sampling strategies.

The presence of dual TCRβ clones. Although ALPHABETR does not identify dual TCRβ clones, we performed simulations to see how the presence of such clones in the parent population affects the ability of ALPHABETR to associate TCRα and TCRβ correctly (Figure 5.11). The presence of dual TCRβ clones at a frequency of 6% increases the false pairing rate by approximately 3% in absolute terms (compare Figure 5.11*D* and Figure 5.3*D*) while not affecting the top and tail depths (Figures 5.11*A–B* and Figures 5.3*A–B*). Since significant levels of dual-TCRβ clones have been shown in only a small number of studies sequencing antigen-specific T cell populations [131, 70], we believe this represents an upper bound on the effect of dual TCRβ clones on the performance of ALPHABETR.

5.3.2 *Testing* ALPHABETR *with tumor-infiltrating lymphocyte sequencing data*

Using simulated data allowed us to assess the performance of ALPHABETR directly using the gold standard of known TCRαβ sequences and under a range of plausible experimental conditions. Despite the wide range of different plausible T cell populations used in the simulations, we wanted to directly demonstrate the utility of ALPHABETR with T cell sequencing data from human samples. So, we illustrate a real-world application by applying ALPHABETR to a published TIL sequencing dataset as described in Section 5.2.2. This dataset had less than 10^6 unique CDR3α and CDR3β sequences and thus likely represents samples from clonally restricted T cell populations since these numbers are far less than the estimated diversity of the naive T cell repertoire. One tumor sampled (Breast 1) yielded only 7 pairs by pairSEQ, and we excluded it from the analysis. We applied ALPHABETR to the chains from each of the remaining 8 tumors in turn and then compared the pairs determined by ALPHABETR to those identified explicitly by pairSEQ (Table 5.5). The true TCR clonotypes are unknown and so our aim was to measure degrees of concordance and conflict between the two methods. In 6 out of 8 tumors, ALPHABETR recovered fewer clones; however we found average concordance rates of 77%, defined as the proportion of the pairs identified by ALPHABETR that were also identified by pairSEQ. Perhaps more strikingly, we also found a very low incidence of conflicting pairs (mean 2%

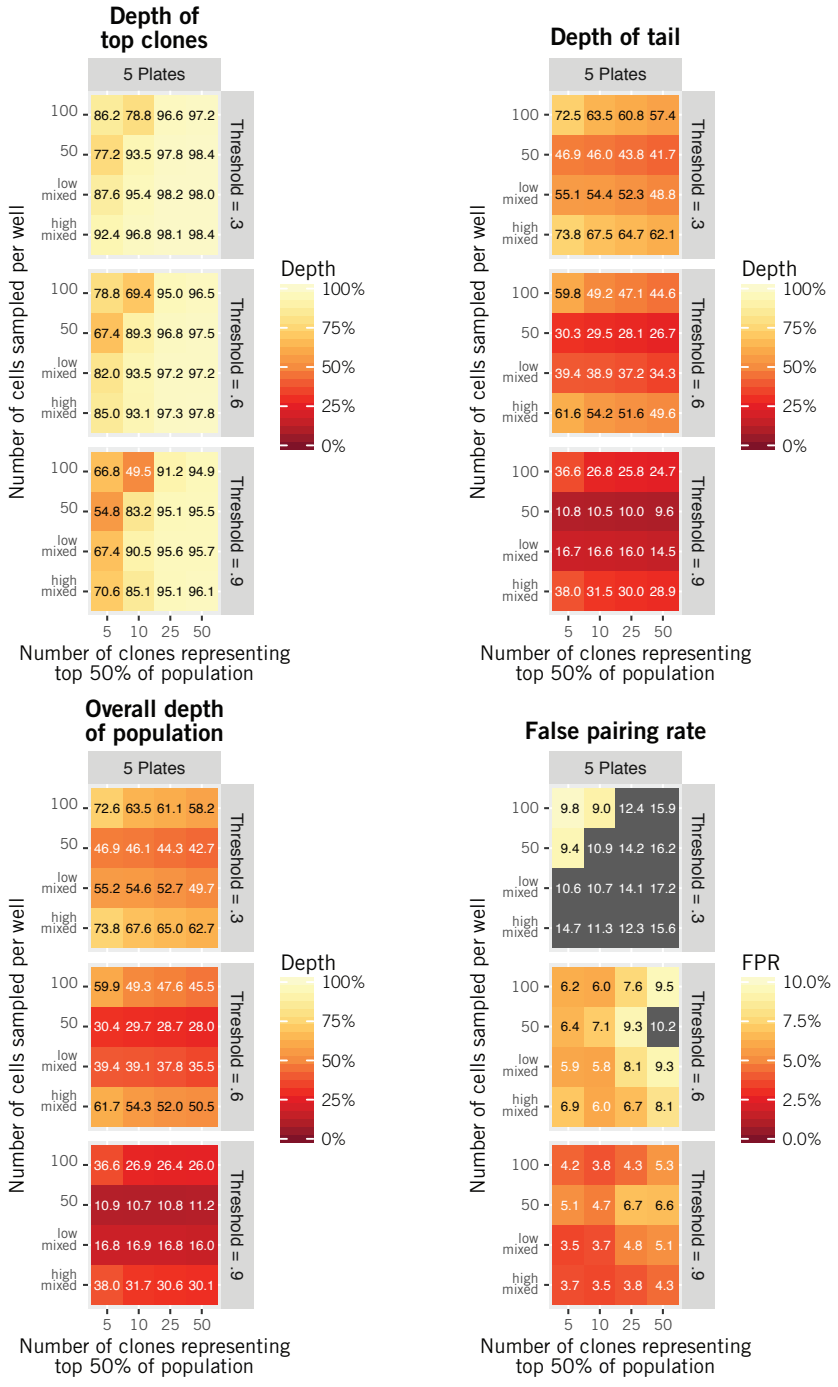


Figure 5.7: Simulations of populations with a high level of sharing. Performance of alphabetr at the high level of chain sharing. The results shown are the averages of 100 simulations.

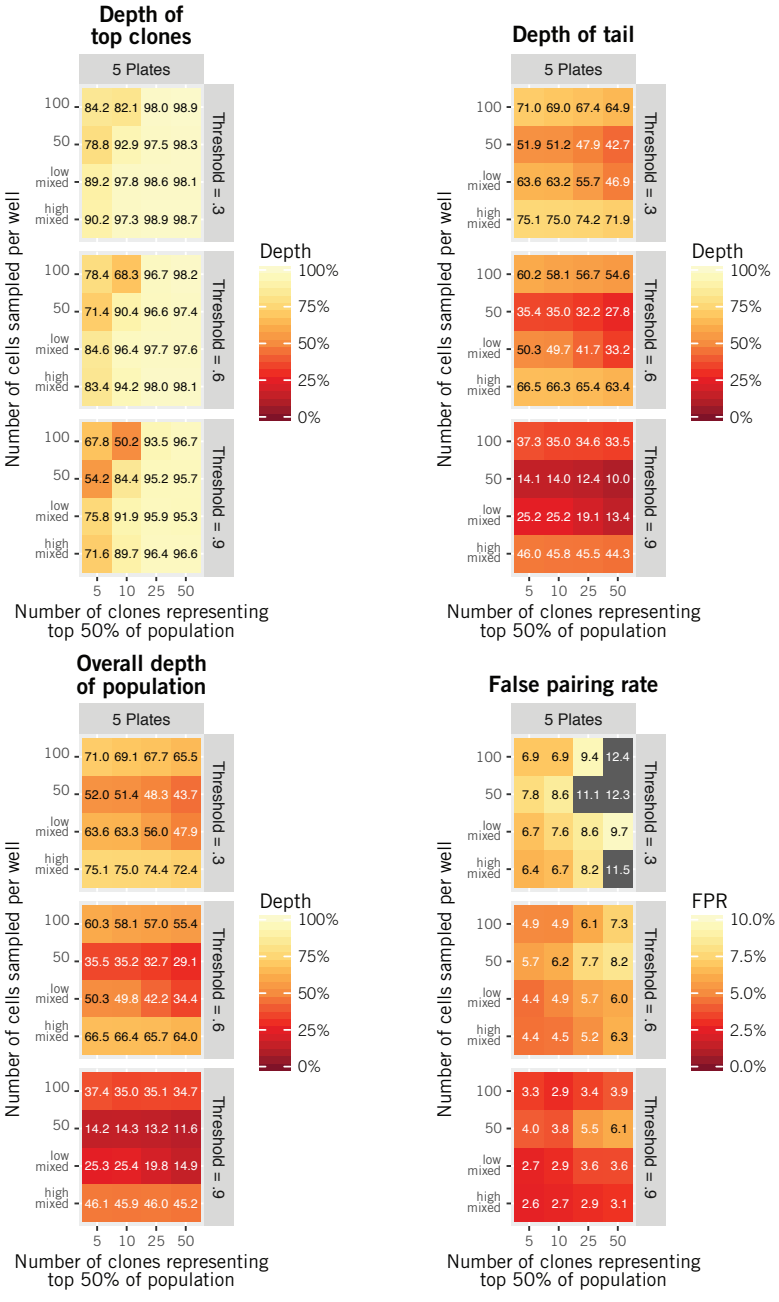


Figure 5.8: Simulations of populations with a low level of sharing. Performance of alphabetr at low level of chain sharing. The results shown are the averages of 100 simulations.

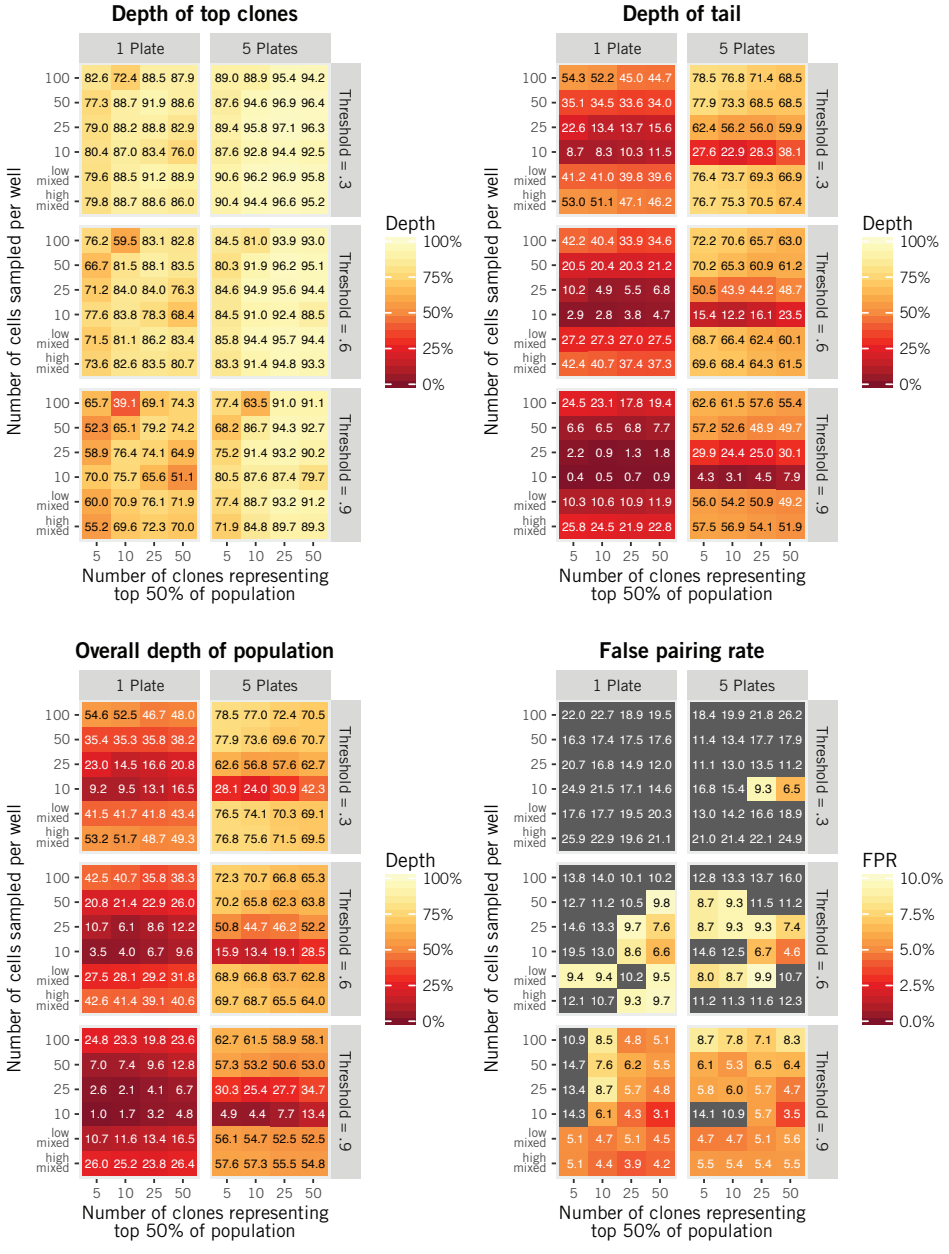


Figure 5.9: Simulations of populations consisting of 500 unique clones. Performance of alphabetr with 500 clones in the parent T cell population. The results shown are the averages of 100 simulations.

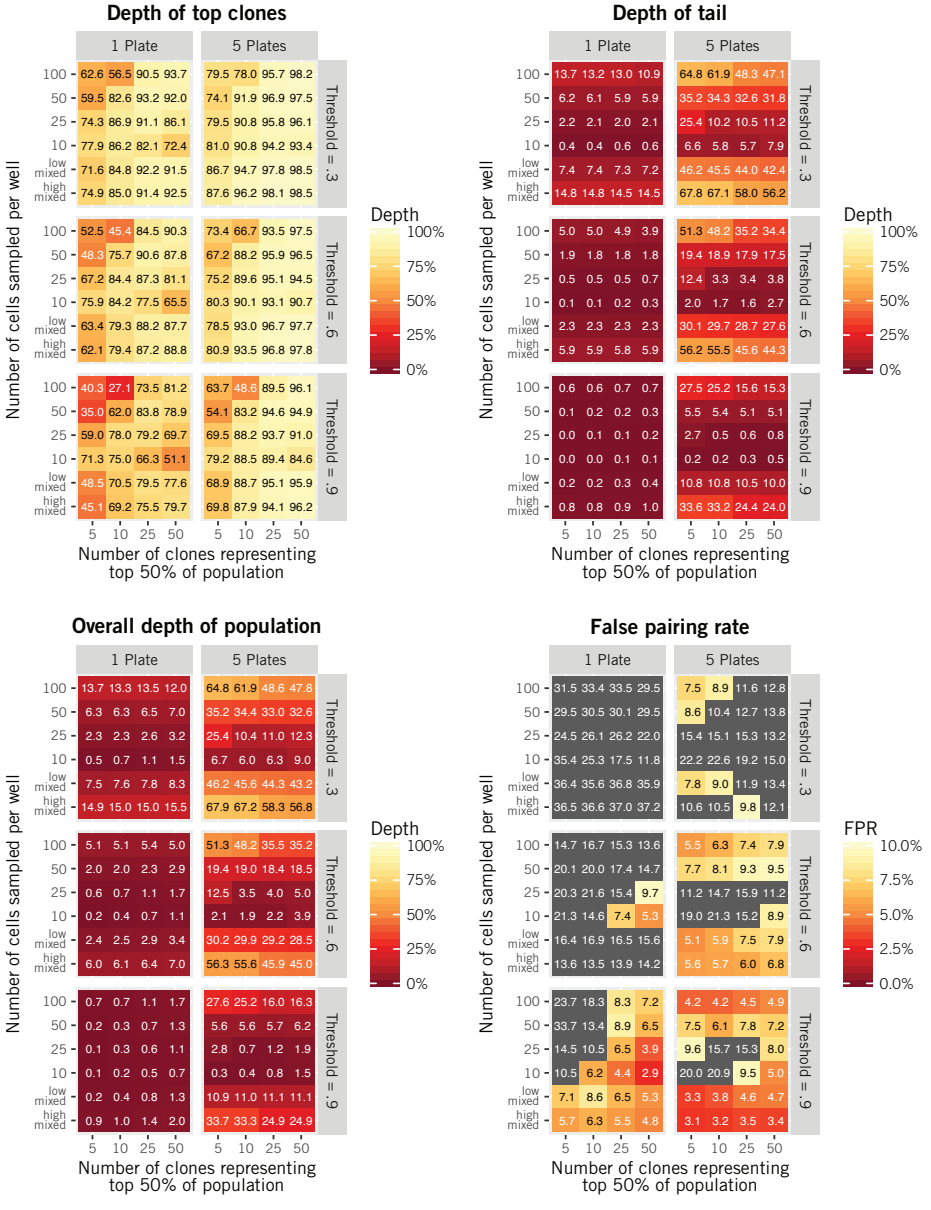


Figure 5.10: Simulations of populations consisting of 3000 unique clones. Performance of alphabetr with 3000 clones in the parent T cell population. The results shown are the averages of 100 simulations.

