Quantitative approaches for profiling the T cell receptor repertoire in human tissues

Boris Grinshpun

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Quantitative approaches for profiling the T cell receptor repertoire in human tissues

Boris Grinshpun

The study of B and T cell receptor repertoires from high throughput sequencing is a recent development that allows for unprecedented resolution and quantification of the adaptive immune response. The immense diversity and long tailed distribution of these repertoires has up until now limited such studies to expanded clonal signatures or to analysis of imprecise signals with limited dynamic range collected by techniques such as radioactive and fluorescent labeling. This thesis presents a number of quantitative methods to characterize the repertoire and examine the questions of sequence diversity and inter-repertoire divergence of T cell repertoires. These approaches attempt to accurately parametrize the inherent distribution of T cell clones drawing from statistical tools derived from ecological literature and information theory.

The methods presented are applied to T cell analyses of various tissue compartments of the human body, including peripheral blood mononucleocytes, thymic tissues, spleen, inguinal lymph nodes, lung lymph nodes and the brain. A number of applications are explored with strong implications for translational use in medicine. Novel insights are made into the mechanism of maintenance and compartmentalization of naïve T cells from human donors of many different ages. Diversity and divergence of the tumor infiltrating sequence repertoire is measured in low grade gliomas and glioblastomas from cancer patients, and

potential sequence based biomarkers are assessed for studying glioma phenotype progression. A careful investigation of the immune response to allogeneic stimulus reveals the effect of HLA on sequence sharing and diversity of the alloresponse, and quantifies for the first time using sequence data the fraction of T cells in a repertoire that are alloreactive.

The use of repertoire sequencing and mathematical models within immunology is a new and emerging concept within the rapidly expanding field of systems immunology and will undoubtedly have a profound impact on the future of immunology research. It is hoped that the tools presented in this thesis will give insight into how to quantitatively explore the breadth and depth of the T cell receptor repertoire, and provide future directions for TCR repertoire analysis.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

My thesis, my publications, my sanity – these are all things that would not exist without the support of mentors, family, and countless friends.

First, I thank my thesis advisor, Dr. Yufeng Shen, for sticking with me these past six years, providing invaluable mentorship and guidance in conducting research and asking the right scientific questions. I also thank him for the numerous opportunities to present my work and to interact and learn from other scientists, through collaborations and at the various conferences I have attended.

I am grateful to my thesis committee – Drs. Raul Rabadan, Peter Sims, Itsik Pe'er, and Donna Farber, for their insights during committee meetings and their willingness to meet in private to discuss my work and provide suggestions and feedback.

I am also thankful to all of my experimental research collaborators for their hard work in generating the data that this entire thesis is about. In particular I wish to thank Dr. Jennifer Sims and Dr. Susan DeWolf for their frequent mentorship and sage advice, as well as their primers on immunology.

All of the above people have taught me to be a better scientist, a more prudent researcher, and a resilient graduate student.

I am eternally and insufficiently grateful to my family. My parents, who have done everything in their power to support me throughout every decision I have ever made, my sister for her concern over my well being, and my grandmother who made sure I was always well fed.

At last, I thank the many friends and labmates who have shared with me the highs and lows of both graduate school and adult life. There are too many of you to name individually, but you know who you are. You guys reminded me how to have fun. Sine qua non.

You are all witnesses to my mistakes, successes, and growth over the course of this PhD.

Thank you all!

Boris Grinshpun
April 2017

Chapter 1

Introduction

The field of immunology began with the realization that infectious diseases have a microbial origin. This germ theory of disease first became accepted in the later half of the 19th century, proposed separately by Louis Pasteur, who isolated the microbe responsible for causing rabies, and by Robert Koch who identified the infectious agents responsible for both anthrax and tuberculosis [49]. In 1885, on the heels of this paradigm shift, Elie Metchnikoff, the discoverer of macrophages, posited the concept of a cellular immune system, the hosts' response to these pathogens. In 1897 his contemporary Paul Ehrlich put forth the theory that specialized molecules, which he called antibodies, targetted and eliminated pathogens with a lock and key specificity. Together Metchnikoff and Ehrlich first described the two branches of immunity, innate and adaptive, for which they shared the Nobel Prize in 1908 [76].

Intially, these ideas of cellular and humoral immunity were at odds with one another. While Metchnikoff arrived at the concept of immunity from the point of view of an evolutionary biologist, as a mechanism by which the cells of an organism distinguish self from non-self, Ehrlich came to his conclusion as a biochemist, postulating that the immune system is an antigen driven "natural selection" process in which cell surface receptors adapt to defend against disease causing agents. In 1957, Sir MacFarlane Burnet, published his

clonal selection theory, in which he combined the two ideas, hypothesizing that the cell was responsible for generating antibodies via a genetic mechanism (which we now know is VDJ recombination), Based on reinfection studies in animals he further introduced the idea that different cells produce different antibodies, and the concepts of immunological memory and antibody maturation [1, 15, 76], asserting that "Antibody production can continue long after the antigen responsible has disappeared from the body" and "Antibody production is a function not only of the cell originally stimulated but of its descendants" [15].

More than fifty years later, it is well understood by most immunologists that the immune system is incredibly complex and carries out a wide range of functions including pathogen recognition and elimination, memory, and self-regulation. New cell types and markers are discovered on a regular basis as anyone who has been to an immunology conference knows. Moreover, the composition of immune cells varies between individuals and across tissue types, and changes over the course of a lifetime. This variability is particularly pronounced in the B and T cells of the adaptive immune system. The enormous sequence space of antigen recognition receptors across these cell types results in specific subsets of the population expanding and differentiating in response to stimulation by antigen, and contracting as the result of competition and regulatory mechanisms designed to maintain homeostasis in a healthy organism.

Despite the gamut of functions and the sheer diversity that comprises the immune system, until recently the study of immunity, just as the rest of biology, has been a largely qualitative science. Geneticists focused on knocking out target genes in mice and other model organisms, biochemists on isolating their desired proteins. Knowledge of the components of the immune system came from histological staining, vaccination, and immunohistochemistry. The large amount of structural, genomic, and proteomic data that can now be generated allows for previously impossible statistical analyses and robust modeling that can identify complex new structures and nonlinear relationships, which are testable in a laboratory setting. [36, 9].

The first published analyses of the receptor repertoire, the focus of this dissertation, were obtained using the canon of biochemical assays including immunopreciptation, gel electrophoresis, and mass spectroscopy, identifying underlying cell surface protein structures, their variable and constant domains, and families of related proteins [2, 61, 80]. By the 1990's spectratyping assays combined with Sanger sequencing provided improved resolution, quantifying distributions of receptor length and the precise composition of nucleotides that encoded the receptor protein complex [3, 70]. Nonetheless, these methods were limited in their throughput, forcing researchers to characterize only select subsets of the full repertoire. High throughput sequence analysis of B and T cell receptor repertoires was realized in 2009 [82, 102], allowing for the first comprehensive screenshots of the diversity and antigen recognition potential of the adaptive immune system. In recent years, high throughput sequencing studies, including profiling of B and T cell receptors, have become a key feature within the growing fields of systems immunology and translational medicine. Many labs around the world now focus on the best way to probe at the diversity and dynamics of the immune response, and companies small and large seek to develop new diagnostic tools and targetted therapies.

This thesis consists of three aims (1) to discuss the goals and challenges of T cell receptor (TCR) repertoire sequencing (2) to describe new quantititative approaches for analysis of sequenced TCR repertoires with general advice and guidelines arrived at from first hand experience and (3) to present novel biological results from studies of T cells in human tissues performed in collaboration with experimental researchers. In the remainder of this introduction I will provide, for motivational purposes, a brief description of some relevant applications and "hot" topics within the growing field of systems immunology in order to elucidate how the field and the approaches described herein will improve our understanding of the immune system and our ability to treat disease in the future.

1.1 "Hot" Topics in Systems Immunology

1.1.1 Vaccination against infectious disease

Among the many scientific achievements of the 20th century have been the eradication of smallpox, and the near eradication of polio. Infectious diseases are no longer the leading cause of death among developed nations. Despite this progress infectious diseases continue to infect and claim many lives every year Table 1.1. There are instances of disease outbreaks like the 2014 spread of ebola virus in Western Africa, which causes deadly hemmorhagic fever and claimed over eleven thousand lives over three years [25], and the 2015 spread of mosquitoe-borne Zika virus which was found to cause microcephaly in newborns.

In developed countries like the United States, the effectiveness of the vaccine against seasonal influenza virus varies from year to year depending on which strain is dominant, and among senior citizens is only 50-60% effective, with significantly reduced efficacy [26]. Pathogens with high antigen variability (e.g. malaria) or that target the immune system itself (e.g. HIV) are also difficult to prevent against with conventional methods. Lastly, there is increasing concern of superbugs with antibiotic resistance developing due to human imposed selective pressure.

An important direction in systems immunology is understanding how the immune system responds to any given infection: recognition properties, downstream molecular signaling, formation of memory, and other complex molecular interactions which can be utilized to develop more effective and robust vaccines [14, 86]. There is further focus on understanding the means by which pathogens disrupt or evade the immune system, and developing novel treatments which undermine these mechanisms, thereby increasing vaccine success rates, as well as recovery and clearance rates for infected individuals [71, 40, 90].

Table 1.1: Global infectious disease statistics for 2015

Disease	Total New Cases	Total Deaths
HIV/AIDS	2.0e6	1.1e6
Tuberculosis	10.4e6	1.4e6
Influenza	3-5e6	2.5-5e5
Hepatitic C	3-4e6**	~7e5
Meningococcal meningitis	2.4e4	1903*

*from 2014, **from 2010 All data compiled from [68]

1.1.2 Treatment of autoimmune disease

Autoimmune diseases result from the immune system reacting to self tissues of the body as though they were foreign agents. B and T cells of the adaptive immune system trigger an immune response that leads to inflammation and tissue damage. Examples of autoimmune diseases include type 1 diabetes, Crohns disease, rheumatoid arthritis, and Grave's disease. In 2012 Approximately 23.5 million people in the United States suffered from at least one of the more than 80 recognized autoimmune disorders [31] and rates have continued to increase among developing countries. In most cases the exact mechanism by which an autoimmune response is triggered is unknown, and such diseases cannot be cured with conventional methods. The best current therapeutic treatments for most autoimmune diseases are targetted immunosuppressants and anti-inflammatory drugs that must be taken for the entire lifetime to suppress this deleterious immune response and reduce risk of subsequent flare ups.

Systems immunology approaches are being utilized to identify the complex pathways and molecular interactions leading to the formation of autoimmunity and to use this information to develop more effective treatments. Recently, pharmaceutical companies have begun to design monoclonal antibodies to target specific components of the autoimmune pathway. Examples include the drugs Remicade (infliximab) and Humira (adalimumab) which have been used with large success in patients with Crohn's disease and rheumatoid arthritis. There is still a great deal of work left to do to identify novel targetted therapies, identify

individuals with increased risk, and perhaps find cures or effective preventative measures for these diseases.

1.1.3 Cancer immunotherapy

In 2013, the journal Science declared cancer immunotherapy the breakthrough of the year [19]. The article described an experimental treatment for metastatic melanoma using a monoclonal antibody that targetted the CTLA-4 receptor on T cells. CTLA-4 functions as a brake on the destructive abilities of T cells. Many cancers, including melanoma produce a ligand that binds the CTLA-4 receptor thereby supressing the protective immune response. While the immune system plays a crucial role in the recognition and destruction of malignant cells, a large number of cancers are able to evade the body's natural defenses by altering their surrounding microenvironment to become hostile to immune cells or to dysregulate the signaling pathways that modulate immune response. For a review on this topic see Vinay, et .al. [100].

Since the Science article was published several more monoclonal antibody treatments, collectively called "checkpoint inhibitors", have been developed to target genes like PD-1 and CTLA-4 and block their ligand binding. Cancer therapy was again the theme of Science journal's 2017 issue for the week of March 17th, highlighting a number of potential therapeutic targets such as the RAS oncogene, PARP1 and 2, and other components involved in cancer signaling pathways [4, 21].

Another approach to cancer immunotherapy is the design of T cells with specific recognition sequences that strongly bind cancer antigens. These cells are grown *in vitro*, and then introduced into the body of the patient. In a 2010 trial, patients with leukemia were treated using T cells with engineered receptor sequences called chimeric antigen receptors (CARs) [19]. A number of other therapies have been tested to treat cancers [See Table 1.2]. In general, these approaches attempt to (1) prevent suppression of the immune response by cancer cells (2) increase tissue infilitration by the immune system, particularly T cells, into

Table 1.2: Current Approaches To Immunotherapy

Type	Mode of Action	Examples
Monoclonal Antibodies	Antibodies are manufactured to downregulate suppressive signaling or stimulate response against a cancer antigen.	– KRAS inactivation – Checkpoint inhibitors
Adoptive T cell transfer	T cells are cultured in a lab and then administered into the patient.	– Naturally occuring host T cells – Genetically modified T cells – Chimeric antigen receptors (CARs)
Cytokines	Immune signaling molecules are introduced to enhance and coordinate immune activity.	– Interferon signaling – IL-2 signaling and related cytokines.
Vaccines	Specific cells or proteins are introduced which prime the immune system against cancer cells.	– Patient tumor derived proteins – Cell line derived proteins – Dendritic cell activation

the tumor and (3) boost the anti-tumor activity of these cells.

Currently, many of these treatments have only been tested in a laboratory setting or are in the clinical trial phase. Among FDA approved drugs, which are largely monoclonal antibody based, success rates vary wildly across individuals and cancer types, with undesirable and sometimes fatal side effects leading to their use only as a last resort. However, a great deal of time and funding continues to be devoted towards identifying new and improved immunotherapeutic approaches. Reseachers have looked at transcriptomic, metabolomic, and epigenetic data to identify pathways of T cell activation and suppression in the tumor microenvironment in search of new drug targets [109, 37, 17, 72]. A plethora of studies in the last thirty years have focused on quantifying T cell subsets among tumor infiltrating lymphocyte populations as a marker of patients' response to cancer [33, 101, 69]. More recently the resolution provided by high throughput sequencing has yielded comprehensive receptor profiling studies that identifed sequence specific attributes among tumor infilitrating T cell populations [51, 88, 83]. The discovery of a cancer associated T cell signature or a

new drug target is perhaps the fastest growing and most lucrative application within the field of systems immunology.

1.1.4 Transplantation

According to the Mayo clinic, the five year success rate for a kidney transplant from a living donor is over 80%, and over 90% for liver transplants [58]. Surgeons are able to perform transplantation for various tissues and organs including heart and intestines, with varying levels of long term success. The biggest risk to the recipient is transplant rejection, where the immune system of the recipient attacks and damages the transplanted tissue. For this reason, transplant donor and recepient must be closely HLA matched (See Chapter 2), and the recepient is required to take immunosupressive drugs for some time after transplantion.

Transplant rejection results from the hosts immune response to the introduced tissue. The immune response to tissues from other members of the same species is known as alloreactivity. Despite the success of transplantation in the hospital setting our knowledge of the human alloresponse is limited. It is not well understood why some patients have lower levels of alloreactivity than others that are similarly HLA matched and on the same drug regimens. Similarly, we do not fully appreciate why certain organs have lower rates of rejection than others. The answer lies in the complex dynamic interplay of the immune system, including the distribution of cell types, variation in cytokine signaling, and individual differences like age, infection history, and repertoire composition. The resolution offered by current technologies allows us to probe these questions, and down the line will lead to the development of better surgical techniques, more effective drugs, and patient tailored drug regimens.

1.2 Thesis Outline

The rest of this thesis will be organized into six chapters:

Background material is presented in Chapter 2 in which all topics relevant to the original research of Chapters 3-6 will be introduced and key concepts and ideas will be explained.

The chapter will be broken down into (1) An overview of $\alpha\beta$–T cell maturation, functional subtypes, and role in the immune response, (2) Discussion of approaches to T cell sequencing and sequence analysis and (3) Overview of important statistical methods and formulas.

Chapter 3 will provide a discussion of the frequency distribution of the TCR repertoire. Using known properties of this distribution, the chapter will present and validate a semi-parametric method for analyzing clonal expansions from stimulated populations. A manuscript, *Quantifying the Size and Diversity of the Human Alloresponse via High-Throughput T Cell Receptor Sequencing*,is currently being prepared.

A study of the naïve repertoire from healthy donor lymphoid tissues is presented in Chapter 4. This is the first comprehensive study of naïve repertoire in healthy lymphoid tissues, looking at the effects that human aging has on thymic output and naïive repertoire diversity, as well as using sequenced repertoire data to investigate comparmentalization and maintenance of the repertoire in the tissues. Results have been published in [94]

Chapter 5 presents analysis of the T cell receptor repertoire in gliomas, and uses methods for analyzing TCR diversity and divergence to uncover previous unseen phenotypes associated with gliomas, as well a potential blood biomarker for tracking glioma status. Results are published in [89].

Chapter 6 quantifies the strength and diversity of the human alloresponse, providing some of the first estimates for the strength of the alloresponse and the role of HLA matching from TCR sequencing data. Results from this study are included in the above mentioned manuscript currently in preparation.

Chapter 7 will present overall conclusions and summarize points of interest. It will also discuss future research directions within the field of TCR repertoire profiling, for which the work presented in this thesis are a starting point.

Chapter 2

Background

This chapter provides an overview of key biological and mathematical concepts that are applied throughout the main chapters of the thesis. Section 2.1 discusses how mature T cells are formed and their effector function within the immune system. Section 2.2 discusses the current technology utilized to profile T cell receptor repertoires, and the benefits and drawbacks of these approaches. Section 2.3 gives an overview of statistical formulas, methods, and considerations applied to the study of diverse repertoires sampled from large populations.

2.1 $\alpha\beta$–T Cell Biology

T cells comprise the cell mediated branch of adaptive immunity. Specialized cell surface receptors function to bind and recognize distinct epitopes with high specificity initiating a robust immune response. The diversity of the $\alpha\beta$-T cell repertoire is estimated to comprise as many as 10^8–10^{11} distinct receptors out of a theoretical maximum 10^{15}–10^{20} [82, 62]. Each T cell commonly presents only a single receptor type, and the identity of this receptor is determined during the maturation phase of the T cell in a process called V(D)J recombination. T cells circulating throughout human

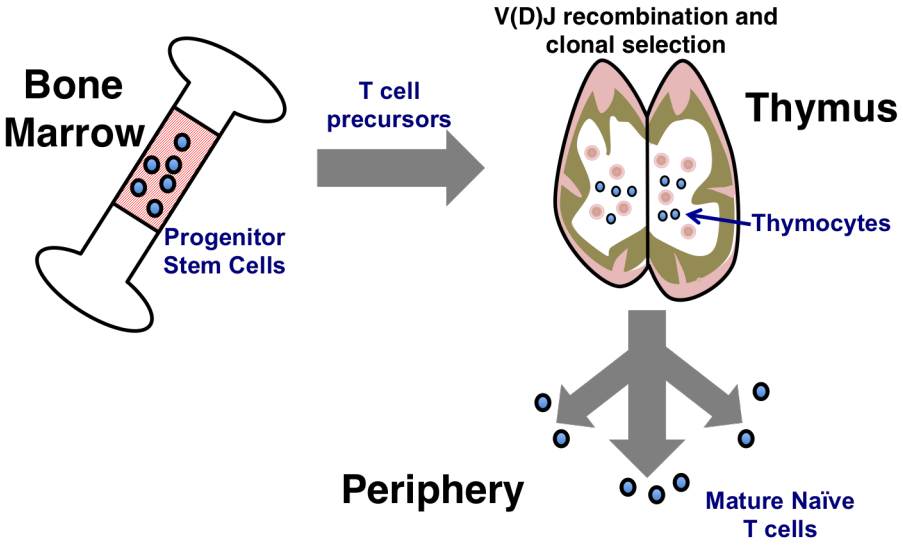

Figure 2.1: T cell progenitors migrate from the bone marrow into the thymus where they undergo V(D)J recombination and clonal selection to form mature naïve T cells.

tissues are capable of recognizing and responding to a nearly unlimited range of foreign peptides.

2.1.1 T cell maturation

All immune cells originate from multipotent hemapoetric stem cells that reside in the bone marrow [66]. Cells destined to become fully mature T cells migrate into the thymus and are referred to as thymocytes. It is within the unique microvenvironment of the thymus that thymocytes activate a transcriptional program which turns them into mature naïve T cells (Figure 2.1). During the maturation step rigorous selection processes remove cells with deleterious or unreactive receptors, leaving the survivors to circulate throughout the tissues of the body and survey for antigens. Thus, the two most crucial components of this transformation are V(D)J recombination and clonal selection.

V(D)J recombination is a cut and paste process in which genetic segments, also called cassettes, are selected and stitched together to form a complete genetic sequence encoding a T cell receptor chain (Figure 2.2). During this process, nucleotide deletions and insertions are introduced at the site of cassette joining, resulting in a great degree of sequence vari-

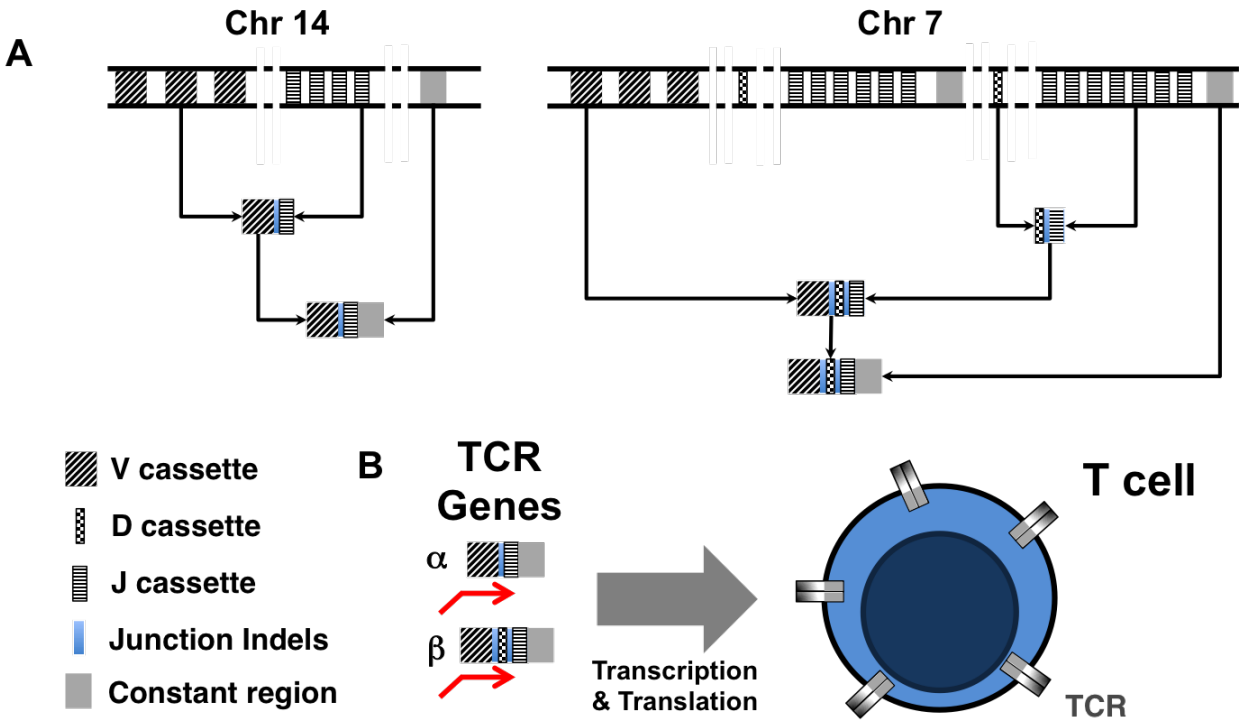

Figure 2.2: A. V(D)J recombination forms the α (left) and β (right) chains of the TCR repertoire. Distinct genetic segments are joined together with insertions and deletions introduced at the recombination junctions. These two events are responsible for the enormous repertoire divercity of T cell receptors. B. Every T cell encodes a distinct receptor heterodimer which is displayed as a protein on the cell surface.

ability in the junctions. The recombination occurs on chromosome 7 to form the β chain, and on chromosome 14 to form the α chain of the receptor heterodimer. Both chromosomes encode many distinct, but evolutionarily related variable (V) cassettes and joining (J) cassettes. The β chain also has two diversity (D) segments [See Table 2.1]. While we have comperehensive annotation for these cassettes in the human and mouse populations [97], it is important to note that this region, and the relatively long V cassette in particular, contain pseudogenes, related sequence families, and distinct allelic differences, so both annotation and receptor profiling techniques continue to be further curated and improved [20]. More recently, there has been increasing interest in identifying the cassette sequences of other animals for use in many applications, including the growing of organs for human transplantation and development of new immunotherapies [43, 79].

Table 2.1: Numbers of human $\alpha\beta$ – T cell receptor cassettes

Element	α (chr 14)			β (chr 7)		
	Always Functional	Never Functional	Allele Dependent	Always Functional	Never Functional	Allele Dependent
Variable (V)	43	9	2	42	29	6
Diversity (D)	0	0	0	2	0	0
Joining (J)	50	11	0	12	1	1

Compiled from [97]

V(D)J recombination, first β then α, is catalyzed by specialized enzymes encoded by recombination-activating genes (RAGs) that bind regions of double stranded DNA known as recombination signal sequences. These bound RAG proteins then bind to each other, creating a DNA hairpin loop that brings into proximity previously distant gene segments. The RAGs then induce DNA cleavage via random single strand nicks, cutting away unused cassettes within the loop. The adjacent gene segments are further processed and ligated together by enzymes involved in DNA repair, with insertions and deletions introduced by terminal deoxynucleotidyl transferase (TdT) [66]. The mature T cell expresses the recombined genes (and many other T cell specific proteins), forming a heterodimeric receptor on the cell surface. In addition to V, D and J gene segments, a constant (C) region is present which serves to tether the protein chains to the T cell surface. The process of clonal selection, which will not be described here in detail, eliminates all T cells that do not express functional T cell receptors or are strongly reactive to self peptides. This ensures that all T cells that exit the thymus and circulate throughout peripheral tissues are able to recognize foreign antigens without attacking the tissues of the body. Only 2-5% of all thymocytes pass clonal selection and exit the thymus as mature naïve T cells [64]. The ability to sample self-peptides of various tissues from all over the human body is a property unique to the thymic microenvironment and the mechanism is still not well understood.

Generation of new T cells decreases with age as the thymus undergoes a process of involution in which functional regions are converted into fatty tissue. The loss of naïve

diversity in human adults due to deterioration of the thymus is discussed in Chapter 4.

2.1.2 T cell recognition

T cells survey the body's tissues, sampling antigens displayed on cell surfaces. When the T cell binds an epitope belonging to a foreign antigen a transcriptional program is turned on, activating effector functions involved in destruction of the antigen source. However, in order for the T cell to be activated it must recognize both the epitope and the self-molecule to which it is bound, called the major histocompatibility complex (MHC), which in humans is encoded on chromosome 6 and is also known as the human leukocyte antigen (HLA). Only if the entire complex is detected will the T cell be activated (Figure 2.3A). Certain types of immune cells, dendritic cells in particular, are known as antigen presenting cells (APCs) and have a large number of MHC-peptide molecules on their cell surface for the purpose of activating T cells. The genetic region encoding the MHC is both polygenic and polymorphic, and as a result there is a great deal of variation of MHC composition across populations. Therefore, it is important to keep in mind that an individual's MHC composition can have a strong impact on how the immune system responds to a threat, and therefore should be taken into account in any study of T cell interactions.

The part of the T cell hetetodimer that recognizes the MHC molecule is located primarily on the V cassette, split across two regions called complementarity determining regions 1 and 2 (CDR1 and CDR2), which are close together in the folded protein structure. The epitope binds to complementarity determining region 3 (CDR3) which includes the end portions of both the V and J cassettes, and the junctions, making it the most diverse of the three and the main target for sequencing. All three CDRs are present on both the α and β chain, producing six hairpin loops in total, two CDR3 loops in the center, and two loops each for CDR1 and CDR2 on the outside, responsible for binding the bound peptide and the MHC molecule respectively [107, 6] (Figure 2.3B).

