Bowties, Barcodes, and DNA Origami;

A Novel Approach for Paired-Chain Immune Receptor Repertoire Analysis

by

Louis Schoettle

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2017 by the
Graduate Supervisory Committee:

Joseph Blattman, Chair
Hao Yan
Yung Chang
Stuart Lindsay

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

There are many biological questions that require single-cell analysis of gene sequences, including analysis of clonally distributed dimeric immunoreceptors on lymphocytes (T cells and B cells) and/or the accumulation of driver/accessory mutations in polyclonal tumors. Lysis of bulk cell populations results in mixing of gene sequences, making it impossible to know which pairs of gene sequences originated from any particular cell and obfuscating analysis of rare sequences within large populations. Although current single-cell sorting technologies can be used to address some of these questions, such approaches are expensive, require specialized equipment, and lack the necessary high-throughput capacity for comprehensive analysis. Water-in-oil emulsion approaches for single cell sorting have been developed but droplet-based single-cell lysis and analysis have proven inefficient and yield high rates of false pairings. Ideally, molecular approaches for linking gene sequences from individual cells could be coupled with next-generation high-throughput sequencing to overcome these obstacles, but conventional approaches for linking gene sequences, such as by transfection with bridging oligonucleotides, result in activation of cellular nucleases that destroy the template, precluding this strategy. Recent advances in the synthesis and fabrication of modular deoxyribonucleic acid (DNA) origami nanostructures have resulted in new possibilities for addressing many current and long-standing scientific and technical challenges in biology and medicine. One exciting application of DNA nanotechnology is the intracellular capture, barcode linkage, and subsequent sequence analysis of multiple messenger RNA (mRNA) targets from individual cells within heterogeneous cell

populations. DNA nanostructures can be transfected into individual cells to capture and protect mRNA for specific expressed genes, and incorporation of origami-specific bowtie-barcodes into the origami nanostructure facilitates pairing and analysis of mRNA from individual cells by high-throughput next-generation sequencing. This approach is highly modular and can be adapted to virtually any two (and possibly more) gene target sequences, and therefore has a wide range of potential applications for analysis of diverse cell populations such as understanding the relationship between different immune cell populations, development of novel immunotherapeutic antibodies, or improving the diagnosis or treatment for a wide variety of cancers.

ACKNOWLEDGMENTS

I would like to first take this opportunity to thank my thesis advisor and mentor, Dr. Joseph Blattman. We began this journey as an idea on the back of a napkin and have overcome more challenges and setbacks than any other endeavor I've experienced thus far in my life. The quote by Thomas Edison comes to mind, "I have not failed 1,000 times. I have successfully discovered 1,000 ways to NOT make a *synthetic DNA nanostructure capable of being transfected into primary immune cells, capturing multiple species of mRNA, and be reisolated and used to provide input material for next generation sequencing*." Joe, thank you for pushing me and believing in me when it seemed like the odds were insurmountable and every time we fixed one problem two more arose in its place. But most importantly, thank you for never letting me quit.

I must also thank Dr. Hao Yan, who has always provided insightful and elegant molecular solutions and allocated the resources to enhance and support this research. In addition, I would like to thank Dr. Yung Chang and Dr. Stuart Lindsay for their input during meetings and presentations, hearing ideas and solutions from each of you vastly improved the design and direction my project has taken over the years. To all current and future graduate students who may come across this dissertation, the knowledge of your committee is one of your most valuable resources for information, without which my project would undoubtedly have stalled in its infancy.

I would like to thank all past and present members of the Blattman and Yan labs that have provided thoughtful discussion, helpful suggestions, and friendship that I truly appreciate. I would like to especially thank Dr. Megan McAffee, Dr. Susan Holechek, Dr.

Xixi Wei, Yu Zhou, Kavita Manhas, and Lauren O'Neil. I must also thank my three former and current undergraduate research assistants Marlene Garcia-Neuer, Morgan Poindexter, and Abigail Reed. The simple fact that you were able to put up with my constant stress and grumpiness nearly exceeds each of your excellence in laboratory research. Each one of you contributed significantly to my work and I am honored to have helped you progress your future in the sciences. I must also thank Dr. Bertram Jacobs and Dr. Brenda Hogue as the Directors of the School of Life Sciences and Microbiology programs, respectively, for their guidance and recommendations as student leader for both faculty and graduate student recruitments.