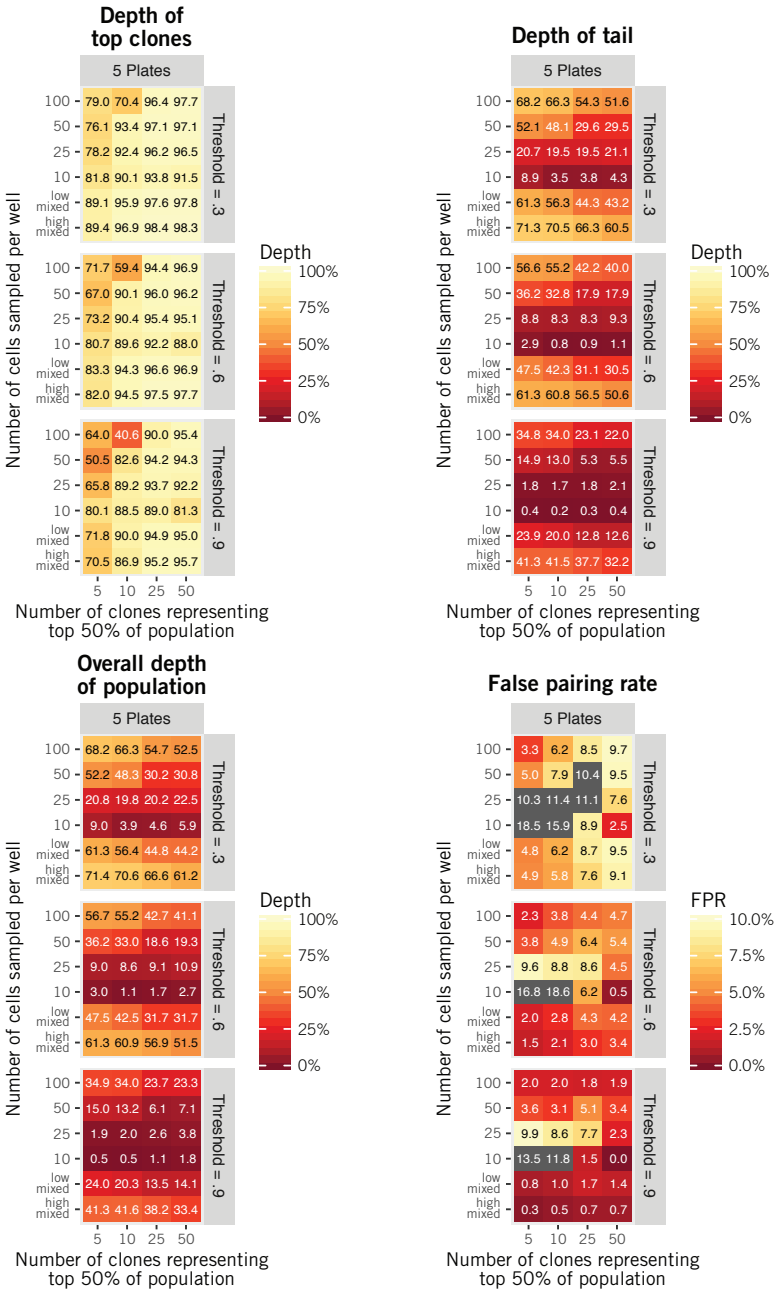
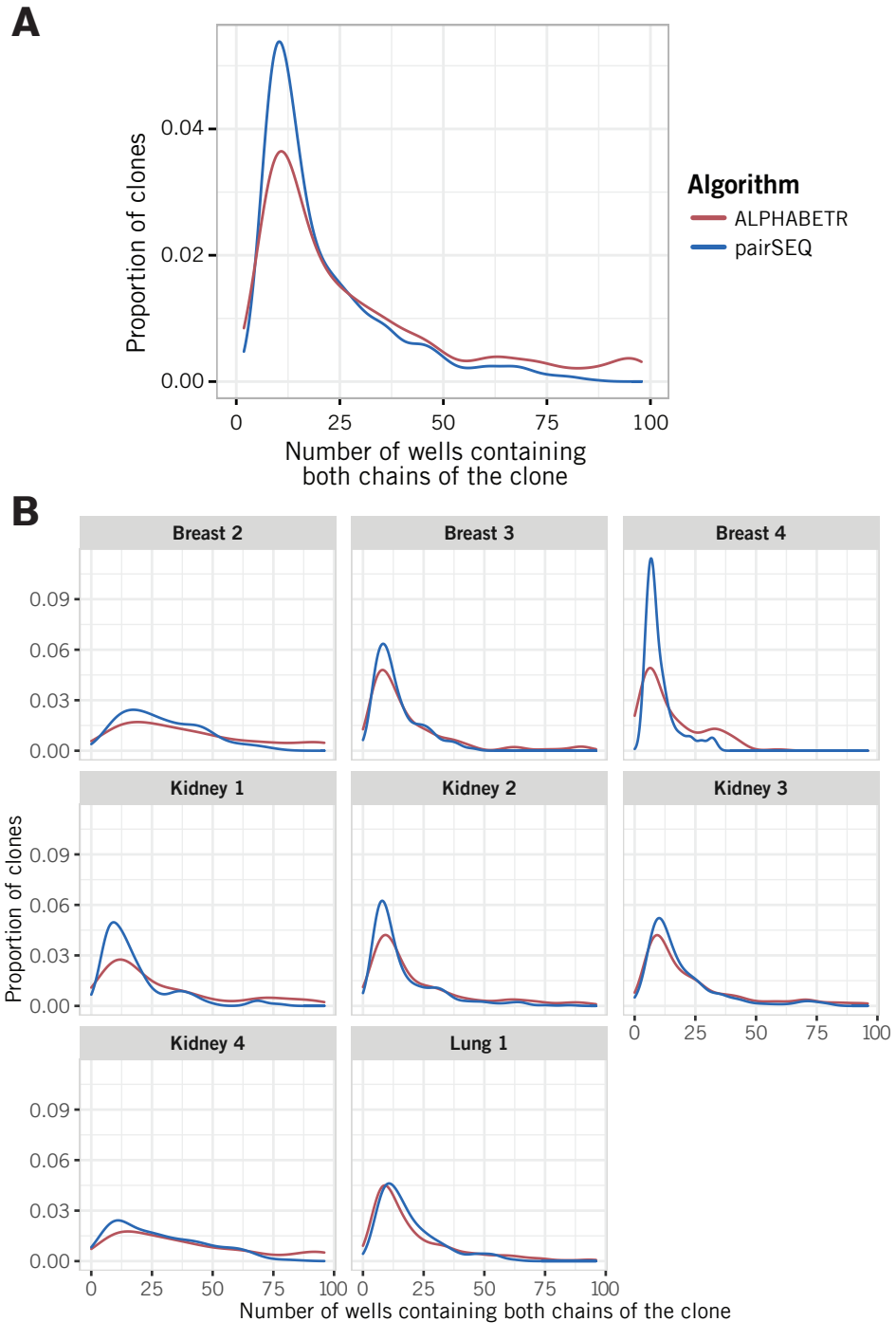


Figure 5.11: Simulations with no dual-TCRβ clones. Performance of alphabetr without dual-CDR3β chains in populations of 2100 clones and medium levels of sharing. The results shown are the averages of 100 simulations.


A
Proportion of clones
0.04
0.02
0.00
0
25
50
75
100
Number of wells containing
both chains of the clone
Algorithm
ALPHABETR
pairSEQ
B
Breast 2
Breast 3
Breast 4
Kidney 1
Kidney 2
Kidney 3
Kidney 4
Lung 1
Proportion of clones
0.09
0.06
0.03
0.00
0
25
50
75
100
Number of wells containing both chains of the clone

Tumor Sample	Number of pairs identified by ALPHABETR	Number of pairs unambiguously identified by pairSEQ	Number of identical identified pairs	Percentage of ALPHABETR pairs agreeing with pairSEQ	Number of conflicting pairs	Number of novel pairs from ALPHABETR
Breast 2	98	85	74	75.5%	0	24
Breast 3	109	129	94	86.2%	1	14
Breast 4	50	85	26	52.0%	1	23
Kidney 1	74	112	58	78.4%	3	13
Kidney 2	145	286	126	86.9%	5	5
Kidney 3	213	282	166	77.9%	8	39
Kidney 4	157	176	131	83.4%	1	25
Lung 1	173	163	124	71.7%	1	48

Table 5.5: Recovery of tumor-infiltrating lymphocyte TCR pairs using alphabetr and data from Reference [1]. The data were processed by associating chains with their tumor sources through exact matching of the CDR3 nucleotide sequences from the mixed tumor samples to the CDR3 libraries obtained from blood samples from each patient. The data were then simplified by selecting only those chains associated with one tumor. We then used alphabetr to identify TCRαβ pairs. The numbers of pairs unambiguously identified by pairSEQ were determined by directly matching nucleotide sequences to the CDR3 libraries, and only those pairs for which both chains could be directly associated with the corresponding tumor sample were included in the analysis.

Figure 5.12: (Facing page) Comparison of well occupancy patterns of the clones identified by alphabetr and by pairSEQ. For each method, TCRαβ pairs identified for all tumor samples were combined to estimate the distribution of the number of wells in which the chains co-appeared. The differences between these distributions indicate the relative efficiency with which the two algorithms identify clones, as a function of their abundance.

across tumors, as a proportion of all pairs identified by ALPHABETR). Conflicts were defined as those determined by the two methods that have only one chain in common. In addition, ALPHABETR identified 5–48 novel pairs in the tumor samples. The high concordance rate and low conflicting rate is particularly encouraging since the sampling strategy used in these data are not ideal for ALPHABETR since it is optimized for pairSEQ. These results all provide additional validation for the accuracy and efficiency of ALPHABETR and indicate that ALPHABETR can be applied to T cell populations of limited clonality such as TIL populations and not just epitope-specific T cell populations.

To compare the abilities of the two algorithms to identify rare or common clones, we stratified the identified αβ chain pairs by the frequency with which they co-appeared in wells. With stringency thresholds greater than 0.7, we find that with a single 96-well plate and a sampling strategy optimized for pairSEQ, ALPHABETR is less efficient at identifying rare clones but identifies clones with moderate to high abundances—for which the TCRα and TCRβ chains co-appear in more than a quarter of the wells—more efficiently, as shown in Figure 5.12*A*. Figure 5.12*B* shows similar patterns when this analysis is broken down by each individual tumor sample. The clones identified by ALPHABETR alone exhibit moderate levels of sharing (CDR3α sharing, mean 16%, range 0-60%; CDR3β sharing, mean 13%, range 4-31%). Of the sharers, an average of 76% share a chain with a clone that was identified by both methods.

Extensive single-cell sequencing is required to achieve equivalent overall depth to ALPHABETR

We wanted to determine if and how implementing ALPHABETR improves upon single-cell methodologies. One approach to assessing this would involving splitting a sample of antigen-specific T cells into two subsamples, perform single-cell sequencing on one subsample, and apply ALPHABETR to the remainder to compare their performance on the same set of parent clones. An alternative approach that we employed is by simulating both scenarios. The advantages of the simulation approach are that it allows us to (*i*) triangulate both methods with the gold-standard of knowing true sequences with complete certainty, which is not possible in practical settings due to dropping of chains and in-frame sequencing errors, and (*ii*) explore levels of single-cell sequencing that are currently prohibitively costly.

We simulated the sequencing of 96 to 9600 single cells sampled from the same synthetic T cell populations used for evaluating ALPHABETR, particularly with the same sequencing errors. Figure 5.13 compares the performance of the two methods for a population of 2100 clones with 5, 10, 25, and 50 clones making up the top 50% by frequency. ALPHABETR was implemented with the five-plate high-mixed sampling strategy and with a stringency threshold $T = 0.6$. Under the conditions used for Figure 5.13, almost double the number of single-cell sequencing runs was required to achieve the

same top depth yielded by ALPHABETR with five plates, and more than 100 plates of single cells are required to approach ALPHABETR's level of recovery of rare clones. With the same clone size distribution, even a single plate analyzed with ALPHABETR yields top depths from 78% to 92%, depending on the threshold parameter used (Figure 5.3*A*), whereas 96 single cells yield a top depth of 60% (Figure 5.13, middle column of plots). Single-cell sequencing will exhibit a false positive rate that is approximately twice the mean of the in-frame error rate, or 4% in our simulations, an accuracy that is comparable to that of ALPHABETR at its most stringent.[5]

5.3.3 *The* ALPHABETR *package*

An open-source implementation of ALPHABETR is available freely on CRAN as an R package, and the most updated development version is available on *github.com/edwardslee/alphabetr*. The package includes a detailed vignette that explains step-by-step usage of the package.

5.4 DISCUSSION

Although high throughput single-cell sequencing approaches are becoming more inexpensive and technically feasible, smaller-scale solutions using frequency-based sampling potentially remain far more economical. Our approach and the commercially available pairSEQ are currently the only two frequency-based pairing technologies described in the literature that can be scaled to identify thousands of TCR sequence pairs without sequencing tens of thousands of single cells. The ALPHABETR approach is the first to our knowledge to directly address the promiscuous nature of CDR3α and

[5] Spurious TCR pairs occur in single-cell sequencing when one or both of the CDR3s are incorrect identified with an in-frame sequencing error. In a single-cell sequencing experiment, if one of the chains are dropped and is identified with a shorten CDR3 with an early stop codon, then these errors are easy to detect and can simply removed from the final list of CDR3 pairs. If a spurious in-frame error occurs on average 2% of the time (as chosen in our simulations), then probability of identifying an incorrect TCR pair is

$$\begin{aligned} &P(\text{CDR3}\alpha \text{ having an in-frame error}) + \\ &P(\text{CDR3}\beta \text{ having an in-frame error}) + \\ &P(\text{both chains having in-frame errors}) = \\ &\quad .02 \times .98 + .98 \times .02 + .02 \times .02 = 3.97\% \end{aligned}$$

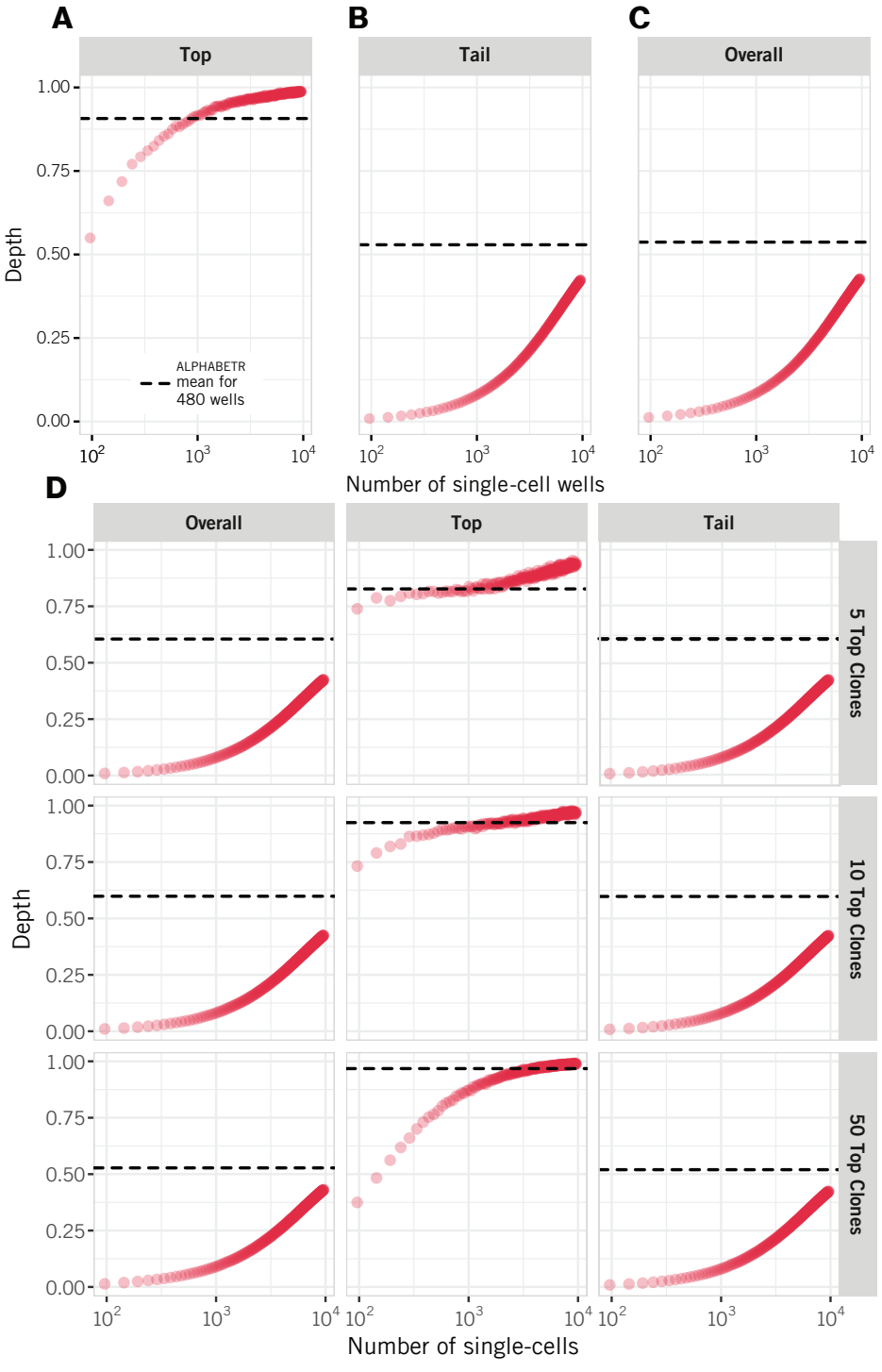

A
B
C
Top
Tail
Overall
1.00
0.75
0.50
0.25
0.00
Depth
ALPHABETR
mean for
480 wells
10^2
10^3
10^4
Number of single-cell wells
D
Overall
Top
Tail
5 Top Clones
10 Top Clones
50 Top Clones
Depth
Number of single-cells

Figure 5.13: (Facing page) Comparison of single-cell approaches and alphabetr. Single-cell sequencing was simulated by sampling from the same populations used to evaluate alphabetr and including both the dropping errors and in-frame sequencing errors. In these simulations, the parent population contains 2100 clones with 25 clones representing the top 50% of the clones ranked by abundance. The results were evaluated for **(A)** top depth, **(B)** tail depth, and **(C)** overall depth. The dashed lines show the mean performance of alphabetr applied to five plates using the high-mixed sampling strategy and a threshold of 0.6 (values taken from Figure 5.3). **(D)** The same simulations were performed for populations with 5, 10, and 50 clones representing the top 50% of the clones ranked by frequency. The single-cell sequencing results are averages of 200 simulations.

CDR3β usage within epitope-specific populations. The combination of ALPHABETR and relatively low-cost sequencing strategies addresses these issues by being capable of handling a wide range of clonal structures—skewed abundances, dual-TCRα clones, sharing of both CDR3α and CDR3β sequences among clones—as well as providing estimates of clonal frequencies.

Single-cell technologies clearly allow the identification of highly-expanded clones within populations. However, most of the currently available technologies cannot sample enough single cells to adequately sample the rare clones of T cell populations. As we have demonstrated, our algorithm offers the potential to both identify these common clones as well as achieve high depths of rarer clones that far exceed those currently possible with reasonable levels of single-cell sequencing. Although TCRβsequencing have been used a proxy for measuring diversity in T cell populations, the presence of a significant number of CDR3α- and CDR3β-sharing clones would cause TCRβ-only sequencing to underestimate TCR diversity. Thus, identifying TCRαβ sequence pairs will be the gold-standard of identifying individual T cell clones and characterizing clonal diversity of T cell populations. In addition, establishing the levels of CDR3α- and CDR3β-sharing within populations details more information about mechanisms of antigen recognition, repertoire diversity, and the efficiency of recruitment into immune responses.

Our analysis shows that distinguishing between CDR3β-sharing and dual-TCRα clones is a difficult challenge in frequency-based pairing schemes. This is because the expected patterns of co-occurrences of the three chains under the two situations are very similar when sequencing samples of a few tens of cells per well since all three chains typically appear in nearly all the wells. The difference in co-occurrence patterns can be made clearer to an extent by sampling very few numbers of cells per well, but this approach

reduces the total sample size of cells sampled, which would sacrifice the depth of identifying rare clones. We might intuitively suppose that upper limit of 30% prevalence of dual-TCRα clones in the naive T cell pool favors dual-TCRα over CDR3β-sharing clones. However, this intuition may not be true in epitope-specific populations. The numbers of precursor naive T cell clones may range 10–1000 cells in mice [52, 141, 117], which we estimate is comparable to or larger than the size of CDR3β-sharing populations exported from the thymus (see Section 6.3.3). If the sharing of a CDR3β among clones confers overlap in their TCR specificities and recruitment into immune responses is efficient, we might expect to see significant levels of CDR3β-sharing within expanded epitope-specific populations. Indeed, as shown in Table 5.1, CDR3β-sharing has been seen to reach levels of up to 25% in responses to influenza epitopes in naive mice [65, 67] and almost 40% in secondary responses [67]. It also occurred at a level of 2% in our analysis of CDR3α and CDR3β usage among $CD8^+$ cells specific for a YFV epitope in a human volunteer. The CDR3βsharing/dual-TCRα ambiguity is therefore a robust feature of epitope-specific responses and is challenging to unravel fully with statistical approaches.

There are at least three ways to address this problem. One solution is to pair ALPHABETR with a set of single-cell samples (e.g. one plate of single cells and four plates of mixed sampling). Since the ambiguity is particularly unclear with abundant CDR3β-sharing and dual-TCRα clones, this limited amount of extra information may serve to resolve the issue. A second approach is to exploit the fact that 30–40% of clones will yield both an in-frame and an out-of-frame CDR3α sequence [65]. Currently, ALPHABETR does not utilize out-of-frame sequences and could be extended to associate clones with their out-of-frame sequences. Clones possessing one in-frame and one out-of-frame CDR3α could then be excluded from the list of dual-TCRα candidates, which would assist in CDR3β-sharing/dual-TCRα discrimination. A third possibility is to extend the algorithm to use the sequence information and make sequence consensus comparison to determine a likeliness of two CDR3α pairing with the same CDR3β. If dealing with epitope-specific populations, we might expect more sequence similarity in the CDR3α in two CDR3β-sharing clones than in a dual-TCRα clone, which have two unlikely related CDR3α sequences; the two TCRα chains rearrange independently and presumably only one is involved in antigen recognition.

Identifying dual-TCRβ clones with a frequency-based pairing approach remains an open problem. Although the frequency of dual-TCRβ clones in T cell populations has been not well characterized, two single-cell sequencing studies have clearly demonstrated that such clones have the potential of existing in T cell populations, albeit in very low frequencies. Since allelic exclusion of the TCRβ chains should be nearly complete, we would not expect dual-TCRβ clones to be a common occurrence and thus ALPHABETR not being able to identify dual-TCRβ clones is not a considerable limitation. And as we have shown in the simulations, the presence of dual-TCRβ clones does not appear significantly affect the accuracy and depth of pairing TCRs.

In practice, one needs a strategy for implementing ALPHABETR on a given sample of T cells with no *a priori* knowledge of the number or size distribution of clones. Assuming the number of available T cells is not limiting, we advocate a high-mixed sampling approach that involves sampling 20-300 cells per well and deals efficiently with a wide range of clonal abundances. With the `alphabetr` package implemented in R, a standard desktop computer with 16 Gb of RAM is able to handle samples from parent distributions of up to 4000 clones. When sampling populations with much fewer clones, lower numbers of cells/well are needed to avoid high false pairing rates. Assuming cell numbers are not limiting, bulk sequencing of the TCRβ chain could be used to gain a rough estimate of the richness of the parent distribution and so indicate when a sparse sampling strategy would be beneficial. In situations where cell numbers are limiting, one approach could be to begin with a single plate of 10 cells/well to obtain a rough lower bound on the richness of the distribution and apply a low or high mixed sampling strategy with the remaining cells from the sample, as appropriate. The single plate of 10 cells/well is then still usable for the pairing process and for frequency estimation.