Several other cell surface signaling molecules are also necessary for successful T cell acti-

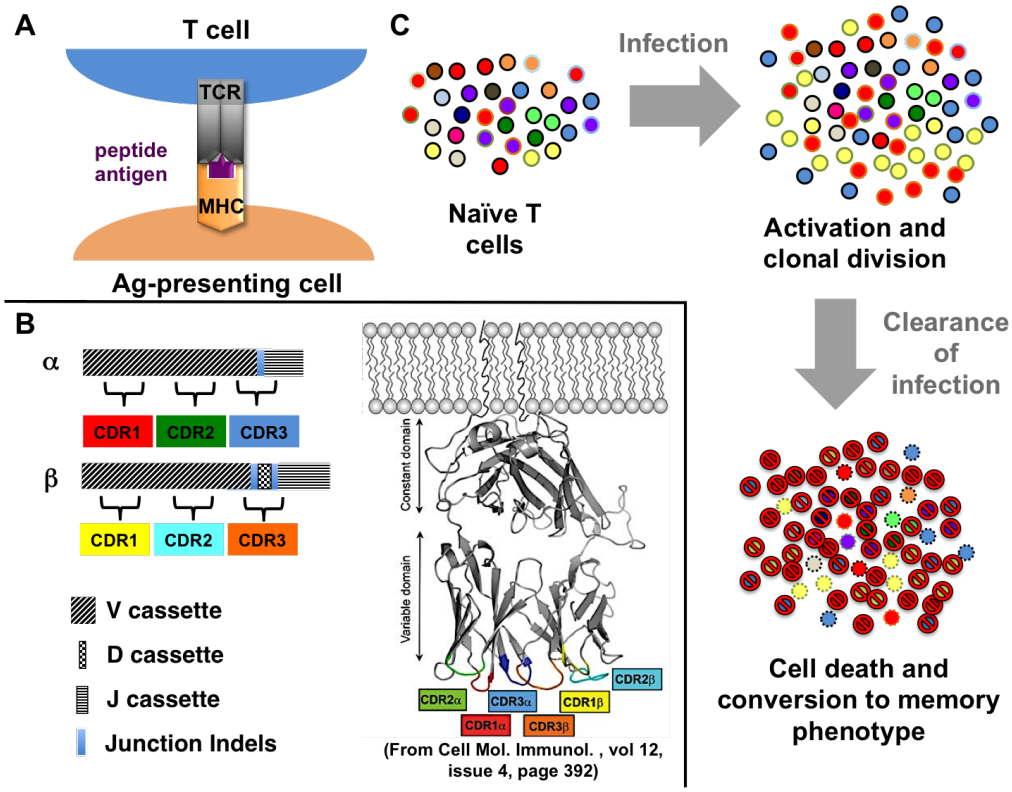

Figure 2.3: A. Peptide antigen is bound to the major histocompatibility complex (MHC) and presented on the surface of antigen presenting cells. Activation of the T cell requires both antigen and MHC to be bound by the TCR. B. Both chains of the TCR encode three hypervariable regions (CDR1, CDR2, CDR3). In the folded protein structure these regions come together at the ends of the heterodimer and are responsible for binding the MHC-peptide complex. The protein structure is reprinted from Figure 1 in [6]. C. Activated T cells undergo rapid clonal expansion. Upon clearance of the infection the majority of activated T cells die off with a small subset converting to a memory phenotype.

vation. The most important of these are the $CD4^+$ and $CD8^+$ receptors, which are discussed in Subsection 2.1.3. Upon activation, the T cell undergoes rapid expansion, which typically peaks 7-15 days after initial antigen stimulation, creating many clones with identical cell surface recognition receptors. For a given disease, multiple different foreign peptides are presented to the T cells. Additionally, antigen recognition does not function like a lock and key mechanism; instead, both the receptor chains and the peptide are flexible. As a result, during immune challenge, different T cells will respond and clonaly expand to varying degrees, producing subpopulations of T cell clones. Once the foreign antigen is cleared, most of

the activated T cells die off and the overall population decreases, in most cases leaving only a small fraction, 5–10%, of remaining activated T cells with a memory subtype (Figure 2.3C) [73].

2.1.3 T cell subsets and functions

There are many different T cell subsets that are defined based on their expression profiles, cell surface proteins, and specialized functions within the immune system. The nuanced and often imprecise differences between T cells makes classification very challenging, and will benefit in the future from careful single cell analysis and computational approaches. For the purposes of this thesis ,the key distinctions are $CD4^+$ vs $CD8^+$, naïve vs effector memory, and circulating vs resident.

All T cells exiting the thymus are antigen-inexperienced naïve cells. They are classified into two distinct categories, $CD4^+$ T cells and $CD8^+$ T cells depending on which of these cell surface receptors is present (double positives exist only in the thymus prior to clonal selection). In addition to the MHC-peptide being bound by TCR, activation of T cells also requires these co-receptors to interact with MHC. $CD4^+$ T cells bind MHC class II, while $CD8^+$ T cells interact with class I. The $CD4^+$ subset of T cells is known as helper T cells and fall into several subclasses, all of which serve the primary role of releasing chemical messengers (cytokines) that activate or suppress components of the immune system including other T cells, B cells, and macrophages. The $CD8^+$ T cells are known as cytotoxic T cells, and function by releasing cytotoxins including perforins and granzymes which lead directly to the death of targetted cells via apoptosis.

Both $CD4^+$ and $CD8^+$ T cell subsets have a memory phenotype created from a small group of activated T cells that remain post immune challenge Figure 2.3C. Memory T cells have a long lifespan with an average half life of 8–15 years [42]. Having previously encountered antigen, these T cells initiate a rapid and much stronger response if the antigen is encountered a second time. There is another subdivision of memory cells into T cell effector

Table 2.2: T cell subsets

Subset	Function	Cell Surface Markers
$CD4^+$	Helper T cells. Cytokine secretion for immune regulation.	$CD4^+$
$CD8^+$	Cytotoxic T cells. Destruction of, infected or otherwise compromised cells, including cancer cells.	$CD8^+$
Naive	Antigen inexperienced cells. Survey the body's tissues. Mostly present in lymphoid tissues.	$CCR7^+$, $CD45RA^+$, $CD45RO^-$
TEM	Cells remaining from previous T cell. response. Found in various peripheral tissues.	$CCR7^-$, $CD45RA^-$, $CD69^-$, $CD103^-$, $CD45RO^+$
TCM	Cells remaining from previous T cell. response. Largely present in lymphoid tissues.	$CCR7^+$, $CD45RA^-$, $CD69^-$, $CD103^-$, $CD45RO^+$
Resident	Non-circulating cells that are resident to a specific tissue site.	$CCR7^-$, $CD45RA^-$, $CD69^+$, $CD103^+$
Circulating	Cells that circulate between tissues.	$CD69^-$, $CD103^-$

Compiled from [32]

memory (TEM) and central memory (TCM) characterized by the expression of homing receptors, which determine the tissues they migrate into, and their circulation between tissue sites [84]. The extent of circulation between sites depends on how the T cells respond to chemokine signaling and can be categorized as tissue resident if they remain in a specific tissue for most of their lifetime, or circulating if they migrate between tissues.

Table 2.3: Numbers of T cells in human tissues

Tissue	Total $CD4^+$($x10^9$)	Total $CD8^+$($x10^9$)
Blood	5	2.5
Lymph nodes	95	38
Spleen	14	21
Thymus	37.5	12.5
Lamina propria of gut and lung	30	15

Data taken with permission of Elsevier, from Table 3 of [35]

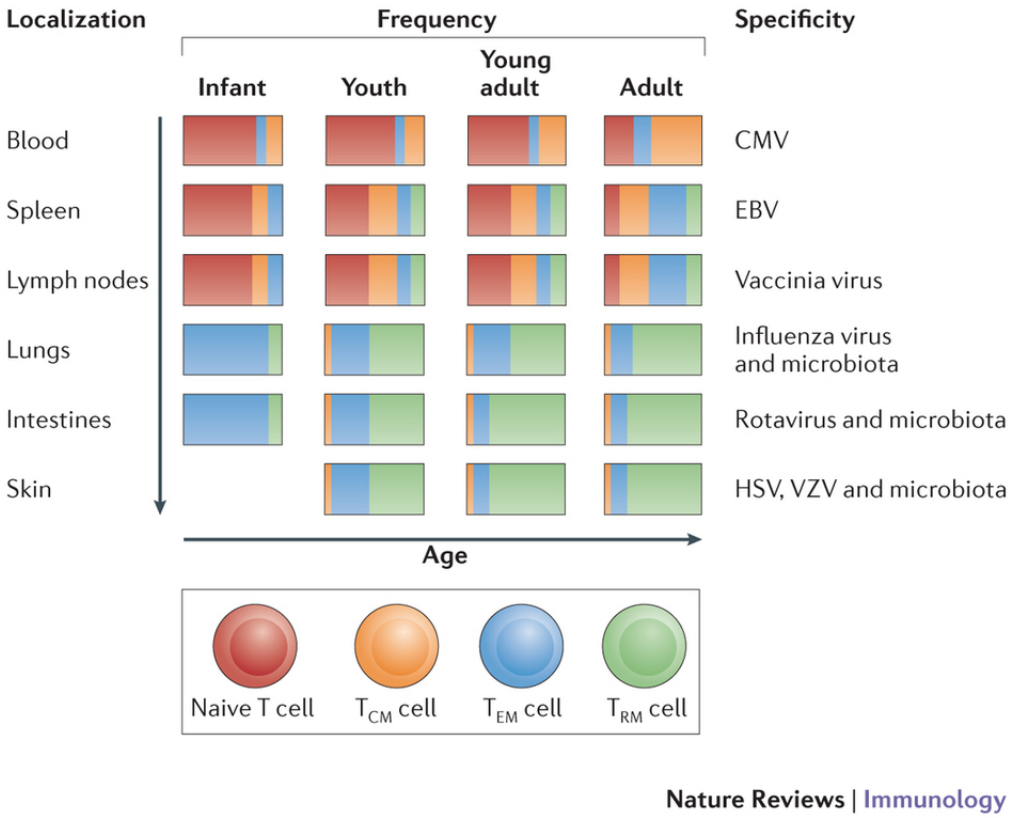

Figure 2.4: Frequencies of naïve and memory T cell subsets in different tissues separated by age group. Naive T cells are present within blood and lymphoid tissues, but are gradually replaced by memory T cells in adulthood. Also indicated are biases in T cell specificity for specific pathogen-derived antigens. Reprinted with permission of Nature Publishing Group, from Figure 4 in [32].

In the laboratory setting T cell subsets are typically identified and sorted into their subsets using cell surface proteins as biomarkers. As mentioned earlier, because gene expression is not a binary process this approach is not error-free, but generally reliable for studying bulk populations of cells. An overview of these subsets and their associated markers is presented in Table 2.2.

In addition to functional differences between T cells, tissues also employ distinct mechanisms for the homing and homeostatic maintenance of these T cell subsets, and the overall prevalence of these populations changes over time [35, 32, 93]. Estimates of overall $CD4^+$, $CD8^+$ T cell numbers in a typical young adult are indicated in Table 2.3 and further subdivision by subtypes is given in Figure 2.4. Thus, studies involving T cell populations must

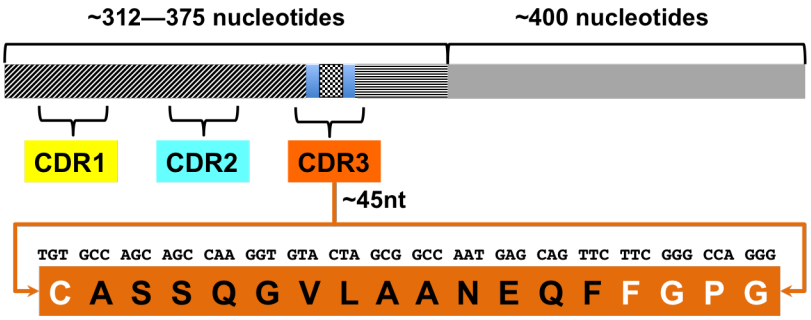

Figure 2.5: Each chain of the T cell receptor is approximately 720 nucleotides (240 amino acids long), with the variable region comprising approximately half the total length. The CDR3 region on average comprises fifteen amino acids and can be identified by a conserved cysteine (C) from the V gene and an FGXG motif from the J gene, where X can vary. Length information comes from the The T Cell Receptor FactsBook [50].

also consider the age of the subject and the location from which samples are collected.

2.2 TCR repertoire sequencing

The goal of TCR sequencing is to uniquely identify the $\alpha\beta$–T cell receptor chains present in a sample and determine its population wide clonal distribution. A way to uniquely define the T cell receptor that captures all elements contributing to its diversity is to use the combination of germline V and J segments, and the non-germline CDR3 region (See Figure 2.5). The non-germline CDR3 region is on average fifteen amino acids in length and defined by a conserved cysteine residue on the V gene segment, and an FGXG motif on the J gene segment, where X can vary. Due to the large diversity of TCRs a comprehensive survey of the population requires targetted amplification approaches to boost the TCR signal among all the other nucleotide fragments present in DNA sequencing. Similarly, an RNA sequencing approach will identify only highly abundant TCR transcripts, typically from active cells, and is therefore only useful in a limited subset of studies focused on populations with large monoclonal or oligoclonal expansions. This section discusses several sequencing approaches that have been developed to comprehensively sequence the bulk TCR repertoire.

2.2.1 Overview of current repertoire sequencing approaches

The 5' RACE (Rapid Amplification of cDNA Ends) system is used to amplify the RNA transcript encoding the TCR. Library preparation begins with isolating the T cells and RNA extraction. The RNA is reverse transcribed into cDNA and specialized sequencing primers are attached to the 3' end (template switching). A specially designed primer corresponding to the constant region of either α or β chains is attached to the 5' end of the cDNA molecule. PCR amplification is carried out from both ends, starting with the sequence template attached to the 3' end, and the primed 5' end on the constant region. This preferentially amplifies and selects the TCR sequences that are then used to prepare a final sequencing library. This approach was used in early TCR sequencing studies including Freeman, et.al [34]. Currently the company Clontech Laboratories, Inc. uses this approach for sequencing of the α and β chains of the TCR.

A second strategy for sequencing the TCR requires design of specialized primers that are complementary to the V and J recombination cassettes. Following primer annealing and extension, sequencing primers are attached to generate the library. This approach can be applied to both cDNA and genomic DNA and primer kits have been developed by several companies including Adaptive Biotechnologies and iRepertoire, Inc.

2.2.2 Comparisons and limitations

There are several considerations that must be taken into account when using these sequencing methodologies. The first is whether to use DNA or cDNA to prepare the sequencing library. Each T cell only contains a single recombined TCR genomic sequence for either chain. As a result, heavy amplification is necessary to create a sufficiently large library, but accurate quantification is possible using unique molecular identifiers to determine the true number of starting molecular templates. Because of the intron present between the variable and constant regions 5'RACE cannot be used for this task. Another difficulty then is to design unique primers for all the known V and J segments, taking into account related subfamilies

and degenerate sequences, and to then correct for PCR errors due to differences in primer annealing rates. Additionally, this approach has been limited to only β chain sequencing, due to "leaky" allelic exclusion of the α chain, which makes it difficult to ensure that the amplified DNA captures all recombined alpha chains.

In contrast, using cDNA produces more genetic material and requires less amplification, and will accurately capture α chain proportions. However, collecting cDNA requires the extra step of using an enzyme to carry out reverse transcription (RT), and the fidelity and processivity of this enzyme must be taken into account. Additionally, transcript numbers will be proportional to T cell activity, and precise quantification is therefore far more challenging in a bulk sequencing experiment. In using cDNA a second consideration is whether to use 5'RACE or VJ priming. Due to degeneracy in V cassette sequences, at least 100 base pairs of the V cassette upstream of the CDR3 must be sequenced for accurate identification, which requires far longer sequence lengths if starting from the constant region. However, this avoids having to develop specialized primers for all V, J cassettes and accounting for differences in annealing rates. 5'RACE can potentially be used to identify new cassettes and alleles since the annealing sequences are not specifically designed for known segments.

For a thorough review see Woodsworth, et.al [106].

2.2.3 Single cell sequencing

Analysis of T cell repertoire is inherently a single cell challenge. Sequencing every T cell individually avoids the major difficulties involved in accurately quantifying CDR3 numbers. It also allows for the α and β chains of a specific T cell to be paired together. Currently, the major challenge of using single cell technology is the time and cost of achieving the throughput required to accurately represent the size and diversity of the repertoire.

2.3 Population statistics

Statistical approaches for T cell receptor repertoire analysis borrow heavily from ecological literature studying the properties of animal and insect populations. In both cases it is necessary to deal with the challenge of having many diverse species of which only a small subset can be observed. There is interest in developing methods to effectively quantify population diversity, make reliable comparisons across samples and environmental conditions, and to develop inference techniques that fill in the gaps in the observed data. While ecological studies must account for imprecision in taxonomic classification, one of the key computational challenges of TCR sequencing is to account for errors from the need to isolate and amplify molecules for sequencing and inherent biases from the different technologies available to accomplish this task as discussed in Subsection 2.2.2. This section discusses the mathematical assumptions and methods used to perform the data analyses described in the remaining chapters of this thesis.

2.3.1 Power Laws

Datasets from several different collaborations that are discussed in further chapters has suggested that the bulk of clonal frequencies in the TCR repertoire is well described by a discrete power law distribution. This agrees with findings from several other quantitative and modeling studies of TCR repertoire [10, 23, 22]. Various types of power law distributions have been observed in a number of natural phenomena including studies of word frequencies, earthquake magnitudes, neuronal avalanches, protein networks, populations of bird and insect species, and other cases in which most observed elements are exceedingly rare [111, 92, 28, 47, 103, 75, 48, 110].

The power law is a long tailed distribution, with several definitons all of which have the property that when plotted on a log-log scale there is a linear relationship between the dependent and independent variable. Zipf's law describes the relationship between the rank of an element and its frequency of occurence in the population, $f = p(r) = \frac{r^{-s}}{\sum r^{-s}}$. A

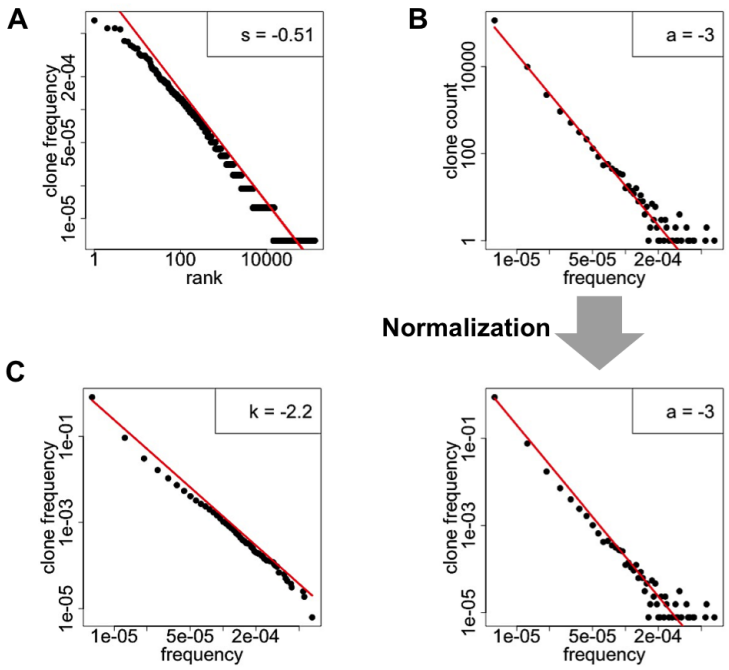

Figure 2.6: A. The Zipf rank distribution, B. The power law distribution before and after normalization. Note that the slope is unaffected by this transformation, and C. The pareto distribution. The linear fits are plotted in red with the slope as indicated.

parametric relationship describes the functional relationship between the number of elements at a given place, time, frequency, etc, $y = f(x) = x^{-\alpha}$ and can be turned into a proper probability distribution by normalization. A related version of the power law distribution is known as the Pareto distribution which describes the cumulative number of such elements $y = P(X > x) = x^{-k}$. Examples of all three with linear fits is shown in Figure 2.6 taken from a TCR repertoire of healthy blood.

Power laws are difficult to study because of their scale free nature. The only useful point estimate for these distributions is the power law exponent, which gives the value of the linear slope. Both the mean and variance are not guaranteed to exist under all values of this slope. In most cases the mean is small and variance is very large, and therefore are not useful for making statistical inferences. Moreover, real data often deviates from the powerlaw at one or both ends of the distribution, further complicating the analyses. However, we use

reasonable estimates of the slope validated by simulations to make conclusions about the TCR repertoire that are presented in Chapter 3.

2.3.2 Measurements of Diversity

The concept of diversity has many different formal definitions, and the correct one to use depends on the application. When studying species in an ecosystem, word usages, TCR sequences, or other large populations, one typically considers the richness (number of distinct elements) and evenness (frequency of observed elements) of the observed sample. Measures of diversity seek to describe both of these quantities with a single metric that provides insight into how these populations are distributed and allows for meaningful comparisons.

A common definition of diversity is the "true" diversity [46], which takes the form,

$$^{q}D = \left(\sum_{i=1}^{s} p_i^q \right)^{(1/(1-q))}$$

where the superscript q is the order of diversity. When $q = 0$ this gives exactly the species richness without any consideration for species evenness. At $q = \infty$ the diversity is defined as the largest observed frequency. Values in between give different weights to the richness and evenness with higher values of q putting more emphasis on larger frequencies in the population. The limit at q=1 is the exponent of Shannon entropy, and strikes an equal balance between richness and evenness. Shannon entropy is often the selected measure of diversity because of this balance and its relationship to other quantities from information theory. It is defined as

$$H = -\sum_{i} p_i \log_2 p_i$$

where p_i is the frequency of clone i in the sample. Even more appealing is the property that entropy can be added for independent elements, $H(X, Y) = H(X) + H(Y)$ if X and Y are independent [67]. Using $q = 1$ returns $\exp(H)$ as the true diversity. Shannon entropy is a measure of information uncertainty, which under the base of $\log_2$ corresponds to the minimum number of binary questions required to precisely select an element from the population based

on its frequency. An entropy H=0 means that there is no information uncertainty and the first guess is always correct, whereas larger entropies correspond to a population where it is more difficult to determine a unique clone within a specified frequency range.

However, it must be kept in mind that smaller values of the exponent q are more heavily affected by sample size and evenness. The maximum entropy is given by $\log_2 N$ where N is the number of elements observed (species richness), and thus two samples cannot be directly compared if their sample sizes differ by a large amount. Therefore, a number of TCR studies use a normalized value of entropy to measure the clonality of the population [24, 18]. Clonality is defined as

$$CL = 1 - \frac{H}{H_{max}}$$

A clonality of 1 indicates no diversity ($H = 0$), while 0 is the maximum possible diversity ($H = H_{max}$). Clonality depends on accurate measurements of maximum entropy, and therefore is still affected by sampling, particularly if most clones are rare and likely not observed in the sample as is the case with the power law. In applying either entropy or clonality it may be useful to perform subsampling or run simulations to ensure minimum bias.

Two final measures of diversity considered for TCR repertoire analysis are the R20 and R50 statistics. This corresponds to the fraction of clones, starting with the largest, that respectively encompasses 20% or 50% of the entire cell population. Smaller values of R20 and R50 correspond to lower diversity, where most cells correspond to a few select clones. In general an R_X measure can be defined as follows:

$$R_X = \left\lceil \frac{N(m_X)}{N(m_100)} \right\rceil$$

where $N(m_X)$ is the number of clones found within the top X% of cells (m).

Python code to compute these diversity measures is available and can be downloaded from GitHub at github.com/ShenLab/TCR/blob/master/simulation/src/diversitymeasures.py. The theory chapters of this thesis will discuss the preferred measures of diversity for exploring a variety of questions within TCR reperotire analysis.

Table 2.4: Measures of true diversity

q	Name	Interpretation
0	Richness	Total number of species
1	Shannon entropy	Equal weights for richness and evenness
2	Simpson index	Probability of observing an element twice
⋮		
∞	Berger-Parker Index	Maximum frequency

2.3.3 Measurements of Divergence

Although two populations may have the same diversity, this does not ensure that they have the same distribution. Two populations can consist of similar frequency values belonging to entirely different clones. In such cases a measure of population divergence is needed to compare how similar two distributions are. The Kullback Leibler (KL) metric from information theory accomplishes this task

$$D_{KL}(p||q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

The KL measures how much information about one distribution, q, is encoded in another, p. Populations that are identically distributed will have $KL = 0$. As with entropy, the KL divergence does not have the same maximum for all distributions. Additionally, the information that p provides about q is not necessarily the same as the information q gives about p. This lack of symmetry or a well defined maximum makes it difficult to compare values of divergence. This metric is best applied when looking at two populations derived from a single larger population. Both subpopulations are then compared relative to the same standard and comparing their divergences has meaning.

One way of normalizing the KL divergence to produce a more meaningful divergence metric is to combine distributions p and q into a new distribution $m = \frac{p+q}{2}$. This Jensen Shannon (JS) divergence is defined as

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)$$

The Jensen Shannon divergence is symmetric and takes on a range of values from 0 (identical distributions) to 1 (maximum difference). A simple calculation also shows that D_{JS} can be expressed in terms of entropies.

$$D_{JS} = H(m) - 0.5\Big[H(p) + H(q)\Big]$$

Taking the square root satisfies the triangle inequality, transforming the Jensen Shannon divergence into a true distance measure (JSD) [30].

Chapter 3

A Semi-parametric Method for Unseen Clones

3.1 Introduction

This chapter presents a method for estimating the total frequency of cells corresponding to the unseen clones in a sample of the TCR repertoire. The average frequency of unseen clones in an expanded population is obtained by obtaining power law parameters to model the distribution of the TCR repertoire. A non-parametric approach is used to estimate the number of such clones, thus resulting in a "semi-parametric" approach for estimating the sum frequency. This method is applied to the analysis of the alloreactive repertoire described in Chapter 6.

3.2 TCR Distribution

While many bulk properties of a repertoire can be well estimated by maximum likelihood measures of diversity and divergence, such as entropy, KL divergence, and others mentioned in Chapter 2, significantly more insight can be gained by identifying the appropriate distribution that describe the data, and estimating its corresponding parameters. Obtaining such a parametric estimate makes it possible to generate new data points and to infer missing or

biased data. It also offers the possiblity of decomposing a complex dataset into distinct clusters of subsets with different parametrizations, and thereby gaining insight into population heterogeneity and dynamics.

In many cases obtaining the correct distribution and its parameters is not possible from the data obtained; however, as discussed in Subsection 2.3.1 the bulk of the TCR distribution is observed to very strongly follow a power law. However, the repertoire also consists of expanded clones which do not follow any easily defined distribution (Figure 3.1A). One possible explanation for this power law nature could have been that PCR and sequencing errors artificially inflated numbers of rare clones; therefore large clones are more likely to produce to spurious singletons (clones of copy number 1) . However, a single nucleotide difference occuring during V(D)J recombination can change the amino acid composition of the CDR3 chain and thereby generate a real clone with distinct binding properties. Using hamming distance, which measures the number of single nucleotide mismatches between two sequences, sequences from expanded clones and from rare clones were compared. Clones with hamming distance of 1 and 2 were removed from the sample. It can be seen that some of the removed rare clones shared sequence similarity with the expanded clones, suggesting that there may exist low levels of sequence error; but, removing these did not change the observed power laws nor their slopes (Figure 3.1B). Thus we conclude that that the T cell sequence repertoire consists of an expanded component and a power law component.

The slope of the power law can then be obtained by a linear fit on the log-log axis. The main remaining challenge is to define where the power law ends and the expanded portion begins. Expanded clones typically have unique copy number, corresponding to a count of 1 on the y-axis of the abundance plot. Therefore, initially the power law slope is defined as the linear fit of the distribution starting from the smallest frequency and ending at the minimum frequency at which a unique clone copy number is observed. When dealing with datasets where the number of unique clones sequenced was very small (<1000), this definition often failed to provide a good fit due to a lack of data. Thereby, a more careful definition was

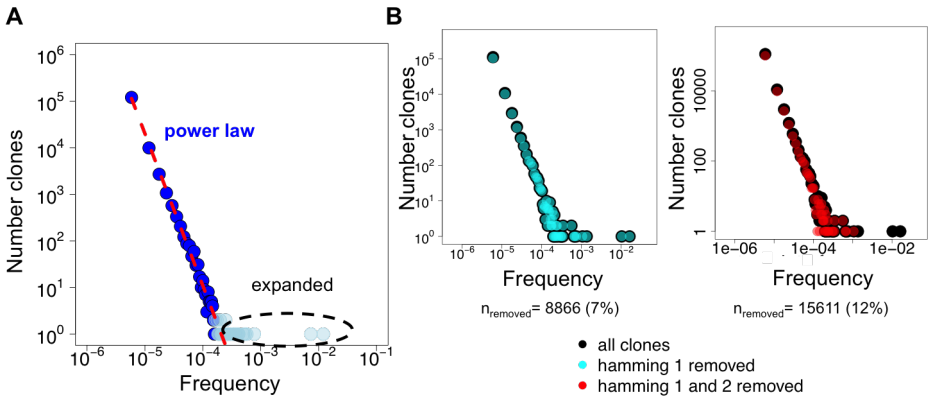

Figure 3.1: A. The typical plot of TCR abundance consists of a powerlaw component and an expanded component. The red line represents a linear fit of the data from which a slope can be obtained. B. Left, in cyan, are clones remaining after all clones with hamming distance of 1 are removed and right, in red are all clones remaining after removing clones with hamming distance of both 1 and 2.

selected, selecting the first two unique clones as the right end of the power law, provided that the second was close to the first on the log scale. A log distance <1 was found to provide the best results.

3.3 The semi-parametric method

The general problem is constructed as follows. Two samples are collected from a large population of T cells. One sample undergoes activation and clonal expansion due to a stimulus while the other remains unchanged. We call these populations unstimulated and stimulated respectively. Both populations are then sequenced (Figure Figure 3.2A). Since most clones are rare, the majority of cloes are captured in only one of the two samples and the overlap between them is small (Figure Figure 3.2B) while the number of stimulated clones not observed in the unstimulated population is much larger. The challenge then is to estimate the true fraction of cells in the unstimulated population that underwent expansion.