I wish to thank all of my family members, loved ones, and extended family. My beautiful daughter Lucy came into my life shortly after beginning my PhD studies and there is no doubt in my mind that without the help and support of her mother and grandparents I would never have succeeded. To Polly, my wonderful girlfriend who has stood by me and supported me over the last two and a half years, your love for knowledge, art, and all things elegant and beautiful has kept me energized and in motion. Regardless of whether things were going good or bad in my life, you have always pushed me to continue climbing the hill. Finally, to my parents Lou and Becki, for their endless encouragement, understanding, and unconditional support that has allowed me to pursue any and all of my dreams. My road has been paved with more bumps, twists, and turns than most, but you have supported me through it all, and for that I can only hope you someday understand how thankful I truly am.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
TCR	T cell receptor
LCMV	Lymphocytic Choriomeningitis Virus
MHC	Major Histocompatibility Complex
BCR	B cell receptor
HIV	Human Immuno Deficiency Virus
HCV	Hepatitis C Virus
LPS	Lipopolysaccharide
PAMP	Pathogen Associated Molecular Pattern
PRR	Pattern recognition receptor
TLR	Toll Like Receptor
RLR	RIG-I like receptor
fMLP	f-methionyl-leucyl-phenylalanyl
NOD-like	Nucleotide-binding oligomerization domain-like
NET	Neutrophil extracellular traps
APC	Antigen presenting cell
ROS	Reactive oxygen species
IFNγ	Interferon gamma
TNFα	Tumor necrosis factor alpha

β_2M	β_2-microglobulin
ER	Endoplasmic reticulum
TAP	Transporter associated with antigen processing
Ii	Invariant chain
T_H1	T helper type 1
T_H2	T helper type 2
K_d	Dissociation constant
k_{off}	Off rate
mAb	Monoclonal antibody
SPR	Surface plasmon resonance
PTK	Protein tyrosine kinase
Src	Proto-oncogene tyrosine-protein
Lck	Lymphocyte-specific protein tyrosine kinase
ITAM	Intracellular tyrosine activation motif
IP_3	Inositol triphosphate
PLC	Phospholipase C
DAG	Diacylglycerol
ZAP-70	ζ-association protein-70
LAT	Linker for the activation of T cells
SH2	Src homology 2
SLP-76	SH2-domain containing leukocyte phosphoprotein-76
GEF	Guanine nucleotide exchange factor

PIP_2	Phosphatidylinositol 4,5 bisphosphate
PKCΘ	Protein kinase C theta
MAPK	Mitogen-associated protein kinase
Erk	Extracellular signal-regulated kinase
STAT	Signal transducer and activator of transcription
AP-1	Activator protein-1
NFκB	Nuclear factor kappa-light-chain-enhancer of activated B cells
IκB	Inhibitor of NFκB
IP_3R	Ca^{2+}-permeable ion channel receptors
CRAC	Ca^{2+} release-activated Ca^{2+}
SOCE	Store-operated Ca^{2+} entry
MEF2	Myocyte-enhancing factor 2
DREAM	Downstream regulatory element antagonist modulator
NFAT	Nuclear factor of activated T cells
IL	Interleukin
F-actin	Filamentous actin
mTOC	Microtubular organizing center
LFA-1	Leukocyte function-associated antigen-1
VLA-4	Very late antigen-4
ICAM	Intracellular adhesion molecule
VCAM	Vascular cell adhesion molecule

PI3K	Phosphoinositide 3-kinase
PIP_3	Phosphatidylinositol (3,4,5) triphosphate
BM	Bone marrow
HSC	Hematopoietic stem cell
NK	Natural Killer
MPP	Multipotent progenitor
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CCR	Chemokine receptor
CCL	Chemokine ligand
TECK	Thymic expressed chemokine
DN	Double negative
DP	Double positive
SP	Single positive
RAG	Recombinase activating gene
cTEC	Cortical thymic epithelial cell
mTEC	Medullary thymic epithelial cell
AIRE	Master transcriptional autoimmune regulator
T_{reg}	Regulatory T cell
HEV	High endothelial venule
pTα	Surrogate α-chain
V	Variable

D	Diversity
J	Joining
CDR	Complementarity determining region
RSS	Recombination signal sequence
nt	Nucleotide
DSB	Double stranded break
P-	Palindromic
TdT	Terminal deoxynucleotide transferase
N-	Non-templated
Pol	Polymerase
HVR	Hypervariable region
TAE	Tris-acetate-EDTA
PCR	Polymerase chain reaction
RT	Reverse Transcription/Reverse Transcribe
PAGE	Polyacrylamide gel electrophoresis
ACK	Ammonium chloride lysis buffer
NMWL	Nominal molecular weight limit
AFM	Atomic force microscopy
FITC	Fluorescein isothiocyanate
TAMRA	5-carboxytetramethylrhodamine
FACS	Fluorescence activated cell sorting
RACE	Rapid amplification of cDNA ends