Finally, we stress the importance of obtaining paired TCR sequences and identifying a representative sample of TCRs of the repertoire of interest. For example, studies that employed TCRβ sequencing have come to erroneous conclusions because of the lack of TCRα information. Cukalac *et al.* found that the sequences of clones responding to the influenza NP_{366} peptide from different mice have identical CDR3β sequences that pair with different CDR3α sequences [67]. If the study had been performed with TCRβ sequencing, the results would suggest that these clones are identical, lead-

ing to the false conclusion that there exists a dominant "public" TCR for D^bNP_{366}-specific cells. Until single-cell sequencing can sample many thousands of clones, ALPHABETR is an important way of obtaining a representative sample of paired TCR sequences of antigen-specific T cell populations.

It should be noted that the advantages of large sample sizes from frequency-based pairing approaches come at the cost of identifying false pairs. pairSEQ attempts to minimize false pairs by estimating a false determining rate while ALPHABETR attempts to constrain the rate of identifying false pairs with the threshold parameter. In research applications, the economical and biological advantage of identifying many thousands of TCR pairs outweighs this disadvantage. However, in clinical applications, false pairs may not be tolerable and perfectly correct pairs might be needed through single-cell sequencing. Frequency-based pairing and single-cell sequencing technologies can complement each other and should be used according to the research question and the application.

While we have framed most of our analysis around the sequencing of epitope-specific populations, ALPHABETR can equally well be applied more generally to T cell populations of restricted and potentially skewed polyclonality, such as tumor infiltrating lymphocytes or T cells extracted from sites of autoimmune responses. It therefore has immediate applications in cancer immunotherapy and other personalized immunomodulatory treatments. Until single cell sequencing becomes more affordable, frequency-based pairing methods provide a rapid and economical means of characterizing the clonal structure of T cell populations.

5.5 DETAILED DESCRIPTION OF THE ALGORITHM

We present a detailed description of the algorithm in Sections 5.5.1–5.5.4 and a description of the simulated immunodominance hierarchies in Section 5.5.5. The equations presented in this chapter are derived in Section 5.5.6.

5.5.1 *Determining αß pairs*

Our approach uses the fact that CDR3α and CDR3β sequences (referred to as α and β chains) from the same clone will tend to appear together in wells. Let N_α be the total number of unique α chains, N_β be the total number of unique β chains, and the α and β chains found in the data set be labeled from

α_1 to α_{N_α} and from β_1 to β_{N_β} respectively. Let W be the number of wells used to collect the sequencing data. The association score between chains α_i and β_j is measured by a score

$$S_{ij} = \sum_{k=1}^{W} \left(\frac{\delta_{ij}^k}{c_\alpha^k} + \frac{\delta_{ij}^k}{c_\beta^k} \right), \tag{5.1}$$

where the wells in the data are labelled from 1, 2, …, W, the numbers of distinct α and β chains in well k are c_α^k and c_β^k respectively, and δ_{ij}^k is 1 if both α_i and β_j are found in well k and 0 otherwise. Equation 5.1 adds the inverse of the total number of α chains found in the well to the inverse of the total number of β chains found in the well and then sums these for all wells where α_i and β_j co-appear. The scaling accounts for the fact that a larger number of unique chains recovered in a well lowers our confidence that a co-occurring α and β pair derive from the same clone.

The algorithm begins by sampling a proportion p_J of the wells in the data without replacement. For all analyses presented here, we used $p_J = 0.75$, which provided a good balance between depth and false pairing rate. The algorithm computes the association scores between every unique α and β chain using Equation 5.1 based on the sampled subset of wells. Let $\mathcal{A}_k$ denote the set of A distinct α chains found in well k, that is,

$$\mathcal{A}_k = \left\{ \alpha_{m_1^k}, \alpha_{m_2^k}, \ldots, \alpha_{m_A^k} \right\},$$

where the $m_i^k \in \{1, \ldots, N_\alpha\}$ are integers that denote the labels of the A TCRα chains found in well k. Similarly, let $\mathcal{B}_k$ denote the set of B distinct β chains found in well k, that is,

$$\mathcal{B}_k = \left\{ \beta_{n_1^k}, \beta_{n_2^k}, \ldots, \beta_{n_B^k} \right\},$$

where the $n_i^k \in \{1, \ldots, N_\beta\}$ subscripts denotes the labels of the B TCRβ chains found in well k. The algorithm solves the following linear assignment

problem using the Hungarian algorithm [136]:

$$\begin{aligned} \text{maximize} \quad & \sum_{\alpha_i \in \mathcal{A}_k} \sum_{\beta_j \in \mathcal{B}_k} S_{ij} x_{ij} \\ \text{subject to} \quad & \sum_{\alpha_i \in \mathcal{A}_k} x_{ij} = 1 \text{ for } \beta_j \in \mathcal{B}_k \\ & \sum_{\beta_j \in \mathcal{B}_k} x_{ij} = 1 \text{ for } \alpha_i \in \mathcal{A}_k \\ & x_{ij} \geq 0, \quad \alpha_i \in \mathcal{B}_k,\ \beta_j \in \mathcal{A}_k, \end{aligned} \tag{5.2}$$

where $x_{ij} = 1$ indicates that α_i and β_j are assigned as a candidate TCR pair and $x_{ij} = 0$ otherwise.

A pair α_i and β_j is defined as an assigned pair of well k if $x_{ij} = 1$ for the solution of Equation 5.2 associated with well k. The number of assignments made for every pair of α and β is recorded as X_{ij}, i.e. X_{ij} equals the number of times $x_{ij} = 1$ from the solutions of Equation 5.2 for each well in the subset.

We then calculate a filter level F that determines the minimum number of assignments required for an assigned candidate pair of α and β chains to be determined as a true TCR pair. This filter level is calculated as the mean of the number of associations of every αβ pair that had at least one association, namely F is the mean of the set

$$\left\{ N(i,j) \ : \ N(i,j) > 0, \ \ i \in 1, 2, \ldots, n_\alpha \ , \ j \in 1, 2, \ldots, n_\beta \right\},$$

where $N(i,j)$ is the number of times $\alpha_i \beta_j$ are assigned to each other and n_α and n_β are the number of unique α and β sequences found across the wells respectively. We then choose the pairs that had more associations than F, i.e. pairs $\alpha_i \beta_j$ such that $X_{ij} > F$. The output of this algorithm is then a list of candidate αβ pairs that may be associated with a T cell clone. At this stage, dual-TCRα cells are not identified; thus a dual-TCRα clone $\alpha_i \alpha_j \beta$ may be represented in this list as one or both of $\alpha_i \beta$ and $\alpha_j \beta$.

The procedure above is performed N_r times on random subsets of the wells (all simulations in this paper use $N_r = 100$), and each replicate yields a list of candidate αβ pairs. We then perform a filtering or consensus step in which only αβ pairings that appear in more than a threshold proportion T of these lists are retained as candidates. This threshold T is chosen by the

Number of top clones	Overall Depth	Top Depth	Tail Depth	False Pairing
5	25.3%	68.1%	25.2%	42.1%
10	25.5%	84.0%	25.3%	42.0%
25	17.6%	87.7%	16.9%	26.0%
50	18.2%	88.5%	88.5%	27.4%

Table 5.6: Simulations without the resampling procedure. Simulations were performed by pairing CDR3α and CDR3β sequences using the information from all wells at the same time instead of pairing using multiple resampled subsets of the wells as described in Section 5.5.1. These simulations show the unsatisfactory results from not using the resampling approach, particularly in the high false pairing rates and low tail depths. 100 simulations were performed for each level of skew in the immunodominance hierarchies.

user of the algorithm. The simulations presented before explore thresholds of $T = 0.3$, $T = 0.6$, and $T = 0.9$.

5.5.2 *Justification of the resampling procedure*

The resampling procedure is performed in order to reduce the false pairing rate and increase the depth of the rare clones. Applying the pairing procedure to all of the wells all at once—namely, pair CDR3α and CDR3β sequences using all of the wells—results in very high false pairing rates with low depth. In order to demonstrate this, we performed simulations (Table 5.6) to show how not resampling hurts the depth and false pairing rates.

Two intuitive reasons explain why resampling improves pairing performance. First, false pairing rates dramatically fall because resampling allows us to depend on consensus across many pairing attempts to find true TCR pairs. Since we are sampling many different subsets of the data, true CDR3α/CDR3β pairs will be paired together many more times than than spurious pairs. Second, resampling allows for the discovery of rare clones by preventing common clones from masking the pairing relationships of the CDR3α and CDR3β sequences of rare clones. Common clones will tend to appear in many wells and thus appear together with rare clones; this can mask the presence of any rare clones. By resampling different subsets of the wells, we allow for the possibility of rare clones to be represented by finding configurations of wells that do not mask the rare clones. Resampling is a standard technique used in many statistical learning algorithms, and we found that it is the crux of the success of ALPHABETR.

5.5.3 *Maximum-likelihood estimation of clonal frequencies*

We use maximum likelihood estimation to infer clonal frequencies based on the number of wells in which a pair of α and β chains both appear. We form a likelihood that models the process of sampling cells from a parent population and the error processes that occur during sequencing.

Let $N = \{n_1, n_2, \ldots, n_s\}$ be the set of s distinct sample sizes (n_i cells per well) in all of the wells and $W = \{w_1, w_2, \ldots, w_s\}$ be a set where w_i represents the number of wells with samples of size n_i cells. Let c_{ij} denote the clone with chains α_i and β_j, and let k_{ij}^l denote the number of wells of sample size n_l cells per well that contain chains α_i and β_j. The observed incidence of clone c_{ij} in the data is then the set $K_{ij} = \{k_{ij}^1, k_{ij}^2, \ldots, k_{ij}^s\}$. The likelihood of the observing the data K_{ij} for clone c_{ij} with a frequency f_{ij} within the population is given by

$$\mathcal{L}(K_{ij} \mid f_{ij}) = \prod_{l=1}^{s} \binom{w_l}{k_{ij}^l} (1 - q_l)^{k_{ij}^l} q_l^{w_l - k_{ij}^l} \tag{5.3}$$

where q_l is the probability of clone c_{ij} not being found in well l and is

$$q_l = (1 - f_{ij})^{n_l} + \sum_{m=1}^{n_l} \left(2\varepsilon^m - \varepsilon^{2m}\right) \binom{n_l}{m} f_{ij}^m (1 - f_{ij})^{n_l - m}. \tag{5.4}$$

Here ε is the average probability that a CDR3 sequence in a cell fails to be amplified and sequenced. For every clone c_{ij}, the algorithm finds the maximum-likelihood estimate $\hat{f}_{ij}$ of Equation 5.3 to estimate its frequency, and 95% confidence intervals are defined by solving for the frequencies $\tilde{f}_{ij}$ that satisfy $\text{loglik}(\tilde{f}_{ij}) = \text{loglik}(\hat{f}_{ij}) - 1.96$. Details of the derivation of Equations 5.3 and 5.4 are given in Section 5.5.6.

This procedure is applied to every αβ pair identified in the first phase of the algorithm. These estimated frequencies are then used to distinguish TCRβ-sharing clone pairs from single TCR clones expressing two TCRα, described in Section 5.5.4.

When a dual-TCRα clone is identified, we revise the frequency estimation as follows. Let $c_{(ij)t}$ denote a clone with chains α_i, α_j, and β_t, and $k^l_{(ij)t}$ denote the number of wells of size n_l that contain chains α_i, α_j, and β_t. The likelihood of the observations given that clone $c_{(ij)t}$ has a frequency $f_{(ij)t} \in (0, 1]$ is

$$\mathcal{L}(\text{observed incidence of clone } c_{(ij)t} \mid f_{(ij)t}) = \prod_{l=1}^{s} \binom{w_l}{k^l_{(ij)t}} (1 - q_l)^{k^l_{(ij)t}} q_l^{w_l - k^l_{(ij)t}} \tag{5.5}$$

where q_l is the probability of clone $c_{(ij)t}$ not being found in well l and is given by

$$q_l = (1 - f_{(ij)t})^{n_l} + \sum_{m=1}^{n_l} \left(3\varepsilon^m - 3\varepsilon^{2m} + \varepsilon^{3m}\right) \binom{n_l}{m} f^m_{(ij)t} (1 - f_{(ij)t})^{n_l - m} \tag{5.6}$$

where ε is the mean drop rate as described above. For every clone $c_{(ij)t}$, the algorithm finds the maximum-likelihood estimate $\hat{f}_{(ij)k}$ for Equation 5.5, and again $\text{loglik}(\tilde{f}_{(ij)t}) = \text{loglik}(\hat{f}_{(ij)t}) - 1.96$ is used to calculate 95% confidence intervals.

5.5.4 *Discriminating between dual-TCRa and shared TCRa chains*

If the algorithm pairs two clones that appear to share a CDR3β (e.g. $\alpha_i\beta$ and $\alpha_j\beta$), we must decide whether this is indeed a CDR3β-sharing pair of clones or that the association derives from one dual-TCRα clone (e.g. $\alpha_i\alpha_j\beta$). To do this, we use the likelihoods of observed co-occurrences of the three chains to assess the relative support for the two alternatives.

Suppose $c_{ij} = (\alpha_i, \beta_j)$ and $c_{kj} = (\alpha_k, \beta_j)$ are two putative clones with a common β chain β_j. We count the number of wells containing all three-way, two-way, and single appearances of the three chains. We then calculate the "full" likelihoods of this pattern of occurrences under two hypotheses:

H_X: c_{ij} and c_{kj} are indeed two β-sharing clones, with frequencies f_{ij} and f_{kj} estimated using Equation 5.3

H_Y: The chains derive from one dual-TCRα clone $c_{(ij)k}$ present at frequency $f_{(ij)k}$, estimated using Equation 5.5.

If the difference between the log-likelihoods under these two hypothesis is greater than or equal to 10, i.e. $\text{loglik}(\text{data}|H_Y) - \text{loglik}(\text{data}|H_X) \geq 10$, we assume the three chains derive from a dual-TCRα clone. Thus, if chains α_i, α_k, and β_j are identified to derive from a dual-TCRα clone, then clones c_{ij} and c_{kj} are removed from the list of TCR pairs and replaced with a dual-TCRα clone $\alpha_i\alpha_k\beta_j$.

These full likelihoods (derived in Section 5.5.6) are computationally intractable for wells with greater than 50 cells due to the need to calculate large multinomial coefficients. The full-likelihood method is therefore only appropriate for estimating frequencies of relatively abundant clones since they are more likely than rare clones to be found in the wells with smaller sample sizes.

As an alternative, we use a more restricted likelihood-based approach for discriminating CDR3β-sharing and dual-TCRα among rare clones, which tend to appear only in wells of larger sample sizes. Let clones $c_{ij} = (\alpha_i, \beta_j)$ and $c_{kj} = (\alpha_k, \beta_j)$ be two clones with a common beta chain β_j, and let f_{ij} and f_{jk} be their respective estimated frequencies. The algorithm calculates the ratio r^j_{ik} of the observed to the expected number of wells in which all three chains from the putative β-sharing pair c_{ij} and c_{jk} co-appear, under the hypothesis that they are indeed two clones and not a dual-TCRα clone.

This results in a set R of ratios

$$R = \left\{ r^j_{ik} = \frac{A(c_{ij}, c_{kj})}{E(c_{ij}, c_{kj})} : i \neq k,\ i,j \in 1,2,\ldots,N_\beta \right\} \tag{5.7}$$

where $A(c_{ij}, c_{kj})$ is the number of times clones c_{ij} and c_{kj} are observed to appear in the same well and N_β is the number of distinct β chains, and the expected number is

$$E(c_{ij}, c_{kj}) = \sum_{l=1}^{s} w_l \left(1 - (1 - f_{ij})^{n_l} - (1 - f_{kj})^{n_l} + (1 - f_{ij} - f_{kj})^{n_l}\right). \tag{5.8}$$

Equations 5.7-5.8 are derived in Section 5.5.6. We then partition the set of ratios R into two groups C_1 and C_2 using k-means clustering, where the mean of ratios of C_1 is greater than the mean of the ratios of C_2. The clones associated with the ratios in C_1 are chosen as dual TCR clones, such that if $r^j_{ik} \in C_1$, then clones c_{ij} and c_{kj} are removed from the list of TCR pairs and replaced with a dual-TCRα clone $\alpha_i\alpha_k\beta_j$.

The frequencies of the dual-TCRα clones identified by these procedures are then estimated using Equation 5.5, and the final output of the algorithm is a list of single and dual-TCRα clones and their clonal frequencies.

5.5.5 *Creation of in silico data sets for validation*

We created synthetic data sets reflecting the properties of antigen-specific T cell populations and sequencing errors. The data sets were sampled from a population of T cell clones where a significant proportion of α and β chains are shared and 10%-30% of clones have dual-TCRα chains (e.g. three clones can have the following chains: $\alpha_i\beta_k$, $\alpha_j\beta_k$, and $\alpha_j\alpha_h\beta_l$). The sharing of CDR3α and CDR3β sequences were set at levels shown in Table 5.1. We determined these levels of sharing by averaging those from the published single-cell data shown in Table 5.1.

The frequencies of the N clones were drawn from a skewed distribution in which n_s clones comprise a proportion p_s of the population and the other $N - N_s$ clones evenly represent $1 - p_s$ of the population. The clone ranked i^{th} in abundance then has frequency f_i where

$$f_i = \begin{cases} f_1 + r\,(i-1) & \text{if } i = 1, 2, \ldots, n_s \\ p_s/(N - n_s) & \text{if } i = n_s + 1, n_s + 2, \ldots, N \end{cases} \tag{5.9}$$

where the frequency of the largest clone f_1 and the step size r are determined by solving the equations

$$\sum_{i=1}^{n_s} f_i = p_s, \qquad f_{n_s} = 1.1 \times \frac{p_s}{N - n_s}. \tag{5.10}$$

The frequency of the smallest clone in the top 50%, f_{n_s}, is set to be 10% higher than the frequency of the clones in the tail. All simulations were based on $p_s = 0.5$. We varied the number of top clones n_s between 5 to 50 to test how skew in the antigen-specific T cell population affects the performance of the algorithm.

In order to make the simulated data more realistic, experimental noise was included in the forms of 'dropped' chain errors and in-frame sequencing errors. Dropped chains are CDR3 sequences that fail to be sequenced due to PCR errors and/or sorting problems, and studies utilizing both single-cell and many-cell techniques have reported average drop rates of 8% to 10% [69, 1]. In the simulations, each clone was assigned a drop rate from a log-normal distribution with a mean of 0.15 and standard deviation of 0.01, and every TCRα and TCRβ chain belonging to that clone was assigned that drop rate. In-frame errors cause a CDR3 sequence to be falsely identified with an incorrect productive nucleotide and/or amino acid sequence. In the simulations, each distinct sequence was assigned an in-frame error rate drawn from a lognormal distribution with a mean of 0.02 and a standard deviation of 0.005. The error model was simulated as follows: when a cell is sampled into a virtual well, each of its chains fails to be sequenced with probability equal to the pre-assigned, clone-specific drop rate. Every surviving chain produces one of three randomly chosen, distinct, and chain-specific false sequences with probability equal to that chain's pre-assigned in-frame error rate.

5.5.6 *Derivations*

Derivation of the frequency estimation likelihood

The maximum likelihood approach for clonal frequency estimation involves modeling how a clone is sampled in the wells of the plates. Since we assume conservatively that sequencing does not give any quantitative information about the number of times a clone is sampled in a well, the data fed to ALPHABETR indicate only whether a given CDR3α or CDR3β sequence is present in each well. Let s denote the number of distinct sample sizes placed in the wells, $\mathcal{N} = \{n_1, n_2, \ldots, n_s\}$ be the set of s distinct sample sizes where n_i is number of cells per well, and $\mathcal{W} = \{w_1, w_2, \ldots, w_s\}$ be the set where w_i represents the number of wells with sample size n_i. Let c_{ij} denote the clone with chains α_i and β_j, and let k_{ij}^l denote the number of wells of size n_l cells per well that contain α_i and β_j.