A power law is fit to the abundance plot of the subset of clones that are observed in both activated and unstimulated populations (Figure Figure 3.2C). The number of such unseen clones is used to draw a horizontal line that intercepts this power law. The x-axis value of

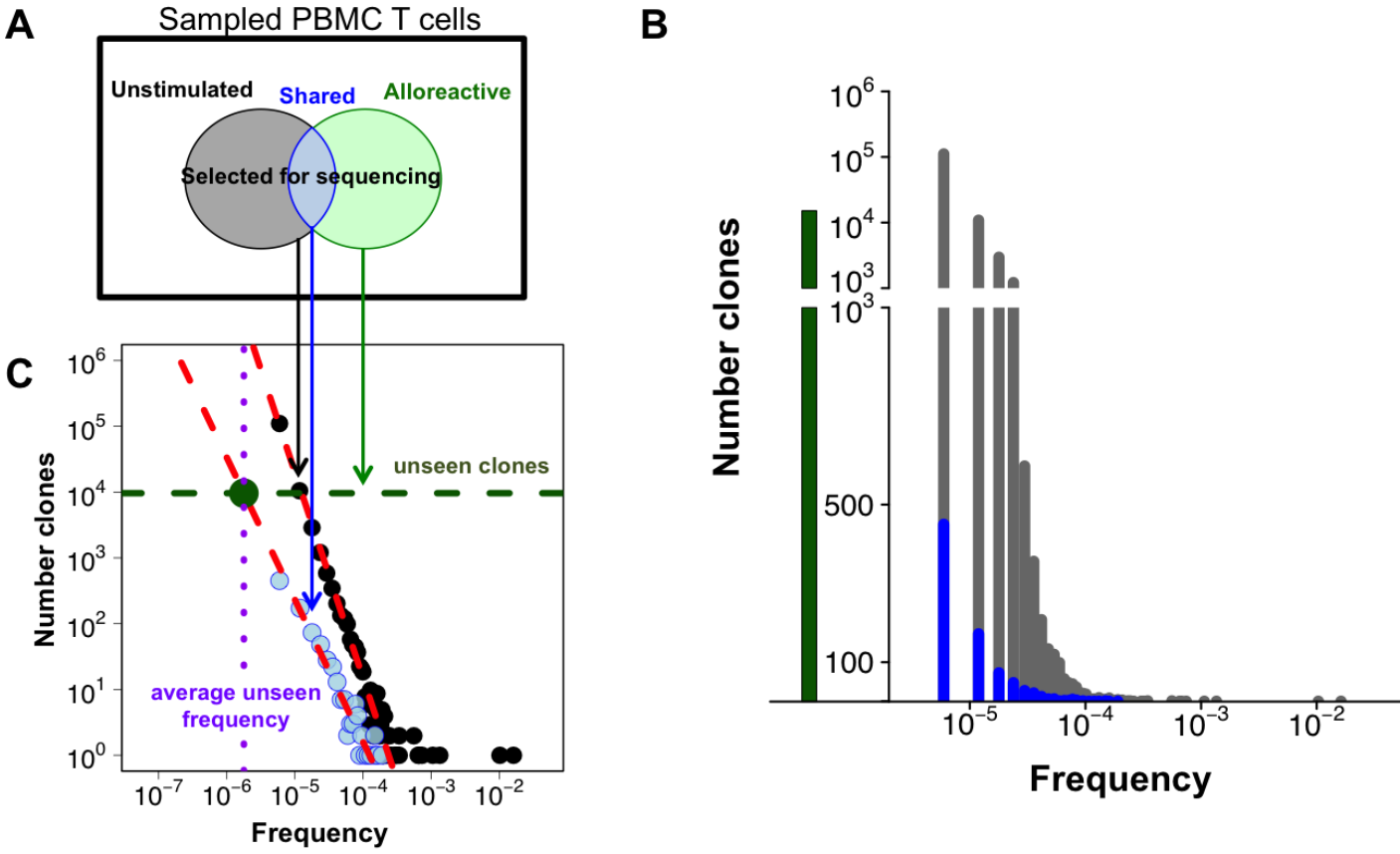

Figure 3.2: A. Two samples are collected from a larger pool of T cells. The unstimulated population is shown in gray, and the alloreactive in green undergoes avtivating stimulation. The two share a small overlap shown in blue. B. A histogram of the abundane distribution of the unstimulated populated and in blue the subset that is observed in the stimulated population. The green bar indicates the number of unseen clones. C. Construction of the semi-parametric method. A power law fit of the shared clone abundance intersects the number of unseen stimulated clone shown by the horizontal green dashed line The x-axis at the point of intersection corresponds to the average clone frequency of the unseen stimulated clones.

the intercept gives the average frequency of an unseen clone.

If we let N represent the number of unshared clones in the stimulated sample, and $x*$ represent their average frequency in the unstimulated population, then for a power law with slope α and K, the value of x^* can be obtained mathematically from the definition of the slope of a line, with values transformed on the log scale:

$$x^* = \exp\left\{\frac{\log(N) - K}{\alpha}\right\}$$

This result arises from redistributing the standard formulation for the slope of a line

$$\alpha = \frac{y_2 - y_1}{x_2 - x_1}$$

and solving for $x_z = x_1$. The resulting unseen frequency is then given by

$$f_{unseen} = N \cdot x^*$$

3.4 Validation by simulation

3.4.1 Method

Validation of the semi-parametric method was performed by taking two samples from a deeply sequenced repertoire. One sample represented the unstimulated population, and a portion of it is was further subsampled to produce the shared population. The other sample is the stimulated population minus the shared piece, corresponding to all the unseen clones. The semi-parametric method is then applied to the shared region, and the resulting frequency of unseen clones is compared to the true unseen frequency of the unshared stimulated population Figure 3.3.

Due to the deviation from power law at large frequencies, which is particularly pronounced among populations undergoing strongly polyclonal expansion, a *de novo* simulation, in which the initial population is generated from a theoretical powerlaw distribution, does not accurately depict the observation of subsampled real data. One future area to explore

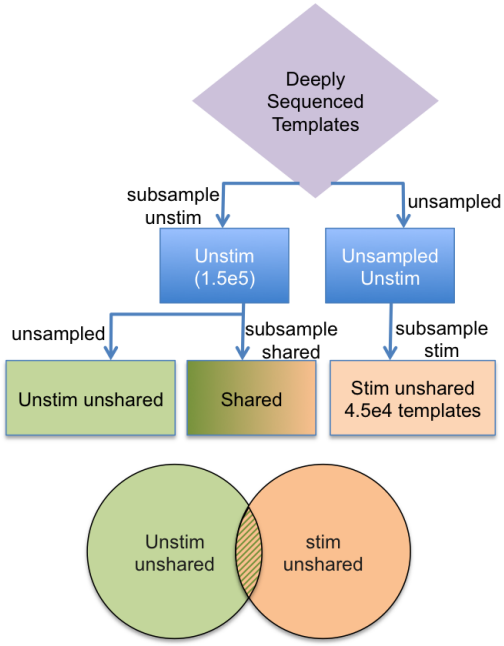

Figure 3.3: Templates from a deeply sequenced dataset, in purple, are subsampled twice. The first subsample (second row, left, in blue) represents an unstimulated population and the second subsample (second row, right, in blue) represents a stimulated population. Subsampled numbers are selected to correspond to our numbers from a typical experiment. The unstimulated sample is futher subsampled to produce a shared portion of clones, and the three resulting populations (third row) are indicated, with unstimulated in green, stimulated in orange, and shared as a mixture of the two. A venn diagram is presented to further highlight the relationship between the three subsampled populations.

would be to construct a mathematical model to accurately describe the process of repertoire expansion, but at present this task has proven a significant challenge. Instead, the initial population from which subsamples are drawn comes from deeply sequenced reference data sets of peripheral blood mononucleocyte (PBMC) derived T cells, downloaded from Adaptive Biotechnologies and publicly accessible under the project name TCRB Time Course. Nine datasets were selected from the same time point, corresponding to three technical replicates from three different subjects.

Samples sequenced by earlier versions of Adaptive's ImmunoSeq platform are available only in read counts which do not lend themselves to a power law fit. In contrast, the samples that are presented in this thesis are obtained from newer versions of the sequencing platform

Table 3.1: Summary of TCRB time course data – Three replicates from three subjects. Number of reads and clones is provided in the downloaded data. Read depth is inferred by simulated annealing and used to compute template numbers and slope values.

Subject	Replicate Id	Templates	Reads	Clones	Power Law Slope	Read Depth
Subject01	110819	628121	25124827	445325	-2.88	40
Subject01	111014	702840	23193700	510910	-2.97	33
Subject01	110915	554826	24412302	417830	-3.06	44
Subject02	110811	272942	15011757	207898	-2.65	55
Subject02	110908	300539	20737164	226771	-2.83	69
Subject02	111006	210599	18111432	155165	-2.83	86
Subject03	110812	271342	21707348	218939	-2.43	80
Subject03	110909	350652	22792374	233694	-2.35	65
Subject03	111007	332762	22461415	210025	-2.46	67.5

which use unique molecular identifiers (UIDs) to produce template information as a proxy for true cell counts. The TCRB time course data therefore had to be converted from reads into template counts, which was accomplished using simulated annealing, discussed in detail in Section 3.6. Each sample contained approximately 300,000 clones on average, with 20 million reads per sample Table 3.1.

3.4.2 Results

Public time course data was converted from reads to templates (Figure 3.4A) and subsampled with replacement using the method summarized in Figure 3.3. The semiparametric method was applied to the shared subset of clones to compute average unseen frequency. Estimated total unseen frequency was computed and compared to the true value of unseen frequency, as calculated from the data. For each of the time course datasets, this procedure was repeated ten times to check that results were independent of sampling. Both estimated values and the true frequencies after sampling remained stable across subsample runs, and the two values were in close agreement. The differences between estimated and true unseen frequency values were of magnitude $10^{-}2$ or less in all cases tested Subsection 3.4.2. Abundance plots from several test cases are included, providing a visual comparison between the true unseen clone frequency and the result of the semi-parametric method (Figure 3.4B).

Table 3.2: Time course subsampling results

Subsample Number	estimated x-intercept	number unseen	estimated frequency unseen	true frequency unseen	Δ(true-estimated)
			Subject01-110819		
1	2.37e-06	40163	9.5e-02	9.7e-02	2.1e-03
2	2.64e-06	40075	0.106	9.7e-02	-8.4e-03
3	2.12e-06	39937	8.5e-02	9.7e-02	1.2e-02
4	2.54e-06	40136	0.101	9.8e-02	-3.9e-03
5	2.23e-06	40093	8.9e-02	9.7e-02	7.8e-03
6	2.22e-06	39980	8.9e-02	9.7e-02	8.0e-03
7	2.16e-06	40157	8.7e-02	9.7e-02	1.0e-02
8	2.34e-06	40067	9.4e-02	9.7e-02	3.5e-03
9	2.25e-06	40252	9.1e-02	9.8e-02	6.9e-03
10	2.26e-06	40115	9.0e-02	9.8e-02	7.4e-03
			Subject01-110915		
1	2.43e-06	39932	9.7e-02	0.103	5.6e-03
2	2.84e-06	39975	0.114	0.103	-1.1e-02
3	2.88e-06	39955	0.115	0.103	-1.2e-02
4	2.34e-06	39919	9.3e-02	0.103	9.5e-03
5	2.44e-06	39975	9.7e-02	0.102	5.0e-03
6	2.91e-06	39876	0.116	0.102	-1.4e-02
7	2.66e-06	39798	0.106	0.102	-4.0e-03
8	2.71e-06	39910	0.108	0.102	-5.9e-03
9	2.41e-06	40022	9.7e-02	0.102	5.7e-03
10	2.26e-06	39745	9.0e-02	0.102	1.2e-02
			Subject02-110908		
1	4.04e-06	31700	0.128	0.127	-9.1e-04
2	4.16e-06	31802	0.132	0.127	-4.8e-03
3	4.18e-06	31618	0.132	0.127	-5.5e-03
4	4.10e-06	31706	0.130	0.128	-2.5e-03
5	4.42e-06	31666	0.140	0.127	-1.3e-02
6	4.46e-06	31674	0.141	0.127	-1.4e-02
7	4.35e-06	31788	0.138	0.128	-1.1e-02
8	4.13e-06	31454	0.130	0.126	-3.6e-03
9	4.06e-06	31669	0.128	0.127	-1.6e-03
10	4.05e-06	31678	0.128	0.127	-1.5e-03
			Subject03-110812		
1	4.43e-06	29626	0.131	0.099	-3.3e-02
2	3.78e-06	30226	0.114	0.100	-1.4e-02
3	3.48e-06	31494	0.110		-5.3e-03
4	3.84e-06	31749	0.122	0.104	-1.7e-02
5	3.49e-06	31491	0.110	0.105	-5.5e-03
6	3.75e-06	31482	0.118	0.104	-1.4e-02
7	3.64e-06	31867	0.116	0.105	-1.1e-02
8	3.59e-06	30958	0.111	0.102	-8.7e-03
9	3.51-06	31976	0.112	0.106	-6.3e-03
10	4.21e-06	30165	0.127	0.100	-2.7e-02

Subsample Number	estimated x-intercept	number unseen	estimated frequency unseen	true frequency unseen	Δ(true-estimated)
			Subject01-111014		
1	2.79e-06	40627	0.113	8.9e-02	-2.4e-02
2	2.14e-06	40600	8.6e-02	8.9e-02	2.6e-03
3	2.09e-06	40614	8.5e-02	8.9e-02	3.7e-03
4	2.88e-06	40605	0.117	8.9e-02	-2.8e-02
5	1.68e-06	40576	0.68e-02	8.9e-02	2.1e-02
6	2.60e-06	40657	0.106	8.9e-02	-1.7e-02
7	9.65e-07	40602	3.9e-2	8.9e-02	5.0e-02
8	1.79e-06	40386	7.2e-2	8.9e-02	1.6e-02
9	2.68e-06	40645	0.109	8.9e-02	-2.0e-02
10	1.84e-06	40716	7.5e-2	8.9e-02	1.4e-02
			Subject02-110811		
1	4.46e-06	30805	0.137	0.131	-6.1e-03
2	4.33e-06	30849	0.133	0.131	-2.3e-03
3	4.25e-06	30880	0.131	0.131	3.12e-05
4	4.49e-06	30796		0.132	-6.7e-03
5	4.41e-06	30968	0.137	0.131	-5.0e-03
6	4.62e-06	31062	0.144	0.132	-1.2e-02
7	4.41e-06	30532	0.135	0.130	-4.2e-03
8	4.39e-06	30504	0.134	0.130	-3.7e-03
9	4.53e-06	31061	0.141	0.132	-8.6e-03
10	4.43e-06	30701	0.136	0.131	-5.0e-03
			Subject02-111006		
1	5.14e-06	27408	0.141	0.148	6.9e-03
2	5.01e-06	27393	0.137	0.147	9.8e-03
3	5.01e-06	27241	0.137	0.146	9.5e-03
4	5.06e-06	27312	0.138	0.147	9.0e-03
5	5.03e-06	27024	0.136	0.146	9.7e-03
6	5.08e-06	27217	0.138	0.147	8.3e-03
7	5.00e-06	27325	0.137	0.147	1.1e-02
8	4.93e-06	27553	0.136	0.149	1.3e-02
9	4.94e-06	27388	0.135	0.147	1.2e-02
10	5.04e-06	27326	0.138	0.147	9.5e-03
			Subject03-110909		
1	3.34e-06	32240	0.108	9.79e-02	-9.7e-03
2	4.09e-06	30936	0.126	9.43e-02	-3.2e-02
3	4.11e-06	32019	0.132	9.73e-02	-3.4e-02
4	3.67e-06	31162	0.114	9.54e-02	-1.9e-02
5	3.34e-06	31680	0.106	9.63e-02	-9.0e-03
6	3.21e-06	31563	0.101	9.59e-02	-5.5e-03
7	3.29e-06	31517	0.104	9.56e-02	-8.1e-02
8	3.74e-06	31379	0.118	9.60e-02	-2.1e-02
9	3.17e-06	32369	0.102	9.85e-02	-4.0e-03
10	3.56e-06	31558	0.112	9.59e-02	-1.7e-02

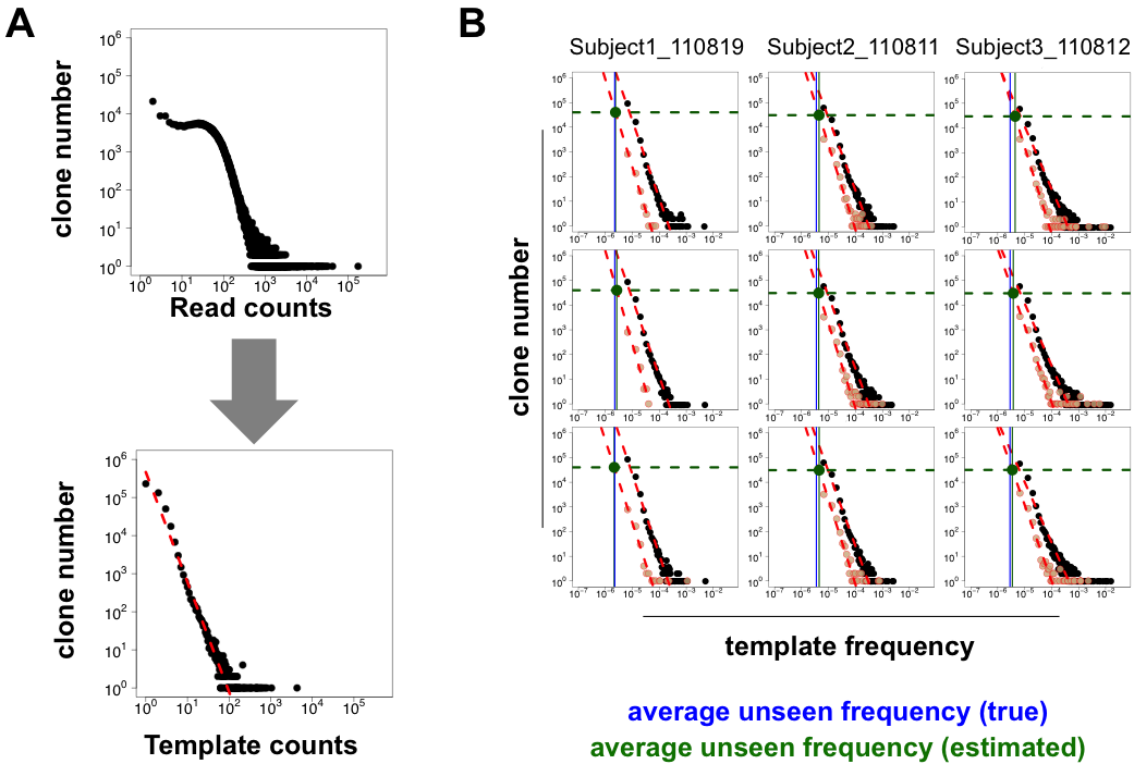

Figure 3.4: A. Conversion from reads to templates for an example TCR repertoire from the time course data set B. Estimated average frequency for unseen clones from three subsampling runs of three time course subjects . Abundance plots for the unstimulated sample are shown in black, with the shared population in orange and a slope fit in red. The number of unseen stimulated clones is displayed as a horizontal green line. The intersection between the slope fit and the unseen clones is indicated by a filled green circle, from which a green line is dropped to the x-axis at the estimated value of the average unseen frequency. The true unseen frequency is shown by the vertical blue line.

3.5 Validation by replicates

The semi-parametric method was applied to a biological replicate, where two samples from the same unstimulated pool underwent polyclonal expansion due to an alloresponse (Figure 3.5A) . The exact method for this experiment is discussed in Chapter 6. The semi-parametric method was applied to each of the two populations of alloreactive clones and the unseen clonal frequency was computed (Figure 3.5B top). Similarly, the unseen clonal frequency was computed for a combined sample with the two replicates pooled into a single

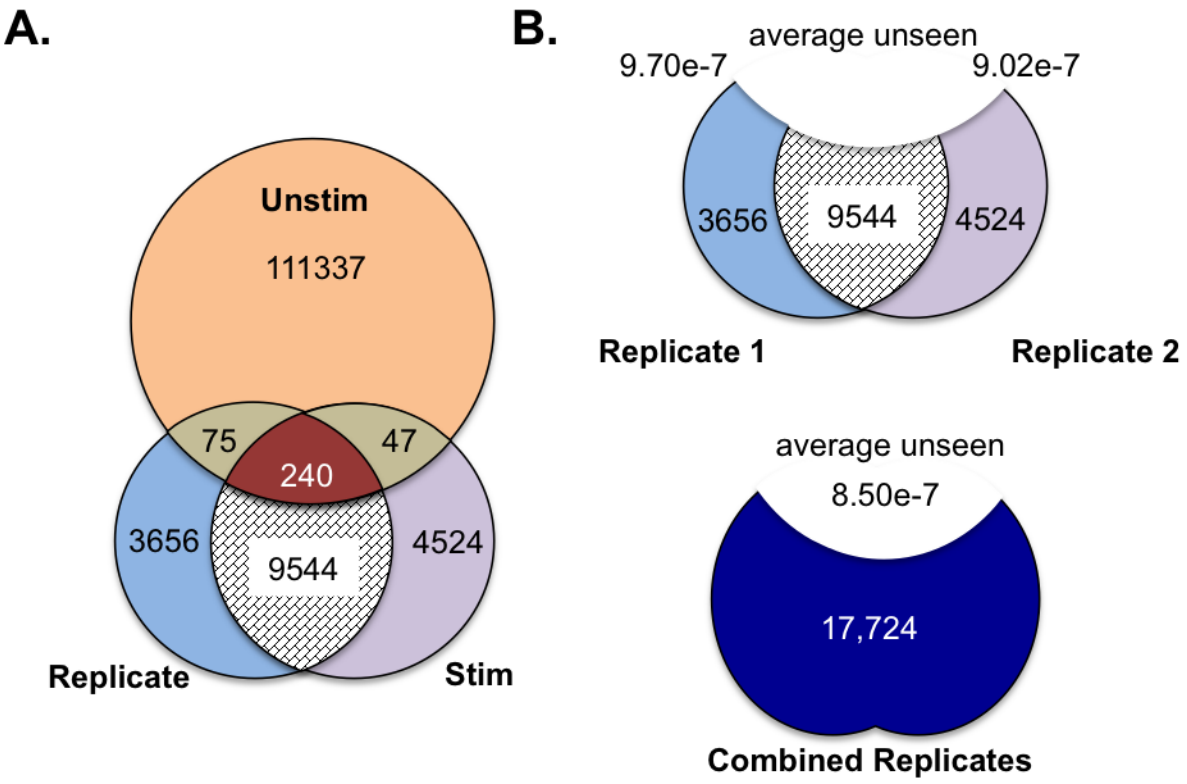

Figure 3.5: A. Venn diagram indicating the overlap between the two alloreactive replicates and corresponding unstimulated sample. B. Clone numbers for shared replicates (top) and clone numbers for the combined sample (bottom).

larger sample (Figure 3.5B bottom). The unseen frequency computation was performed as follows:

$$f_{replicate1+replicate2} = 8.50e-07 * 17724 = 0.015$$

$$f_{combined} = 9.7e-07 * 3656 + 9.02e-07 * 4524 + (9.70e-07 + 9.02e-07)(9544/2) = 0.0166$$

where the average unseen frequency in the overlap between the two alloreactive samples was averaged. Adding the two replicates individually produced an average unseen clone frequency of 0.0166, compared to 0.0151 from the combined sample, an overestimate of only 0.15% demonstrating that there is little loss of accuracy due to subsampling, and validating the efficacy of the semi-parametric method.

3.6 Obtaining template counts by simulated annealing

Starting with an initially power law distributed repertoire the PCR amplification step to generate reads was mathematically described by the gamma-Poisson mixture. In the gamma-Poisson mixture, the cell count of every clone is sampled from a Poisson distribution, where the rate, λ, obeys the gamma distribution. For a known sampling rate, the probability of finding a clone consisting of n cells is described by:

$$P(\lambda)P(N=n|\lambda) = \overbrace{\left[\frac{\alpha^{\beta}}{\Gamma(\beta)}\lambda^{\beta-1}e^{-\alpha\lambda}\right]}^{\Gamma(\lambda|\alpha,\beta)}\overbrace{\left[\frac{e^{-\lambda}\lambda^{n}}{n!}\right]}^{Poiss(n|\lambda)}$$

In the limiting case of the gamma-Poisson mixture, where λ is unknown, we can take an integral to obtain the average probability independent of the sampling rate. Evaluating the integral produces the relationship

$$P(N=n) = \int_0^{\infty} P(\lambda)P(N=n|\lambda)d\lambda = \binom{n+\beta-1}{n}\left(\frac{\alpha}{\alpha+1}\right)^{\beta}\left(\frac{1}{\alpha+1}\right)^{n}$$

which describes a negative binomial. The average sampling rate for read counts is related to average sampling rate for cell counts by the average amplification depth, μ=λ D. Then, the negative binomial mean and variance for selecting each clone from a larger sample is given by the following relationships:

$$\mu = \lambda D$$

$$\sigma^2 = \mu + \frac{1}{s}\mu^2$$

where s describes overdispersion from the Poisson. This negative binomial approximation of read depth is frequently chosen for analysis of mean and variance from read counts collected from RNA sequencing experiments [56].

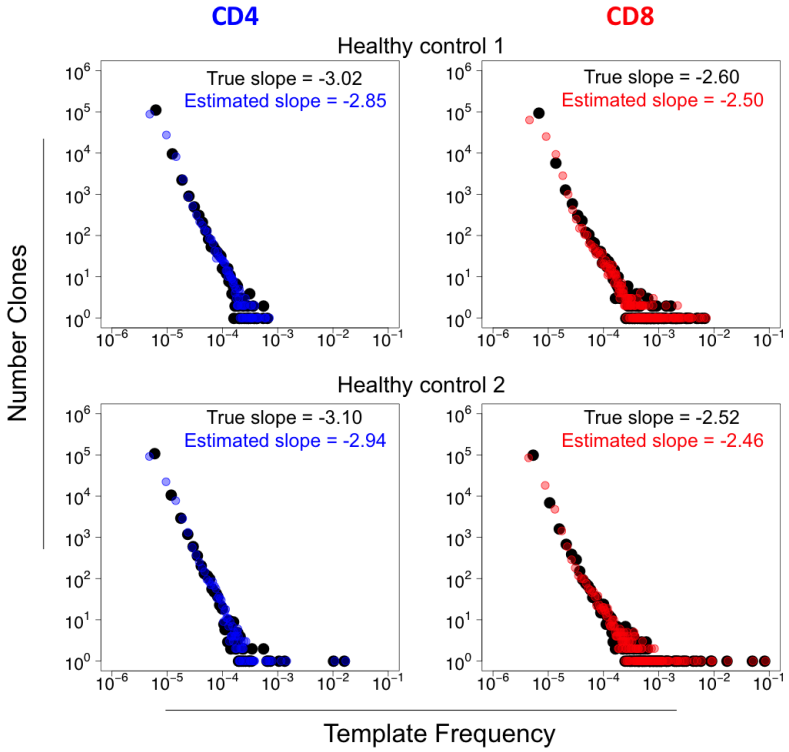

Figure 3.6: CD4$^+$ and CD8$^+$ samples of TCR from two individual healthy PBMC controls with known template numbers are shown alongside fits obtained from read counts for those same samples. True template abundance is depicted in black, while counts obtained by starting with reads and running the simulated annealing procedure are colored in blue, for CD4$^+$ and in red for CD8$^+$ samples. Template slope and estimed slope from simulated annealing are presented in each plot.

The simulated annealing algorithm belongs to the Markov Chain Monte Carlo family of methods for finding locally optimal parameters. It consists of doing a random walk through the space of possible parameter values and evaluating a "goodness-of-fit" function that provides a readout for the success of the optimization. As the value of the readout improves the step size for the random walk becomes smaller, allowing estimated of paramater values to be fine tuned. The procedure is repeated multiple times, to allow different local optima to be captured, with the best parameters corresponding to the best optimum across runs.

For converting reads (R) to templates (T) , the key parameters were depth, D, and

overdispersion, s. At each step of optimization templates numbers were computed as follows:

$$T = \left\lceil R/D \right\rceil$$

The power law slope was estimated from the templates as discussed in Section 3.2. A discrete power law with this slope was then simulated containing roughly the same number of clones as in the observed data set. Reads were sampled from the negative binomial with mean and variance given by the equations above. Finally KL divergence, $D_{KL}(p||q)$, was used as the goodness-of-fit function, computed between the simulated distribution, p, and the true distribution, q. The values of D and s were allowed to change according to a random walk with predefined step size, with the step accepted if D_{KL} decreased and a probability of acceptance $\exp(\frac{L_{prev}-L_{new}}{T})$ if it increased, where T serves as a temperature parameter. The starting value of D_{KL} was recorded, and the temperature parameter T and step sizes were decreased each time a new computed D_{KL} fell below an experimentally determined threshhold of $\frac{D_{KL}^{new}}{D_{KL}^{init}} \leq 0.75$. This divergence was then recorded as the new "initial value" and the optimization continued until convergence, with a burn in period of 1000 steps. Convergence required the change in D_{KL} to be less than $1e-4$ for 100 consecutive steps. Values of D_{KL} varied by sample but were generally on the order of $1e-2$.

The simulated annealing algorithm was tested on samples for which both read and template data was available to validate its usefulness and produced reasonable estimates of template repertoire and slope. Figure 3.6 shows the results from these test cases.