NGS	Next generation sequencing
CAR	Chimeric antigen receptor
BC	Barcode
wt	Wild type

CHAPTER 1

INTRODUCTION

Historical review of T cell biology

Whether he realized it or not, Thucydides observed the phenomenon of immunity nearly 2500 years ago. Observing soldiers who survived disease during the Peloponnesian War, he discovered that these survivors rarely if ever succumbed to the same disease [1]. While the concept of vaccinations and immunological memory would not be defined for millennia, the seed was laid for the further analysis of how one's body protects itself from disease. Advancements in the study and practice of what would be come known as the field of immunology progressed in China and India in the 11th century as these countries battled small pox epidemics. However, it wasn't until the mid 19th century and the discovery that microorganisms were responsible for illness that the science of immunology was formally established. Furthermore, most of the initial experimental studies involved humoral immunity and the biochemical characterization of antibodies [2].

Some of the earliest pioneering work demonstrating CD8 T cell pathology came from James Murphy between 1911 and 1926 [3, 4]. Murphy suggested that it was a subset of cells (lymphocytes), rather than antibodies that were causing allograft rejection. Murphy demonstrated that injection of adult lymphoid tissue near implanted tumor cells in chick embryos led to a rapid destruction of the tumor graft. However, it would be decades before CD8 T cells were fully understood in terms of both immune protection

and graft rejection as leading immunologists of the day fully admitted that nothing of importance was known about lymphocytes [2].

In the mid 1950s, it was well known that while stable levels of lymphocytes remained in the blood, the thoracic duct drained nearly ten times that number back into the circulatory system each day. Numerous hypotheses tried to account for this discrepancy including that other cells degraded lymphocytes, or that they died in peripheral tissues. In 1964, James Gowans published a seminal paper illustrating how lymphocytes continually recirculate from the lymph system to the circulatory system via the thoracic duct and back [5]. Once lymphocyte migration patterns and kinetics were understood, studies outlining lymphocyte roles in graft rejection, differentiation into antibody producing cells, and immunological memory were rapidly performed [6-8].

While the seminal work of Gowans et al. laid the foundation for our current understanding of T cell biology, during his work it was postulated that all small lymphocytes were homogenous in terms of their potential. In the early 1960s Jacques Miller began to identify the role of the Thymus in lymphocyte development. Performing thymectomies, Miller's research demonstrated the existence of at least two major classes of lymphocytes: antibody-producing cells derived from the bone marrow, and a second class of cells derived from the thymus [9-12]. It was these pioneering studies from Gowans and Miller that began research into the cellular aspect of immunology.

By the early 1970s it had become widely accepted that lymphocytes derived from the thymus (T cells) were functionally distinct from their antibody-producing counterparts derived from the bone marrow (B cells). Initial experiments into T cell

function demonstrated the cytotoxicity of T cells; additionally, the removal of B cells from lymphocyte fractions did not alleviate cytotoxicity [13, 14]. The initial killing assays demonstrated a high level of specificity that would eventually manifest itself as the discovery of clonal antigen receptors on T cells. While the discovery of T cell receptor (TCR) clonality occurred in the late 1960s [15, 16], it would take more than a decade before an accurate understanding of T cell antigen recognition was established.

In1974, Peter Doherty and Rolf Zinkernagel accomplished the next major discovery in T cell antigen recognition by demonstrating that T cells obtained from Lymphocytic Choriomeningitis Virus (LCMV) infected mice would only kill target cells that expressed the same Major Histocompatibility Complex (MHC) proteins [17]. Differing opinions on how the MHC was involved with T cell antigen recognition revolved around two main hypotheses. The two-receptor model was based on the hypothesis that T cell antigen recognition involved two interactions, the first between the T cell and the viral antigen, and the second between the T cell and the MHC. This hypothesis was rejected however, when Zinkernagel and Doherty demonstrated that MHCs were complexed with viral antigens and therefore T cell antigen recognition occurred via a single T cell receptor (TCR) and an MHC:antigen complex [18].