The likelihood of clone $\alpha_i\beta_j$ appearing in k_{ij}^l wells of sample size n_l cells for $l = 1, \ldots, s$ is the probability of the clone being sampled in k_{ij}^l out of the w_l possible wells, which is

$$\begin{aligned} P\big(k_{ij}^l \text{ wells} \mid \text{ frequency } f_{ij}\big) = \\ \binom{w_l}{k_{ij}^l} P\big(\text{clone sampled in well of size } n_l\big)^{k_{ij}^l} \times \\ \Big(1 - P\big(\text{clone sampled in well of size } n_l\big)\Big)^{w_l - k_{ij}^l} \end{aligned} \tag{5.11}$$

We define $q_l = 1 - P(\text{clone sampled in well of size } n_l)$, which the probability of clone $\alpha_i\beta$ not being sampled in a well of size n_l. We then rewrite Equation 5.11 as

$$P\big(k_{ij}^l \text{ wells} \mid \text{frequency } f_{ij}\big) = \binom{w_l}{k_{ij}^l} (1 - q_l)^{k_{ij}^l} q_l^{\,w_l - k_{ij}^l} \tag{5.12}$$

Since the k_{ij}^l appearances are independent, the probability of observing is determined by summing Eq (5.11) for all sample sizes and is given by

$$P(k_{ij}^1, k_{ij}^2, \ldots, k_{ij}^s \text{ wells} \mid \text{frequency } f_{ij}) = \prod_{l=1}^{s} \binom{w_l}{k_{ij}^l} (1 - q_l)^{k_{ij}^l} q_l^{\,w_l - k_{ij}^l}, \tag{5.13}$$

which is Equation 5.3.

The probability q_l is calculated by adding the probabilities of all of the events that would result in the clone not being sampled in the well, which are:

- A: clone $\alpha_i\beta_j$ not being sampled at all
- B: clone $\alpha_i\beta_j$ is sampled $m \leq n_l$ times (resulting in m copies of chains α_i and β_j), all m copies of α_i are dropped, and at least 1 copy of β_j is not dropped
- C: clone $\alpha_i\beta_j$ is sampled $m \leq n_l$ times, at least 1 copy of α_i chain is not dropped, and all m copies of β_j are dropped
- D: clone $\alpha_i\beta_j$ is sampled $m \leq n_l$ times, and all m copies of α_i and m copies of β_j are dropped.

The probability of event A is given by

$$P(A) = (1 - f_{ij})^{n_l} \tag{5.14}$$

Events B and C are symmetric, and the probability of these events is the probability that the clone will be sampled $m \leq n_l$ times multiplied by the probability of dropping all of one of the chains and not dropping at least one of the other chains. This is given by

$$P(B) = P(C) = \sum_{m=1}^{n_l} \binom{n_l}{m} (f_{ij})^m (1 - f_{ij})^{n_l - m} \varepsilon^m (1 - \varepsilon^m) \tag{5.15}$$

where ε^m is the probability of dropping m of one of the component chains. The probability of event D is derived similarly:

$$P(D) = \sum_{m=1}^{n_l} \binom{n_l}{m} (f_{ij})^m (1 - f_{ij})^{n_l - m} \varepsilon^m \varepsilon^m. \tag{5.16}$$

Summing these yields Equation 5.4.

The likelihood for dual clones is obtained in a similar fashion, where q_l is calculated by summing the probabilities of the clone not being sampled at all and of being sampled but dropping one, two, or all three of the clone's chains.

Derivation of k-means approach

The ratios in calculated in Equation 5.7 involve the expected number of wells in which the three chains $\alpha_i\alpha_j\beta$ co-occur under the assumption that they derive from two CDR3β-sharing clones $\alpha_i\beta$ and $\alpha_j\beta$. Let c_{ij} and c_{kj} be two clones that share β_j. Let A_{ij}^l and A_{kj}^l denote the events of sampling clones c_{ij} and c_{kj} in a well of size n_l cells respectively and $A_{ij}^{l^C}$ and $A_{kj}^{l^C}$ denote the complement of these events. The probability of sampling both clones in a well of n_l cells is then

$$\begin{aligned} P(A_{ij}^l \cap A_{kj}^l) &= 1 - P(A_{ij}^{l^C} \cup A_{kj}^{l^C}) \\ &= 1 - \left(P(A_{ij}^{l^C}) + P(A_{kj}^{l^C}) - P(A_{ij}^{l^C} \cap A_{kj}^{l^C})\right) \end{aligned} \tag{5.17}$$

In calculating Eq (5.17), including the effect of stochastic dropping of chains results in large multinomial coefficients that cannot be computed efficiently for wells with larger sample sizes (approximately ≥ 50 cells per well). Heuristically, however, neglecting the drop rate has no impact on discrimination of CDR3β-sharing and dual-TCRα clones. We then have

$$P(A_{ij}^{l^C}) = (1 - f_{ij})^{n_l} \tag{5.18}$$

$$P(A_{ik}^{l^C}) = (1 - f_{kj})^{n_l} \tag{5.19}$$

$$P\left(A_{ij}^{l^C} \cap A_{kj}^{l^C}\right) = \left(1 - (f_{ij} + f_{kj})\right)^{n_l} \tag{5.20}$$

By substituting Eqs (5.18)-(5.20) into Eq (5.17) and multiplying by w_l (the total number of wells of size n_i), we obtain the expected number of wells of size n_i that contain both clones:

$$E(c_{ij}, c_{kj}) = w_l\Big(1 - (1 - f_{ij})^{n_l} - \big(1 - f_{kj}\big)^{n_l} + \big(1 - f_{ij} - f_{kj}\big)^{n_l}\Big). \tag{5.21}$$

By summing this quantity over all wells and sample sizes, we obtain Equation 5.8, which forms the denominator of the ratios in R (Equation 5.7). With inclusion of the drop rate, this ratio should be close to 1 for a true β-sharing pair; neglecting dropping shifts this ratio to higher values, as seen in the left-hand cluster in Figure 5.6, but discrimination of CDR3β-sharing and dual-TCRα is still possible. The computational limitation on the calculation of multinomial coefficients does not exist for wells with smaller sample sizes, and so likelihoods can be directly calculated for clones that appear in these smaller wells, which explore in the next section.

Derivations of the full likelihoods

Discriminating between relatively abundant CDR3β-sharing clones and dual-TCRα clones requires comparing the likelihoods of the data under these two hypotheses. Let $\alpha_q\beta$ and $\alpha_r\beta$ be a pair of candidate TCRs that share the same βchain with frequencies f_q and f_r respectively. Given the data have wells of s distinct sample sizes, we record the numbers of wells of each sample size that contain all three chains (α_q, α_r, β) or contain only two of the three.

For $i \in \{1, 2, \ldots, s\}$, let

- k_i^1 = the number of wells of sample size n_i containing chains β and α_q only
- k_i^2 = be the number of wells of sample size n_i containing chains β and α_r only
- k_i^3 = be the number of wells of sample size n_i containing chains α_q and α_r only
- k_i^d = be the number of wells of sample size n_i containing all three chains β, α_q, and α_r
- $k_i^o = w_i - k_i^1 - k_i^2 - k_i^3 - k_i^d$ be the number of wells of sample size n_i that contain none of the chains or only one of the three chains.

For chains a, b, and c, let W_{abc}^i, W_{ab}^i, W_a^i, and $W_\emptyset^i$ denote the events of finding exactly chains a, b, c, finding exactly chains a and b, finding exactly chain a, and finding none of the chains in a well of sample size n_i respectively. As before, let w_i denote the number of wells with sample size n_i cells per well, and let s be the number of distinct sample sizes.

The likelihood of observing the data

$$\mathcal{K} = \Big\{k_i^1, k_i^2, k_i^3, k_i^d, k_i^o : k = 1, 2, \ldots, s\Big\}$$

under the hypothesis that $\alpha_q\beta$ and $\alpha_r\beta$ represent two β-sharing clones is

$$\begin{aligned}&\mathcal{L}\big(\mathcal{K}|\text{clone } \alpha_q\beta \text{ with frequency } f_q, \text{ clone } \alpha_r\beta \text{ with frequency } f_r\big) = \\ &\prod_{i=1}^{s} \frac{w_i!}{k_i^1!k_i^2!k_i^3!k_i^d!k_i^o!} P(W_{\alpha_q\beta}^i)^{k_i^1} P(W_{\alpha_r\beta}^i)^{k_i^2} P(W_{\alpha_q\alpha_r}^i)^{k_i^3} P(W_{\alpha_q\alpha_r\beta}^i)^{k_i^d} \times \\ &P(W_{\alpha_q}^i \cup W_{\alpha_r}^i \cup W_\beta^i \cup W_\emptyset^i)^{k_i^o}\end{aligned} \tag{5.22}$$

where

$$P(W^i_{\alpha_q\beta}) = \sum_{k=1}^{n_i} \binom{n_i}{k} f_q^k \left(1-f_q-f_r\right)^{n_i-k} \left(1-\varepsilon^k\right)^2 +$$
$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right)^2 \varepsilon^{n_2} +$$
$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right) \varepsilon^{n_1} \left(1-\varepsilon^{n_2}\right) \varepsilon^{n_2}$$

$$P(W^i_{\alpha_r\beta}) = \sum_{k=1}^{n_i} \binom{n_i}{k} f_r^k \left(1-f_q-f_r\right)^{n_i-k} \left(1-\varepsilon^k\right)^2 +$$
$$\sum_{n_2=1}^{n_i-1}\sum_{n_1=1}^{n_i-n_2} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_r^{n_1} f_q^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_2}\right)^2 \varepsilon^{n_1} +$$
$$\sum_{n_2=1}^{n_i-1}\sum_{n_1=1}^{n_i-n_2} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_r^{n_1} f_q^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right) \varepsilon^{n_1} \left(1-\varepsilon^{n_2}\right) \varepsilon^{n_2}$$

$$P(W^i_{\alpha_q\alpha_r}) = \sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \varepsilon^{n_1} \left(1-\varepsilon^{n_1}\right) \varepsilon^{n_2} \left(1-\varepsilon^{n_2}\right)$$

$$P(W^i_{\alpha_q\alpha_r\beta}) = \sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \varepsilon^{n_1} \left(1-\varepsilon^{n_1}\right) \left(1-\varepsilon^{n_2}\right)^2 +$$
$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_1-f_2\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right)^2 \left(1-\varepsilon^{n_2}\right)^2 +$$
$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\left(n_i-n_1-n_2\right)!} f_q^{n_1} f_r^{n_2} \left(1-f_q-f_r\right)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right)^2 \varepsilon^{n_2} \left(1-\varepsilon^{n_2}\right)$$

$$P(W^i_{\alpha_q} \cup W^i_{\alpha_r} \cup W^i_{\beta} \cup W^i_{\emptyset}) = 1 - P(W^i_{\alpha_q\beta}) - P(W^i_{\alpha_r\beta}) - P(W^i_{\alpha_q\alpha_r}) - P(W^i_{\alpha_q\alpha_r\beta})$$

The likelihood of observing the data $\mathcal{K} = \left\{k_i^1, k_i^2, k_i^3, k_i^d, k_i^o : k = 1, 2, \ldots, s\right\}$ under the hypothesis that $\alpha_q\beta$ and $\alpha_r\beta$ represent one dual TCRα clone $\alpha_q\alpha_r\beta$ is

$$\mathcal{L}(\mathcal{K}|\text{clone } \alpha_q\alpha_r\beta \text{ with freq } f_d) =$$
$$\prod_{i=1}^{s} \frac{w_i!}{k_i^1!k_i^2!k_i^3!k_i^d!k_i^o!} P(W^i_{\alpha_q\beta})^{k_i^1} P(W^i_{\alpha_r\beta})^{k_i^2} P(W^i_{\alpha_q\alpha_r})^{k_i^3} P(W^i_{\alpha_q\alpha_r\beta})^{k_i^d}$$
$$P(W^i_{\alpha_q} \cup W^i_{\alpha_r} \cup W^i_{\beta} \cup W^i_{\emptyset})^{k_i^o}$$

where the terms are given by

$$P(W^i_{\alpha_q\beta}) = P(W^i_{\alpha_r\beta}) = P(W^i_{\alpha_q\alpha_r}) = \sum_{k=1}^{n_i} \binom{n_i}{k} f_d^k (1-f_d)^{n_i-k}\, \varepsilon^k \left(1-\varepsilon^k\right)^2$$

$$P(W^i_{\alpha_q\alpha_r\beta})^{k_i^d} = \sum_{k=1}^{n_i} \binom{n_i}{k} f_d^k (1-f_d)^{n_i-k} \left(1-\varepsilon^k\right)^3$$

$$P(W^i_{\alpha_q} \cup W^i_{\alpha_r} \cup W^i_{\beta} \cup W^i_{\emptyset}) = 1 - P(W^i_{\alpha_q\beta}) - P(W^i_{\alpha_r\beta}) - P(W^i_{\alpha_q\alpha_r}) - P(W^i_{\alpha_q\alpha_r\beta})\,.$$

The derivation of $P(W^i_{\alpha_q\beta})$ term for the two β-sharing clone hypothesis is shown below, and the other terms can be derived in a similar fashion. We begin by writing down the events that would result in a well containing the chains α_q and β exactly (illustrated in Figure 5.14):

- A_i: clone $\alpha_q\beta$ is sampled in the well w_i and at least 1 α_q and β not dropped from clone $\alpha_q\beta$, clone $\alpha_r\beta$ not sampled
- B_i: clone $\alpha_q\beta$ is sampled in the well w_i and at least 1 α_q and β not dropped from clone $\alpha_q\beta$, clone $\alpha_r\beta$ is sampled and all α_r are dropped and at least 1 β not dropped from clone $\alpha_r\beta$
- C_i: clone $\alpha_q\beta$ is sampled in the well w_i and at least 1 α_q not dropped and all β dropped from clone $\alpha_q\beta$, clone $\alpha_r\beta$ is sampled and all α_r are dropped and at least one β not dropped

For event A_i, we first calculate the probability of the two independent events of (*i*) sampling clone $\alpha_q\beta$ without sampling clone $\alpha_r\beta$ and (*ii*) not dropping all of the α_q and β chains of the sampled $\alpha_q\beta$ clone. For the former, we calculate the probability of sampling clone $\alpha_q\beta_i$ from 1 to n_i times while not sampling $\alpha_r\beta_i$. Each of n_i cells in the well has a probability of f_q of being clone $\alpha_q\beta_i$ and a probability $1-f_q-f_r$ of not being clone $\alpha_q\beta_i$ or $\alpha_r\beta_i$, which follows a binomial distribution. For the latter, each chain has a probability ε of being dropped, so if i cells are sampled as clone $\alpha_q\beta$, then the probability of dropping all α_q chains from those i cells is $1-\varepsilon^i$ (and similarly $1-\varepsilon^i$ for the β chains. Combining these, we get

$$\sum_{k=1}^{n_i} \binom{n_i}{k} f_q^k (1-f_q-f_r)^{n_i-k} \left(1-\varepsilon^k\right)^2$$

We then calculate the probability of (*i*) sampling clone $\alpha_q\beta$ n_1 times, sampling clone $\alpha_r\beta$ n_2 times, and sampling any other clone $n_i-n_1-n_2$ times, (*ii*) not dropping all of the α_q and β chains of the sampled $\alpha_q\beta$ cells, and (*iii*) dropping all of the α_r and β chains of the sampled $\alpha_r\beta$ cells. This is a multinomial distribution of the three sampling events multiplied by the probability of not dropping all of the chains of clone $\alpha_q\beta$ while dropping all of the chains of $\alpha_r\beta$, which is

$$\sum_{n_1=1}^{n_i-1} \sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1!n_2!\,(n_i-n_1-n_2)!} f_q^{n_1} f_r^{n_2} (1-f_q-f_r)^{n_i-n_1-n_2} \left(1-\varepsilon^{n_1}\right)^2 \varepsilon^{2n_2}$$

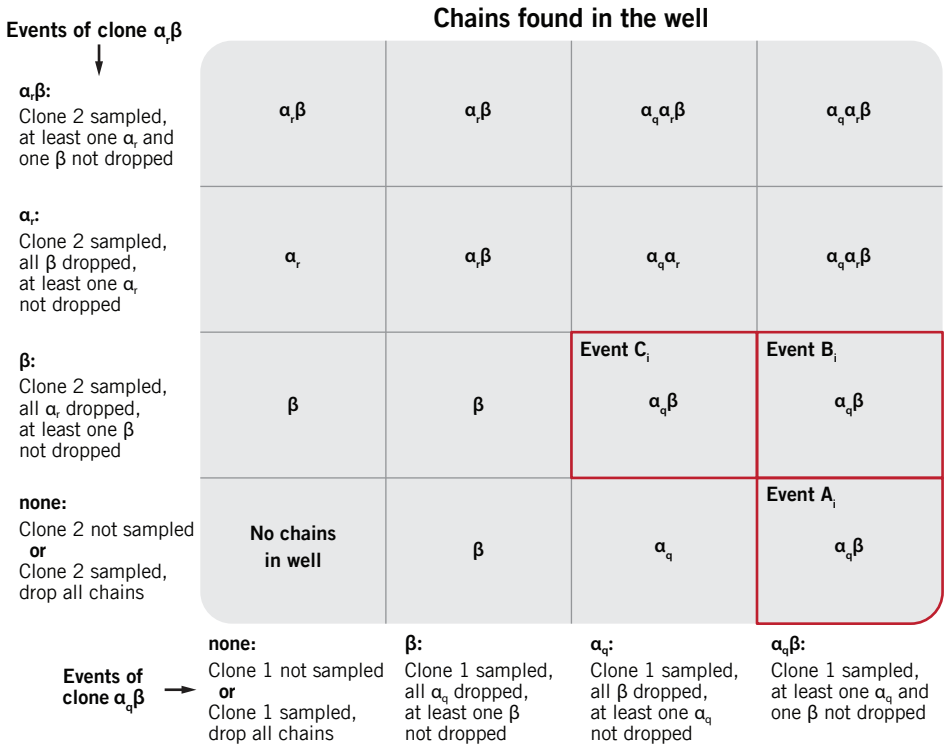


Figure 5.14: Sample space for calculating likelihoods of two- and three-way co-occurrences of chains under the hypotheses of CDR3β-sharing. The events labeled in red represent all of the possible ways a well could contain the chains α_q and β from two β-sharing clones $\alpha_q\beta$ and $\alpha_r\beta$.

Adding these two together, we get

$$P(W^i_{\alpha_q\beta}) = \sum_{k=1}^{n_i} \binom{n_i}{k} f_q^k (1 - f_q - f_r)^{n_i - k} \left(1 - \varepsilon^k\right)^2 + \sum_{n_1=1}^{n_i-1} \sum_{n_2=1}^{n_i-n_1} \frac{n_i!}{n_1! n_2! (n_i - n_1 - n_2)!} f_q^{n_1} f_r^{n_2} (1 - f_q - f_r)^{n_i - n_1 - n_2} \left(1 - \varepsilon^{n_1}\right)^2 \varepsilon^{2n_2}$$

For event B, we calculate the probability of n_1 cells being sampled as $\alpha_q\beta$, n_2 cells being sampled as $\alpha_r\beta$, and $n_i - n_1 - n_2$ cells being sampled as neither of the two clones, where $n_1 \geq 1, n_2 \geq 1, n_1 + n_2 \leq n$. This looks like a multinomial distribution of three events with probabilities f_q, f_r, and $1 - f_q - f_r$ occurring n_1, n_2, and $n_i - n_1 - n_2$ times. This is multiplied by the probability of not dropping all of the chains from the n_1 cells of clone $\alpha_q\beta$, not dropping all of the β chains from the n_2 cells of clone $\alpha_r\beta$, and dropping all α_q chains from the n_2 cells

of clone $\alpha_r\beta$. Then

$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1}\frac{n_i!}{n_1!n_2!\,(n_i-n_1-n_2)!}f_q^{n_1}f_r^{n_2}\,(1-f_q-f_r)^{n_i-n_1-n_2}\,(1-\varepsilon^{n_1})^2\,\varepsilon^{n_2}\,(1-\varepsilon^{n_2}) \quad (5.23)$$

For event C, we calculate a similar multinomial probability as above and multiply it by the probability of dropping all β chains from the n_1 cells of clone $\alpha_q\beta$, not dropping all α_q chains from the n_1 cells of clone $\alpha_q\beta$, dropping all of the α_r chains from n_2 cells of clone $\alpha_r\beta$, and not dropping all of the β chains from the n_2 cells of clone $\alpha_r\beta$. From this,

$$\sum_{n_1=1}^{n_i-1}\sum_{n_2=1}^{n_i-n_1}\frac{n_i!}{n_1!n_2!\,(n_i-n_1-n_2)!}f_q^{n_1}f_r^{n_2}\,(1-f_q-f_r)^{n_i-n_1-n_2}\,(1-\varepsilon^{n_1})\,\varepsilon^{n_1}\,(1-\varepsilon^{n_2})\,\varepsilon^{n_2} \quad (5.24)$$

Since $P(W^i_{\alpha_q\beta}) = P(A_i) + P(B_i) + P(C_i)$, we add all three expressions to obtain the term as stated above.