3.7 Conclusion

We use the power law to describe the bulk distribution of a TCR repertoire. While modeling of the TCR repertoire is a complex task because typical point estimates like the mean and variance are not useful in analysis of power laws, the slope of the power law is powerful tool for investigating questions of repertoire size and diversity, but requires precise measurements of template counts. Using this template information where available, and estimating it from reads where required, a semi-parametric method was found to accurately measure the

unseen frequency of cells from a pair of samples. The validity of this sample was tested by analysis of subsamples taken from deeply sequenced TCR data and by investigating stability among sample replicates. The use of the semi-parametric method makes it possible to answer a question often asked by immunologists and highly relevant for development of new therapies, namely the frequency of a repertoire that is expanded when undergoing olgiclonal or polyclonal expansion due to antigen stimulation.

3.8 Discussion

Quantitative analysis of both B and T cell repertoires can be greatly improved by obtaining good parameter estimates to fitt the correct distribution. Several papers have proposed power law fitting of the TCR repertoire [10, 23, 22], but there is virtually no published literature to utilize such a fit. Much of the difficulty comes from the fact that, until recently, there was a lack of accurate template data and observations were reliant upon rank based distributions like Zipf law, rather than counts based estimates. Without a sufficiently large dataset the accuracy of power law fitting is further reduced and methods are difficult to validate. This work utilized the recent switch from read information to template counts, and converted reads to templates via simulated annealing where necessary, with template counts as the golden standard used to obtain optimal parameters for performing the converstion. While TCR repertoire analysis is inherently a single cell challenge, in the next few years bulk amplicon sequencing will continue to be the standard for capturing repertoire size and diversity, and as this technology continues to improve, parametric methods will become increasingly more useful.

For the work discussed in this chapter, several improvements can be made. While the power law fits obtained here were sufficient for making comparisons between subjects, a more systematic approach for power law fitting, such as by using expectation maximization, may yield more accurate estimates with less depedence on ad-hoc cut-off criteria. This is a significant challenge, because neither the mean nor variance are well defined for power laws,

and the expanded portion does not appear to follow any well-described distribution and contains only a few clones. Additionally, for unsorted repertoires, we obtained an average slope, but there are likely multiple distinct distributions present, corresponding to different TCR subsets. Ideally, with a sufficiently large dataset of TCR repertoire cell abundance, it may be possible both to differentiate distinct populations of T cells, and to accurately describe their distributions, allowing for deep insights into repertoire size, diversity, and divergence to be made from the repertoire alone.

Chapter 4

Long-term maintenance of human naïve T cells through in situ homeostasis in lymphoid tissue sites

4.1 Introduction

Subsection 2.1.1 gave an overview of T cell development in the thymus that generates the mature naïve repertoire. Most naïve T cells are found in the spleen and lymph nodes whereas the most accessible compartment in humans is blood, of which the naïve T cell repertoire represents only 2-3% of the total peripheral blood mononucleocyte (PBMC) derived T cell population [35]. Most of our understanding of naïve T cell maintenance and diversity comes from mouse studies where the lymphoid organs are readily accessible in the laboratory. It is reasonable to consider that the difference in lifespan and size of the mouse suggests differences in naïve T cell maintenance as compared to humans. Indeed, naïve cell maintenance in mice has been found to be largely driven by thymic output, whereas in humans the evidence points to a homeostatic process of maintenance within peripheral sites [11]. The naïve population in humans is also known to decline with age as a result of thymic involution and

increasing numbers of memory T cells, with high frequency circulating clones in $CD4^+$ and $CD8^+$ subsets [95, 94]. Thus, there is great interest in understanding how the fraction of naïve reprertoire within lymphoid tissues changes over the lifespan of a typical human adult, and to gain further insight into this maintenance mechanism. This chapter presents an in depth analysis of the naïve T cell repertoire from human lymphoid tissues obtained from organ donors through collaboration with the organ procurement organization for the New York metropolitan area (LiveOnNY). Experimental work was performed by several different people in the lab of Dr. Donna Farber, but the majority of the analysis and figures were performed by Joseph Thome, PhD. and Brahma Kumar, while statistical analysis of TCR repertoire forms the focus of this thesis. For completeness, both components of the research are described. The results discussed have been published in the Science Immunology journal [94].

4.2 Experimental analysis of naïve T cells over multiple age ranges

This section describes experimental results from work performed by collaborators, but is presented here in order to give a complete story. Naïve T cells were collected from over 70 donors aged from 2 months to 73 years from many different tissues including thymus, spleen (SP), inguinal lymph nodes (ILN), lung lymph nodes (LLN) and mesenteric lymph nodes (MLN).

4.2.1 Analysis of thymic function

Thymic involution with age was looked at by histological analysis, using hematoxylin and eosin staining to look at Hassal corpuscles, a structure associated with functional thymic activity. Younger donors had larger numbers of Hassal corupscles present at high density, compared with significantly fewer Hassal corpuscles of larger size in adults (Figure 4.2A,B).

Thymic activity was further tracked by looking at numbers of FACS sorted double pos-

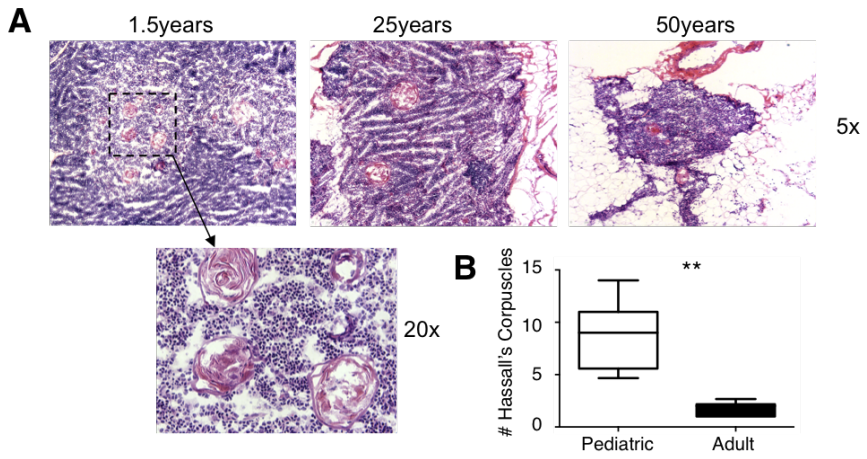

Figure 4.1: Staining of Hassal corpuscles. A. Histological staining for Hassal corpuscles in thymic tissue at 10x magnification (top) and 40x (bottom). The pink circular structures are Hassal corpuscles. B. The average number of these structures in pediatric donors (<2 years old) compared with adults [94]. Images taken by Joseph Thome, PhD.

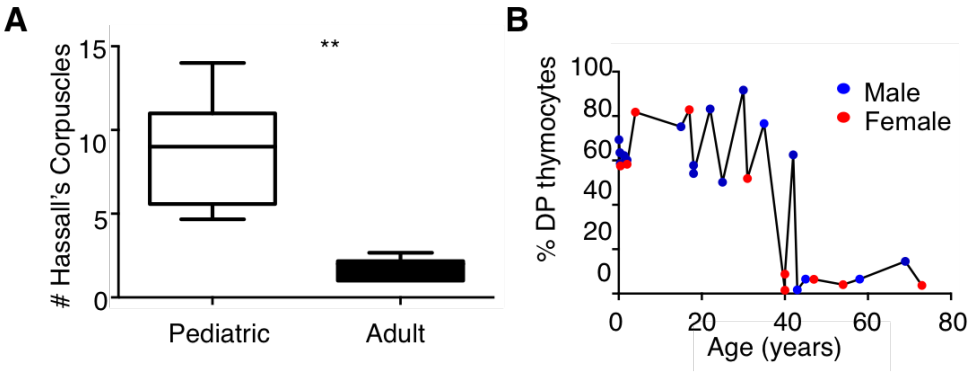

Figure 4.2: A. Representative flow cytometry plots showing numbers of DP thymocytes for different ages B. DP thymocyte percentage across all donors separated by age [94]. Figure by Joseph Thome, PhD.

itive (DP) $CD4^+$ and $CD8^+$ thymocytes from functional thymic tissue. As discussed in Chapter 2, DP thymocytes are present in the thymus prior to clonal selection. The highest frequencies of DP thymocytes were found in younger donors, and lower frequencies in older donors (Figure 4.2A,B). In both male and female donors the number of DP thymocytes declined sharply after 40 years of age.

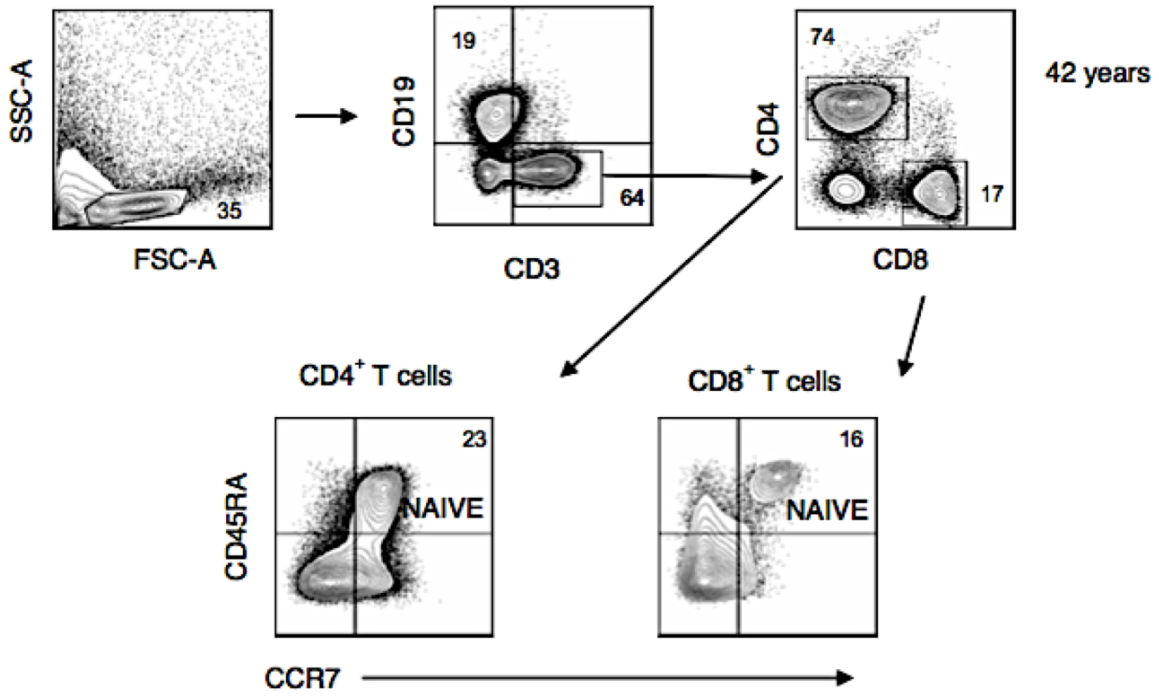

Figure 4.3: Gating strategy for naïve T cells [94]. Figure by Joseph Thome, PhD.

4.2.2 Changes in naïve T cell numbers in lymphoid tissues

Fraction of naïve T cells in the different lymphoid compartments was analyzed by flow cytometry. The gating strategy is presented in Figure 4.3 showing numbers of $CD4^+$ and $CD8^+$ T cells sorted for naïve subsets based on cell surface markers discussed in Table 2.2.

Consistent with Figure 2.4, at all ages most naïve T cells are found in the blood, spleen and lymph nodes, with few cells of this phenotype in the intestinal tissues (jejunum, ileum, and colon). However, the fraction of naïve T cells is seen to decrease with age. The largest numbers, as high as 80% of all T cells, have the naïve phenotype among samples taken from pediatric donors. In contrast this drops to 40% or less for $CD4^+$ and 60% or less for $CD8^+$ in donors over 40 years of age (See Figure 4.4).

This decrease in naïve T cell fraction with age is particularly evident when plotting each individual as a separate point stratified by age. By age 40, the splenic repertoire drops to

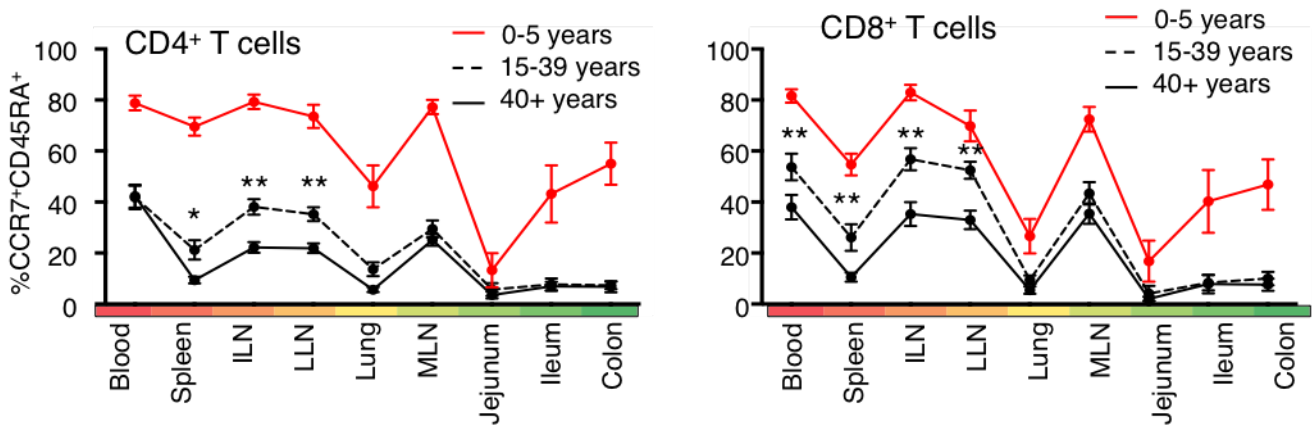

Figure 4.4: Mean percentages of naïve T cells at various tissues sites by age [94]. Figure by Brahma Kumar and Joseph Thome, PhD.

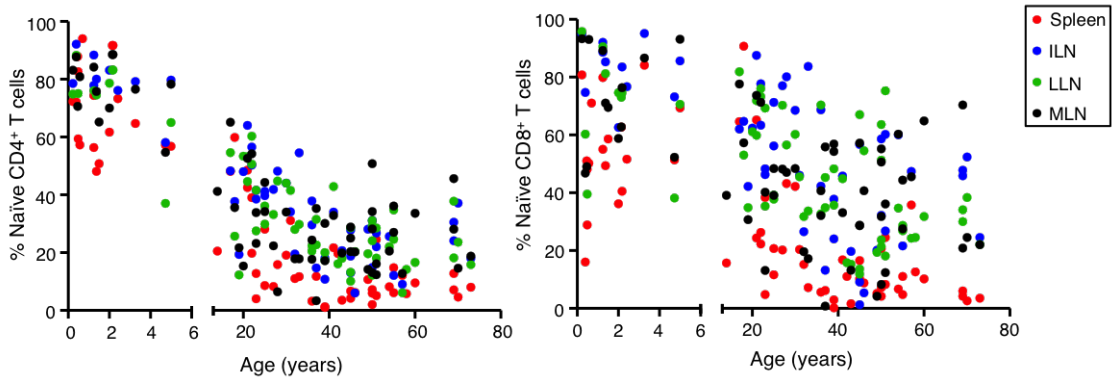

Figure 4.5: Individual percentages of naïve T cells across tissue sites by age [94]. Figure by Brahma Kumar and Joseph Thome, PhD.

nearly zero, recapitulating the loss of thymic output. The spleen serves as a filter for blood, giving evidence that while the repertoire is maintained at low levels in lymphoid tissue sites, the blood and spleen repertoire is largely devoid of this phenotype in the older age group (See Figure 4.5).

4.2.3 Thymic output and naïve T cell function

Histological staining of the thymus, sorting of DP thymocytes, and quantification of naïve T cell numbers all point to a loss of diversity in adults, becoming particularly pronounced after age 40. Another way to assess these waning T cell numbers in the tissue sites is to directly

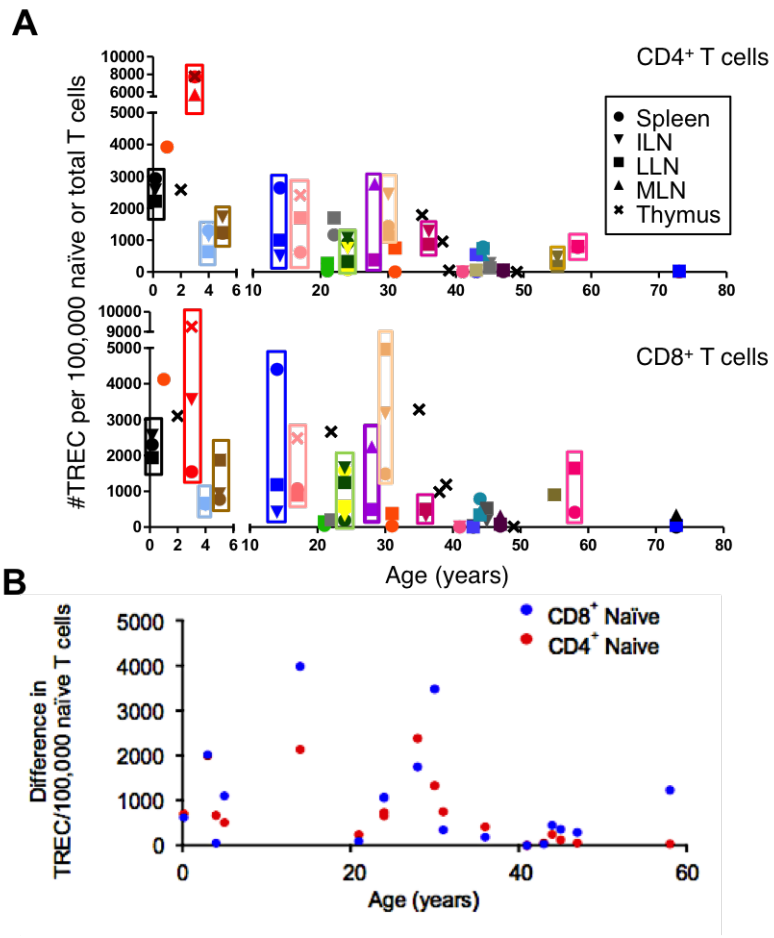

Figure 4.6: A. Decrease of TREC levels with age for forty donors. B. Greatest difference between tissues for each donor for $CD4^+$ and $CD8^+$ subtypes [94].Figure by Joseph Thome, PhD. and Gregory Sempowski, PhD.

quantify thymic output by production of recent thymic emigrants (RTEs) – naïve T that have recently entered the peripheral tissues from the thymus. These can be quantified using TCR excision circles (TRECs) which are remnants of V(D)J rearrangement that are diluted out over multiple round of cell division that occur as part of the T cell maintenance process in the tissues. Analysis of TREC numbers similarly shows the drop of with age, becoming particularly pronounced by age 40. Although there were no significant differences in $CD4^+$ vs $CD8^+$, younger individuals showed greater variation of TREC levels in different lymphoid tissue sites, suggesting potential differences in RTE seeding and naïve T cell maintenance (Figure 4.6).

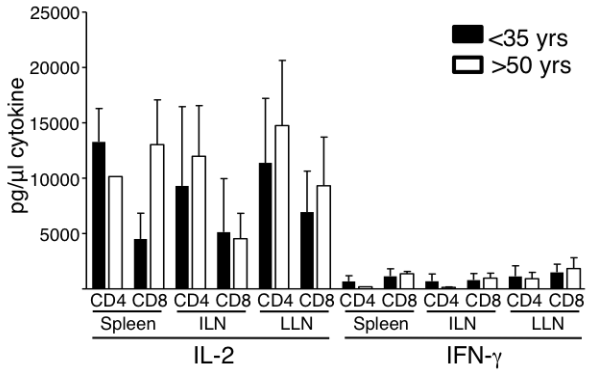

Figure 4.7: Levels of IL-2 and IFN-γ for three tissue sites (spleen, lung lymph node, inguinal lymph node), for sorted CD4$^+$ and CD8$^+$ naïve T cells. Donors <35 years of age are in white and >50 years of age are in black.

In light of the decreasing output of new naïve T cells and overall decrease in naïve T cell fraction, the functonality of the repertoire was studied to determine whether the functional phenotype had features of effector memory (TEM) T cells in older individuals. This was done by assessing their levels of cytokine production when activated with anti-CD3/CD28/CD2 beads. While antigen experienced TEM cells produce large numbers of INF-γ, IL-4 and IL-10 upon activation, antigen inexperienced naïve T cells produced predominantly IL-2 with low levels of these other cytokines. The results, summarized in Figure 4.7, indicate that the cells obtained from the older donors correspond to a strongly naïve phenotype.

4.3 Statistical analysis of sequence data

Following experimental quantification of T cells a subset of donor tissues were sequenced by Adaptive Biotechnologies. For these donors β chains of both naïve and effector memory T cells from spleen, lung lymph nodes and inguinal lymph node tissues were sequenced for CD4$^+$ and CD8$^+$ subsets. These samples are summarized in Table 4.1 and can be downloaded http://adaptivebiotech.com/pub/Farber-2016-SciImmunol. This section describes the statistical analysis of TCR repertoire diversity and divergence in human tissues to further investigate the compartmentalization and maintenance of the naïve repertoire.

Table 4.1: Sequenced donor TCR data

			CD4+					CD8+				
	Donor	Tissue	Reads	Clones	Entropy	Clonality	Simpson index	Reads	Clones	Entropy	Clonality	Simpson index
NAÏVE DONORS	D99	ILN	512666	16626	13.59	0.030	9.84e-5	520125	12720	13.22	0.030	1.29e-4
	D100	ILN	649404	23000	14.02	0.032	7.43e-5	332291	13843	13.26	0.036	1.31e-4
	D115	ILN						314150	5363	11.81	0.046	3.66e-4
		LLN						551707	13466	13.10	0.045	1.54e-4
		SP						332714	3756	11.20	0.057	5.92e-4
	D125	ILN	314186	7678	12.39	0.040	2.3e-4	270314	9165	12.65	0.039	1.96e-4
		LLN	227204	2150	10.49	0.053	9.2e-4	318147	13916	13.00	0.055	1.79e-4
		SP	140702	4716	11.53	0.055	4.7e-4	131403	2646	10.69	0.060	8.44e-4
	D127	ILN	454773	8792	12.41	0.053	2.5e-4	459948	6861	12.18	0.044	2.85e-4
		LLN						780510	16531	13.42	0.043	1.23e-4
		SP	875487	18953	13.57	0.045	1.1e-4	643360	20679	13.73	0.043	1.00e-4
	D139	ILN	654506	8666	12.39	0.053	2.6e-4	160460	5160	11.44	0.073	5.31e-4
	D141	ILN						710430	9587	12.29	0.071	1.32e-3
	D149	ILN	866857	57973	15.44	0.024	2.75e-5					
		SP	1091481	49286	15.11	0.031	3.57e-5	409254	44597	14.95	0.032	4.18e-5
	D194	ILN	939723	48041	15.26	0.019	2.90e-5	93814	2767	11.12	0.027	5.20-4
		LLN	501600	54184	15.25	0.030	3.27e-5	49448	3061	11.19	0.034	5.14e-4
		SP	267521	9014	12.81	0.025	1.6e-4	289309	7729	11.80	0.086	5.41e-3
	D200	ILN	87753	2836	11.14	0.028	5.1e-4	35805	892	9.43	0.038	1.86e-3
		LLN						362810	19556	13.71	0.038	1.96e-4
		SP	46383	1248	9.90	0.038	1.3e-3	51088	2471	10.72	0.049	1.40e-3
	D201	ILN	732402	50051	15.30	0.020	2.85e-5	801288	49967	15.39	0.020	2.91e-5
		LLN	2221562	13605	16.71	0.20	1.15e-5	2346619	120293	16.46	0.25	1.49e-5
		SP	795274	51203	15.32	0.021	2.84e-5	639176	53431	15.26	0.028	3.35e-5
	D238	ILN	690140	25427	14.22	0.028	6.42e-5	106063	3404	11.11	0.053	6.60e-4
		LLN	400712	19777	13.63	0.045	1.1e-4	29611	3416	10.15	0.136	3.00e-3
		SP	458654	29482	14.26	0.040	7.03e-5	199030	16270	12.98	0.072	2.61e-4
TEM DONORS	D72	ILN	6979048	26661	14.0	0.045	1.05e-4	4056501	7677	10.82	0.162	2.53e-3
		LLN	6952247	29265	14.19	0.044	8.76e-5	6657759	8757	9.55	0.270	3.72e-2
		SP	6525612	34202	14.10	0.064	2.15e-4	4144747	8903	10.56	0.195	3.75e-3
	D73	ILN	1923946	32134	14.18	0.052	1.34e-4	3230208	5816	8.03	0.358	4.89e-2
		LLN	2321345	31774	14.14	0.054	1.04e-4	2225157	8320	9.45	0.275	9.27e-3
		SP	2548563	24251	13.14	0.098	6.80e-4	4117270	5011	8.12	0.339	1.65e-2
	D76	ILN	2510322	33112	14.21	0.053	4.47e-4	3669324	14137	10.45	0.242	1.16e-2
		LLN	2791277	29359	13.93	0.061	7.53e-4	4058300	12870	9.79	0.283	2.00-2
		SP	2283091	22988	13.13	0.094	8.10e-4	967815	6278	8.64	0.315	2.29e-2
	D79	ILN	5349212	36240	14.43	0.047	1.68e-3	2165991	5237	9.76	0.210	1.01e-2
		LLN	3103785	27749	14.12	0.044	1.10e-3	2024061	6815	10.83	0.149	3.11e-3
		SP	5665441	29299	13.86	0.066	4.38e-3	9104434	8657	9.94	0.240	1.08e-2
	D86	ILN	2137677	24432	13.18	0.096	1.02e-4	2537491	8518	9.22	0.294	8.88e03
		LLN	2411597	22134	12.91	0.106	1.59e-4	1967586	6659	8.52	0.330	1.71e-2
		SP	1416897	18515	12.20	0.140	4.04e-3	2558884	6246	8.70	0.310	1.12e-2
	D125	ILN	914326	43415	14.69	0.0468	2.25e-4	241762	2031	8.57	0.220	1.18e-2
		LLN	947864	39781	14.61	0.044	1.32e-4	225814	3534	9.99	0.152	6.05e-3
		SP	573616	12510	12.10	0.111	3.75e-3	469019	5653	9.27	0.257	9.59e-3
	D137	LLN	1816358	58631	14.95	0.056	8.73e-5	1802110	23775	10.96	0.246	1.01e-2
		SP	1768448	43160	12.12	0.213	0.036	2553091	17209	9.84	0.301	9.39e-3
	D194	ILN	553353	26516	13.91	0.0533	1.44e-4	49814	1311	9.24	0.108	5.19e-3
		LLN	129564	12694	12.29	0.0985	5.38e-4	109116	3401	10.07	0.141	4.66e-3
		SP	150399	5100	11.53	0.0635	8.23e-4	98979	1672	8.87	0.172	1.35e-2
	D200	ILN	116335	3118	11.21	0.0345	5.57e-4					
		LLN						121286	3090	9.57	0.175	5.10e-3
		SP	200698	7868	12.13	0.0624	5.32e-4	184427	4092	9.72	0.190	4.84e-3
	D201	ILN	488793	23776	13.64	0.0617	5.97e-4	453061	4757	10.00	0.181	8.28e-3
		LLN	761650	32269	13.97	0.0670	5.13e-4	723093	19458	11.80	0.172	3.70e-3
		SP	818385	28198	13.31	0.100	3.93e-4	773821	13735	10.17	0.260	1.16e-2

4.3.1 Decrease in T cell diversity over lifetime

For all sequenced donors the nucleotide diversity was computed using Simpson index (Figure 4.8A). The abundance distribution indicated that the naïve repertoire did not contain large expansions. Therefore, for the number of reads collected, a higher order of true diversity was most informative for comparing the different T cell subsets. The Simpson index decreased significantly with age for the $CD4^+$ subset of the naïve repertoire, with significant loss of diversity in many of the donors after age 40, consistent with the loss of T cell fraction and thymic output previously described. A similar, though less pronounced trend was found in $CD8^+$ naïve T cells, again consistent with the somewhat smaller differences observed in T cell fraction. However, the $CD8^+$ repertoire had lower diversity overall at all ages, with the least diversity among older donors.

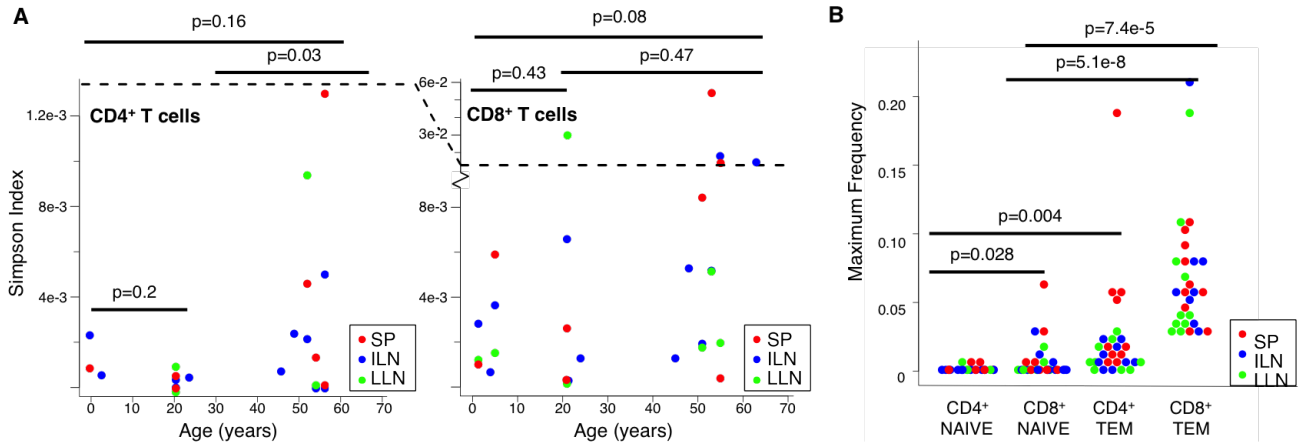

Figure 4.8: A. Repertoire diversity quantified by Simpson index for $CD4^+$ (left) and $CD8^+$ (right) T cells separated by age and tissue. B. Maximum clonal frequencies for all donors and tissues separated by T cell subset. P-values were computed using the t-test [94].