Although somewhat similar in genetic organization, the B cell antigen receptor (BCR) was characterized well before a complete understanding of its counterpart in T cells was understood. Somatic rearrangement of BCR genes explained the phenomenon of how such a wide diversity of antigens could be recognized by receptors encoded by so few genes. However, characterization of the TCR eluded scientists due to the inability of

TCRs to bind free antigen and the fact that T cells did not secrete their receptors as B cells did. It wasn't until the early 1980s when Allison *et al.* generated clonotypic antibodies against T cell clones permitting analysis of the TCR [19, 20]. The initial biochemical analysis of the TCR revealed it to be very similar to the B cell receptor; a heterodimeric receptor containing both constant and variable regions.

Characterization of the TCR complex led way to further studies to outline antigen processing and presentation for T cell recognition. Shortly after Allison and colleagues characterized the TCR, in 1983 Shimonkevitz, Marrack, Kappler and Grey revealed that antigen must first be fragmented into short peptides for TCR recognition [21] and two years later Townsend and Yewdell *et al.* demonstrated T cell recognition of intracellular influenza proteins and therefore the existence of a mechanism to transport and present antigenic peptides on the surface of cells [22, 23]. It was these pioneering studies that firmly established the one-receptor hypothesis of T cell antigen recognition.

Concurrently with the research on understanding antigen processing and MHC presentation, studies were being performed to understand the T cell receptor genetic locus. Numerous groups had begun identifying the composition of the TCR in that it consisted of a dimer covalently linked by disulfide bridges. Additionally, each of the two monomers contained both constant and variable regions [24-26]. These studies helped propel two separate groups to independently clone the TCR and allowed them to outline the genetic organization of the beta locus of the TCR [27-29]. Once the genes were characterized a more complete understanding of the basis for constant and variable

regions of the TCR was used to provide an initial explanation for the vast heterogeneity of an individual's T cell repertoire.

Although a detailed structural and genetic picture of the TCR had been elucidated, how the T cell interacted with the antigen:MHC complex was still unclear. It wasn't until 1987 when Bjorkman and Wiley were able to crystallize Class I MHC and thus provide its structure [30] that the deep, negatively charged, MHC antigen-binding groove was visualized. Realization that antigenic peptide was presented to T cells bound to a groove on the surface of MHC molecules resolved any remaining questions about the one-receptor hypothesis. Bjorkman and colleagues used the crystal structure of the MHC to propose that TCR engage both the antigenic peptide as well as surface residues of the MHC [31]. They also discovered that the majority of polymorphisms between different MHC corresponded to either the peptide-binding site or to TCR contact residues, further evaluation of which would eventually yield the mechanisms behind T cell MHC restriction and allele specificity [2].

With this new knowledge of MHC:antigen presentation, several groups began to outline the CD8 co-receptor's role in antigen recognition. In the late 1980s, it was demonstrated in a number of different ways that CD8 increases the avidity of T cell antigen recognition through binding interactions with Class I MHC complexes [32-34]. It wouldn't be until eight years later in 1996 that David Garboczi would finally crystalize and thus provide the first complete picture of the TCR:peptide:MHC complex [35].

Once a clear picture of how T cells recognized and thus responded to antigen was determined, the next decade would see huge strides in a more complete understanding of

all aspects in T cell biology. Specifically, the roles of perforin/granzyme and the Fas ligand in T cell cytotoxicity, proteosome degradation of antigen and the numerous chaperone proteins involved with antigen processing, positive and negative selective processes delineated over stages of thymic development, classifying numerous types of memory T cells, and the impact of chronic infections on T cell function would all be elucidated [2]. Advancements in biochemical assays such as ELISPOT and intracellular cytokine staining allowed for the quantification of antigen-specific T cell subsets during various infections [36]. Additionally, in 1996 David Altman would develop MHC tetramers, capable of being loaded with various peptides that could be used to analyze and/or quantitate antigen-specific subsets of T cells from polyclonal populations [37].

The late 1990s and early 2000s would expound upon the discoveries of the early 90s and utilizing the new technologies provided by Altman and others a more thorough understanding of memory T cells and T cell differentiation would elucidate differences in central and effector memory T cell subsets [38]. Additionally it was discovered that CD4 T cells have a major role in CD8 T cell differentiation [39] as well as during chronic infections leading to T cell exhaustion and/or deletion [40]. These studies (and others) have laid the groundwork for current research into diagnostics and therapeutics for chronic diseases plaguing our world today.