We assessed empirically that if the difference between the logarithms of the likelihoods in Eq (5.23) and Eq (5.22) is greater than or equal to 10, then chains α_q, α_r, and β should be assumed to comprise a dual-TCRα clone. As noted before, the multinomial coefficients contained in these equations are computationally limiting for wells of large sample sizes, and so calculations of these likelihoods include only wells of sample sizes less than 50 cells per well. Since these wells are most likely to contain the common clones, this approach is applicable to distinguishing common CDR3β-sharing and dual-TCRα clones.

CHAPTER 6

Identifying paired TCR sequences with the stable matching problem

6.1 Introduction to matching problems

Before ALPHABETR was developed, we imagined a process that would match TCRα and TCRβ chains together to find a set of "optimal" partnerships for all chains that would result in the identification of correct TCR pairs. This idea appeared analogous to a class of algorithms used to solve problems called matching under preferences [142, 143]. The agents involved in these matching processes have preferences for each other, and the algorithms assign a globally optimal set of partnerships for all agents.

These algorithms have been applied to many important problems, such as matching kidney patients to donors and assigning students to universities [142]. A well-known application of these algorithms is used in the National Resident Matching Program (NRMP), a process that pairs graduating medical students to residency training programs in the United States.[1] The introduction of the NRMP transformed a chaotic and competitive process into a formal, methodical procedure for finding a set of matchings that is 'optimal' for both the trainees and the hospitals. Before the NRMP, hospitals would compete with each other for desirable candidates by offering positions as early as possible while applicants would delay accepting positions as late as possible in hopes of getting better offers. This resulted in mayhem that included contracts being made up to two years before the start of the training programs and offers that would expire as soon as 12 hours to coerce trainees to accept. In the NRMP, medical students rank their preferred training programs, the hospitals rank their preference of students, and the NRMP algorithm determines an optimal matching that is given to all participants at the same time.

The classic version of these problems is known as the *stable marriage problem*, which attempts to find one-to-one matches between two equally-sized groups that is optimal for all participants. In its original formulation,

[1] Recently graduated medical students who are working in training programs (called residencies) are called *residents*, a term signifying trainees who are approximately equivalent to junior doctors in the United Kingdom who are undertaking their first few years of training.

the stable marriage problem matches a group of men to a group of women, where each man has an ordered list of his preferences of all of the women and each woman has an ordered list of her preferences of all of the men. An algorithm called the Gale-Shapley algorithm finds an optimal solution by finding a *stable* matching. Defining stability forms the foundation of these algorithms; in the stable marriage problem, a stable matching is one that is not unstable, and an unstable matching occurs when a pair of unmatched participants can undermine the matching by choosing each other over their matched partners. A surprising but elegant result of the Gale-Shapley algorithm is it guarantees at least one stable matching for any instance of the stable marriage problem.

The hospital-residents (HR) problem generalizes the stable marriage problem and, and hospital-oriented Gale-Shapley (HGS) algorithm solves the matching problem for graduating medical students and hospitals. Hospitals specify the maximum number of residency spots and thus multiple residents can be matched to a single hospital. Residents rank only their preferred hospitals in their preference lists, and hospitals rank only their preferred residents. These partial preference lists contrast with the stable marriage problem, where each person has to rank every person of the other group. The HGS algorithm also always guarantees at least one stable matching (dependent on an appropriate definition of stability defined later).

We adapted the same data used for ALPHABETR to draw an analogy between the hospital-residents problem and matching TCRβ and TCRα chains. The chains have partial preference lists for each other based on their co-occurrences, and the possibility of matching multiple partners theoretically allows for the identification of dual-TCRα clones. We find that our adaptation works only the case where CDR3α-sharing, CDRβ-sharing, and dual TCRs do not occur in the repertoire, but we discuss the potential for these algorithms to be extended for pairing TCRs of naive T cell populations.

We will begin by overviewing the details of the stable marriage problem and the hospital-residents problem in Sections 6.1.1–6.1.2. We then describe how the sequencing data is processed as input for the hospital-residents problem in Section 6.2. Sections 6.3.1–6.3.3 describe simulations that test the accuracy and precision of this approach and what its limitations are. Finally, Section 6.4 will wrap up the chapter with a discussion about possible extensions of these algorithms.

6.1.1 *Gale-Shapley algorithm and the stable marriage problem*

The stable marriage problem attempts to find a stable matching between a group of n men and a group of n women. Each person has a strictly ordered preference list of all the members of the opposite group. If person A prefers b over c, then b precedes c on the preference list of person A. A matching M is a one-to-one correspondence between the group of men and the group of women; namely, each man is matched with exactly one woman and each woman is matched with exactly one man. If man m and woman w are matched together in M, then we denote this as $m = p_M(w)$ (read as m is the partner of w in matching M) and $w = p_M(m)$. A matching is unstable if there exist a blocking pair, which is a man m and woman w who are unmatched in M but prefer each other over their partners under M (in our notation, m prefers w over $p_M(m)$ and w prefers m over $p_M(w)$).

For example, consider the following instance of the stable marriage problem of 2 men and 2 women with these preference lists:

- Woman 1: Man 2, Man 1
- Woman 2: Man 2, Man 1
- Man 1: Woman 1, Woman 2
- Man 2: Woman 2, Woman 1

Suppose we have a matching M_u in which

- Man 1 to Woman 2 are matched
- Man 2 to Woman 1 are matched

Man 2 and Woman 2 are a blocking pair since they would prefer each other over their matched partners. Man 1 and Woman 1 are not a blocking pair because although Man 1 would prefer Woman 1 over his partner Woman 2, Woman 1 is happy to be matched to Man 2 over Man 1. Because of the blocking pair Man 2 and Woman 2, M_u is an unstable matching. Instead, the matching M_s

- Man 1 to Woman 1
- Man 2 to Woman 2

is a stable matching with no blocking pairs.

The fundamental theorem proved by Gale and Shapley states that there is always at least one stable matching for every instance of the stable marriage problem [144]. The proof is given by the Gale-Shapley algorithm, which always provides a stable matching for any instance of the stable marriage problem. The algorithm is imagined as a series of "proposals" from the men to the women, where the men propose to the women in the order of their preference lists and the women accepts or rejects depending on the desirability of the proposing man. Each person has a state of being either free or engaged, and a woman will always accept a proposal if she is free. She is always in a state of engaged after that first engagement, but she chooses another fiancé if a more desirable man proposes.

The first man proposes to the woman highest on his preference list, and she accepts since she is free. The next man then proposes to the woman highest on his preference list. If she free, then she accepts, but if she is engaged, she accepts the proposal only if the proposing man is preferable over to her current fiancé. If the latter happens, then the man with the broken engagement is considered free again, and he then proposes to the next woman on his preference list. This continues until everybody is engaged and results in a stable matching.

The Gale-Shapely algorithm is shown in the following pseudocode [143]:

```
set each person to be free
while some man m is free
    w = first woman on m's list to whom m has not yet proposed
    if w is free
        assign m and w to be engaged to each other
    else
        if w prefers m to her fiancé m′
            assign m and w to be engaged
            set m′ to be free
        else
            w rejects m (and m remains free)
```

The algorithm presented here is the man-oriented version, which results in the man-optimal matching where every man is matched with the best partner he can obtain in any stable matching and every woman is matched with the worst partner she can obtain in any stable matching. It also always terminates in the same stable matching regardless of the order in which the men propose. These results are summarized and proved in Reference [143].

6.1.2 *Hospital-residents problem*

The hospital-residents problem is an asymmetric extension of the stable marriage problem where the members of one group can have multiple partners and the members of the other group have exactly one partner. In the application of matching residents to hospitals, residents are matched to one hospital (or not at all), and hospitals choose the number of available spots for residents. This problem is more general in that the number of residents does not necessarily equal the total number of spots available from all hospitals, multiple residents can be matched to a hospital, and preference lists need not contain every single hospital or resident. A hospital said to be acceptable to a resident if the hospital is in the resident's preference list (and vice versa).

A matching M is then a mapping from the set of residents to the set of hospitals where the number of residents does not exceed the number of spots provided by each hospital. The definition of stability needs to be generalized for this problem, and so a matching is unstable if there exists a resident r and hospital h such that the following three conditions hold:

1. h is acceptable to r and r is acceptable to h
2. either r is unmatched or r prefers h to their assigned hospital
3. either h has open spots in the matching or h prefers r to at least one of its assigned residents

This definition describes all of the ways a blocking pair r and h can unravel a hospital-residents matching. First, r and h must be on each others' preference list to even be a blocking pair. In order for the resident r to undermine a matching by partnering with another hospital h, the resident must not be matched at all (in which the resident will then take any willing hospital, particularly hospital h), or the resident must have a hospital h that is preferable

over their matched hospital. If that hospital h has an open spot, then r can undermine the matching by partnering with h. Or if that hospital h prefers r over any of its assigned residents, then r and h again can undermine the matching by ditching that assigned resident for resident r.

The hospital-oriented Gale-Shapley (HGS) algorithm provides a stable matching that favors the preferences of the hospitals. The hospitals play a role analogous to the men of the Gale-Shapley algorithm and offer a place to residents in preference order until all of its available spots are filled or until no more acceptable residents remain on its list. The residents accept an offer if they are "free" and accept new offers only if the proposing hospital is preferable over its current hospital. The HGS algorithm is guaranteed to terminate in a stable matching.

Before the algorithm begins, only mutually acceptable hospitals and residents are found in preference lists. So, if resident r does not have hospital h on their list, then r is removed from h's list. We use the term *assignment* instead of engagement in the hospital-resident context, and a hospital is said to be *undersubscribed* if it is assigned fewer residents than the number of spots available.

The HGS algorithm gives the hospitals the role of the men in the Gale-Shapley algorithm for the stable marriage problem. Each hospital offers a spot to residents in the order of its preference lists until all of its spots are filled or until there are no available residents acceptable to the hospital. When a resident is assigned to a hospital, the resident trims their preference list by removing hospitals less preferable to their assigned hospital. Trimming down the preference list is possible because a resident will never be assigned to a hospital worse than its assigned one, which is analogous to the women in the Gale-Shapley, who refuse men not preferable to their engaged man and break an engagement only if a more preferable man proposes. When a hospital is removed from a resident's preference list, the same hospital removes that resident from its own preference list (if the resident was on its preference list at all). This procedure is all summarized in the following pseudocode [143]:

```
set each resident to be free
set each hospital to be totally unsubscribed
while (some hospital h is undersubscribed) AND
  (h's list contains a resident r not assigned to h):
    r = first such resident on h's list
    if r is already assigned to hospital h'
        break the provisional assignment of r to h'
    assign r to h
    for every hospital ĥ less preferable than h on r's list
        remove ĥ and r from each other's lists
```

The HR algorithm terminates with a stable matching that is not dependent on the order of the hospitals used in the assignments.

6.2 FROM HOSPITAL-RESIDENTS TO T CELL RECEPTORS

We saw a possible application of the HGS algorithm for pairing CDR3α and CDR3β sequences from the sequencing data obtained using the same experimental setup utilized by ALPHABETR. The sequencing data and the co-occurrences between α chains and β chains serve as a basis for creating preference lists. The level of association between chain α_i and β_j is measured by the score used by ALPHABETR defined in Equation (5.1), and then preference lists for each chain are created by ranking these scores in descending order.

Each unique sequence α_i has a preference list of

$$P_{\alpha_i} = \left(\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_n}\right)$$

where $S_{ik_j} \geq S_{ik_{j+1}} > 0$ for $j = 1, \ldots, n-1$, and n is the number of unique β chains that co-occur with α_i.

Each unique sequence β_j has a preference list of

$$P_{\beta_k} = \left(\alpha_{l_1}, \alpha_{l_2}, \ldots, \alpha_{l_m}\right)$$

where $S_{kl_j} \geq S_{kl_{j+1}} > 0$ for $j = 1, \ldots, n-1$ and m is the number of unique α chains that co-occur with β_j. In the case where there are ties in the scores, the chains with those scores are randomly ordered within their spots on the pref-

erence list.[2] These preference lists are the inputs into the hospital-oriented HR algorithm, where α chains play the role of the residents and β chains play the role of the hospitals. We specify capacities of 1 for each β chain, which is equivalent of hospitals having exactly one spot. This prevents the discovery of dual-TCRα clones and chain sharing. In order to be able to identify dual-TCRα clones, we would need some procedure to determine which CDR3β sequences derive from dual-TCRα clones and which are shared among many clones. This information would then be an input to the HGS algorithm, and as we will discuss later, this is an open problem that cannot be solve with currently available algorithms. Thus, we can at best discover one of the two chains of a dual-TCRα expressing clone and cannot identify CDR3α- and CDR3β-sharing clones.

We provide an implementation of the HGS algorithm in the `matchmaker` package, which was used for the simulations shown in this chapter.[3]

6.3 RESULTS

6.3.1 *The HGS algorithm shows high depth and low false pairing rates in populations with no shared chains.*

We first tested the HGS algorithm approach by simulating data with the same skewed immunodominance hierarchies used in the ALPHABETR simulations, the 5-plate "high mixed" sampling strategy (Table 5.3), and lognormal dropping errors with either (*i*) no in-frame sequencing errors or (*ii*) in-frame sequencing errors identical to those used in the ALPHABETR simulations. The simulations used T cell populations containing no shared chains and either (*iii*) 0% or (*iv*) 15% of clones expressing dual TCRα chains.

Figures 6.1–6.4 show the results of the simulations using the HGS algorithm. The left columns show the results of simulations of populations with no dual-TCRα clones, and the right columns show results of simulations of populations with 15% dual-TCRα clones. The bottom row of plots simulated in-frame sequencing errors while the top row simulated no in-frame

[2] Algorithms that deal with ties within the preference lists exist, but the different kinds of stability that can be defined for these problems suffer a lack of interpretability for our sequencing application, and the computational requirements of these algorithms add another difficult hurdle for their use.

[3] Available on `github.com/edwardslee/matchmaker`

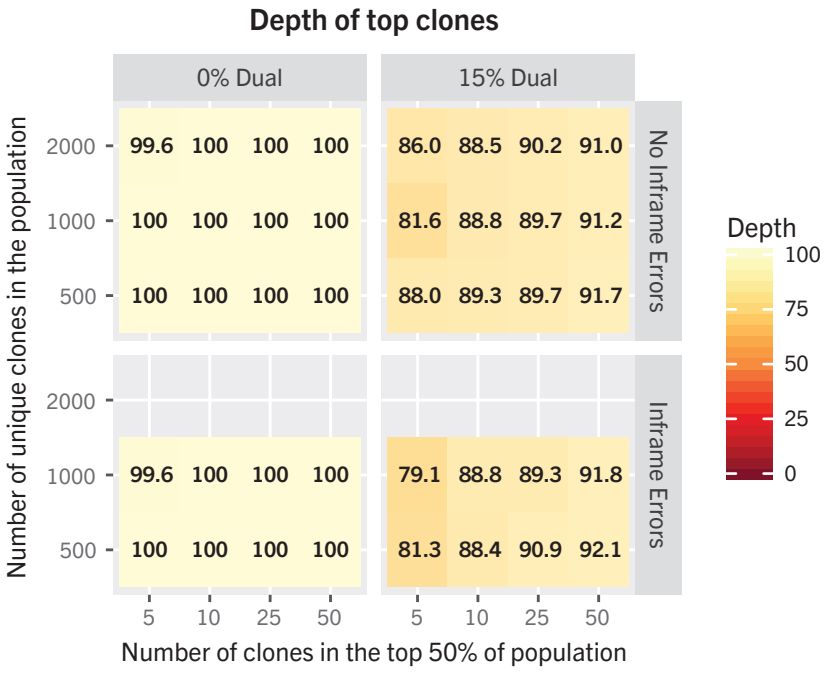


Figure 6.1: Depth of the common clones representing the top 50% of the population in frequency. We performed simulations using the HGS algorithm on populations with 500, 100, and 2000 clones with 0% or 15% dual-TCRα clones and with in-frame or without in-frame sequencing errors (averaged over 100 simulations). The HGS algorithm is able to identify practically all common clones in populations with 0% dual-TCRα clones, and 79–92% of common clones in populations with 15% dual-TCRα clones. We see no discernible effect on top depths with the presence of in-frame sequencing errors.

sequencing errors. Figure 6.1 shows high depth of the top clones in all scenarios of errors. The HGS algorithm is able to identify practically all common clones when populations have no dual-TCRα clones. For populations with 15% dual-TCRα clones, top depths ranged from 81.6% to 91.7% in populations with no in-frame sequencing errors and ranged from 79.1% to 92.1% in populations with in-frame sequencing errors. The majority of the decrease in top depth seems to be due to the fact that the HGS algorithm cannot identify both TCRs of a dual-TCRα clones, so automatically the algorithm cannot achieve 100% depth. Simulations of populations with 2000 unique clones and in-frame sequencing errors were not performed due to computational limitations.

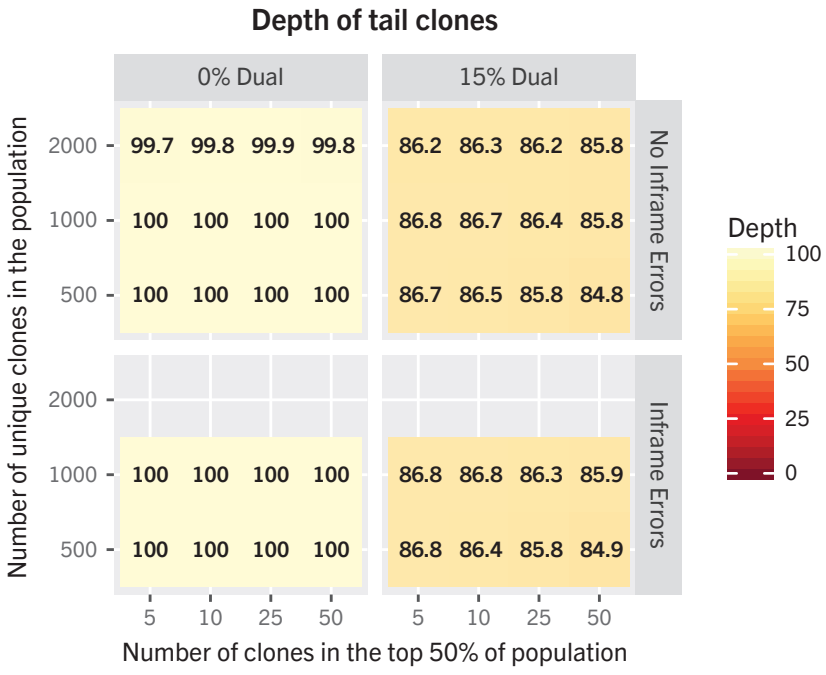


Figure 6.2: Depth of the rare clones representing the bottom 50% of the population in frequency. We performed simulations using the HGS algorithm on populations with 500, 100, and 2000 clones with 0% or 15% dual-TCRα clones and with in-frame or without in-frame sequencing errors (averaged over 100 simulations). The HGS algorithm is able to identify practically all rare clones in populations with 0% dual-TCRα clones, and ~85% of rare clones in populations with 15% dual-TCRα clones. We see no effect on tail depths with the presence of in-frame sequencing errors.

Similar patterns can be seen in the tail depths in Figure 6.2. With no dual-TCRα clones, practically all simulations result in 100% recovery of rare clones. With the presence of dual-TCRα clones, tail depths were approximately 86%, which again reflects the fact that the algorithm cannot identify both chains of dual-TCRα clones. Figure 6.3 shows the same pattern in overall depth of all clones. Figure 6.4 shows the HGS algorithm makes very few mistakes with false pairing rates $\leq$ 1%.