Diversity in both repertoires was further analyzed by maximum frequency and compared to the TEM repertoire (Figure 4.8B), showing much larger expansions in TEM subsets compared to naïve, as expected based on the large clone sizes found in [95]. Interestingly,

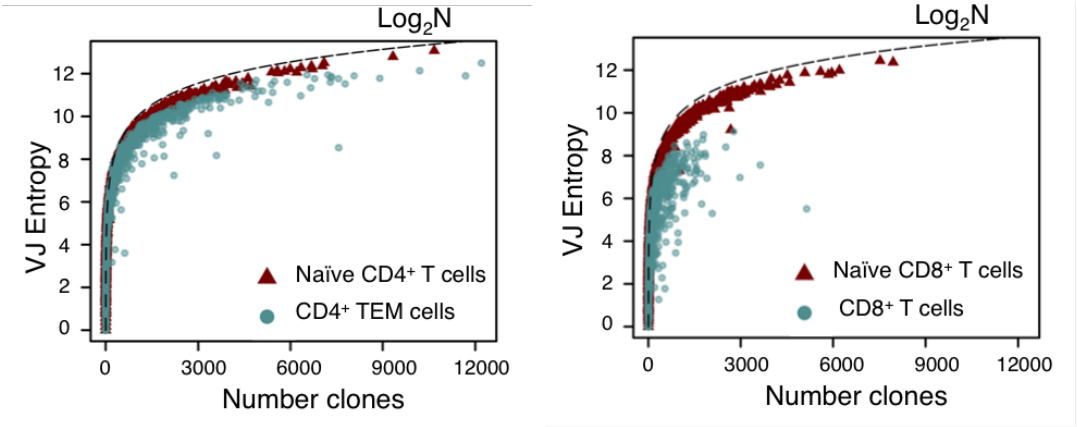

Figure 4.9: Shannon entropy of VJ usage $CD4^+$ (left) and $CD8^+$ (right) for each cassette pair plotted vs the number of clones generated from those cassettes. The maximum diversity is indicated by the black dashed curve [94].

slightly higher maximum frequency values were also found in $CD8^+$ T cells of the naïve repertoire as compared with $CD4^+$.

VJ diversity was looked at using Shannon entropy. This metric is the the best to use for looking at VJ cassette usage because the space of VJ pairs is many orders of magnitude smaller than the sequence space of CDR3s, thus producing more accurate frequency values, and because the distribution has more evenness. The VJ entropy was lower for TEM than for naïve T cells for both $CD4^+$ and $CD8^+$ subtypes, indicating a less diverse repertoire of VJ usage. The naïve diversity was close to the maximum possible entropy, whereas TEM shows presence of VJ cassettes from large expansions, particularly evident among the $CD8^+$ subset.

4.3.2 Analysis of clonal overlap between tissues

Having quantified the loss of naïve T cell diversity with age an analysis of clonal overlap was done to determine how the cells are maintained between distinct lymphoid tisue sites. The overlap among the TEM subsets was used as a point of comparison. An analysis of the top 1000 clonal sequences from all tissues showed a striking lack of overlap among naïve T cells at all ages and in both $CD4^+$ and $CD8^+$ subsets. In contrast there was significantly more

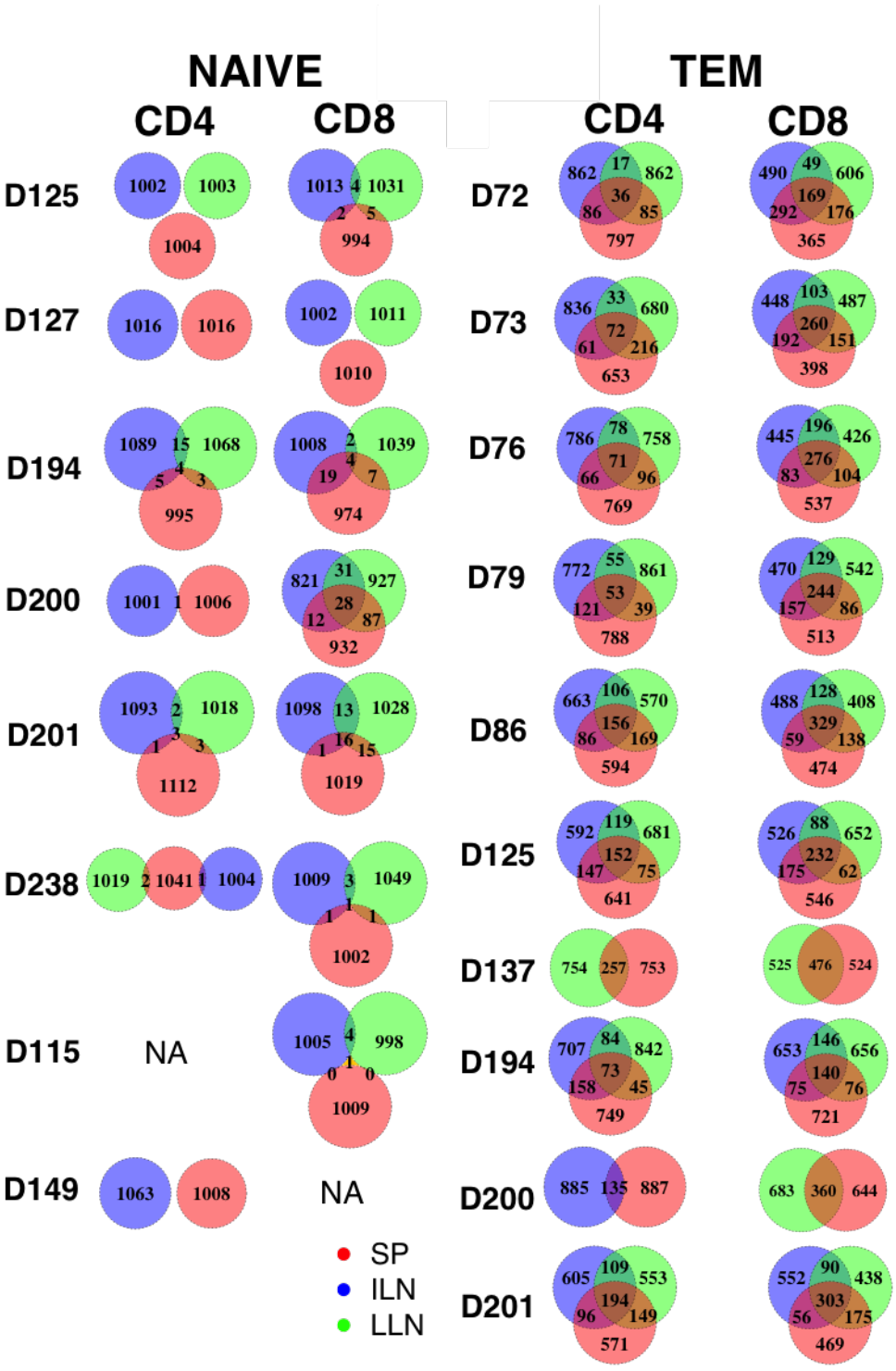

Figure 4.10: Inter-tissue sequence overlap among top 1000 clones [94].

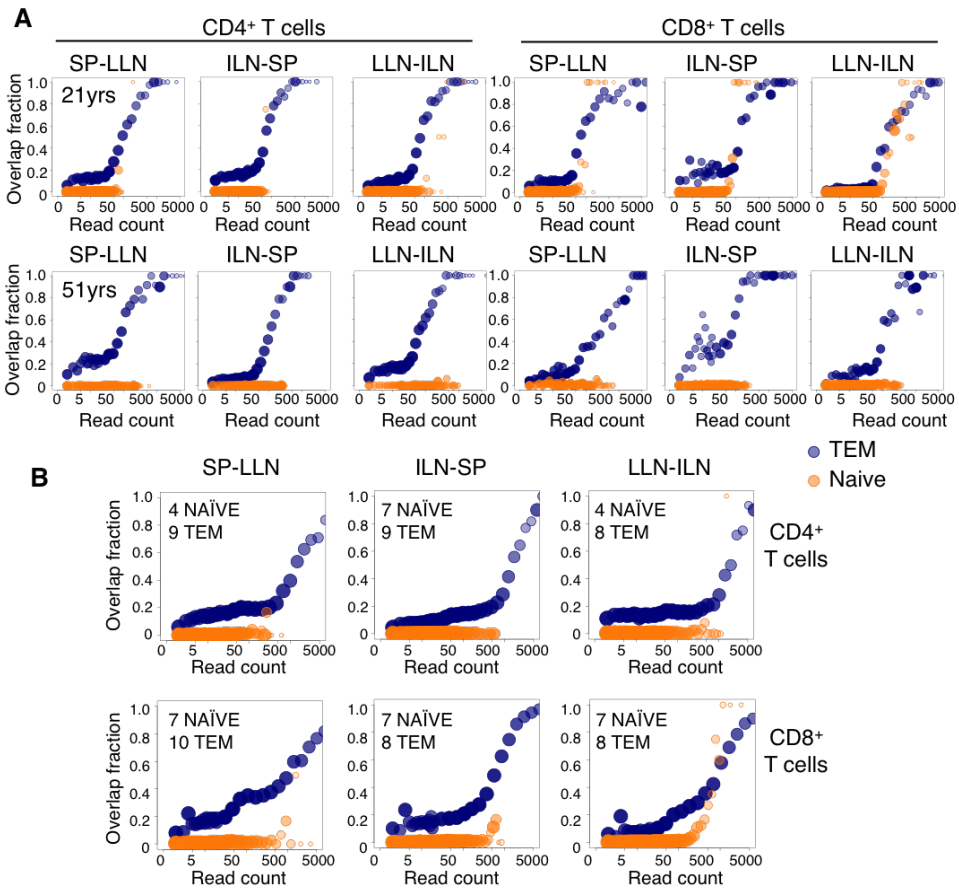

Figure 4.11: Overlap as a function of read count. Each plot represents a pair of tissues X-Y, and represents the fraction of overlap of clones in tissue Y with a subset of clones at a given read count in tissue X. The x-axis is binned and the transparency and size of the circles are proportional to the number of clones in each bin. Two representative donors of disparate ages, (21 and 51 years old) are shown in A. , while all donors are pooled together for B. Naïve T cells are showin on orange, and TEM in blue. [94]

sequence overlap between tissues within the TEM repertoire (Figure 4.10). The evidence therefore strongly suggests that regardless of thymic activity, functional naïve T cells have significantly less circulation between tissues compared to memory, in contrast with prior expectation.

Most clones are rare and many of the unseen shared clones may occur at low numbers, especially since TEM frequencies are typically higher overall. The overlap was therefore further quantified by read count to see if overlap is frequency dependent. In this case TEM still shared significantly more overlap at all read counts (Figure 4.11), with the trend more

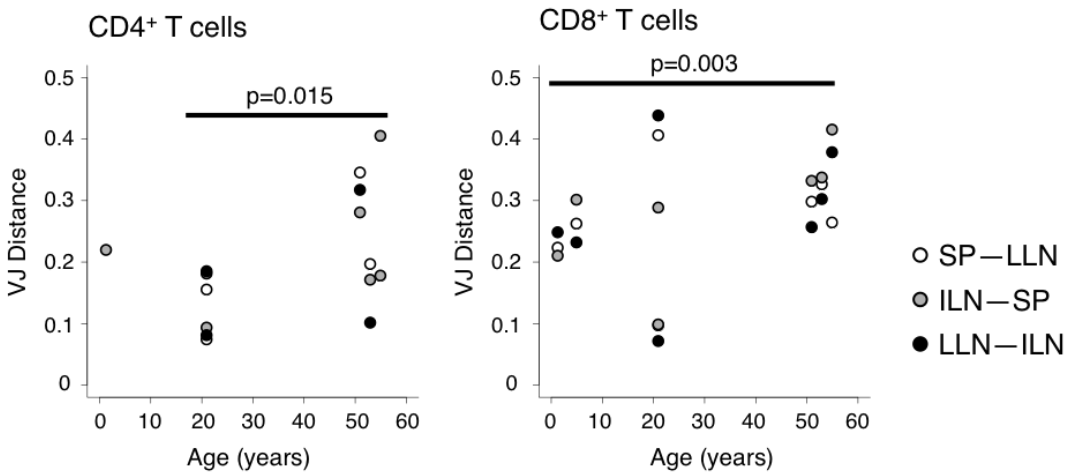

Figure 4.12: Inter-tissue VJ distance as a function of donor age for $CD4^+$ and $CD8^+$ subtypes. Each point represents the distance for a distinct pair of tissues as indicated. P-values are computed by a Wilcoxon rank test [94].

obvious among the $CD4^+$ subset. This was further confirmed using replicate overlap as a baseline.

4.3.3 Site specific maintenance of the naïve repertoire

Lack of overlap was further measured by VJ usage, using the Jensen Shannon Distance (JSD). The lack of tissue sharing, and depleted seeding of new cells from thymic output suggested the presence of a tissue specific maintenance mechanism for naïve T cells. We observed a significant increase of inter-tissue JSD for the $CD4^+$ subset and a similar trend between pediatric and older donors in $CD8^+$, suggesting that this may indeed be the case (Figure 4.12).

4.4 Methods

Detailed methods are provided in Thome, Grinshpun, et.al. [94]

4.4.1 Organ tissue acquisition and experimental analysis

The research presented here focused on capturing an accurate representation of the aïve T cell receptor repertoire from lymphoid tissue sites. Therefore all tissues were obtained from diseased (brain-dead) human donors at the time of organ procurement for transplantation. Donors were selected to be free of chronic disease and cancers, and tested negative for Hepatitis B and C, and HIV. All organs were flushed with cold preservation solution after extraction. Additional thymic tissue was collected by cardiothoracic surgeons during perdiatric cardiac surgery. A range of ages was selected from 2 months of age to 73 years to ensure accuracy for investigating longitudinal changes.

Histological analysis of thymic tissue was performed from slices of stained and cryopreserved tissue, with Hassal corpuscles counted from three separate tissue sections at 10x magnification. T cells from each tissue were sorted into $CD4^+$ and $CD8^+$ subtypes by flow cytometry and analyzed for cytokine content by cytometric bead array. TRECS numbers were obtained from sorted cells by real-time PCR.

All TCRβ sequencing was performed by Adaptive Biotechnologies using their proprietary ImmunoSeq platform [81], obtained from DNA in sorted cells.

4.4.2 Statistical Analysis of TCR receptor repertoire

Sequenced CDR3 sequences were filtered and selected for productive sequences as indicated by the Adaptive processing pipeline. Nucleotide, amino acid, V and J gene, and read data was used for all the analysis. At the time the analysis was done, few of the datasets provided template information, and performing conversion from reads to templates was not readily available. Clones were further filtered by sorting error, under the assumptions that a particular CDR3 sequence belongs to either the $CD4^+$ or $CD8^+$ subset. A minimum two-fold difference was required between $CD4^+$ and $CD8^+$ to identify a clone as belong to one or the other subset. Ambiguous clones were discarded. In total, $<0.4\%$ of clones in naïve samples were discarded and $<1.5\%$ in TEM samples, consistent with the roughly 99% accuracy

during cell sorting.

Analysis of clonal overlap for top clones looked at the top 1000 clones by read count, thereby producing slightly more than 1000 clones in those cases where the smallest read number was shared by multiple distinct sequences. Analysis of overlap fraction by read count used replicate data from both naïve and TEM cells as a baseline for clonal overlap. The overlap fraction was computed with the denominator determined by the clonal abundance of the sample on the x-axis at the specified read count. Clones were binned on the $\log_{10}$ scale to avoid inflated overlap fractions due to low clone counts among larger frequencies.

Statistical analyses utilized Simpson index, entropy, and Jensen Shannon distance, as described in Chapter 2.

4.5 Conclusion

This is a novel study that comprehensively investigated the naïve T cell repertoire within human lymphoid tissues. Thymic function and output was assessed, as well as age related change in naïve T cell fraction. The data showed that there was virtually no thymic output in healthy individuals over age 40, while a small fraction of functionally naïve T cells was homeostatically maintained in the tissues. Studies of diversity and divergence of TCR sequence repertoire further confirmed these findings, and also established that naïve T cells have significantly less inter-tissue sharing when compared to TEM. Slightly higher diversity and sharing was found in the $CD8^+$ subsets as compared to $CD4^+$, but in both cases the evidence suggests tissue specific maintenance in the lymph nodes. This understanding of how the naïve repertoire is affected by aging is an important aspect of human immunity to consider and explore further when designing vaccines and developing new immunotherapies.

4.6 Discussion

One of the key challenges in performing this analysis was determining the appropriate methods for diversity analysis. The naïve repertoire contains few clonal expansions and the

samples obtained were often small (only a few thousand clones). Measurements of entropy and clonality remained useful when looking at VJ combinations, where combinations are limited, and frequency calculations reliable. However, these measures were not stable for the nucleotide counts available. In such a situation, the best approach is to use the minimum value of true diversity, q, that provides an informative measure of the data. The Simpson index, at q=2, provided a reliable measure that captured trends in the repertoire diversity across different age groups, and indicated that for older donors there were increased numbers of small clonal expansions . Measures such as maximum frequency can capture such differences, but are not able to provide such a complete explanation for the observed trends.

The Jensen Shannon divergence was used as a measure of divergence in VJ cassette analysis, but not in nucleotide analysis. This again is due to the decrease in theoretical diversity of VJ pairings compared to CDR3. JSD assumes that both datasets derive from a larger distribution in which all elements are represented. Due to the uncertain dynamics of naïve T cell seeding in the lymphoid tissues, as indicated by the lack of overlap, this measure could not be reliably applied even to tissues from the same donor. All of these considerations are described to highight how imperative it is that sample size, distribution, and the assumptions inherent in use of the statistical methods are considered prior to running data analysis of TCR repertoires and other highly diverse populations.

Chapter 5

Diversity and Divergence of the glioma infiltrating T cell repertoire

5.1 Introduction

Gliomas are the most common type of brain and spinal cancers, arising from the transformation of glial cells. The most severe and aggressive of these is glioblastoma (GBM), which accounts for more than half of diagnoses and typically has poor rates of survival (1-2 years with treatment) [77]. Although prognosis is typically done base on histological analysis of biopsies, it is now known that the molecular origins of the disease vary, with distinct glial cells of origin and affected gene pathways [74, 99, 45]. Typical treatment involves surgical tumor resection followed by chemotherapy. However, despite all of these treatment protocols, glioblastoma survival rates remain poor, in large part because GBMs employ a number of strategies that suppress the immune response [78]. Researchers seek to understand the exact mechanisms involved in immunosuppression and to develop novel immunotherapeutic treatments that target the dysregulated pathways [44].

This chapter outlines the application of diversity and divergence methods previously described in Subsection 2.3.2 to the study of the tumor infiltrating lymphocyte (TILs) sequence repertoire of T cells. Tumor tissue and blood samples were obtained from hospital biopsies of

Figure 5.1: Image of a GBM tumor taken from [13].

glioma patients as well as from healthy controls. The resulting analysis uncovered previously undescribed phenotypes of the disease, using only TCR sequence data. TCR sequences from blood were found that are associated with these phenotypes and can potentially serve as immunological markers for studying glioma progression. The results are published in [89].

5.2 Preparation and sequencing of the T cell repertoire

This section summarizes the experimental methods by which T cells were acquired from PBMC and tumor tissues, and the computational tools developed to extract VJ usage and CDR3 sequence information from raw sequence data. All experimental work was done by collaborators.

5.2.1 T cell collection and sequencing

RNA sequencing libraries for both α and β receptor chains were prepared from peripheral blood mononucleocytes (PBMCs) and cryofrozen tumor tissue, obtained from glioma patients in the hospital. PBMCs were collected from several healthy individuals as well as healthy brain tissue from non-glioma related biopsies. A total of three non-neoplastic (NN), three low grade glioma (LGG) and 8 GBM samples in total were analyzed in this study and raw sequence files are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79338.

A sample summary is provided in Table 5.1

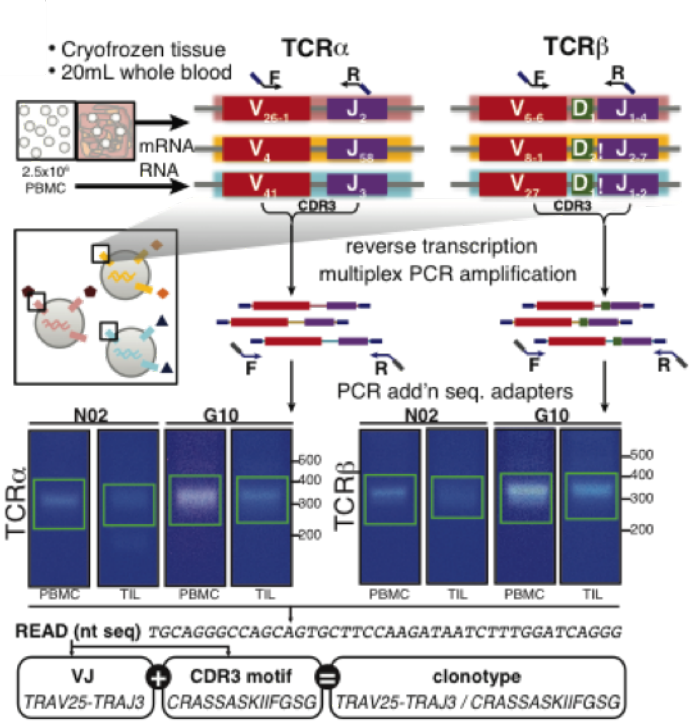

Figure 5.2: Library preparation for TCR sequencing of glioma samples. RNA is extracted from PBMC and tumor tissue with specialized primers used to amplify the α and β TCR chains. V and J cassette identity and CDR3 sequence are determined using a computational pipeline to uniquely define a T cell clone [89].

Primers for V and J cassettes were obtained from the iRepertoire sequencing kit, and reverse transcription and amplification were performed using a kit from Qiagen. Sequencing was done at the Columbia Genome Center on an Illumina MiSeq, generating paired end reads which overlapped the nucleotide sequence belonging to the CDR3. Unlike in previous chapters T cell data was not sorted into $CD4^+$ and $CD8^+$ data, but both receptor chains were sequenced.

5.2.2 CDR3 identification

Sequenced reads were merged using FLASH 1.2.11, which resolved mismatches based on sequence quality score. The merged sequences were mapped to the human GRCh37 reference genome using the Burrows-Wheeler Aligner (bwa-mem). A complete recombined TCR chain

Table 5.1: Sequenced LGG, GBM, and NN patient data

		PBMC				TIL			
	ID	Unique CDR3	Reads	Clones	VJ pairs	Unique CDR3	Reads	Clones	VJ pairs
TCRα	**N01**	358919	14164782	517570	2161	6790	2953925	8369	1002
	N02	300671	16166316	416182	2117	2395	727218	2824	441
	N03	181201	6247481	254769	2028	288	4157	290	195
	L04	26324	5871277	29035	1200	6598	848196	7740	1081
	L05	173601	11785693	213297	2022	11409	2135680	12151	1156
	L06	61527	1525449	72414	1917	6713	111665	7436	1420
	G07	195466	6891503	256062	2040	14397	2102587	17215	1679
	G08	229633	3251968	294383	2060	10244	548877	14109	1652
	G09	230166	4318029	305613	2077	7449	289283	8147	1472
	G10	237445	4208474	316193	2115	16276	581924	19960	1782
	G11	220503	3473415	279029	2044	11891	994193	14931	1661
	G12	156073	4381706	218793	2057	32724	777599	41309	1915
	G13	28887	2588839	37778	1800	6134	290766	7418	1362
	G14	27321	2530088	32503	1681	6732	611661	8897	1511
TCRβ	**N01**	479582	10840682	984226	660	3720	597886	5971	368
	N02	261484	14506770	515287	667	4638	533406	6140	368
	N03	139687	4846289	279874	642	5465	2086134	8106	390
	L04	55640	12542641	71676	573	10271	1396407	14498	456
	L05	89601	11954418	112704	591	12730	2072898	17530	485
	L06	7795	14843	8889	456	6034	161583	9480	504
	G07	230577	12706006	430302	650	16516	500294	23496	543
	G08	298761	3145154	540100	652	10857	823759	19895	546
	G09	202923	2896028	370845	648	6101	193705	9069	487
	G10	462931	11283991	979855	664	18606	615132	34293	576
	G11	244552	3324140	408173	640	14655	1259883	27190	562
	G12	138056	4203412	261142	643	40613	1157226	80834	614
	G13	37387	1919386	53294	589	2532	75313	4351	419
	G14	16864	2515617	24505	534	9595	514210	20567	563

sequence would be mapped twice, corresponding to the V and J genes in the unrecombined genome. After V,J identification, the CDR3 region was selected using *in silico* translation based on conserved sequence motifs (See Figure 2.5). An overview of library preparation and CDR3 identification is shown in Figure 5.2

5.3 Analysis of TCR repertoire diversity

Clonal diversity was measured by Shannon entropy in order to make use of the diversity of independent components being separable. The total diversity was partitioned into two part: H_{VJ}, the component produced by VJ usage and therefore representative of T cell generation, and H_{Δ}, the VJ idependent component of the CDR3 amino acid sequence which is tied to the antigen response of activated T cells. The total clonal diversity is related to these components by the following expression:

$$\boxed{H_{clonotypes} = H_{VJ} + H_{\Delta}}$$

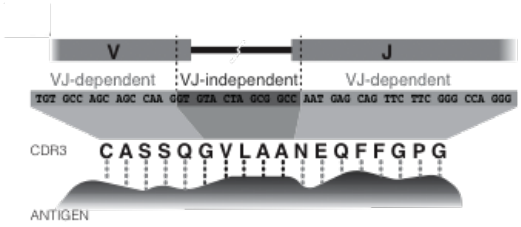

Figure 5.3: Visual representation of clonal diversity being separated into VJ dependent and VJ independent components [89].

The extend of VJ expansion was found to vary significantly between individuals, in both PBMC and TIL. Many of the GBM samples exhibited highly expanded VJ pairs, as indicated by the thick color ribbons in the Circos plots Figure 5.5, but the identity of these pairs differed between individuals. However, glioma patients had a consistently greater VJ-independent H_{Δ} component, consistent with antigen driven activation. The Circos plots were generated using code that can be found at https://github.com/bgrinshpun/CircosVJ.

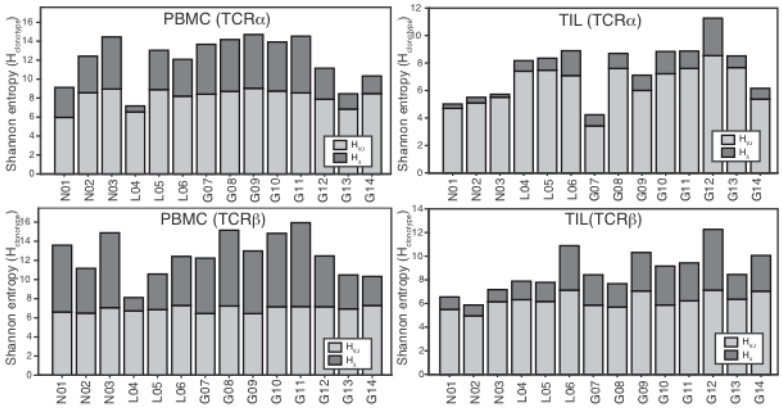

Figure 5.4: Separation of clonotype entropy into components by patient [89]

VJ cassette diversity was also quantified using clonality. GBM was most diverse within the TIL population fo both VJ dependent and independent components, suggesting a polyclonal antigenic response. However, non-neoplastic patients showed the lowest diversity in TIL, consistent with the brain being an immunologically protected organ (See Figure 5.6).