While Murphy's pioneering studies of what would later come to be identified as T cells began in the early 1900s, it has only been in the last 20 years that a complete understanding of how T cells recognize and thus respond to antigen has been elucidated. Furthermore, while the discovery of many of our successful vaccinations preceded our

understanding of the immune system, advancement in our understanding of T cell structure and function may be necessary before we are able to develop protective vaccines against the many deadly pathogens (HIV, HCV, malaria) that still plague our world. Future studies should also be directed towards a more thorough evaluation of the effects of and relationships between antigen density and persistence, CD4 T cell help via co-stimulation and cytokine secretion, and other factors involved in CD8 T cell activation [41]. Our knowledge of T cells has been slowly evolving over the last hundred years, however remarkable progress has been seen over the last few decades. It goes without saying that many of these recent advancements have coincided with the development of new and more precise biochemical tools and assays. We should strive to continue our development of new molecular tools, as future discoveries in T cell biology will undoubtedly depend on such advancements.

The innate immune system is the first line of defense

All multicellular eukaryotic organisms have evolved highly complex and intricate systems of immunity. Innate immunity is the most universal aspect of these systems, as most living organisms survive solely based on the actions of an innate immune system. Only vertebrates have developed adaptive immune systems (discussed later). Regardless of which branches of the immune system an organism has, all types of immunity must be represented by three core elements: the ability to sense and thus respond to an ever increasing diversity of pathogens, neutralizing or eliminating these pathogens from the

host, and maintaining the ability to discern self from non-self and thus sparing host tissues from the detrimental effects of eliminating the pathogens [42].

The innate immune system can be broken into two main branches, the afferent or "sensing" branch and the efferent or "effector" branch. While some cells specialize in one branch or the other, most innate immune cells will have roles in both. Additionally, there are both cellular and humoral components of the innate immune system and once again each component will have various roles in the two branches of the system [42].

The first major priority of the innate immune system is to sense the initial onset of an infection. While vastly different in both approach and specificity compared to the adaptive immune system, the innate immune system recognizes foreign substances using a series of receptors and soluble components that alert cells to an infection. Some of the earliest recorded understanding of this fact came in the late 1800s and early 1900s when scientists discovered the disease-causing toxin from gangrenous infections "Sepsin" [43, 44]. Later renamed endotoxin and now commonly referred to as lipopolysaccharide (LPS), several decades would pass before LPS would be chemically characterized [45, 46] and even longer before it was recognized as the major component to Gram-negative bacterial cell walls [46]. Concurrent with the discovery of LPS as an immune system activator, many other microbial components were discovered to trigger immune responses as well (i.e. peptidoglycan, lipoteichoic acid, double stranded RNA, etc) [42]. These "non-self" antigens act as triggers upon innate system recognition for the activation of the innate immune response.

The innate immune system varies from the adaptive immune system in a number of ways, including pathogen recognition. While the adaptive immune system has evolved to develop repertoires of highly unique receptors for the recognition of virtually any invading pathogen, the innate immune system relies on a limited series of receptors, each dedicated to recognition of a class of broadly conserved Pathogen Associated Molecular Patterns (PAMPs). These broadly recognizing receptors, referred to as pattern recognition receptors (PRRs), have evolved to distinguish ligands specific and indispensable to microbes while maintaining indifference to host tissue [47]. One of the first classes of these receptors to be recognized is the Toll Like Receptors (TLRs) [48]. Thirteen different TLRs have been discovered to date (eleven found in humans) [49-51], each of which is believed to confer the ability to recognize a different PAMP or class of PAMPs. Additionally, some TLRs form hetero- or homo-dimers with other TLRs (i.e. TLR2/TLR6) while others work as monomers for antigen recognition. Initial recognition of bacteria, fungi, protozoa and viruses can all be accomplished through various ligand:TLR interactions [52, 53]. Similar to TLRs, numerous other broadly recognizing innate immune receptors have evolved including RIG-I like receptors (RLRs), f-methionyl-leucyl-phenylalanyl (fMLP) receptors, nucleotide-binding oligomerization domain-like (NOD-like) receptors, and others. Each of these receptor types sense microbial products and lead to effector responses via intracellular secondary messenger system signaling cascades.

While the ability of the innate immune system to sense an infection is paramount in the cascade of events leading to immune protection, its ability to kill microbes is

equally important. Although the adaptive immune response is the primary means by which vertebrates clear an infection, the innate immune response acts as our first line of defense, and must slow the progression of the infection down long enough for activation of the adaptive immune response to occur. For this purpose, innate immune cells have developed a number of chemical means by which to kill invading pathogens.