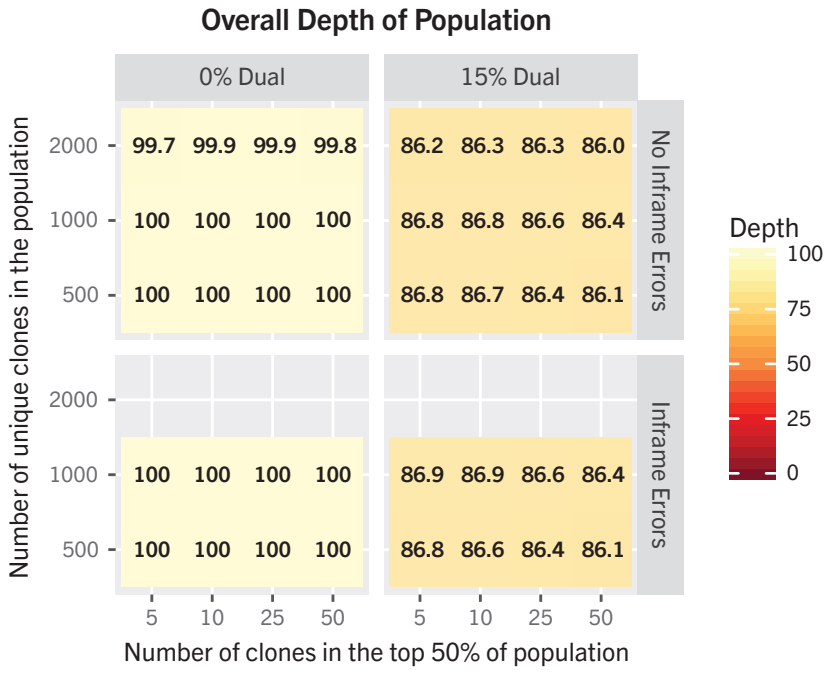


Figure 6.3: Depth of all clones in the simulated populations. We performed simulations using the HGS algorithm on populations with 500, 100, and 2000 clones with 0% or 15% dual-TCRα clones and with in-frame or without in-frame sequencing errors (averaged over 100 simulations). The overall depth performance of the HGS algorithm shows the same patterns for the top and tail depths.

6.3.2 *The HGS algorithm cannot reliably pair TCR sequences in populations with CD3a- and CDRß-sharing*

Although we could not identify TCRs from clones sharing CDR3α and CDR3β sequences, we wanted to see the effect on depth and false pairing rates when chain sharing occurs in the sampled population. We repeated the simulations with populations exhibiting sharing at the "medium level" described in Table 5.1 and 15% dual-TCRα clones with dropping errors and no in-frame sequencing errors. In Figure 6.5*A–C*, we see that the simulations resulted in lower top, tail, and overall depths. More strikingly, Figure 6.5*D* shows that the false pairing rates jumped to 17.9–21.6%, which is unacceptable for accurate pairing. Thus, the HGS algorithm is unable to handle all of the features found in epitope-specific T cell populations.

6.3.3 *Samples of T cells from the naive pool are unlikely to exhibit significant levels of shared chains*

Although the HGS algorithm cannot handle populations that exhibit the sharing of CDR3α and CDR3β sequences, its ability to reliably identify TCR pairs in populations with no chain-sharing begs the question of whether it could have utility in TCR sequencing analysis. Here, we show that samples of the naive αβ T cell repertoire from standard blood samples from humans will not exhibit significant CDR3β sharing.

Let B be the number of unique TCRβ chains found in the naive αβ T cell repertoire. Supposed we collect N αβ T cells in a sample of human blood, assuming $N \ll 10^{11}$, the total naive T cell population size. How many unique β chains U_β are expected to be found in a sample of N naive T cells? This problem is equivalent to randomly sampling with replacement N times from

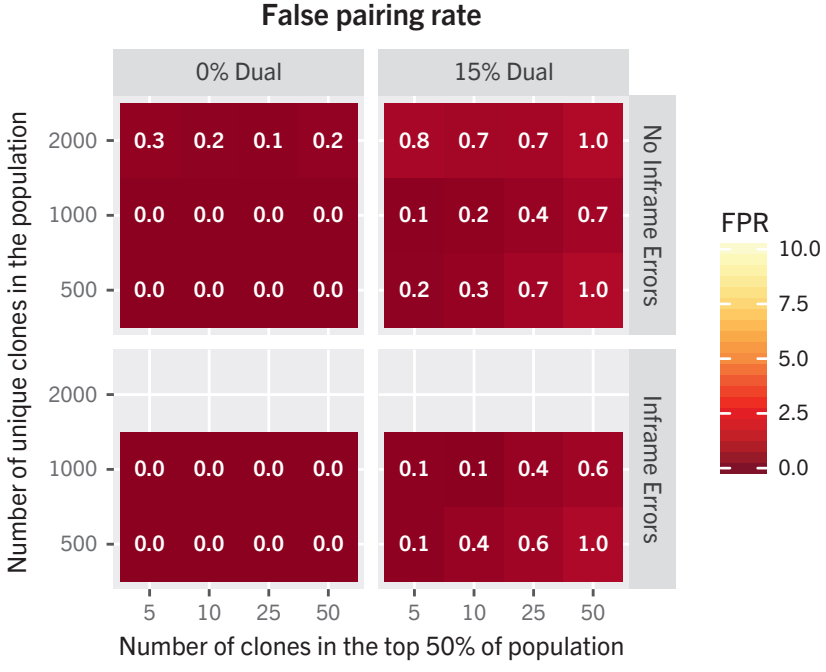


Figure 6.4: False pairing rates in the simulated populations. We performed simulations using the HGS algorithm on populations with 500, 100, and 2000 clones with 0% or 15% dual-TCRα clones and with in-frame or without in-frame sequencing errors (averaged over 100 simulations). The false pairing rates of the HGS algorithm are very low in all simulated conditions, never going above 1%.

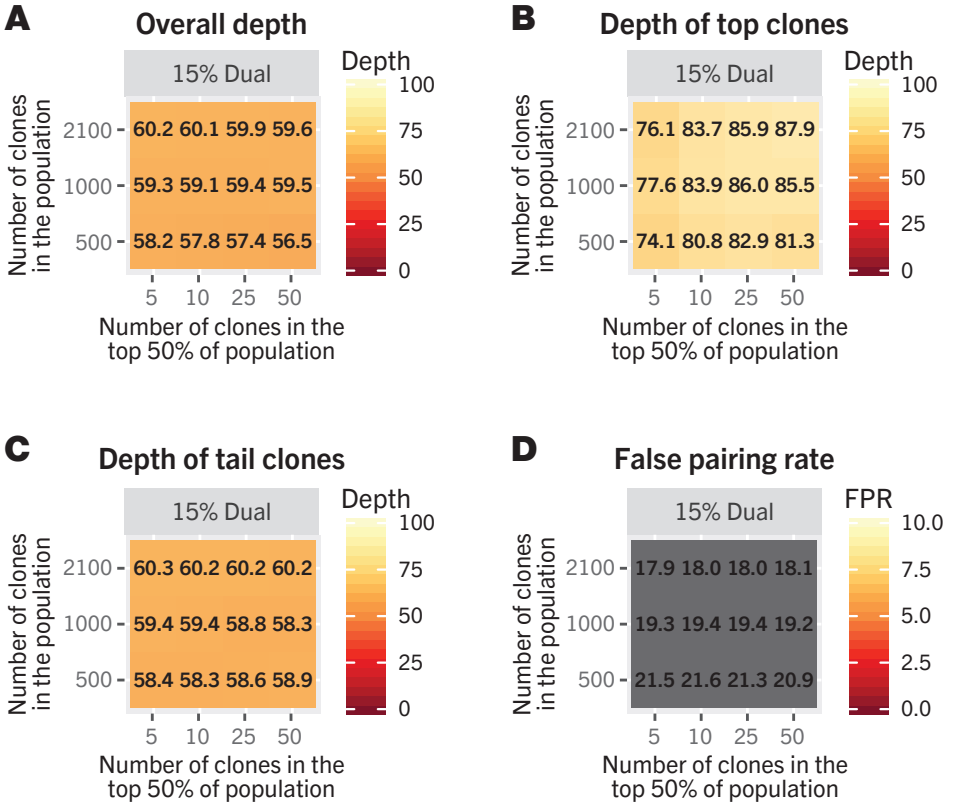


Figure 6.5: Results from simulations of populations exhibiting medium levels of sharing. Simulations were performed to test the effect on depth and false pairing rates due to the presence of CDR3α- and CDR3β-sharing. The simulated experiments had dropping chain errors and no in-frame sequencing errors. The number of clones in the populations were varied (500, 1000, and 2000), and different levels of skew were tested by changing the number of clones in the top 50% of the population by clonal frequency. **(A)** The overall depth of all clones in the population. **(B)** The depth of the common clones that make up the top 50% of the population by frequency. **(C)** The depth of the rare clones that make up the bottom 50% of the population by frequency. **(D)** The false pairing rates associated with each set of simulations. These results are the average of 100 simulations.

a set of B different β chains labeled $\beta_1, \beta_2, \ldots, \beta_B$ and asking how many of them appear in that sample on average. This is straightforward to calculate if one assumes that each distinct β chain is present roughly at equal frequency within the naive T cell pool. For any non-uniform distribution of naive TCRβ clone sizes [145, 146], the following calculation provides a lower bound on the sample size required to achieve a given level of coverage of the repertoire.

Let I_i be a random variable equal to 1 if chain β_i appears in the sample at least once and 0 otherwise. $P(I_i = 0)$ is $\left((B-1)/B\right)^N$, and so its expected value $E(I_i)$ is $1 - \left((B-1)/B\right)^N$. The expected number of different β chains in the sample is then

$$U_\beta = E\left(\sum_{i=1}^{B} I_i\right) = \sum_{i=1}^{B} E(I_i) = B\left(1 - \left(\frac{B-1}{B}\right)^N\right). \qquad (6.1)$$

Similarly, let J_i be a random variable that is equal to 1 if clone i appears only once in the sample, and zero otherwise. $P(J_i = 1)$ is $(N/B)(1 - 1/B)^{N-1}$, and so the expected number of chains that appear only once is $B \cdot E(J_i) = N(1 - 1/B)^{N-1}$. Figure 6.6 shows the dependence of these quantities on sample size using the current best estimate of B in humans, 10^8 [20]. Since $1/B$ is very small, we expected the number of chains recovered from N cells to be approximately N.

We can then estimate typical TCRβ clone size distributions obtained from bulk sequencing of T cells recovered from human blood samples. One milliliter of human blood yields typically 0.5×10^6 to 1.8×10^6 αβ T cells, of which 0.1×10^6 to 0.8×10^6 are naive $CD4^+$ T cells and 0.03×10^6 to 0.2×10^6 are naive $CD8^+$ T cells. A standard 10 mL blood sample will then contain between 10^6 and 10^7 naive αβ T cells. Equation 6.1 predicts that this will yield 1–10% of the β chain repertoire, with correspondingly 99.5% and 95% of the identified chains will have come from a sample size of a single cell of the clonotypes and thus a majority of the β chains come from one cell. In a 50 mL sample, we expect to cover between 5% and 40% of the β chain repertoire, with correspondingly 98% and 80% of the chains deriving from only one cell. We would expect to see very little chain sharing in the T cells sampled in a sample of blood. Thus, the HGS algorithm potentially could be used in T cells obtained from a standard 10 mL blood sample. In fact,

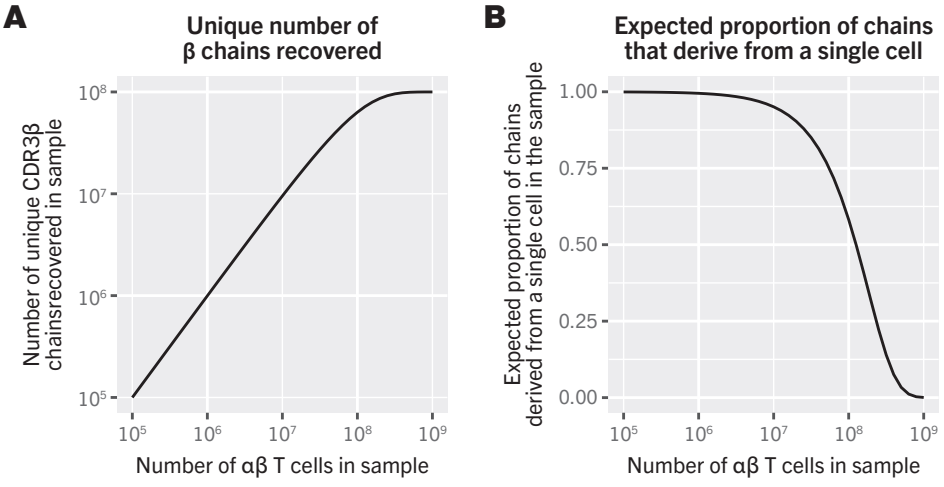


Figure 6.6: Expected number of unique β chains recovered from a large population with a repertoire of 10^8 different CDR3β chains, for varying sample sizes (assuming sampling with replacement). **(A)** The number of unique CDR3β expected to be recovered in a given number of sampled T cells from the naive pool. **(B)** The number of chains we expect to come from exactly one T cell given a fixed number of T cells that are sampled.

because any given clone would be sampled only a few times, preference lists would be very short and would allow the HGS algorithm to efficiently find stable matchings.

A practical limitation to using the HR algorithm for naive populations is the long computational time needed for solving large instances of the hospital-residents problem. The HGS algorithm determines stable resident-hospital pairs in $O(m)$ time, where m is the number of acceptable resident-hospital pairs. This means that the time required by the computer to solve an instance of the hospital-residents problem increase linearly with the number of acceptable pairs, given that the implementation uses appropriate data structures.[4] Since a reasonable or large sample of blood from humans might contain tens or hundreds of thousands of unique CDR3α and CDR3β sequences, an application of the HGS algorithm could take an unreasonable about of computing time to determine αβ pairs. Our simulations indicate

[4]The `matchmaker` package that we have written is not guaranteed to be this efficient, and our testing shows that our implementation is most likely slower than $O(m)$ time. Other less cleverly named R packages such as the `matchingR` and `matchMarkets` packages may provide more efficient implementations.

that the HGS algorithm could accurate identify TCR pairs, but computational advances will need to be made in order for the algorithm to be practical for sequencing T cells from the naive repertoire.

6.4 DISCUSSION

Matching algorithms have the potential to efficiently and accurately identify paired TCR sequences without using single-cell approaches. We demonstrated that in special situations, the HR algorithm outperforms ALPHABETR with very high depths and very low false pairing rates. However, we also showed that the HGS algorithm fails to work for populations with shared chains, a feature that is prominent in many epitope-specific T cell populations. Thus, the HGS algorithm in its current form cannot be used for epitope-specific T cell populations.

The simulated populations used in Section 6.3.1 do reflect the features of a sample of T cells obtained from the naive repertoire, and the success of the HGS algorithm in these simulations points to a possible application for sequencing T cells from the naive repertoire. However, the HGS algorithm is practically not useable at the scale needed for sequencing the many tens or hundreds of thousands of CDR3 sequences found in samples of naive T cells. Advances in the algorithm and in computing may allow for the identification of pairs at this scale in the future. Another limitation of the HGS algorithm is the inability to identify dual-TCRα clones. In order to do so, we would need to allow the CDR3β sequences of the the dual-TCRα clones to pair with two CDRα sequences (equivalent to a hospital having two spots for residents). This would require *a priori* knowledge of which CDR3β sequences belong to dual-TCRα clones. Of course, we would like the pairing algorithms themselves to discover dual-TCRα clones, and this dilemma is an open problem that is not straightforward to solve. A similar issue occurs for discovering shared chains; we need to know beforehand how many clones share a given CDR3α or CDR3β chain as an input to the HGS algorithm. Although the HGS algorithm pairs TCR sequences successfully in limited special cases, we do not believe that the algorithm in its current form is usable for epitope-specific populations nor practical for naive T cell populations without computational and/or algorithmic advances.

The definition of stability in the application of TCR sequence pairs has a slightly unnatural interpretation. What does it mean for a set of CDR3α and CDR3β sequences to form a stable matching? In the context of stable marriages and hospital/residents, stability is a desired property since no unmatched pair can undermine a matching by choosing each other rather than their matched partners. In the CDR3α/CDR3β setting, a stable matching of TCR pairs does not have an analogous type of interpretation. What does it mean for a β chain to be matched with its best α chain partner such that they cannot undermine the whole matching? Although our simulations show that using stability provides good results, this issue does suggest that other approaches to matching problems might be more appropriate for our sequencing problem. One approach might involve matching problems that utilize *cardinal* utilities rather than the ordinal preferences used the stable marriage problem and the hospital-residents problem; in our problem, this approach would use the association scores directly instead of converting them into ranked preference lists. Such problems like the Assignment problem and maximum weight matching in graphs could have extensions that pair TCR sequences in epitope-specific populations. These problems would use the association scores directly rather than converting these to ordinal preference lists and find a set of pairings that optimizes for the scores in some way.[5]

Both chapters 5 and 6 demonstrate the undeniable advantage of frequency-based pairing techniques: thousands of paired TCRs can be identified without large single-cell sequencing experiments. ALPHABETR is able to identify many common clones and over a thousand rare clones of an epitope-specific population with an experiment using five 96-well plates, whereas an equivalent single-cell experiment would identify at best 480 TCRs and we estimate in practice would recover on the order of 100 TCRs. The HGS algorithm shows the potential to identify thousands of clones at the same scale as well. We believe that fully characterizing T cell repertoires requires paired sequencing information and the discovery of both common and rare clonotypes. Current single-cell technologies are too cost-prohibitive to achieve

[5]The conversion from scores to ordinal preferences is not necessarily a problem since this approach has been used successfully in many applications. For example, junior doctors are matched to hospitals in Scotland by calculating scores for the junior doctors from exam marks and application assessments and converting them to ordinal preferences.

these goals, and so frequency-based approaches can provide answers to our questions about repertoires until single-cell sequencing becomes more economical and ubiquitous.

CHAPTER 7

Discussion and Conclusions

The protection conferred by adaptive immunity results in part from a large population of T cells with a vast repertoire of TCRs. In this thesis, we have combined experimental and mathematical modeling techniques to probe these properties of T cell populations. We set out to quantify the number of T cells and its subsets in the SLOs, to characterize the development of these T cells from neonates to adults, and to develop tools to quantify αβ T cell clonotypes and measure their frequencies.

Identifying T cell clonotypes with TCR sequencing has become an increasingly important tool in understanding the role and function of T cells in infections and in diseases [147, 148, 62]. The community has been actively working to standardize and improve the pipeline for TCR sequencing: this includes standardizing file formats [149], creating software for processing and aligning the raw reads [150, 151, 102, 152, 153, 154], and developing new experimental protocols for sequencing samples of T cells (described in Section 1.5). In Chapters 5–6, we described experimental and statistical approaches that can identify thousands of paired TCR sequences without having to resort to single-cell sequencing. Both approaches—ALPHABETR and the application of the hospital-oriented Gale-Shapley algorithm—pair CDR3α and CDR3β sequences obtained from multiple bulk-sequenced subsamples of T cells from the parent population of interest. We designed ALPHABETR to identify TCRs from antigen-specific populations and to identify TCRs with shared CDR3α or shared CDR3β sequences. Chain sharing is found in many antigen-specific repertoires (Table 5.1) and has not been explicitly considered by other frequency-based pairing approaches. The HGS algorithm identifies paired TCR sequences very efficiently in special cases but fails when a significant degree of chain-sharing occurs in the repertoire of interest. We demonstrated that the performance of ALPHABETR with 480 subsamples outperforms sequencing of several thousands of single cells (Figure 5.13). Although the newest single-cell technologies are starting to scale to tens and hundreds of thousands of single cells (demonstrated in a preprint paper [155]), we believe that frequency-based pairing is currently the most cost-effective and efficient way to identify paired TCRα and TCRβ sequences. We believe these techniques could provide much insight in the

context of cancer and in vaccine design by identifying important clonotypes for designing effective immunotherapies and vaccines.