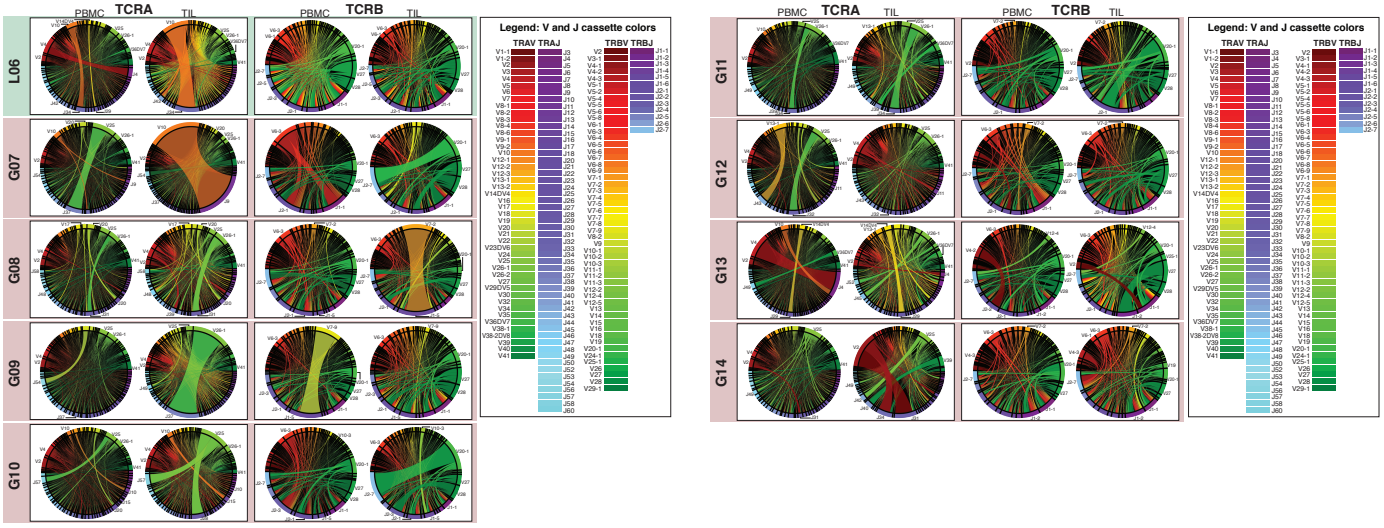

Figure 5.5: Circos plots of VJ usage [89]

5.4 Analysis of TCR repertoire divergence

T cell repertoire from PBMCs and TILs within the same individual was compared using Jensen Shannon distance. Once again a VJ dependent and VJ independent components were

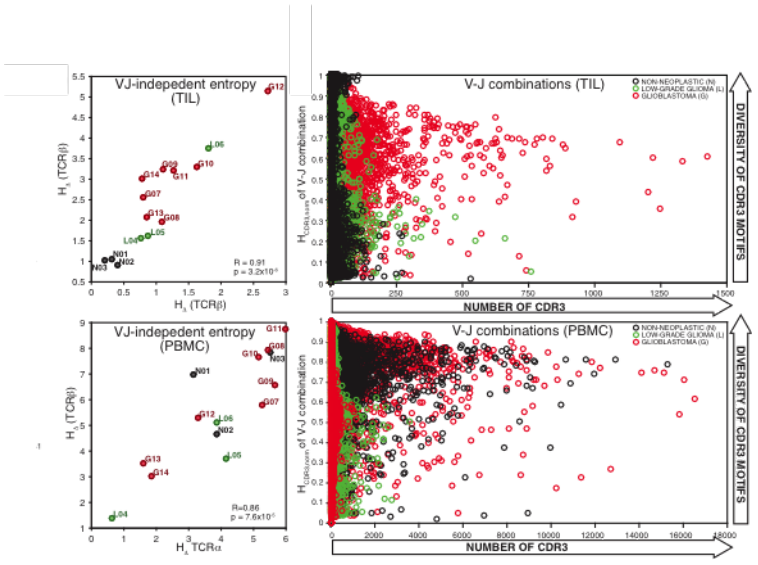

Figure 5.6: Entropy for VJ-independent components and VJ cassette combinations [89]

computed. Recall that Jensensen Shannon divergence can be expressed as a combination of entropies, thus allowing for the following additional relationship:

$$JS_{clonotype}(PBMC|TIL) = JS_{VJ}(PBMC|TIL) + JS_{\Delta}(PBMC|TIL)$$

The Jensen Shannon distance is then acquired by taking the square root $JSM = \sqrt{JS}$.

In order to account for the larger number of T cells available in the blood, T cells derived from PBMCs were subsampled to the size of the brain repertoire (PBMC') to obtain a corrected (corr) value for the true diversity given by the following relationship:

$$JSM_{\Delta,corr}(PBMC|TIL) = JSM_{\Delta}(PBMC|TIL) - JSM_{\Delta}(PBMC|PBMC')$$

The resulting corrected distances are summarized in Figure 5.8 for average α and β chains. The deviation of the VJ-independent Δ distance from zero quantifies how the TIL and PBMC distance compares to the expectation for a complete PBMC repertoire. Striking differences in distance were observed, with all non-neoplastic samples falling below expectation, suggesting minimal impact of antigen response to overall divergence, whereas LGG

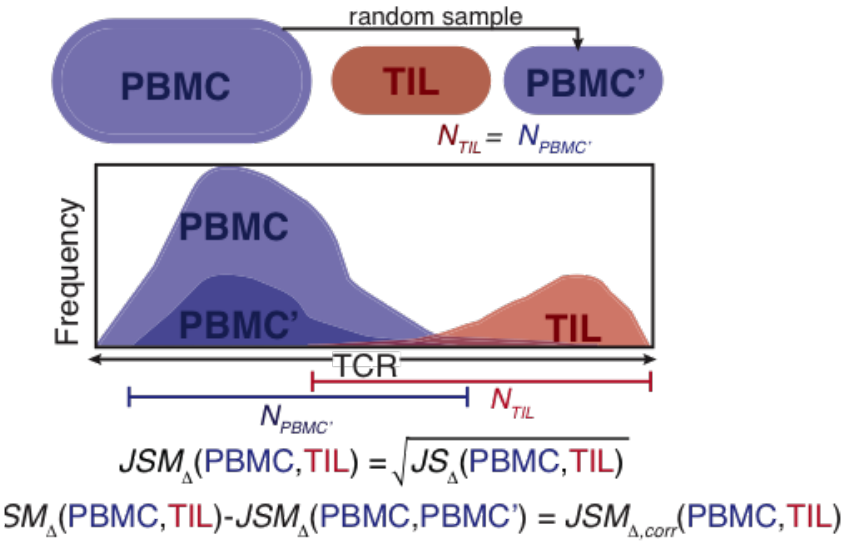

Figure 5.7: Visual representation of ΔJSM calculation. PBMC is subsampled to the same size as the TIL repertoire and the property of JSM as a distance metric is used to compute a final diversity score [89].

samples were significantly above expectation corresponding to a divergence strongly influenced by neoantigen driven T cell activation. On the other hand GBM varied between patients, with several of the samples having Δ divergence more closely resembling LGGs vs others more closely resembling the NN repertoire, indicating phenotypic differences among different GBM tumors. The $JSM_{\Delta,corr}(PBMC|TIL)$ of GBMs was not correlated with white blood cell count or steroid use, and is therefore unlikely to be due to lymphopenia or therapy based immunosuppresion. Thus, separating the Jensen Shannon distance into its VJ-driven and VJ-independent components serves as a means of separating active from inactive T cell populations in the brain.

5.5 A public PBMC repertoire is associated with TIL divergence

Although strong signatures of diversity and divergence can be associated with the GBM repertoire, surgery is required to obtain brain tissues for this purpose. Finding a biomarker

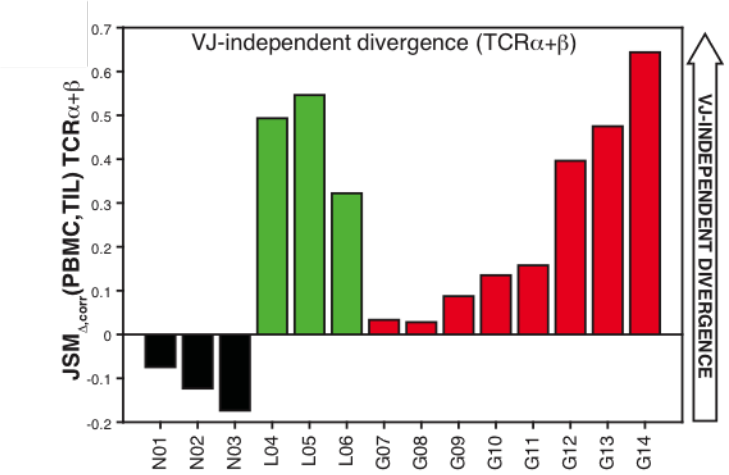

Figure 5.8: VJ independent JSM across samples [89].

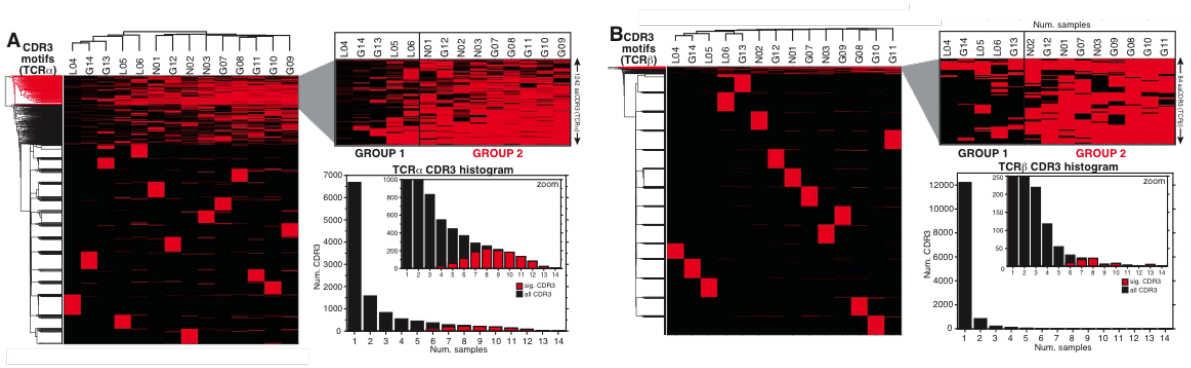

Figure 5.9: Heatmap of observed signature clones for α chain (panel A) and β chain (panel B) with clones ordered by hierarchical clustering. Red boxes indicate that the clone is observed in the indicated samples. Signature clones are shown in the zoomed in region to the right of each heatmap (1242 for TCRα) and (84 for TCRβ). The histograms represent an abundance for the unique clones shared by the number of samples indicated on the x-axis [89].

by which to track GBM status in the blood would provide a means to monitor the cancer in a non-invasive way. To explore the possibility for such a means of cancer tracking, top TCR clones by frequency were looked at across all patients PBMC samples from a total collection of 11638 α chains and 13,561 β chains. For both chains a set of "signature" clones was identified that separated patients into groups, 1326 combined total for α and β chain (Figure Figure 5.9).

Strong positive correlation across samples was observed between the fraction of these

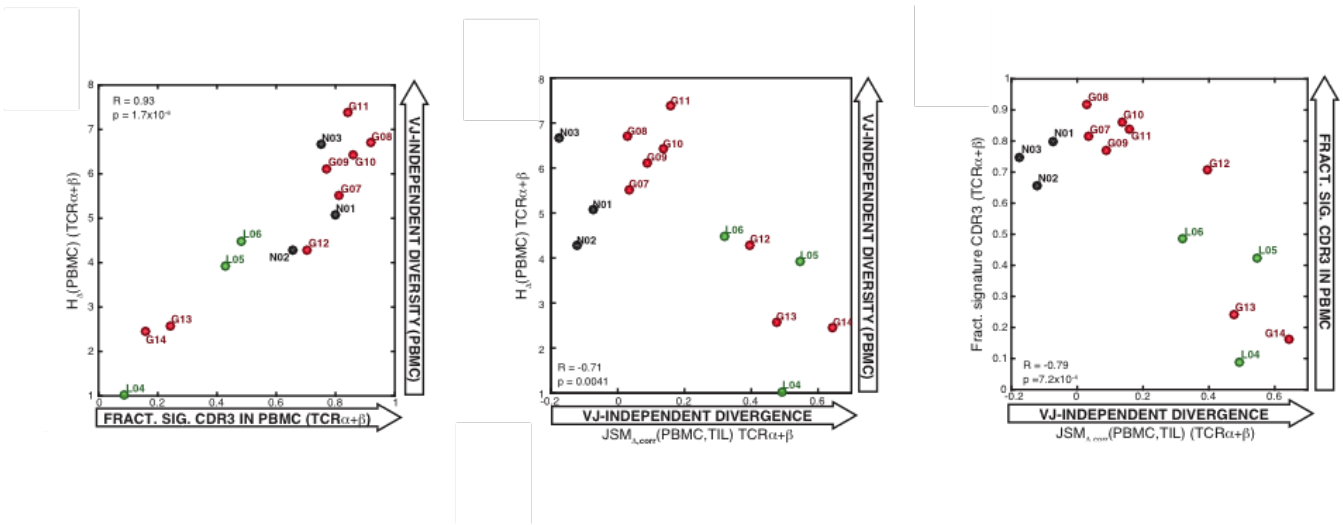

Figure 5.10: Left: VJ-independent PBMC diversity vs signature CDR3 fraction, Middle: Divergence vs VJ-independent PBMC diversity Right: Signature CDR3 fraction vs VJ-independent PBMC diversity [89].

CDR3s present in the sample and the VJ-independent entropy based diversity of the blood. There was also strong negative correlation between this diversity and the JSM quantified VJ-independent divergence between TIL and PBMC. Putting these two correlations together, a final result was obtained that showed a strong negative correlation between the fraction signature CDR3s in the blood and the $\Delta_{JSM}(TIL, PBMC)$. Most of the GBM samples showed high signature CDR3 fraction and low divergence, similar to that of NN samples, while two of the samples were more like those of the LGG which had low CDR3 fraction and high divergence. This suggests that the microenvironment of many GBM samples phenotypically resembles that of NN samples rather than the more robust immune response of LGG.

5.6 Comparison with previous studies

5.6.1 Signature clones among healthy PBMC samples

Both TCR α and β chain information was downloaded from six additional healthy human PBMC populations described in [112]. While this study defined CDR3 clones by their full

FGXG amino acid motif, most other studies truncate the definition of a CDR3 to include only the first phenylalanine (F) residue. The resulting numbers of unique amino acid clones were 1,241 for the α chain and 74 for β.

Presence of signature clones was assessed in these samples and used to recluster the data (Figure 5.11). Among these samples 5/6 of samples clustered with the NN patients among α chain sequences and all of the samples were clustered with NN for β chain sequences, indicating the effectiveness of signature clones in separating healthy individuals from the glioma phenotype.

5.6.2 Viral reactive clones

Signature clones were compared with previously identified public clones from [12] that were found to be shared by many individuals with diverse repertoire profiles. Only β-chain information was available; however, 62/74 (83%) of the signature clones were found among this

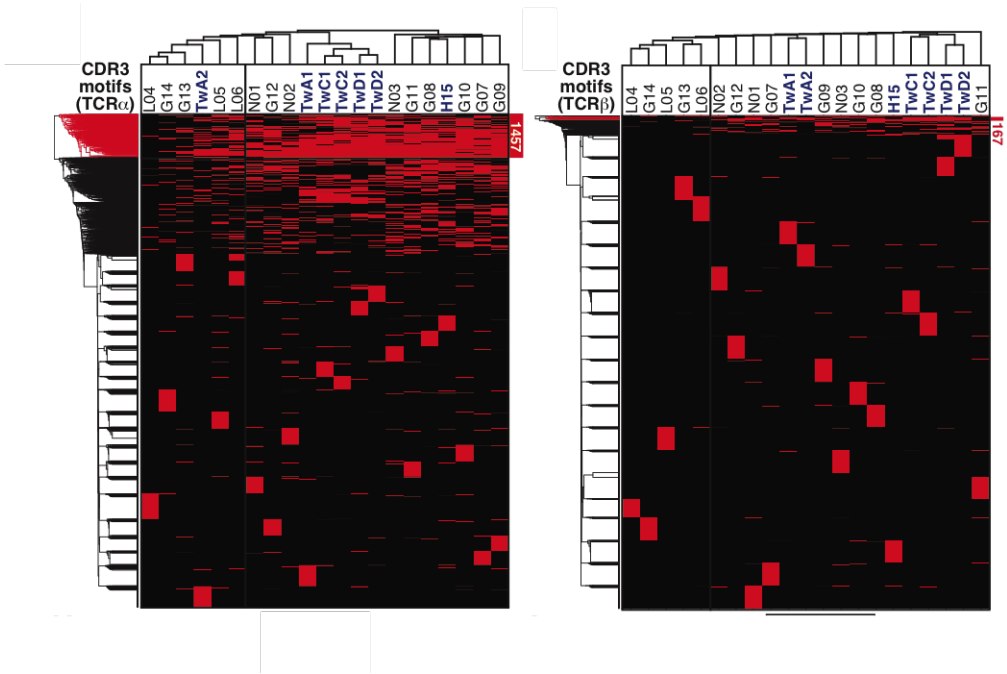

Figure 5.11: Heatmaps reclustered with additional healthy samples indicated in bold blue font. Signature clones effectively clustered healthy samples into the group that contains the non-neoplastic samples for α-chain (left) and β-chain right [89].

list, suggesting that there may be a link between public clone use and glioma phenotype.

A similar analysis was performed for viral reactive clones for a number of common pathogens including *Clostridium tetani*, *Candida albicans*, *Mycobacterium tuberculosis*, HSV, CMV EBV and influenza, compiled from [7, 98, 52, 59, 108, 96, 8, 55, 91]. Once again, only β chain information was available. Low levels of enrichment were found for *Clostridium tetani* (8.1% among signature vs 0.5% among non-signature), *Candida albicans* (2.7% vs 0.08%) and *Mycobacterium tuberculosis* (27% vs 1%). A point of interest that may warrant further consideration is that all of these enrichments were for non-viral pathogens.

None of these specific clones were able to cluster patients by cancer status as effectively as the complete set of β chain clones obtained in this study, indicating that the complete set of signature clones more comprehensively describes the glioma associated repertoire.

Signature clonal overlap with healthy PBMC, public, and viral clones is summarized in Figure 5.12.

5.7 Conclusion

PBMC derived T cells and tumor infilitrating lymphocytes from glioma biopsies were sequenced from RNA to study the diversity and divergence of the immune tumor response. A novel approach was applied separating VJ-dependent and VJ-independent components of the repertoire. Diversity and divergence were quantified for each of these components

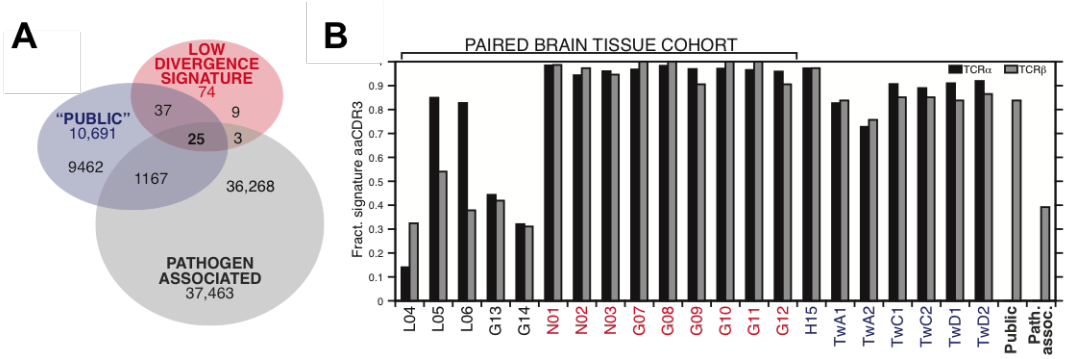

Figure 5.12: A. Venn diagram of overlap between signature clones, viral-associated clones and public clones. B. Fraction of signature clones identified for each sample [89].

to identify differences between a healthy repertoire, represented by the non-neoplastic patients, and the glioma patients, including six with glioblastoma multiforme. In particular, the analysis focused on the VJ-independent component which is indicative of antigen response rather than aspects of T cell maturation and brain specific compartmentalization. TILs from GBM exhibited higher VJ-independent diversity and divergence. CDR3 clones shared among PBMC samples from glioblastoma patients were found to be predictive of the degree of divergence, thereby defining a potential biomarker for studying tumor progression.

5.8 Discussion

There is great interest in identifying components of the T cell response that can provide insight into diagnosis and/or prognosis of cancer. High throughput repertoire analysis of the TCR repertoire has the potential to identify specific T cells involved in immune regulation and neoantigen binding. These can serve as markers may become the targets for novel immunotherapies and tracking of disease progression.

In this study, we have shown that the combination of CDR3 amino acid and VJ usage is a powerful method for analyzing diversity in response to disease. With blood serving as a reservoir for T cells from different tissues in the body, the Jensen Shannon distance provides a way to directly compare distinct tissue sites while controlling for differences in sample size. Ideally, a TCR repertoire study will reveals information that is testable from a sample of blood, thus avoiding the need for more invasive tissue extraction except in cases where surgery may be required.

This work also highlights the importance of capturing the often neglected α-chain of the TCR. The α-chain has more potential for diversity due to greater V-J cassette pairings, and as shown in this study, it is this aspect of the repertoire that may uncover a disease signature. More likely, the exact combination of the two chains is more important than any single chain and hopefully as technology improves the analysis of TCR $\alpha--\beta$ pairing in disease will become a staple of TCR repertoire profiling studies and reveal complex relationships that

single chain analysis is unable to resolve.

Finally, most studies as well as those presented in some of the other chapters, where the data was not sequenced and processed in-house, provide a single VJ pair for each nucleotide sequence. What we observe here, and in fact what should be expected given how the repertoire is generated, is that multiple VJ cassettes can produce identical CDR3 sequences. These differences in TCR generation may manifest as HLA-specific effects in disease response, as factors that affect the signaling or cofactors, or perhaps even components of the protein structure which have downstream effect on the shape of the CDR3 during antigen binding. While more work is required to truly explore the relationship between VJ choice and CDR3 binding potential for identical nucleotide sequences, there is still much to learn about how TCR recognition and activation occurs on the amino acid binding level, and it is certainly an area that ought to be explored further.

Chapter 6

Diversity of the human alloresponse

6.1 Introduction

Alloreactivity is the robust immune response to the tissues of another organism of the same species. In 1944 Peter Medawar first showed in experiments on rabbits that rejection of skin allografts was initiated by the immune system [60]. Work by MacFarlane Burnet in colonial marine forms and flowering plants further hilighted that the immune system is involved in self-recognition of tissues [16]. Failure of thymectomized mice to reject grafts first proved that the T cell response was the basis for immune rejection [63, 41]. It is now understood that self-tolerance and allorecognition rely heavily on the MHC molecule, in addition to the peptide it presents to the T cell [87].

An obvious benefit from understanding the mechanism behind alloreactivity is the ability to succesfully transplant tissues and organs between people. Many hypotheses exist for why the immune system exhibits alloreactivity, but it is unlikely that this response came about as a result of an evolutionary benefit that it provided [24]. However, it is well known that T cells generated in the thymus undergo clonal selection to prevent autoreactivity, as dicussed in Subsection 2.1.2. This mechanism is only effective for the HLA genes encoded by that individual, which leaves open the possibility for mature T cells that leave the thymus to be reactive to isoforms of HLA from other individuals [38, 57, 85]. For this reason, transplants

performed in the hospital require careful HLA matching between donor and recepient in addition to regimens of immunosupressive drugs. Current estimates of the strength of this response suggest that roughly between 1-10% of T cells are alloreactive [87, 104, 5, 105, 53] but these numbers come from radioactive labeling and dilution assays, which rely on signals with limited dynamic range and sensitivity.

This chapter describes a comprehensive study of the alloresponse from high throughput sequencing of the T cell receptor repertoire. Measures of diversity and divergence are used to compare healthy repertoires to alloreactive ones, to compare the effect of T cell stimulators with differing levels of HLA matching with the recipient, and to describe a novel method to accurately quantify the fraction of alloreactive T cells by using data from sequencing.

6.2 Experimental procedure and data processing

6.2.1 Stimulating T cell alloreacivity by MLR

To investigate the alloreactive repertoire PBMC samples were collected from healthy individuals. A mixed lymphocyte reaction (MLR) was performed, in which an alloresponse from a healthy individual (the responder) was evoked due to stimulation by an irradiated sample from another individual (Figure 6.1A). The responding repertoire was labeled by CFSE staining, while the irradiated sample was stained with a violet dye. Irradiated T cells were unable to proliferate, whereas proliferation of responder T cells was tracked by dilution of CFSE. T cells were sorted into $CD4^+$ and $CD8^+$ subsets, generating four populations – two from an unstimulated responder sample, and two from the expanded cells within the CFSE stained population. The cells were then sent to Adaptive Biotechnologies for β-chain sequencing (Figure 6.1B). All MLR experiments were performed by collaborators.

6.2.2 Identifying alloreactive clones

Although stimulated clones were identified on the basis of their CFSE signal, the distributions of stimulated vs unstimulated T cells is not perfectly separable. Therefore there is low

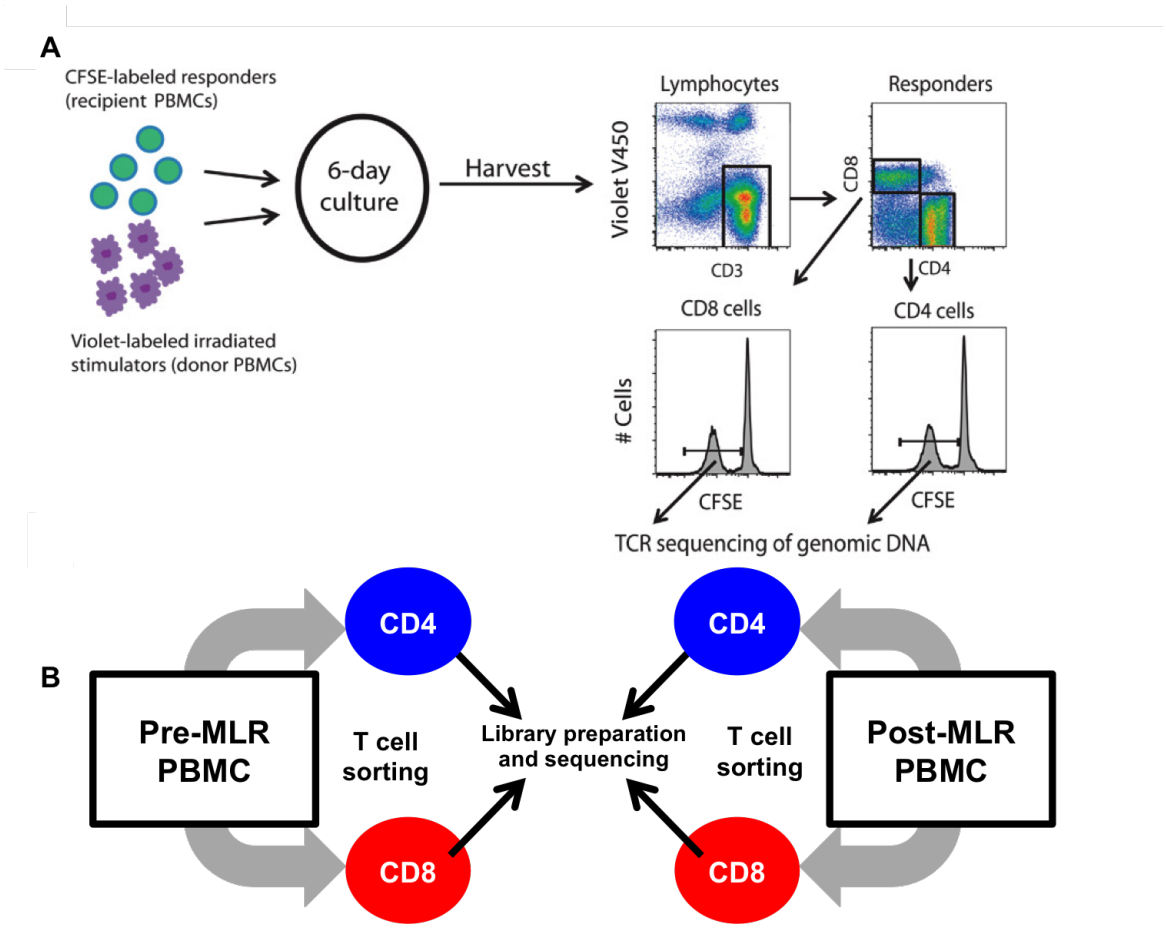

Figure 6.1: A. Description of the mixed lymphocyte reaction reprinted from Figure 1 of [54] with permission from AAAS. B. Overview of sequenced samples.

level error in classifying low frequency clones as alloreactive. Unstimulated and alloreactive T cell frequencies were plotted against each other to identify a cutoff threshhold for alloreactivity. K-means clustering was used to separate the clones into either of two populations, true stimulated vs error (Figure 6.2). Multiple cutoff criteria were test, from 1-5x, to define the best ratio of stimulated to unstimulated clones. Across samples, the best separator for defining allreactivity was a 2x threshhold. Clones with frequency below 1e-5 could not be well resolved by any linear separator, and therefore were not included in further analysis of alloreactivity. Approximately 0.05-2% of clones were removed in this way (Figure Figure 6.3). A summary of all healthy control samples used in the analysis of alloreactivity are summarized in Table 6.1.

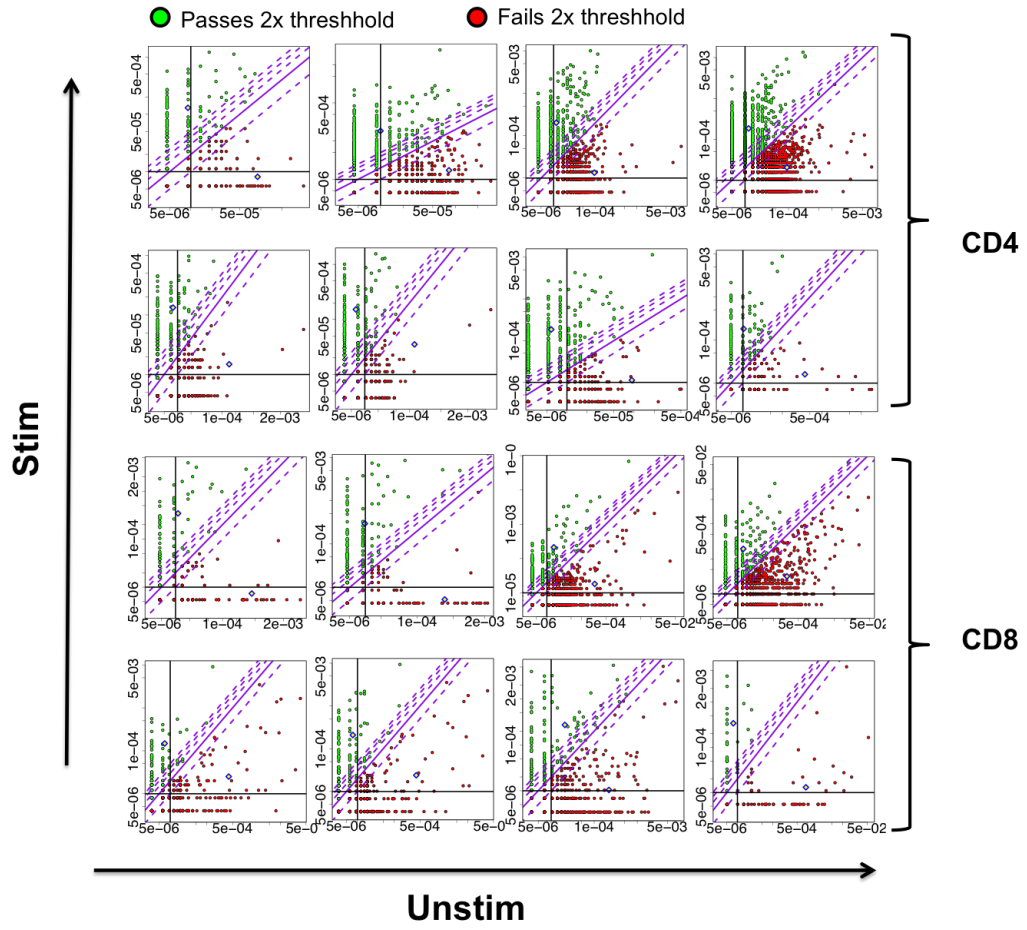

Figure 6.2: $CD4^+$ and $CD8^+$ clones are clustered by k-means to determine whether they correspond to the alloreactive population or to CFSE error. Red clones fail the 2x criteria, green clones pass the 2x criteria. Cluster centers are shown as yellow diamonds with a blue border. The 2x separation is given by the solid purple line, while 1x,3x,4x,and 5x thresholds are indicated by dashed purple lines. Frequency cutoffs of 1e-5 are indicated by vertical and horizontal black lines.

6.3 Comparison of unstimulated and alloreactive populations

6.3.1 Number of sequenced unique clones

Although little is known about the size of the alloreactive repertoire, it has been shown to be highly polyclonal and is dominated by high abundance clones [29, 54, 27, 39]. From the

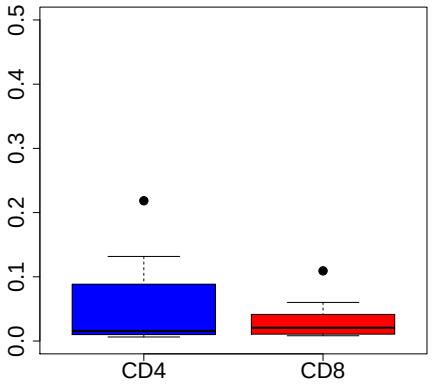

Figure 6.3: The fraction of all stimulated clones removed by 2x criteria for $CD4^+$ and $CD8^+$ subsets.

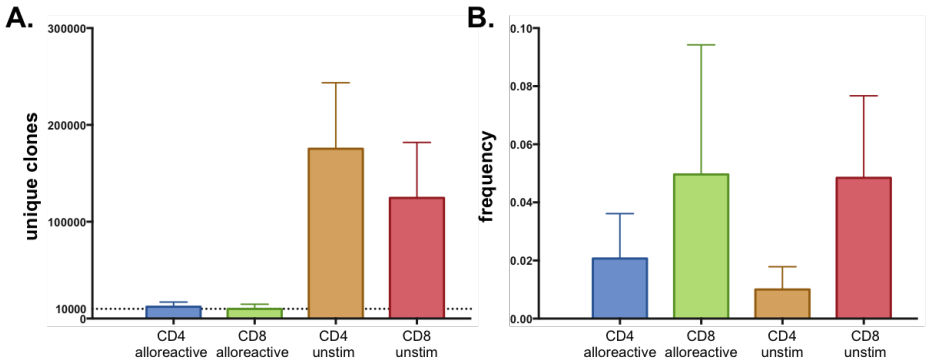

Figure 6.4: A. Total number of clones collected from both alloreactive and unstimulated populations and B. Maximum clonal frequency observed in each population.

roughly 2e5 templates and 1200ng of DNA that were sequenced for each sample, an order of magnitude fewer clones were identified in stimulated populations. Total clone numbers of stimulated clones were in a similar range among all sequenced samples, suggesting similar levels of expansion across individuals. The maximum clonal frequency of alloreactive samples was also several times greater than those of the healthy controls Figure 6.4. Both of these results verify that the alloresponse is very robust.

6.3.2 Analysis of CDR3 length and VJ usage

As described in the introduction, the alloresponse depends on the foreign MHC as much as the bound peptide. The MHC region is recognized by the CDR1 and CDR2 sequences on the V gene (See Subsection 2.1.2). This could serve to restrict the usage of VJ pairs among alloreactive cells. Similarly, the TCR interaction with foreign MHC may impose limitations on the CDR3-peptide bond by affecting flexibility, hydrophobicity, or overall charge, which could manifest as a change in the CDR3 length distribution. While in mice no differences in CDR3 length were identified [65], no such analysis has been performed for the human alloresponse.

VJ usage frequencies and amino acid CDR3 length frequencies were compared between unstimulated CDR3 clones and the alloreactive set. The CDR3 lengths were approximately normally distributed, with a maximum at 15 amino acids, and a simple Mann-Whitney test showed no significant difference between the two populations for either $CD4^+$ or $CD8^+$ subsets (Figure 6.5A,B). The VJ usage was quantified by Jensen Shannon divergence, which produced low values for all pairs of samples, indicating very similar frequency distributions between unstimulated and alloreactive populations (Figure 6.5C). This was further visualized using Circos plots (Figure 6.5D). The finding that no significant differences exist between alloreactive and unstimulated T cells in a repertoire, either in CDR3 length or VJ usage, suggests that the alloreactive repertoire consists of typical T cell clones, and is not biased towards specific clonal subsets.

6.4 Quantifying the diversity of the alloresponse

Diversity of the allorepertoire was quantified using clonality and power law slope. Analysis of clonality between unstimulated and alloreactive populations (Figure 6.6A–C) showed that while CD8 was more clonal among the unstimulated repertoires, this difference is significantly less pronounced in the allorepertoire, directly quantifying the strong clonal expansions suggested by the numbers in Figure 6.4. The increase in clonality in $CD4^+$ was much larger than

Table 6.1: Sequenced allorepertoire data

CD4$^+$

	Sample	Templates	Reads	Clones	Max Frequency	Entropy	Clonality	Simpson index	R20	Power law slope
UNSTIMULATED	HC10	159558	1978371	128729	6.6e-4	17	0.016	1.70e-5	0.061	-3.02
	HC18_2013	167929	1840148	256486	0.016	16	0.031	3.9e-4	0.044	-3.10
	HC18_2014	170095	1911625	135679	0.012	17	0.024	2.1e-4	0.055	-3.26
	HC19_2013	156782	1792963	101900	0.01	16	0.047	3.0e-4	0.022	-2.85
	HC19_2014	173109	2048183	235343	6.10e-3	16	0.034	1.1e-4	0.032	-2.82
	HC27	155801	3347128	118885	0.024	16	0.055	1.2e-3	0.029	-3.15
	HC42	186452	6067138	153049	0.011	17	0.015	1.6e-4	0.1	-3.26
	HC46	313980	5385131	272763	0.00061	18	9.7e-3	6.80e-6	0.079	-3.33
ALLOREACTIVE (2x cutoff)	HC10v12	58378	1503929	8031	0.004	12	0.085	0.00052	0.02	-1.82
	HC10v19	100743	4123428	17839	0.053	12	0.12	3.4e-3	4.7e-3	-2.12
	HC18v10_2013	86689	1708345	10459	0.024	11	0.17	2.6e-3	2.5e-3	-1.86
	HC18v10_2014	77775	1723476	10627	0.012	11	0.15	1.2e-3	6.4e-3	-1.72
	HC19v10_2013	95997	1704180	14231	0.024	11	0.18	3.0e-3	1.8e-4	-2.00
	HC19v10_2014	98901	1866837	10000	0.028	11	0.19	3.6e-3	2e-3	-1.79
	HC27v19	38377	2311113	4202	0.032	10	0.13	2.8e-3	9.5e-3	-1.66
	HC27v22	76200	3545830	7676	0.01	11	0.11	8.6e-4	0.012	-1.67
	HC42v39	55825	1219490	15770	0.0039	13	0.062	2.7e-4	0.1	-2.28
	HC42v44	91920	4081911	19627	0.0045	13	0.081	3.2e-4	0.013	-2.08
	HC46v39	98626	2952273	15989	0.032	12	0.14	2.1e-3	3.3e-3	-2.10

CD8$^+$

	Sample	Templates	Reads	Clones	Max Frequency	Entropy	Clonality	Simpson index	R20	Power law slope
UNSTIMULATED	HC10	145317	2190139	103790	6.8e-3	16	0.062	3.2e-4	7.7e-3	-2.60
	HC18_2013	186863	2031525	223393	0.084	14	0.16	0.01	1.3e-4	-2.52
	HC18_2014	146545	1884748	96190	0.051	15	0.11	4.8e-3	1.1e-3	-2.57
	HC19_2013	159353	1997443	67061	0.063	13	0.22	7.7e-3	1.2e-4	-2.26
	HC19_2014	164461	1837111	186503	0.033	14	0.13	2.4e-3	5.0e-4	-2.47
	HC27	131201	3849423	80015	0.082	14	0.17	8.9e-3	1.2e-4	-2.66
	HC42	217945	3974379	157235	0.051	15	0.12	5.1e-3	2.1e-4	-2.92
	HC46	121123	3310782	82758	0.017	15	0.091	1.1e-3	2.1e-3	-2.49
ALLOREACTIVE (2x cutoff)	HC10v12	62742	2342752	6053	0.11	11	0.12	1.1e-3	0.013	-1.54
	HC10v19	107715	3560271	9804	0.018	11	0.15	1.9e-3	5e-3	-1.58
	HC18v10_2013	1808921	13487	0.15	0.016	11	0.2	0.027	2.2e-3	-1.95
	HC18v10_2014	1873991	13514	0.088	0.012	12	0.15	8.5e-3	5.6e-3	-1.86
	HC19v10_2013	1841385	6717	0.1	0.01	10	0.2	0.014	2.1e-3	-1.53
	HC19v10_2014	2156621	5864	0.046	6.10e-3	10	0.2	5.6e-3	2.9e-3	-1.51
	HC27v19	79335	2726653	5479	0.047	11	0.15	3.2e-3	0.011	-1.38
	HC27v22	77236	2742252	6124	8.7e-4	11	0.12	9.1e-4	0.016	-1.41
	HC42v39	74952	3376373	15240	0.029	13	0.084	1.2e-3	0.017	-1.97
	HC42v44	117969	1621686	19613	.023	13	0.097	9.5e-4	0.013	-2.04
	HC46v39	96988	3741423	8564	0.02	11	0.17	1.8e-3	5e-3	-1.74

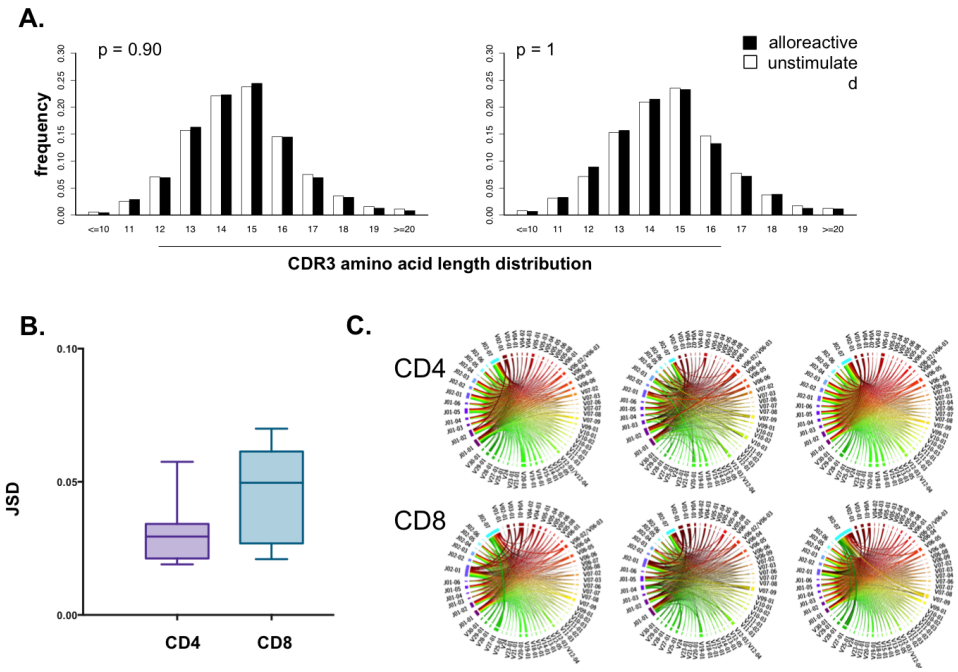

Figure 6.5: A. Comparison of CDR3 length frequencies in alloreactive and unstimulated populations for $CD4^+$. (left) $CD8^+$ (right). Jensen Shannon divergence for VJ usage between unstimulated and alloreactive samples. C. Circos plots of VJ usage. Ribbons are drawn between each V-gene on the right side (shades of red, yellow, and green) and J-gene on the left side (shades of blue and purple). The thickness of the ribborn is proportional to the usage frequency of a given combination.

that in $CD8^+$, but $CD8^+$ nonetheless contained slightly higher frequency top clones in the alloreactive population (Figure 6.6D). However, the most distinguishing feature among the samples was the power law slope, calculated as described in Section 3.2. The slopes showed clearly that $CD8^+$ subsets are less diverse than $CD4^+$ subsets, and alloreactive populations are less diverse than unstimulated populations. Additionally, the average difference in slope between unstimulated populations was greater than between alloreactive populations, again suggesting that $CD4^+$ subsets expand most strongly during the alloresponse and is similar to $CD8^+$ in overall diversity (Figure 6.6E). This result provides evidence for the robustness of the alloresponse at a greater resolution than previous studies, and highlights the benefits of using the slope of the T cell receptor repertoire to look at diversity.

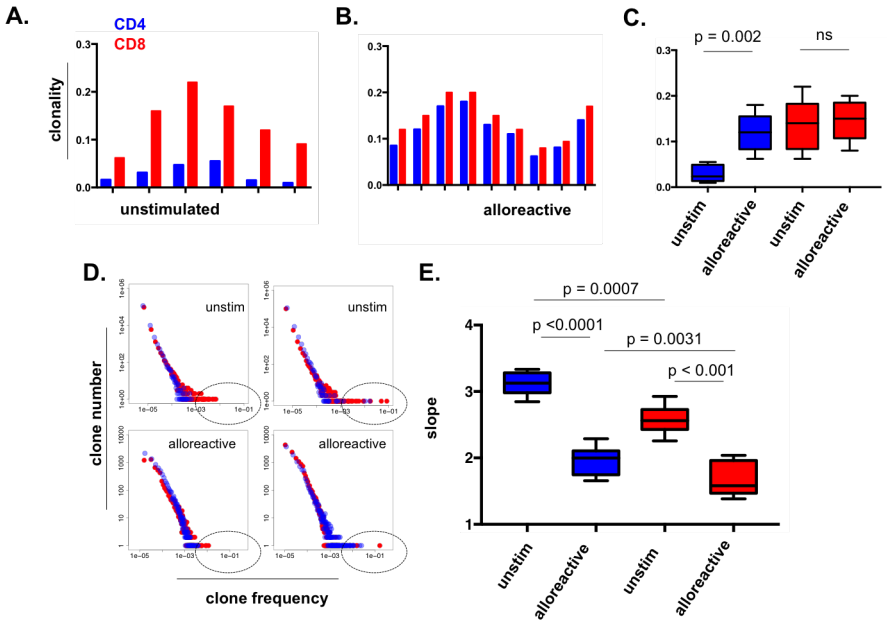

Figure 6.6: Diversity of $CD4^+$ and $CD8^+$ T cell subsets was quantified using clonality for both unstimulated (A) and alloreactive (B) populations. Alloreactive and unstimulated populations are compared directly in (C). Abundance plots of unstimulated and alloreactive repertoires are plotted in (D) with expanded clones indicated by the dotted ellipses. Boxplot of powerlaw slopes is shown in (E). All $CD4^+$ clones are in blue, and $CD8^+$ in red.

6.5 Allospecificity of the alloreactive repertoire and the role of HLA

It is not known whether a repertoire responding to two different stimulators will activate the same or different T cell clones. For three of the samples (HC10, HC27 and HC42) the alloresponse was measured for two different stimulators. Frequencies of the top 100 alloreactive clones were plotted and for all three cases very few clones were found to overlap, suggesting that alloreactivity is not limited to a specific set of clones (Figure 6.7A). Furthermore the Jensen Shannon divergence was calculated for the nucleotide sequences of the two stimulated populations, and compared across the three samples which varied in their degree of HLA mismatch (Figure 6.7B). HC27 was completely mismatched for both HLA Class I and HLA Class II, while HC10 had 2/6 Class I match and HC42 had 2/6 Class I match and

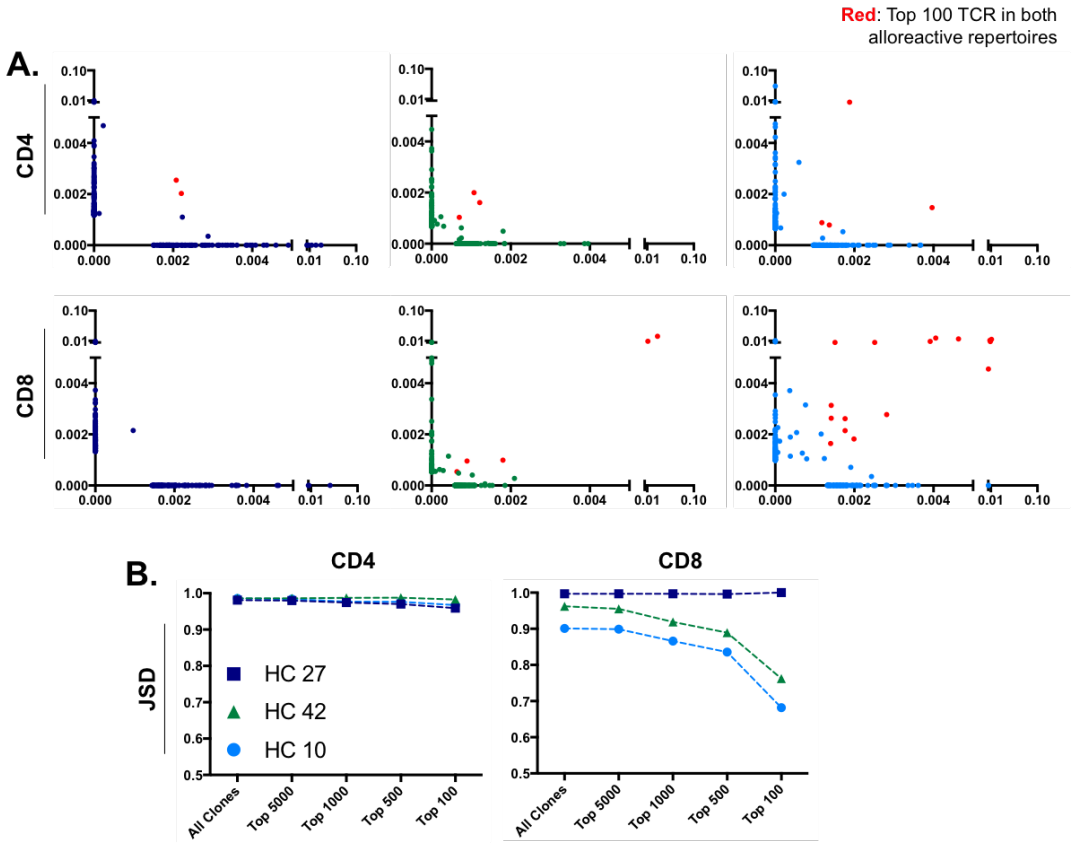

Figure 6.7: A. Overlap of alloreactive clones from two stimulators of the same healthy control population, for three samples (HC10,HC27,HC42). B. Jensen Shannon divergence of nucleotide sequences between stimulated populations for top clones as indicated.

1/6 Class II match. The amount of mismatch directly corresponded to the Jensen Shannon divergence, particularly among top clones, with greater mismatch corresponding to higher values of divergence. This is direct evidence showing that the more highly HLA matched the two samples are the greater the overlap among reactive clones.

To further validate that clonal sharing among stimulated populations was dependent on HLA mismatch, four additional sequenced datasets were analyzed from combined kidney and bone marrow transplants (CKBMTs) collected in [54]. Prior to this analysis reads were converted to templates as discussed in Section 3.4. All four unstimulated repertoires were stimulated by a matched transplant donor. A second sample from the same repertoire was stimulated in a heavily mismatched MLR experiment. The reactive clones TCR population

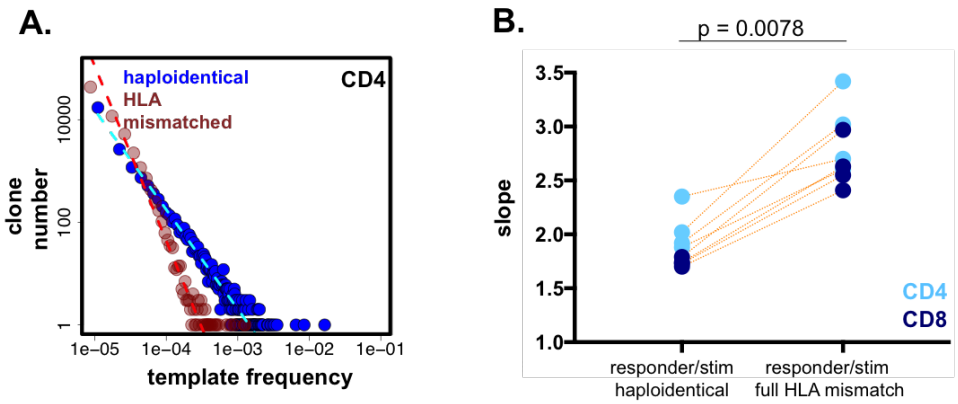

Figure 6.8: A. Representative abundance plot showing the steeper and therefore more diverse power law distribution from a response to an HLA mismatched repertoire compared to a shallower power law for a haploidentical transplant. B. Difference in repertoire slope for all CKBMT samples. Dashed lines connect stimulator pairs for the same subject. P-value obtained by Wilcoxon test.

after transplantation were compared to reactive clones of the mismatched population. In all cases the mismatched sample had greater alloreactivity as indicated by the lower slope of the abundance distribution (Figure 6.8A,B) in both $CD4^+$ and $CD8^+$ subsets.

6.6 Frequency of the alloreactive repertoire

Identifying the fraction of cells within an unstimulated population has so far not been addressed by TCR repertoire sequencing, and was attempted for the first time in this study. Where clones from the alloreactive population were also sampled in the unstimulated population the frequency was used as observed, providing an upper bound for those clones. The majority of alloreactive clones are rare and were therefore unobserved as shown in Figure 6.9A. Using the semi-parametric method first outlined in Chapter 3, the frequency of these clones was estimated and used to compute the total fraction of the original collected sample that was alloreactive Figure 6.9B. The overall frequency of alloreactive cells fluctuated between 1-7% for $CD4^+$ T cells and 0.5-4% for $CD8^+$ T cells. This result seems to indicate that previous estimates of many as 10% of the repertoire being alloreactive are overestimates. Additionally it indicates that the $CD4^+$ is more alloreactive than $CD8^+$,

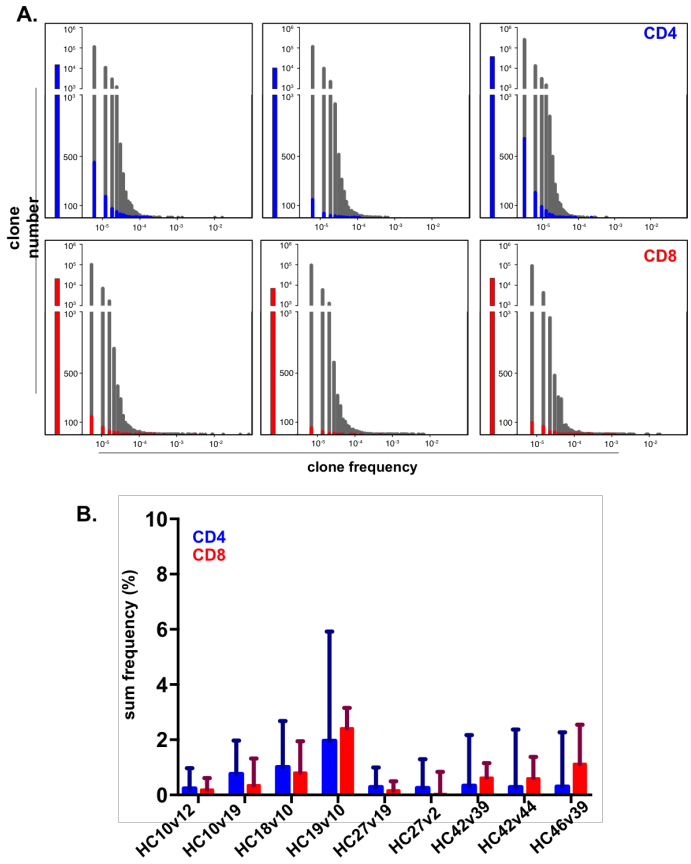

Figure 6.9: A. Histograms of clonal abundance. Instimulated populations are colored in gray. The subset of clones found in both alloreactive and unstimulated populations are indicated in blue for $CD4^+$ and red for $CD8^+$. Barplots on the left indicate number of alloreactive clones not found in the unstimulated population. B. Sum frequency estimate combining observed frequencies shown by bar plots, and additional unseen frequency presented by error bars.

consistent with the greater change in observed $CD4^+$ diversity.

6.7 Conclusions

This chapter presented a novel study to comprehensively investigate the alloreactive response through analysis of nucleotide sequence, VJ cassette usage and CDR3 length. Measures derived from information theory including clonality and mutual information were used to investigate changes in repertoire diversity and constraints on repertoire use among alloreactive clones. The role of slope calculations in comparing TCR subsets was demonstrated. The importance of HLA matching was also explored, finding that stimulators which were

more closely matched stimulated similar groups of T cell clones. Finally, an estimate of the fraction of expanded TCRs in the alloresponse from high throughput sequencing studies was obtained. These results offer early insights into the nature of the alloresponse and computational methods for studying repertoires of large activated T cell populations.

Chapter 7

Conclusions and future work

Quantitative approaches for analyzing diversity and divergence of T cell repertoire have been explored and applied to a number of applications, including the maintenance of naïve repertoire in healthy lymphoid tissues (Chapter 4), the distribution of TCRs in gliomas (Chapter 5), and the strength of the alloresponse (Chapter 6). A semiparametric method was introduced to quantify the fraction of expanded clones from sampled populations (Chapter 3), and was used to measure the strength of the polyclonal alloresponse. This thesis shows that a great deal can be learned from comprehensive sequencing of the TCR repertoire that cannot be obtained through biochemical approaches. The insights gained from these analyses and the tools presented have significant implications for development of immunotherapeutics.

There are many further directions for improving our understanding of the T cell immune response. All of these studies can benefit from continued collection of samples and replicates to better quantify the variability among the human population. Additionally, both the naïve T cell and glioma work would benefit from longitudinal studies thereby minimizing noise from individual differences in repertoire. Finally, most studies have only looked at the β-chain of the repertoire; however, analysis of α-chain may reveal further associations of specific T cells as markers of immune response. Much like single nucleotide polymorphisms (SNPs) can be used to identify physical traits and make health predictions, in the future

it may be possible to use TCR repertoire to develop methods for diagnosis and treatment tailored to individual patients.

On the functional side, TCR data can be combined with RNA sequencing and single cell sequencing experiments to better understand T cell effector pathways and differential changes in activity during tissue maintenance or antigen induced activation. On the structural side, careful analysis of TCR-protein interactions can lead to development of T cells specially designed to respond to cancer neoantigens or to predict alloreactivity among T cell sequences. Additional interesting questions to consider are whether TCR repertoire can be used to infer MHC genotype, potential for autoimmunity and allergic response, drug responses, and virtually any other response that depends on the immune system.