The diversity of the αβ T cell repertoire is distributed across roughly 10^8 T cells in mice, resulting in approximately 100–1000 T cells in the naive repertoire that can respond to a given antigen [52, 2]. We set out to more accurately characterize the numbers of αβ T cells and their subsets in the secondary lymphoid organs of the mouse. In Chapter 4, we used thoracic duct cannulations to drain recirculating T cells from SLOs and and then used T cell counts of the collected lymph and the SLOs to calculate total numbers. This approach was based on a simple idea that assuming free and constant recirculation, all SLOs would lose the same proportion of cells and that this loss would be represented by the number of cells collected by the cannulation (Figure 4.2). The data showed that SLOs were affected differently by the TD cannulation (Figures 4.5–4.6), presumably due to different lymphatic connections, and this resulted in too many unknown parameters, which prevented us from estimating T cell numbers accurately. In addition, we discovered -that the stress from the surgical procedure, in addition to the cannulation itself, caused a strikingly profound loss of cells in lymph nodes. Although previous studies have shown changes in T cell counts and proportions [120, 119, 122, 121], we showed that the loss of cells occurs across many different subsets of T cells and B cells. These findings may impact the interpretation of studies that use invasive procedures. Procedures such as intravital imaging or lymphadenectomies potentially cause T cell and B cell numbers to decrease and may fundamentally change immune responses. We characterized these changes in our thoracic duct cannulation system, but further work will be needed to see if the magnitude of changes that we observed generalizes to other surgical procedures. Future studies could explore the effect of different anesthetics, different dosages of anesthetic, and different duration of anesthesia and/or surgery.

Finally, we used mathematical modeling to characterize the development of $CD4^+$ T cells from the DP2 thymocyte stage to the peripheral naive $CD4^+$ T cell compartment and the development of DP3 thymocyte stage through to the export of SP8 thymocytes. By confronting models with cell count and Ki67 expression data in thymocyte subsets and peripheral T cell subsets in neonatal mice, we were able to characterize the shifting dynamics of T cell development in the thymus and the periphery. We elucidated the pat-

tern of differentiation in SP4 and SP8 thymocytes—particularly, how immature SP4/SP8 thymocytes differentiate to mature SP4/SP8 thymocytes—and found evidence of changing per capita division rates in both the SP4 and SP8 compartments in the neonatal mice. We were also particularly interested in determining whether lymphopenia-induced proliferation plays an important role in the development of naive $CD4^+$ T cell populations. Although the neonatal mouse can support LIP [46], we showed that a simple mathematical model could explain the dynamics of mature SP4 thymocyte and naive $CD4^+$ T cell compartments from soon after birth without recourse to LIP; high Ki67 expression in naive cells early in life is largely inherited expression from transiently high levels of intrathymic proliferation. Thus, our work supports the conclusion that the development of a healthy naive $CD4^+$ T cell compartment results mainly from the thymic output of T cells during ontogeny without a significant amount of peripheral expansion. Further work will be performed to (*i*) quantify how the distribution of αβ T cell clone sizes emerging from the thymus changes with age, (*ii*) perform a similar analysis to assess whether LIP occurs in naive $CD8^+$ T cells during development, and (*iii*) characterize the development of $CD4^+$ and $CD8^+$ memory T cells in neonates.

In this thesis, we employed a wide range of mathematical and statistical techniques in the analysis and interpretation of our data. We used ODEs and Bayesian analysis to describe the dynamics of T cell ontogeny in detail, and the conclusions made regarding the cannulation experiments derived from a simple experimental idea and elementary algebra. The TCR pairing study attempted to extend existing algorithms as well as develop new ones and employed maximum likelihood estimation. In all cases, the quantitative tools were crucial in helping us to encode and understand the immunology underlying our data. By combining these quantitative approaches with careful experimental design, we were able to make richer conclusions and draw more information from the data that would not have been possible with experiments alone.

Bibliography

[1] Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-Throughput Pairing of T Cell Receptor α and β Sequences. Science Translational Medicine. 2015;7(301):301ra131–301ra131. doi:10.1126/scitranslmed.aac5624.

[2] Blattman JN, Antia R, Sourdive DJD, Wang X, Kaech SM, Murali-Krishna K, et al. Estimating the Precursor Frequency of Naive Antigen-Specific CD8 T Cells. The Journal of Experimental Medicine. 2002;195(5):657–664.

[3] Murali-Krishna K, Altman JD, Suresh M, Sourdive DJD, Zajac AJ, Miller JD, et al. Counting Antigen-Specific CD8 T Cells: A Reevaluation of Bystander Activation during Viral Infection. Immunity. 1998;8(2):177–187. doi:10.1016/S1074-7613(00)80470-7.

[4] Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. Structure of the Complex between Human T-Cell Receptor, Viral Peptide and HLA-A2. Nature. 1996;384(6605):134–141. doi:10.1038/384134a0.

[5] Garcia KC, Degano M, Stanfield RL, Brunmark A, Jackson MR, Peterson PA, et al. An Alphabeta T Cell Receptor Structure at 2.5 A and Its Orientation in the TCR-MHC Complex. Science (New York, NY). 1996;274(5285):209–219.

[6] Bentley GA, Boulot G, Karjalainen K, Mariuzza RA. Crystal Structure of the Beta Chain of a T Cell Antigen Receptor. Science (New York, NY). 1995;267(5206):1984–1987.

[7] van der Merwe PA, Dushek O. Mechanisms for T Cell Receptor Triggering. Nature Reviews Immunology. 2010;11(1):47–55. doi:10.1038/nri2887.

[8] Li Y, Yin Y, Mariuzza RA. Structural and Biophysical Insights into the Role of CD4 and CD8 in T Cell Activation. Frontiers in Immunology. 2013;4. doi:10.3389/fimmu.2013.00206.

[9] Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. Annual Review of Immunology. 2015;33(1):169–200. doi:10.1146/annurev-immunol-032414-112334.

[10] Blackwell TK, Alt FW. Molecular Characterization of the Lymphoid V(D)J Recombination Activity. The Journal of Biological Chemistry. 1989;264(18):10327–10330.

[11] Lieber MR. The Polymerases for V(D)J Recombination. Immunity. 2006;25(1):7–9. doi:10.1016/j.immuni.2006.07.007.

[12] Giudicelli V, Chaume D, Marie-Paule L. IMGT/GENE-DB: A Comprehensive Database for Human and Mouse Immunoglobulin and T Cell Receptor Genes. Nucleic Acids Research. 2004;33(Database issue):D256–D261. doi:10.1093/nar/gki010.

[13] Gilfillan S, Dierich A, Lemeur M, Benoist C, Mathis D. Mice Lacking TdT: Mature Animals with an Immature Lymphocyte Repertoire. Science. 1993;261(5125):1175–1178. doi:10.1126/science.8356452.

[14] Komori T, Okada A, Stewart V, Alt FW. Lack of N Regions in Antigen Receptor Variable Region Genes of TdT-Deficient Lymphocytes. Science (New York, NY). 1993;261(5125):1171–1175.

[15] Davis MM, Bjorkman PJ. T-Cell Antigen Receptor Genes and T-Cell Recognition. Nature. 1988;334(6181):395–402. doi:10.1038/334395a0.

[16] Laydon DJ, Bangham CRM, Asquith B. Estimating T-Cell Repertoire Diversity: Limitations of Classical Estimators and a New Approach. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370(1675):20140291. doi:10.1098/rstb.2014.0291.

[17] Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A Direct Estimate of the Human Alphabeta T Cell Receptor Diversity. Science (New York, NY). 1999;286(5441):958–961.

[18] Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive Assessment of T-Cell Receptor -Chain Diversity in T Cells. Blood. 2009;114(19):4099–4107. doi:10.1182/blood-2009-04-217604.

[19] Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-Cell Repertoire Sequencing of Human Peripheral Blood Samples Reveals Signatures of Antigen Selection and a Directly Measured Repertoire Size of at Least 1 Million Clonotypes. Genome Research. 2011;21(5):790–797. doi:10.1101/gr.115428.110.

[20] Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and Clonal Selection in the Human T-Cell Repertoire. Proceedings of the National Academy of Sciences. 2014;111(36):13139–13144. doi:10.1073/pnas.1409155111.

[21] Yin Y, Mariuzza RA. The Multiple Mechanisms of T Cell Receptor Cross-Reactivity. Immunity. 2009;31(6):849–851. doi:10.1016/j.immuni.2009.12.002.

[22] Macdonald WA, Chen Z, Gras S, Archbold JK, Tynan FE, Clements CS, et al. T Cell Allorecognition via Molecular Mimicry. Immunity. 2009;31(6):897–908. doi:10.1016/j.immuni.2009.09.025.

[23] Borbulevych OY, Piepenbrink KH, Gloor BE, Scott DR, Sommese RF, Cole DK, et al. T Cell Receptor Cross-Reactivity Directed by Antigen-Dependent Tuning of Peptide-MHC Molecular Flexibility. Immunity. 2009;31(6):885–896. doi:10.1016/j.immuni.2009.11.003.

[24] Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size Estimate of the TCR Repertoire of Naive Mouse Splenocytes. The Journal of Immunology. 2000;164(11):5782–5787. doi:10.4049/jimmunol.164.11.5782.

[25] Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, et al. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. The Journal of Immunology. 2014;192(6):2689–2698. doi:10.4049/jimmunol.1302064.

[26] Fagnoni FF, Vescovini R, Passeri G, Bologna G, Pedrazzoni M, Lavagetto G, et al. Shortage of Circulating Naive CD8(+) T Cells Provides New Insights on Immunodeficiency in Aging. Blood. 2000;95(9):2860–2868.

[27] Nikolich-Žugich J, Slifka MK, Messaoudi I. The Many Important Facets of T-Cell Repertoire Diversity. Nature Reviews Immunology. 2004;4(2):123–132. doi:10.1038/nri1292.

[28] DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley PH. Human T Cell Receptor Occurrence Patterns Encode Immune History, Genetic Background, and Receptor Specificity. 2018;doi:10.1101/313106.

[29] Krueger A, Ziętara N, Łyszkiewicz M. T Cell Development by the Numbers. Trends in Immunology. 2017;38(2):128–139. doi:10.1016/j.it.2016.10.007.

[30] Starr TK, Jameson SC, Hogquist KA. Positive and Negative Selection of T Cells. Annual Review of Immunology. 2003;21(1):139–176. doi:10.1146/annurev.immunol.21.120601.141107.

[31] Lind EF, Prockop SE, Porritt HE, Petrie HT. Mapping Precursor Movement through the Postnatal Thymus Reveals Specific Microenvironments Supporting Defined Stages of Early Lymphoid Development. The Journal of Experimental Medicine. 2001;194(2):127–134.

[32] Hayday AC, Barber DF, Douglas N, Hoffman ES. Signals Involved in Gamma/Delta T Cell versus Alpha/Beta T Cell Lineage Commitment. Seminars in Immunology. 1999;11(4):239–249. doi:10.1006/smim.1999.0180.

[33] Borowski C, Li X, Aifantis I, Gounari F, von Boehmer H. Pre-TCRα and TCRα Are Not Interchangeable Partners of TCRβ during T Lymphocyte Development. The Journal of Experimental Medicine. 2004;199(5):607–615. doi:10.1084/jem.20031973.

[34] Dudley EC, Petrie HT, Shah LM, Owen MJ, Hayday AC. T Cell Receptor Beta Chain Gene Rearrangement and Selection during Thymocyte Development in Adult Mice. Immunity. 1994;1(2):83–93.

[35] Hoffman ES, Passoni L, Crompton T, Leu TM, Schatz DG, Koff A, et al. Productive T-Cell Receptor Beta-Chain Gene Rearrangement: Coincident Regulation of Cell Cycle and Clonality during Development in Vivo. Genes & Development. 1996;10(8):948–962. doi:10.1101/gad.10.8.948.

[36] Falk I. Proliferation Kinetics Associated with T Cell Receptor-Beta Chain Selection of Fetal Murine Thymocytes. Journal of Experimental Medicine. 1996;184(6):2327–2340. doi:10.1084/jem.184.6.2327.

[37] Pénit C, Vasseur F. Expansion of Mature Thymocyte Subsets before Emigration to the Periphery. Journal of Immunology (Baltimore, Md: 1950). 1997;159(10):4848–4856.

[38] Sinclair C, Bains I, Yates AJ, Seddon B. Asymmetric Thymocyte Death Underlies the CD4:CD8 T-Cell Ratio in the Adaptive Immune System. Proceedings of the National Academy of Sciences. 2013;110(31):E2905–E2914. doi:10.1073/pnas.1304859110.

[39] Merkenschlager M, Graf D, Lovatt M, Bommhardt U, Zamoyska R, Fisher AG. How Many Thymocytes Audition for Selection? The Journal of Experimental Medicine. 1997;186(7):1149–1158. doi:10.1084/jem.186.7.1149.

[40] van Meerwijk JPM, Marguerat S, Lees RK, Germain RN, Fowlkes BJ, MacDonald HR. Quantitative Impact of Thymic Clonal Deletion on the T Cell Repertoire. The Journal of Experimental Medicine. 1997;185(3):377–384. doi:10.1084/jem.185.3.377.

[41] Saini M, Sinclair C, Marshall D, Tolaini M, Sakaguchi S, Seddon B. Regulation of Zap70 Expression During Thymocyte Development Enables Temporal Separation of CD4 and CD8 Repertoire Selection at Different Signaling Thresholds. Science Signaling. 2010;3(114):ra23–ra23. doi:10.1126/scisignal.2000702.

[42] Weinreich MA, Hogquist KA. Thymic Emigration: When and How T Cells Leave Home. The Journal of Immunology. 2008;181(4):2265–2270. doi:10.4049/jimmunol.181.4.2265.

[43] McCaughtry TM, Wilken MS, Hogquist KA. Thymic Emigration Revisited. The Journal of Experimental Medicine. 2007;204(11):2513–2520. doi:10.1084/jem.20070601.

[44] Zachariah MA, Cyster JG. Neural Crest-Derived Pericytes Promote Egress of Mature Thymocytes at the Corticomedullary Junction. Science. 2010;328(5982):1129–1135. doi:10.1126/science.1188222.

[45] Modigliani Y, Coutinho G, Burlen-Defranoux O, Coutinho A, Bandeira A. Differential Contribution of Thymic Outputs and Peripheral Expansion in the Development of Peripheral T Cell Pools. European Journal of Immunology. 1994;24(5):1223–1227. doi:10.1002/eji.1830240533.

[46] Min B, McHugh R, Sempowski GD, Mackall C, Foucras G, Paul WE. Neonates Support Lymphopenia-Induced Proliferation. Immunity. 2003;18(1):131–140.

[47] Surh CD, Sprent J. Homeostasis of Naive and Memory T Cells. Immunity. 2008;29(6):848–862. doi:10.1016/j.immuni.2008.11.002.

[48] Tilney NL. Patterns of Lymphatic Drainage in the Adult Laboratory Rat. Journal of Anatomy. 1971;109(Pt 3):369–383.

[49] Gilroy AM, MacPherson BR, Ross LM, Schuenke M, Schulte E, Schumacher U. Atlas of Anatomy. 2nd ed. New York, NY: Thieme Medical Publishers; 2012.

[50] Kawashima Y, Sugimura M, Hwang YC, Kudo N. The Lymph System in Mice. Japanese Journal of Veterinary Research. 1964;12(4):69–78.

[51] Jenkins MK, Moon JJ. The Role of Naive T Cell Precursor Frequency and Recruitment in Dictating Immune Response Magnitude. Journal of Immunology (Baltimore, Md: 1950). 2012;188(9):4135–4140. doi:10.4049/jimmunol.1102661.

[52] Obar JJ, Khanna KM, Lefrançois L. Endogenous Naive CD8+ T Cell Precursor Frequency Regulates Primary and Memory Responses to Infection. Immunity. 2008;28(6):859–869. doi:10.1016/j.immuni.2008.04.010.

[53] Westermann J, Ehlers EM, Exton MS, Kaiser M, Bode U. Migration of Naive, Effector and Memory T Cells: Implications for the Regulation of Immune Responses. Immunological Reviews. 2001;184(1):20–37. doi:10.1034/j.1600-065x.2001.1840103.x.

[54] Mackay CR. Naive and Memory T Cells Show Distinct Pathways of Lymphocyte Recirculation. Journal of Experimental Medicine. 1990;171(3):801–817. doi:10.1084/jem.171.3.801.

[55] Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T Cell Subsets, Migration Patterns, and Tissue Residence. Annual Review of Immunology. 2013;31(1):137–161. doi:10.1146/annurev-immunol-032712-095954.

[56] Gowans JL, Knight EJ. The Route of Re-Circulation of Lymphocytes in the Rat. Proceedings of the Royal Society B: Biological Sciences. 1964;159(975):257–282. doi:10.1098/rspb.1964.0001.

[57] Pilch H, Hohn H, Freitag K, Neukirch C, Necker A, Haddad P, et al. Improved Assessment of T-Cell Receptor (TCR) VB Repertoire in Clinical Specimens: Combination of TCR-CDR3 Spectratyping with Flow Cytometry-Based TCR VB Frequency Analysis. Clinical and Vaccine Immunology. 2002;9(2):257–266. doi:10.1128/CDLI.9.2.257-266.2002.

[58] Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The Past, Present, and Future of Immune Repertoire Biology – The Rise of Next-Generation Repertoire Analysis. Frontiers in Immunology. 2013;4. doi:10.3389/fimmu.2013.00413.

[59] Fozza C, Barraqueddu F, Corda G, Contini S, Virdis P, Dore F, et al. Study of the T-Cell Receptor Repertoire by CDR3 Spectratyping. Journal of Immunological Methods. 2017;440:1–11. doi:10.1016/j.jim.2016.11.001.

[60] Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, et al. High-Throughput Sequencing of T-Cell Receptors Reveals a Homogeneous Repertoire of Tumour-Infiltrating Lymphocytes in Ovarian Cancer: Tumour-Restricted and Homogeneous TILs in Ovarian Cancer. The Journal of Pathology. 2013;231(4):433–440. doi:10.1002/path.4260.

[61] Robert L, Tsoi J, Wang X, Emerson R, Homet B, Chodon T, et al. CTLA4 Blockade Broadens the Peripheral T-Cell Receptor Repertoire. Clinical Cancer Research. 2014;20(9):2424–2432. doi:10.1158/1078-0432.CCR-13-2648.

[62] DeWitt WS, Emerson RO, Lindau P, Vignali M, Snyder TM, Desmarais C, et al. Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection. Journal of Virology. 2015;89(8):4517–4526. doi:10.1128/JVI.03474-14.

[63] Clemente MJ, Przychodzen B, Jerez A, Dienes BE, Afable MG, Husseinzadeh H, et al. Deep Sequencing of the T-Cell Receptor Repertoire in CD8+ T-Large Granular Lymphocyte Leukemia Identifies Signature Landscapes. Blood. 2013;122(25):4077–4085. doi:10.1182/blood-2013-05-506386.

[64] Calis JJA, Rosenberg BR. Characterizing Immune Repertoires by High Throughput Sequencing: Strategies and Applications. Trends in Immunology. 2014;35(12):581–590. doi:10.1016/j.it.2014.09.004.

[65] Dash P, McClaren JL, Oguin TH, Rothwell W, Todd B, Morris MY, et al. Paired Analysis of TCRα and TCRβ Chains at the Single-Cell Level in Mice. Journal of Clinical Investigation. 2011;121(1):288–295. doi:10.1172/JCI44752.

[66] Kim SM, Bhonsle L, Besgen P, Nickel J, Backes A, Held K, et al. Analysis of the Paired TCR α- and β-Chains of Single Human T Cells. PLoS ONE. 2012;7(5):e37338. doi:10.1371/journal.pone.0037338.

[67] Cukalac T, Kan WT, Dash P, Guan J, Quinn KM, Gras S, et al. Paired TCR$\alpha\beta$ Analysis of Virus-Specific CD8 $^+$ T Cells Exposes Diversity in a Previously Defined 'Narrow' Repertoire. Immunology and Cell Biology. 2015;93(9):804–814. doi:10.1038/icb.2015.44.

[68] Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H. Single-Cell Based High-Throughput Sequencing of Full-Length Immunoglobulin Heavy and Light Chain Genes: New Technology. European Journal of Immunology. 2014;44(2):597–603. doi:10.1002/eji.201343917.

[69] Han A, Glanville J, Hansmann L, Davis MM. Linking T-Cell Receptor Sequence to Functional Phenotype at the Single-Cell Level. Nature Biotechnology. 2014;32(7):684–692. doi:10.1038/nbt.2938.

[70] Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T Cell Fate and Clonality Inference from Single-Cell Transcriptomes. Nature Methods. 2016;13(4):329–332. doi:10.1038/nmeth.3800.

[71] Lindeman I, Emerton G, Sollid LM, Teichmann S, Stubbington MJT. BraCeR: Reconstruction of B-Cell Receptor Sequences and Clonality Inference from Single-Cell RNA-Sequencing. 2017; p. –. doi:10.1101/185504.