Finally, a great deal of focus has recently been placed on the human microbiome and its interactions with the immune system. The microbiome is also extremely diverse and variable across tissue sites. Studies of diversity and divergence can be used to look at the effect of the microbiome on the immune system and on T cell responses.

The study of the immune responses through the use of sequencing is still in its early stages and TCR studies will continue to improve and impact our understanding of human evolution and diversity, as well as our understanding of the immune response in various diseases. This knowledge in turn will impact our ability to treat diseases on an individual basis.

Bibliography

[1] G. Ada, "The enunciation and impact of macfarlane burnet's clonal selection theory of acquired immunity", *Immunology and Cell Biology Journal*, vol. 86, pp. 116–118, 2008. DOI: 10.1038/sj.icb.7100156.

[2] J. P. Allison, B. W. McIntyre, and D. Bloch, "Pillars article: Tumor-specific antigen of murine t-lymphoma defined with monoclonal antibody", *The Journal of Immunology*, vol. 174, pp. 1144–1151, 1982.

[3] P. T. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky, "A direct estimate of the human $\alpha\beta$ t cell receptor diversity", *Science*, vol. 286, pp. 958–961, 1999. DOI: 10.1126/science.286.5441.958.

[4] A. Ashworth and L. J. Christopher, "Parp inhibitors: Synthetic lethality in the clinic", *Nature*, vol. 355, pp. 1152–1168, 2017. DOI: 10.1126/science.aam7344.

[5] R. Atkins and W. Ford, "Early cellular events in a systemic graft-vs.-host reaction. i. the migration of responding and nonresponding donor lymphocytes", *The Journal of Experimental Medicine*, vol. 141, pp. 664–680, 1975. DOI: 10.1084/jem.141.3.664.

[6] M. Attaf, E. Huseby, and A. K. Sewell, "$\alpha\beta$ t cell receptors as predictors of health and disease", *Cellular & Molecular Immunology*, vol. 12, pp. 391–399, 2015. DOI: 10.1038/cmi.2014.134.

[7] S. Becattini, D. Latorre, F. Mele, M. Foglierini, C. De Gregorio, and A. B. Cassotta, "Functional heterogeneity of human memory cd4+ t cell clones primed by pathogens or vaccines", *Science*, vol. 347, pp. 400–406, 2015. DOI: 10.1126/science.1260668.

[8] S. Bhaduri-McIntosh, M. J. Rotenberg, B. Gardner, M. Robert, and G. Miller, "Repertoire and frequency of immune cells reactive to epstein-barr virus–derived autologous lymphoblastoid cell lines", *Blood*, vol. 111, pp. 1334–1343, 2008. DOI: 10.1182/blood-2007-07-101907.

[9] C. J. Blohmke, D. O'Connor, and A. J. Pollard, "The use of systems biology and immunological big data to guide vaccine development", *Genome Biology*, vol. 7, no. 114, 2015. DOI: 10.1186/s13073-015-0236-1.

[10] O. V. Bolkhovskaya, D. Y. Zorin, and M. V. Invanchenko, "Assessing t cell clonal size distribution: A non-parametric approach", *Plos One*, vol. 9, e108658, 2014. DOI: 10.1371/journal.pone.0108658.

[11] I. den Braber, T. Mugwagwa, N. Vrisekoop, *et al.*, "Maintenance of peripheral naive t cells is sustained by thymus output in mice but not humans.", *Immunity*, vol. 36, pp. 288–297, 2012. DOI: 10.1016/j.immuni.2012.02.006.

[12] O. V. Britanova, E. V. Putintseva, M. Shugay, *et al.*, "Age-related decrease in tcr repertoire diversity measured with deep and normalized sequence profiling", *The Journal of Immunology*, vol. 192, pp. 2689–2698, 2014. DOI: 10.4049/jimmunol.1302064.

[13] J. Bruce, B. Kennedy, R. Shephard, F. Talavera, R. McKenna, and J. Harris, "Glioblastoma multiforme [monograph on the internet]". (visited on 2017).

[14] L. Buonaguro and B. Pulendran, "Immunogenomics and systems biology of vaccines", *Immunological Reviews*, vol. 239, pp. 197–208, 2011. DOI: doi:10.1111/j.1600-065X.2010.00971.x..

[15] M. Burnet, *The Generation of Diversity*. The MacMillan Company, 1941.

[16] M. F. Burnet, ""self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity", *Nature*, vol. 232, pp. 230–235, 1971. DOI: 10.1038/232230a0.

[17] C. Chang, Q. J., D. O'Sullivan, *et al.*, "Metabolic competition in the tumor microenvironment is a driver of cancer progression", *Cell*, vol. 162, pp. 1229–1241, 2015. DOI: 10.1016/j.cell.2015.08.016.

[18] Z. A. Cooper, D. T. Frederick, V. R. Juneja, R. J. Sullivan, D. P. Lawrence, A. Piris, A. H. Sharpe, D. E. Fisher, K. T. Flaherty, and J. A. Wargo, "Braf inhibition is associated with increased clonality in tumor-infiltrating lymphocytes", *Oncoimmunology*, vol. 2, e26615, 2013. DOI: 10.4161/onci.26615.

[19] J. Couzin-Frankel, "Cancer immunotherapy", *Science*, vol. 342, pp. 1432–1433, 2013. DOI: 10.1126/science.342.6165.1432.

[20] J. Dean, R. O. Emerson, M. Vignali, A. M. Sherwood, M. J. Rieder, C. S. Carlson, and R. S. Harlan, "Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci", *Genome Medicine*, vol. 7, no. 123, 2015. DOI: 10.1186/s13073-015-0238-z.

[21] C. J. Der and B. Papke, "Drugging ras: Know the enemy", *Science*, vol. 355, pp. 1158–1163, 2017. DOI: 10.1126/science.aam7622.

[22] J. Desponds, T. Mora, and A. M. Walczak, "Fluctuating fitness shapes the clone-size distribution of immune repertoires", *Proceedings of the National Academy of Sciences*, vol. 113, pp. 274–279, 2015. DOI: 10.1073/pnas.1512977112.

[23] W. DeWitt, P. Lindau, T. Snyder, M. Vignali, R. Emerson, and H. Robins, "Replicate immunosequencing as a robust probe of b cell repertoire diversity", 2014. arXiv: 1410.0350 [q-bio.QM].

[24] S. DeWolf, Y. Shen, and M. Sykes, "A new window into the human alloresponse", *Transplantation*, vol. 100, pp. 1639–1649, 2016. DOI: 10.1097/TP.0000000000001064.

[25] C. for Disease Control and Prevention, "2014-2016 ebola outbreak in west africa", 2016. (visited on 03/27/2017).

[26] ——, "Vaccine effectiveness - how well does the flu vaccine work?", 2017. (visited on 03/28/2017).

[27] G. A. Dos Reis and E. M. Shevach, "The syngeneic mixed leukocyte reaction represents polyclonal activation of antigen-specific t lymphocytes with receptors for self-ia antigens", *The Journal of Immmunology*, vol. 127, pp. 2456–2460, 1981.

[28] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did shakespeare know?", *Biometrika*, vol. 63, pp. 435–447, 1976. DOI: 10.1093/biomet/63.3.435.

[29] R. O. Emerson, J. M. Matthew, I. M. Konieczna, H. S. Robins, and J. R. Leventhal, "Defining the alloreactive t cell repertoire using high-throughput sequencing of mixed lymphocyte reaction culture", *Plos One*, vol. 9, e111943, 2014. DOI: 10.1371/journal.pone.0111943.

[30] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions", *IEEE Transactions on Information Theory*, vol. 49, pp. 1858–1860, 2003. DOI: 10.1109/TIT.2003.813506.

[31] N. I. of Environmental and H. Sciences, "Autoimmune diseases", 2012. (visited on 03/28/2017).

[32] D. L. Farber, N. A. Yudanin, and N. P. Restifo, "Human memory t cells: Generation, compartmentalization and homeostasis", *Nature Reviews Immunology*, vol. 14, pp. 24–35, 2014. DOI: 10.1038/nri3567.

[33] P. E. Fecci, D. A. Mitchell, J. F. Whitesides, W. Xie, A. H. Friedman, G. E. Arher, J. E. I. Herndon, D. D. Bigner, G. Dranoff, and J. H. Sampson, "Increased regulatory t-cell fraction amidst a diminished cd4 compartment explains cellular immune defects

in patients with malignant glioma", *Cancer Cell*, vol. 66, no. 6, 2006. DOI: 10.1158/0008-5472.CAN-05-3773.

[34] D. J. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt, "Profiling the t-cell receptor beta-chain repertoire by massively parallel sequencing", *Genome Research*, vol. 19, pp. 1817–1824, 2009. DOI: 10.1101/gr.092924.109.

[35] V. V. Ganusov and R. J. De Boer, "Do most lymphocytes in humans really reside in the gut?", *Trends in Immunology*, vol. 28, pp. 514–518, 2007. DOI: 10.1016/j.it.2007.08.009.

[36] R. N. Germain, M. Meier-Schellersheim, A. Nita-Lazar, and I. D. Fraser, "Systems biology in immunology: A computational modeling perspective", *The Annual Review of Immunology*, vol. 29, pp. 527–585, 2011. DOI: 10.1146/annurev-immunol-030409-101317.

[37] M. Giordano, C. Henin, J. Maurizio, *et al.*, "Molecular profiling of cd8 t cells in autochthonous melanoma identifies maf as driver of exhaustion", *The EMBO Journal*, vol. 34, pp. 2042–2058, 2015. DOI: 10.15252/embj.201490786.

[38] A. D. Griesemer, E. C. Sorenson, and M. A. Hardy, "The role of the thymus in tolerance", *Transplantation*, vol. 90, pp. 465–474, 2010. DOI: 10.1097/TP.0b013e3181e7e54f.

[39] M. Guillet, F. Sebille, and J.-P. Soulillou, "Tcr usage in naive and committed alloreactive cells: Implications for the understanding of tcr biases in transplantation", *Current Opinion in Immunology*, vol. 13, pp. 566–571, 2001. DOI: 10.1016/S0952-7915(00)00260-0.

[40] M. Hale, T. Mesojednik, G. Rommano Ibarra, J. Sahni, B. A., K. Sommer, A. Scharenberg, D. Rawlings, and T. Wagner, "Engineering hiv-resistant, anti-hiv chimeric antigen receptor t cells", *Molecular Therapy*, vol. 25, pp. 570–579, 2017. DOI: 10.1016/j.ymthe.2016.12.023.

[41] B. Hall, D. S., and B. Roser, "The cellular basis of allograft rejection in vivo. i. the cellular requirements for first-set rejection of heart grafts", *The Journal of Experimental Medicine*, vol. 148, pp. 878–889, 1978. DOI: 10.1084/jem.148.4.878.

[42] E. Hammarlund, M. W. Lewis, S. G. Hansen, L. I. Strelow, J. A. Nelson, G. J. Sexton, J. M. Hanifin, and M. K. Slifka, "Duration of antiviral immunity after smallpox vaccination", *Nature Medicine*, vol. 9, pp. 1131–1137, 2003. DOI: 10.1038/nm917.

[43] M. Hidalgo, F. Amant, A. V. Biankin, *et al.*, "Patient-derived xenograft models: An emerging platform for translational cancer research", *Cancer Discovery*, vol. 4, pp. 998–1013, 2014. DOI: 10.1158/2159-8290.CD-14-0001.

[44] B. Huang, H. Zhang, L. Gu, B. Ye, Z. Jian, C. Stary, and X. Xiong, "Advances in immunotherapy for glioblastoma multiforme", *Journal of Immunology Research*, 3597613, 2016. DOI: 10.1155/2017/3597613.

[45] A. E. Ivliev, P. A. 't Hoen, and M. G. Sergeeva, "Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and sprouty signaling in glioma", *Cancer Research*, vol. 70, no. 24, 2010. DOI: 10.1158/0008-5472.CAN-10-2465.

[46] L. Jost, "Entropy and diversity", *OIKOS*, vol. 113, pp. 363–375, 2006. DOI: 10.1111/j.2006.0030-1299.14714.x.

[47] T. H. Keitt and E. H. Stanley, "Dynamics of north american breeding bird populations", *Nature*, vol. 393, pp. 257–260, 1998. DOI: 10.1038/30478.

[48] A. Klaus, S. Yu, and D. Plenz, "Statistical analyses support power law distributionsfound in neuronal avalanches", *Plos One*, vol. 6, e19779, 2011. DOI: 10.1371/journal.pone.0019779.

[49] P. de Kruif, *Microbe Hunters*. Harcourt Brace Jovanovich, 1926, ISBN: 9780156027779.

[50] M.-P. Lefrank and G. Lefranc, *The T cell receptor Factbook*. Academic Press, London UK, 2001, ISBN: 9780124413528.

[51] B. Li, T. Li, J.-C. Pignon, *et al.*, "Landscape of tumor-infiltrating t cell repertoire of human cancers", *Nature Genetics*, vol. 48, pp. 725–732, 2016. DOI: doi:10.1038/ng.3581.

[52] A. Lim, L. Trautmann, M.-A. Peyrat, C. Couedel, F. Davodeau, F. Romagné, P. Kourilsky, and M. Bonneville, "Frequent contribution of t cell clonotypes with public tcr features to the chronic response against a dominant ebv-derived epitope: Application to direct detection of their molecular imprint on the human peripheral t cell repertoire", *The Journal of Immunology*, vol. 165, pp. 2001–2011, 2000. DOI: 10.4049/jimmunol.165.4.2001.

[53] K. F. Lindahl and D. B. Wilson, "Histocompatibility antigen-activated cytotoxic t lymphocytes. i. estimates of the absolute frequency of killer cells generated in vitro", *The Journal of Experimental Medicine*, vol. 145, pp. 500–507, 1977. DOI: 10.1084/jem.145.3.500.

[54] ——, "Histocompatibility antigen-activated cytotoxic t lymphocytes. i. estimates of the absolute frequency of killer cells generated in vitro", *The Journal of Experimental Medicine*, vol. 145, pp. 500–507, 2015. DOI: 10.1084/jem.145.3.500.

[55] A. Lossius, J. N. Johansen, F. Vartdal, H. Robins, B. Jurate Saltyte, T. Holmoy, and J. Olweus, "High-throughput sequencing of tcr repertoires in multiple sclerosis reveals intrathecal enrichment of ebv-reactive cd8$^+$ t cells", *European Journal of Immunology*, vol. 44, pp. 3439–3452, 2014. DOI: 10.1002/eji.201444662.

[56] M. Love, S. Anders, and W. Huber, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2", *Genome Biology*, vol. 15, p. 550, 2014. DOI: 10.1186/s13059-014-0550-8.

[57] P. Marrack and J. Kappler, "T cells can distinguish between allogeneic major histocompatibility complex products on different cell types", *Nature*, vol. 332, pp. 840–843, 1988. DOI: 10.1038/332840a0.

[58] MayoClinic, 2017.

[59] J. McCluskey, C. Kanaan, and M. Diviney, "Nomenclature and serology of hla class i and class ii alleles", *Current Protocols in Immunology*, vol. 52, A.1S.1–A.1S.8. 2003. DOI: 10.1002/0471142735.ima01s52.

[60] P. B. Medawar, "The behaviour and fate of skin autografts and skin homografts in rabbits", *Journal of Anatomy*, vol. 78, pp. 176–199, 1944.

[61] S. C. Meuer, K. A. Fitzgerald, R. E. Hussey, J. C. Hodgdon, S. F. Schlossman, and E. L. Reinherz, "Clonotypic structures involved in antigen-specific human t cell function. relationship to the t3 molecular complex", *The Journal of Experimental Medicine*, vol. 157, pp. 705–719, 1983. DOI: 10.1084/jem.157.2.705.

[62] J. J. Miles, D. C. Douek, and D. A. Price, "Bias in the $\alpha\beta$ t-cell repertoire: Implications for disease pathogenesis and vaccination", *Immunology and Cell Biology*, vol. 89, pp. 375–387, 2011. DOI: 10.1038/icb.2010.139.

[63] J. Miller, "Effect of neonatal thymectomy on the immunological responsiveness of the mouse", *Proceedings of the Royal Society B*, vol. 156, pp. 415–248, 1962. DOI: 10.1098/rspb.1962.0048.

[64] C. Molina-París and G. Lythe, *Mathematical Models and Immune Cell Biology*. Springer, 2011, ISBN: 9781441977243.

[65] G. P. Morris, P. P. Ni, and P. M. Allen, "Alloreactivity is limited by the endogenous peptide repertoire", *Proceedings of the National Academy of Sciences*, vol. 108, pp. 3695–3700, 2011. DOI: 10.1073/pnas.1017015108.

[66] K. P. Murphy, *Janeway's Immunobiology*. Garland Science, Taylor & Francis Group, LLC, 2012, ISBN: 9780815342434.

[67] A. Murugan, T. Mora, A. M. Walczak, and C. J. J. Callan, "Statistical inference of the generation probability of t-cell receptors from sequence repertoires", *Proceedings of the National Academy of Sciences*, vol. 109, pp. 16 161–16 166, 2012. DOI: 10.1073/pnas.1212755109.

[68] W. H. Organization, "2014–2016 ebola outbreak in west africa", 2017. (visited on 03/27/2017).

[69] F. Pagès, A. Kirilovsky, B. Mlecnik, *et al.*, "In situ cytotoxic and memory t cells predict outcome in patients with early-stage colorectal cancer", *Journal of Clinical Oncology*, vol. 27, no. 35, 2009. DOI: 10.1200/JCO.2008.19.6147.

[70] C. Pannetier, M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky, "The sizes of the cdr3 hypervariable regions of the murine t-cell receptor β chains vary as a function of the recombined germ-line segments", *Proceedings of the National Academy of Sciences*, vol. 90, pp. 4319–4323, 1993.

[71] N. Pardi, A. Secreto, X. Shan, *et al.*, "Administration of nucleoside-modified mrna encoding broadly neutralizing antibody protects humanized mice from hiv-1 challenge", *Nature Communications*, vol. 8, no. 14630, 2017. DOI: 10.1038/ncomms14630.

[72] D. Peng, I. Kryczek, N. Nagarsheth, *et al.*, "Epigenetic silencing of th1-type chemokines shapes tumour immunity and immunotherapy", *Nature*, vol. 527, pp. 249–253, 2015. DOI: 10.1038/nature15520.

[73] N. D. Pennock, J. T. White, E. W. Cross, E. E. Cheney, B. A. Tamburini, and R. M. Kedl, "T cell responses: Naïve to memory and everything in between", *Advances in Physiology Education*, vol. 37, pp. 273–283, 2013. DOI: 10.1152/advan.00066.2013.

[74] H. S. Phillips, S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, and L. Soroceanu, "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis", *Cancer Cell*, vol. 9, pp. 157–173, 2006. DOI: 10.1016/j.ccr.2006.02.019.

[75] V. Pisarenko and D. Sornette, "Characterization of the frequency of extreme earthquake events by the generalized pareto distribution", *Pure and applied geophysics*, vol. 160, pp. 2343–2364, 2003. DOI: 10.1007/s00024-003-2397-x.

[76] S. H. Podolsky and A. I. Tauber, *The Generation of Diversity*. Harvard University Press, 1997, ISBN: 9780674001824.

[77] M. Preusser, S. de Ribaupierre, and A. Wöhrer, "Current concepts and management of glioblastoma", *Anals of Neurology*, vol. 70, pp. 9–21, 2011. DOI: 10.1002/ana.22425.

[78] S.-M. Razavi, K. E. Lee, B. E. Jin, P. S. Aujla, S. Gholamin, and G. Li, "Immune evasion strategies of glioblastoma", *Frontiers in Surgery*, vol. 3, no. 11, 2016. DOI: 10.3389/fsurg.2016.00011.

[79] S. Reardon, "New life for pig-to-human transplants", *Nature*, vol. 527, pp. 152–154, 2015. DOI: doi:10.1038/527152a.

[80] E. L. Reinherz, S. C. Meuer, K. A. Fitzgerald, R. E. Hussey, J. C. Hogdon, O. Acuto, and S. F. Schlossman, "Comparison of t3-associated 49- and 43-kilodalton cell surface molecules on individual human t-cell clones: Evidence for peptide variability in t-cell receptor structures", *Proceedings of the National Academy of Sciences*, vol. 80, pp. 4104–4108, 1983.

[81] H. Robins, "Immunosequencing: Applications of immune repertoire deep sequencing", *Current Opinion Immunology*, vol. 5, pp. 646–652, 2013. DOI: 10.1016/j.coi.2013.09.017.

[82] H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, "Comprehensive assessment of t-cell receptor β-chain diversity in $\alpha\beta$ t cells", *Blood*, vol. 19, pp. 4099–4107, 2009. DOI: 10.1182/blood-2009-04-217604.

[83] H. S. Robins, N. G. Ericson, J. Guenthoer, K. C. O'Briant, M. Tewari, C. W. Drescher, and J. H. Bielas, "Digital quantification of tumor infiltrating lymphocytes", *Science Translational Medicine*, vol. 5, no. 214, 2013. DOI: 10.1126/scitranslmed.3007247.

[84] F. Sallusto, J. Geginat, and A. Lanzavecchia, "Central memory and effector memory t cell subsets: Function, generation, and maintenance", *Annual Review of Immunology*, vol. 2, pp. 745–763, 2004. DOI: 10.1146/annurev.immunol.22.012703.104702.

[85] H. Schild, O. Rotzschke, and H. Rammensee, "Limit of t cell tolerance to self proteins by peptide presentation", *Science*, vol. 247, pp. 1587–1589, 1990. DOI: 10.1126/science.2321019.

[86] B. Schubert and O. Kohlbacher, "Designing string-of-beads vaccines with optimal spacers", *Genome Medicine*, vol. 8, no. 9, 2016. DOI: 10.1186/s13073-016-0263-6.

[87] L. Sherman and S. Chattopadhyay, "The molecular basis of allorecognition", *Annual Review of Immunology*, vol. 11, pp. 385–402, 1993. DOI: 10.1146/annurev.iy.11.040193.002125.

[88] A. M. Sherwood, R. O. Emerson, D. Scherer, *et al.*, "Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of t cell receptor sequences that differ from the t cells in adjacent mucosal tissue.", *Cancer Immunology, Immunotherapy*, vol. 62, pp. 1453–1461, 2013. DOI: 10.1007/s00262-013-1446-2.

[89] J. S. Sims, B. Grinshpun, Y. Feng, T. H. Ung, J. A. Neira, J. L. Samanamud, P. Canoll, Y. Shen, P. A. Sims, and J. N. Bruce, "Diversity and divergence of the glioma-infiltrating t-cell receptor repertoire", *Proceedings of the National Academy of Sciences*, vol. 112, E3529–E3537, 2016. DOI: 10.1073/pnas.1601012113.

[90] N. B. Strauli and R. D. Hernandez, "Statistical inference of a convergent antibody repertoire response to influenza vaccine", *Genome Medicine*, vol. 8, no. 60, 2016. DOI: 10.1186/s13073-016-0314-z.

[91] Y. Suessmuth, R. Mukherjee, B. Watkins, *et al.*, "Cmv reactivation drives posttransplant t-cell reconstitution and results in defects in the underlying tcrβ repertoire", *Blood*, vol. 125, pp. 3835–3850, 2015. DOI: 10.1182/blood-2015-03-631853.

[92] L. Taylor, "Aggregation, variance, and the mean", *Nature*, vol. 189, pp. 732–735, 1961. DOI: 10.1038/189732a0.

[93] J. J. C. Thome, N. A. Yudanin, Y. Ohmura, M. Kubota, B. Grinshpun, T. Sathaliyawala, T. Kato, H. Lerner, Y. Shen, and D. L. Farber, "Spatial map of human t cell compartmentalization and maintenance over decades of life", *Cell*, vol. 159, pp. 814–828, 2014. DOI: 10.1016/j.cell.2014.10.026.

[94] J. J. Thome, B. Grinshpun, B. V. Kumar, M. Kubota, Y. Ohmura, H. Lerner, G. D. Sempowski, Y. Shen, and D. L. Farber, "Long-term maintenance of human naïve t cells through in situ homeostasis in lymphoid tissue sites", *Science Immunology*, vol. 1, no. 6, 2016. DOI: 10.1126/sciimmunol.aah6506.

[95] J. J. Thome, N. Yudanin, Y. Ohmura, M. Kubota, B. Grinshpun, T. Sathaliyawala, K. Tomoaki, H. Lerner, Y. Shen, and D. L. Farber, "Spatial map of human t cell compartmentalization and maintenance over decades of life", *Cell*, vol. 159, pp. 814–828, 2014. DOI: 10.1016/j.cell.2014.10.026.

[96] G. Tosato and J. I. Cohen, "Generation of epstein-barr virus (ebv)–immortalized b cell lines", *Current Protocols in Immunology*, vol. 76, pp. 7.22.1–7.22.4, 2007. DOI: 10.1002/0471142735.im0722s76.

[97] G. V., C. D., and L. M.-P., "Imgt/gene-db: A comprehensive database for human and mouse immunoglobulin and t cell receptor genes", *Nucleic Acids Research*, vol. 33, pp. 256–261, 2005. DOI: 10.1093/nar/gki010.

[98] V. Venturi, H. Y. Chin, T. E. Asher, *et al.*, "Tcr β-chain sharing in human cd8$^+$ t cell responses to cytomegalovirus and ebv", *The Journal of Immunology*, vol. 181, pp. 7853–7862, 2008. DOI: 10.4049/jimmunol.181.11.7853.

[99] R. G. Verhaak, K. A. Hoadley, E. Purdom, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in

pdgfra, idh1, egfr, and nf1", *Cancer Cell*, vol. 17, pp. 98–110, 2010. DOI: 10.1016/j.ccr.2009.12.020.

[100] D. S. Vinay, E. P. Ryan, G. Pawelec, *et al.*, "Immune evasion in cancer: Mechanistic basis and therapeutic strategies.", *Seminars in Cancer Biology*, vol. 35, S185–S198, 2015. DOI: 10.1016/j.semcancer.2015.03.004.

[101] A. Waziri, B. Killory, A. T. I. Ogden, P. Canoll, R. C. Anderson, S. C. Kent, D. E. Anderson, and J. N. Bruce, "Preferential in situ cd4+cd56+ t cell activation and expansion within human glioblastoma", *The Journal of Immunology*, vol. 180, pp. 7673–7680, 2008. DOI: 10.4049/jimmunol.180.11.7673.

[102] J. Weinstein, N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake, "High-throughput sequencing of the zebrafish antibody repertoire", *Science*, vol. 324, pp. 807–810, 2009. DOI: 10.1126/science.1170020.

[103] G. B. West, "The origin of universal scaling laws in biology", *Physica A.*, vol. 263, pp. 104–113, 1999.

[104] F. W.L. and R. Atkins, "The proportion of lymphocytes capable of recognizing strong transplantation antigens in vivo", *Advances in Experimental Medicine and Biology*, vol. 29, pp. 255–262, 1973. DOI: 10.1007/978-1-4615-9017-0_37.

[105] F. W.L., S. Simmonds, and R. Atkins, "Early cellular events in a systemic graft-vs.-host reaction. ii. autoradiographic estimates of the frequency of donor lymphocytes which respond to each ag-b-determined antigenic complex", *The Journal of Experimental Medicine*, vol. 141, pp. 681–696, 1975. DOI: 10.1084/jem.141.3.681.

[106] D. J. Woodsworth, M. Castellarin, and R. A. Holt, "Sequence analysis of t-cell repertoires in health and disease", *Genome Medicine*, vol. 5, no. 98, 2013. DOI: 10.1186/gm502.

[107] K. W. Wucherpfennig, E. Gagnon, M. J. Call, E. S. Huseby, and M. E. Call, "Structural biology of the t-cell receptor: Insights into receptor assembly, ligand recognition, and initiation of signaling", *Cold Spring Harbor Perspectives in Biology*, vol. 2, a005140, 2010. DOI: 10.1101/cshperspect.a005140.

[108] K. W. Wucherpfennig and J. L. Strominger, "Molecular mimicry in t cell-mediated autoimmunity: Viral peptides activate human t cell clones specific for myelin basic protein", *Cell*, vol. 80, pp. 695–705, 1995. DOI: 10.1016/0092-8674(95)90348-8.

[109] X. Yang, X. Zhang, M. L. Fu, R. R. Weichselbaum, T. F. Gajewski, Y. Guo, and F. Yang-Xin, "Targeting the tumor microenvironment with interferon-β bridges innate and adaptive immune responses", *Cancer Cell*, vol. 25, pp. 37–48, 2014. DOI: 10.1016/j.ccr.2013.12.004.

[110] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesen, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, and M. Vidal, "High-quality binary protein interaction map of the yeast interactome network", *Science*, vol. 322, pp. 104–110, 2008. DOI: 10.1126/science.1158684.

[111] U. Yule, "A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s.", *Philosophical Transactions of the Royal Society B*, vol. 213, pp. 402–410, 1925. DOI: 10.1098/rstb.1925.0002.

[112] I. V. Zvyagin, M. V. Pogorelyy, M. E. Ivanonva, *et al.*, "Distinctive properties of identical twins' tcr repertoires revealed by high-throughput sequencing", *Proceedings of the National Academy of Sciences*, vol. 111, pp. 5980–5985, 2014. DOI: 10.1073/pnas.1319389111.