[72] DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-Throughput Sequencing of the Paired Human Immunoglobulin Heavy and Light Chain Repertoire. Nature Biotechnology. 2013;31(2):166–169. doi:10.1038/nbt.2492.

[73] Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-Cell Receptor Chains via Emulsion PCR: New Technology. European Journal of Immunology. 2013;43(9):2507–2515. doi:10.1002/eji.201343453.

[74] DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-Depth Determination and Analysis of the Human Paired Heavy- and Light-Chain Antibody Repertoire. Nature Medicine. 2015;21(1):86–91. doi:10.1038/nm.3743.

[75] McDaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G. Ultra-High-Throughput Sequencing of the Immune Receptor Repertoire from Millions of Lymphocytes. Nature Protocols. 2016;11(3):429–442. doi:10.1038/nprot.2016.024.

[76] Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal Antibodies Isolated without Screening by Analyzing the Variable-Gene Repertoire of Plasma Cells. Nature Biotechnology. 2010;28(9):965–969. doi:10.1038/nbt.1673.

[77] Lee ES, Thomas PG, Mold JE, Yates AJ. Identifying T Cell Receptors from High-Throughput Sequencing: Dealing with Promiscuity in TCRα and TCRβ Pairing. PLOS Computational Biology. 2017;13(1):e1005313. doi:10.1371/journal.pcbi.1005313.

[78] Hogan T, Gossel G, Yates AJ, Seddon B. Temporal Fate Mapping Reveals Age-Linked Heterogeneity in Naive T Lymphocytes in Mice. Proceedings of the National Academy of Sciences. 2015;112(50):E6917–E6926. doi:10.1073/pnas.1517246112.

[79] Gossel G, Hogan T, Cownden D, Seddon B, Yates AJ. Memory CD4 T Cell Subsets Are Kinetically Heterogeneous and Replenished from Naive T Cells at High Levels. eLife. 2017;6. doi:10.7554/eLife.23013.

[80] den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, Bregje de Boer A, et al. Maintenance of Peripheral Naive T Cells Is Sustained by Thymus Output in Mice but Not Humans. Immunity. 2012;36(2):288–297. doi:10.1016/j.immuni.2012.02.006.

[81] Vrisekoop N, den Braber I, de Boer AB, Ruiter AFC, Ackermans MT, van der Crabben SN, et al. Sparse Production but Preferential Incorporation of Recently Produced Naive T Cells in the Human Peripheral Pool. Proceedings of the National Academy of Sciences. 2008;105(16):6115–6120. doi:10.1073/pnas.0709713105.

[82] Westera L, Drylewicz J, den Braber I, Mugwagwa T, van der Maas I, Kwast L, et al. Closing the Gap between T-Cell Life Span Estimates from Stable Isotope-Labeling Studies in Mice and Humans. Blood. 2013;122(13):2205–2212. doi:10.1182/blood-2013-03-488411.

[83] McCune JM, Hanley MB, Cesar D, Halvorsen R, Hoh R, Schmidt D, et al. Factors Influencing T-Cell Turnover in HIV-1–seropositive Patients. Journal of Clinical Investigation. 2000;105(5):R1–R8. doi:10.1172/JCI8647.

[84] Hellerstein MK, Hoh RA, Hanley MB, Cesar D, Lee D, Neese RA, et al. Subpopulations of Long-Lived and Short-Lived T Cells in Advanced HIV-1 Infection. Journal of Clinical Investigation. 2003;112(6):956–966. doi:10.1172/JCI17533.

[85] Macallan DC, Asquith B, Irvine AJ, Wallace DL, Worth A, Ghattas H, et al. Measurement and Modeling of Human T Cell Kinetics. European Journal of Immunology. 2003;33(8):2316–2326. doi:10.1002/eji.200323763.

[86] Macallan DC, Wallace D, Zhang Y, de Lara C, Worth AT, Ghattas H, et al. Rapid Turnover of Effector–Memory CD4 $^+$ T Cells in Healthy Humans. The Journal of Experimental Medicine. 2004;200(2):255–260. doi:10.1084/jem.20040341.

[87] Wallace DL, Zhang Y, Ghattas H, Worth A, Irvine A, Bennett AR, et al. Direct Measurement of T Cell Subset Kinetics in Vivo in Elderly Men and Women. Journal of Immunology (Baltimore, Md: 1950). 2004;173(3):1787–1794.

[88] Ribeiro R. Modeling Deuterated Glucose Labeling of T-Lymphocytes. Bulletin of Mathematical Biology. 2002;64(2):385–405. doi:10.1006/bulm.2001.0282.

[89] De Boer RJ, Perelson AS, Ribeiro RM. Modelling Deuterium Labelling of Lymphocytes with Temporal and/or Kinetic Heterogeneity. Journal of The Royal Society Interface. 2012;9(74):2191–2200. doi:10.1098/rsif.2012.0149.

[90] De Boer RJ, Ganusov VV, Milutinović D, Hodgkin PD, Perelson AS. Estimating Lymphocyte Division and Death Rates from CFSE Data. Bulletin of Mathematical Biology. 2006;68(5):1011–1031. doi:10.1007/s11538-006-9094-8.

[91] Choo DK, Murali-Krishna K, Anita R, Ahmed R. Homeostatic Turnover of Virus-Specific Memory CD8 T Cells Occurs Stochastically and Is Independent of CD4 T Cell Help. The Journal of Immunology. 2010;185(6):3436–3444. doi:10.4049/jimmunol.1001421.

[92] Buchholz VR, Flossdorf M, Hensel I, Kretschmer L, Weissbrich B, Graf P, et al. Disparate Individual Fates Compose Robust CD8+ T Cell Immunity. Science. 2013;340(6132):630–635. doi:10.1126/science.1235454.

[93] Asquith B, Debacq C, Macallan DC, Willems L, Bangham CRM. Lymphocyte Kinetics: The Interpretation of Labelling Data. Trends in Immunology. 2002;23(12):596–601.

[94] De Boer RJ, Perelson AS. Quantifying T Lymphocyte Turnover. Journal of Theoretical Biology. 2013;327:45–87. doi:10.1016/j.jtbi.2012.12.025.

[95] Ganusov VV, De Boer RJ. A Mechanistic Model for Bromodeoxyuridine Dilution Naturally Explains Labelling Data of Self-Renewing T Cell Populations. Journal of the Royal Society, Interface. 2013;10(78):20120617. doi:10.1098/rsif.2012.0617.

[96] Asquith B, Borghans JAM. Modelling Lymphocyte Dynamics In Vivo. In: Molina-París C, Lythe G, editors. Mathematical Models and Immune Cell Biology. New York: Springer; 2011.

[97] McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. No. 122 in Chapman & Hall/CRC texts in statistical science series. Boca Raton: CRC Press/Taylor & Francis Group; 2016.

[98] Ganusov VV. Strong Inference in Mathematical Modeling: A Method for Robust Science in the Twenty-First Century. Frontiers in Microbiology. 2016;7. doi:10.3389/fmicb.2016.01131.

[99] Reinius B, Mold JE, Ramsköld D, Deng Q, Johnsson P, Michaëlsson J, et al. Analysis of Allelic Expression Patterns in Clonal Somatic Cells by Single-Cell RNA-Seq. Nature Genetics. 2016;48(11):1430–1435. doi:10.1038/ng.3678.

[100] Picelli S, Björklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells. Nature Methods. 2013;10(11):1096–1098. doi:10.1038/nmeth.2639.

[101] Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects. Genome Research. 2014;24(12):2033–2040. doi:10.1101/gr.177881.114.

[102] Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, et al. MiTCR: Software for T-Cell Receptor Sequencing Data Analysis. Nature Methods. 2013;10(9):813–814. doi:10.1038/nmeth.2555.

[103] Team RC. R: A Language and Environment for Statistical Computing; 2017.

[104] Berger JO, Berry DA. Statistical Analysis and the Illusion of Objectivity. American Scientist. 1988;76(2):159–165.

[105] Gelman A. Prior Information, Not Prior Belief; 2015.

[106] Gelman A. Bayesian Data Analysis. Third edition ed. Chapman & Hall/CRC texts in statistical science. Boca Raton: CRC Press; 2014.

[107] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. *Stan* : A Probabilistic Programming Language. Journal of Statistical Software. 2017;76(1). doi:10.18637/jss.v076.i01.

[108] Vehtari A, Gelman A, Gabry J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. Statistics and Computing. 2017;27(5):1413–1432. doi:10.1007/s11222-016-9696-4.

[109] Piironen J, Vehtari A. Comparison of Bayesian Predictive Methods for Model Selection. Statistics and Computing. 2017;27(3):711–735. doi:10.1007/s11222-016-9649-y.

[110] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York, NY: Springer; 2010.

[111] Yates A, Saini M, Mathiot A, Seddon B. Mathematical Modeling Reveals the Biological Program Regulating Lymphopenia-Induced Proliferation. The Journal of Immunology. 2008;180(3):1414–1422. doi:10.4049/jimmunol.180.3.1414.

[112] White JT, Cross EW, Kedl RM. Antigen-Inexperienced Memory CD8+ T Cells: Where They Come from and Why We Need Them. Nature Reviews Immunology. 2017;17(6):391–400. doi:10.1038/nri.2017.34.

[113] Haluszczak C, Akue AD, Hamilton SE, Johnson LDS, Pujanauski L, Teodorovic L, et al. The Antigen-Specific CD8 $^{+}$ T Cell Repertoire in Unimmunized Mice Includes Memory Phenotype Cells Bearing Markers of Homeostatic Expansion. The Journal of Experimental Medicine. 2009;206(2):435–448. doi:10.1084/jem.20081829.

[114] Wyss L, Stadinski BD, King CG, Schallenberg S, McCarthy NI, Lee JY, et al. Affinity for Self Antigen Selects Treg Cells with Distinct Functional Properties. Nature Immunology. 2016;17(9):1093–1101. doi:10.1038/ni.3522.

[115] Le Campion A, Bourgeois C, Lambolez F, Martin B, Leaument S, Dautigny N, et al. Naive T Cells Proliferate Strongly in Neonatal Mice in Response to Self-Peptide/Self-MHC Complexes. Proceedings of the National Academy of Sciences. 2002;99(7):4538–4543. doi:10.1073/pnas.062621699.

[116] Ganusov VV, De Boer RJ. Do Most Lymphocytes in Humans Really Reside in the Gut? Trends in Immunology. 2007;28(12):514–518. doi:10.1016/j.it.2007.08.009.

[117] Jenkins MK, Chu HH, McLachlan JB, Moon JJ. On the Composition of the Preimmune Repertoire of T Cells Specific for Peptide–Major Histocompatibility Complex Ligands. Annual Review of Immunology. 2010;28(1):275–294. doi:10.1146/annurev-immunol-030409-101253.

[118] Lee M, Mandl JN, Germain RN, Yates AJ. The Race for the Prize: T-Cell Trafficking Strategies for Optimal Surveillance. Blood. 2012;120(7):1432–1438. doi:10.1182/blood-2012-04-424655.

[119] Ogawa K, Hirai M, Katsube T, Murayama M, Hamaguchi K, Shimakawa T, et al. Suppression of Cellular Immunity by Surgical Stress. Surgery. 2000;127(3):329–336. doi:10.1067/msy.2000.103498.

[120] Inada T, Yamanouchi Y, Jomura S, Sakamoto S, Takahashi M, Kambara T, et al. Effect of Propofol and Isoflurane Anaesthesia on the Immune Response to Surgery. Anaesthesia. 2004;59(10):954–959. doi:10.1111/j.1365-2044.2004.03837.x.

[121] Vasileiou I, Xanthos T, Koudouna E, Perrea D, Klonaris C, Katsargyris A, et al. Propofol: A Review of Its Non-Anaesthetic Effects. European Journal of Pharmacology. 2009;605(1-3):1–8. doi:10.1016/j.ejphar.2009.01.007.

[122] Kim M, Kim M, Kim N, D'Agati VD, Emala CW, Lee HT. Isoflurane Mediates Protection from Renal Ischemia-Reperfusion Injury via Sphingosine Kinase and Sphingosine-1-Phosphate-Dependent Pathways. American Journal of Physiology-Renal Physiology. 2007;293(6):F1827–F1835. doi:10.1152/ajprenal.00290.2007.

[123] Majumdar S, Deobagkar-Lele M, Adiga V, Raghavan A, Wadhwa N, Ahmed SM, et al. Differential Susceptibility and Maturation of Thymocyte Subsets during Salmonella Typhimurium Infection: Insights on the Roles of Glucocorticoids and Interferon-Gamma. Scientific Reports. 2017;7:40793. doi:10.1038/srep40793.

[124] Puffer RC, Murphy M, Maloney P, Kor D, Nassr A, Freedman B, et al. Increased Total Anesthetic Time Leads to Higher Rates of Surgical Site Infections in Spinal Fusions:. SPINE. 2017;42(11):E687–E690. doi:10.1097/BRS.0000000000001920.

[125] Tsai PS, Hsu CS, Fan YC, Huang CJ. General Anaesthesia Is Associated with Increased Risk of Surgical Site Infection after Caesarean Delivery Compared with Neuraxial Anaesthesia: A Population-Based Study. British Journal of Anaesthesia. 2011;107(5):757–761. doi:10.1093/bja/aer262.

[126] Bancroft T, Du C, Nettleton D. Estimation of False Discovery Rate Using Sequential Permutation p -Values: Sequential Permutation p -Values. Biometrics. 2013;69(1):1–7. doi:10.1111/j.1541-0420.2012.01825.x.

[127] Padovan E, Casorati G, Dellabona P, Meyer S, Brockhaus M, Lanzavecchia A. Expression of Two T Cell Receptor Alpha Chains: Dual Receptor T Cells. Science (New York, NY). 1993;262(5132):422–424.

[128] Elliott JI, Altmann DM. Dual T Cell Receptor Alpha Chain T Cells in Autoimmunity. The Journal of Experimental Medicine. 1995;182(4):953–959.

[129] Tuovinen H, Salminen JT, Arstila TP. Most Human Thymic and Peripheral-Blood CD4+CD25+ Regulatory T Cells Express 2 T-Cell Receptors. Blood. 2006;108(13):4063–4070. doi:10.1182/blood-2006-04-016105.

[130] He X, Janeway CA, Levine M, Robinson E, Preston-Hurlburt P, Viret C, et al. Dual Receptor T Cells Extend the Immune Repertoire for Foreign Antigens. Nature Immunology. 2002;3(2):127–134. doi:10.1038/ni751.

[131] Eltahla AA, Rizzetto S, Pirozyan MR, Betz-Stablein BD, Venturi V, Kedzierska K, et al. Linking the T Cell Receptor to the Single Cell Transcriptome in Antigen-Specific Human T Cells. Immunology and Cell Biology. 2016;94(6):604–611. doi:10.1038/icb.2016.16.

[132] Egerton M, Scollay R, Shortman K. Kinetics of Mature T-Cell Development in the Thymus. Proceedings of the National Academy of Sciences of the United States of America. 1990;87(7):2579–2582.

[133] Huesmann M, Scott B, Kisielow P, von Boehmer H. Kinetics and Efficacy of Positive Selection in the Thymus of Normal and T Cell Receptor Transgenic Mice. Cell. 1991;66(3):533–540.

[134] Thomas-Vaslin V, Altes HK, de Boer RJ, Klatzmann D. Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. Journal of Immunology (Baltimore, Md: 1950). 2008;180(4):2240–2250.

[135] Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A Mechanism for TCR Sharing between T Cell Subsets and Individuals Revealed by Pyrosequencing. Journal of Immunology (Baltimore, Md: 1950). 2011;186(7):4285–4294. doi:10.4049/jimmunol.1003898.

[136] Kuhn HW. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly. 1955;2(1-2):83–97. doi:10.1002/nav.3800020109.

[137] Callan MF, Annels N, Steven N, Tan L, Wilson J, McMichael AJ, et al. T Cell Selection during the Evolution of CD8+ T Cell Memory in Vivo. European Journal of Immunology. 1998;28(12):4382–4390. doi:10.1002/(SICI)1521-4141(199812)28:12<4382::AID-IMMU4382>3.0.CO;2-Z.

[138] Silins SL, Cross SM, Krauer KG, Moss DJ, Schmidt CW, Misko IS. A Functional Link for Major TCR Expansions in Healthy Adults Caused by Persistent Epstein-Barr Virus Infection. The Journal of Clinical Investigation. 1998;102(8):1551–1558. doi:10.1172/JCI4225.

[139] Waldrop SL, Davis KA, Maino VC, Picker LJ. Normal Human CD4+ Memory T Cells Display Broad Heterogeneity in Their Activation Threshold for Cytokine Synthesis. Journal of Immunology (Baltimore, Md: 1950). 1998;161(10):5284–5295.

[140] Sester M, Sester U, Gärtner B, Kubuschok B, Girndt M, Meyerhans A, et al. Sustained High Frequencies of Specific CD4 T Cells Restricted to a Single Persistent Virus. Journal of Virology. 2002;76(8):3748–3755.

[141] Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, et al. Naive CD4(+) T Cell Frequency Varies for Different Epitopes and Predicts Repertoire Diversity and Response Magnitude. Immunity. 2007;27(2):203–213. doi:10.1016/j.immuni.2007.07.007.

[142] Manlove D. Algorithmics of Matching under Preferences. No. vol. 2 in Series on theoretical computer science. Singapore ; Hackensack, NJ: World Scientific Pub. Co; 2013.

[143] Gusfield D, Irving RW. The Stable Marriage Problem: Structure and Algorithms. Foundations of computing. Cambridge, Mass: MIT Press; 1989.

[144] Gale D, Shapley LS. College Admissions and the Stability of Marriage. The American Mathematical Monthly. 1962;69(1):9. doi:10.2307/2312726.

[145] Murugan A, Mora T, Walczak AM, Callan CG. Statistical Inference of the Generation Probability of T-Cell Receptors from Sequence Repertoires. Proceedings of the National Academy of Sciences. 2012;109(40):16161–16166. doi:10.1073/pnas.1212755109.

[146] Desponds J, Mora T, Walczak AM. Fluctuating Fitness Shapes the Clone-Size Distribution of Immune Repertoires. Proceedings of the National Academy of Sciences. 2016;113(2):274–279. doi:10.1073/pnas.1512977112.

[147] Hammerbacher J, Snyder A. Informatics for Cancer Immunotherapy. Annals of Oncology. 2017;28(suppl_12):xii56–xii73. doi:10.1093/annonc/mdx682.

[148] Wong GK, Heather JM, Barmettler S, Cobbold M. Immune Dysregulation in Immunodeficiency Disorders: The Role of T-Cell Receptor Sequencing. Journal of Autoimmunity. 2017;80:1–9. doi:10.1016/j.jaut.2017.04.002.

[149] Zhang L, Cham J, Paciorek A, Trager J, Sheikh N, Fong L. 3D: Diversity, Dynamics, Differential Testing – a Proposed Pipeline for Analysis of next-Generation Sequencing T Cell Repertoire Data. BMC Bioinformatics. 2017;18(1). doi:10.1186/s12859-017-1544-9.

[150] Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In: Christiansen FT, Tait BD, editors. Immunogenetics. vol. 882. Totowa, NJ: Humana Press; 2012. p. 569–604.

[151] Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: A Tool for Fast, Efficient Gene Assignment in T-Cell Receptor Sequences Using a Finite State Machine. Bioinformatics. 2013;29(5):542–550. doi:10.1093/bioinformatics/btt004.

[152] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: Software for Comprehensive Adaptive Immunity Profiling. Nature Methods. 2015;12(5):380–381. doi:10.1038/nmeth.3364.

[153] Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K, et al. IMSEQ—a Fast and Error Aware Approach to Immunogenetic Sequence Analysis. Bioinformatics. 2015;31(18):2963–2971. doi:10.1093/bioinformatics/btv309.

[154] Gerritsen B, Pandit A, Andeweg AC, de Boer RJ. RTCR: A Pipeline for Complete and Accurate Recovery of T Cell Repertoires from High Throughput Sequencing Data. Bioinformatics. 2016;32(20):3098–3106. doi:10.1093/bioinformatics/btw339.

[155] Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Briggs AW, Vigneault F, et al. T-Cell Receptor $A\beta$ Chain Pairing Is Associated with CD4 and CD8 Lineage Specification. 2018; p. –. doi:10.1101/293852.