In vertebrates, innate immunity relies heavily on cells derived from the myeloid lineage. Macrophages and dendritic cells are two classes of mononuclear phagocytes highly efficient at T cell antigen presentation (discussed later), and are therefore excellent at phago- or pinocytosis of extracellular compounds and lysosomal degradation. While dendritic cells have virtually no anti-microbial capability, macrophages are capable of phagocytosing and killing whole microbes. Neutrophils, basophils and eosinophils are three classes of polymorphonuclear cells each with killing ability. Neutrophils are the most abundant innate cell type and can perform 4 main anti-microbial functions: phagocytosis, release of reactive oxygen species, degranulation of anti-microbial proteins and enzymes, and secretion of their DNA as a web of fibers trapping and killing microbes referred to as neutrophil extracellular traps (NETs) [54]. Eosinophils and basophils are more restricted and rely mainly on the secretion of different anti-microbial granules. Numerous other innate immune cell types exist, each of which may utilize one or many of the previously mentioned anti-microbial effector responses, however all innate immune cells perform an additional role over the course of the immune response, the release of cytokines to direct other innate immune cells or for the activation of adaptive immune cells.

Antigen processing by the innate immune system

Professional antigen presenting cells (APCs) such as dendritic cells and macrophages play a key role in T cell immunology. Unlike B cells, whose antigen receptors are capable of interacting with the intact native structure of antigens, T cells require proteolytic degradation of antigens into small peptides and then presentation of these antigens on the surface of cells via MHC complexes. A unique characteristic of MHC proteins is that they are both polygenic and polymorphic. Humans have three genes that express classical MHC class I proteins, HLA-A, HLA-B and HLA-C, and three genes that express classical MHC class II proteins, HLA-DR, HLA-DQ and HLA-DP [55]. Intracellular infections can lead to the buildup of foreign cytosolic proteins, proteosomal degradation of these proteins followed by MHC I presentation allows for activation of CD8 T cells. Additionally, APCs are constantly surveying our extracellular environment by sampling exogenous matter (including microbes) through phagocytosis, pinocytosis, and receptor-mediated endocytosis [56]. Following endosomal lysis of these ingested antigens, extracellular proteins are usually loaded onto MHC II complexes for activation of CD4 T cells. More recently discovered however, is the process of cross-presentation by MHC I, which involves extracellular antigens from phagocytosed particles or proteins assembled with MHC I. While more restricted in the cell types capable of this process (specific classes of dendritic cells only), cross-presentation is an extremely important aspect of antigen presentation [57, 58].

Originally discovered in 1908 by Elie Metchnikov [59], phagocytosis was originally thought to be solely a mechanism of microbial killing by the innate immune

system. The role of phagocytosis for MHC II presentation of extracellular antigens (and for cross presentation by MHC I) has since been well defined. Phagocytosis is a process involving the remodeling of actin to form a bowl around the engulfed particle that will eventually close to form a phagosome [60, 61]. Following phagosome formation, the vesicles will interact with lysosomal vesicles from the endocytic system to form phagolysosomes. While the events leading up to formation and the specific properties of any given phagolysosomes may differ by phagocytic cell type [62], maturation of the phagolysosomes is highly regulated by innate signaling pathways initiated by the engulfment of exogenous particles [63]. Particles in the phagolysosome are rapidly degraded by a highly acidic environment (pH ≈ 4.5-5), reactive oxygen species (ROS), proteases, and many other antimicrobial proteins and peptides [60].

The processing of intracellular pathogens from viral or intracellular bacterial infections involves the use of a specialized proteolytic machine called the immunoproteosome. Short peptides produced by immunoproteosomal cleavage will then be loaded onto MHC I complexes for activation of CD8 T cells. The immunoproteosome is derived from the ubiquitously expressed proteosome (26S proteosome) who functions to degrade ubiquitinated proteins [64]. Conversion of the proteosome to the immunoproteosome is triggered by inflammatory cytokines (IFNγ and TNFα) that may be released under inflammatory conditions such as during an infection, autoimmune disease, or allergy [65, 66]. The basic proteosome complex itself is a large barrel shaped protein complex composed of a catalytic 20S core and two 19S regulatory caps [67]. The inner subunits of the catalytic core confer caspase, trypsin, and chymotrypsin like

cleavage capability [68], and the 19S cap of the proteosome recognizes ubiquitinated proteins and transfers them into the core for degradation. However, upon inflammatory cytokine stimulation, five of the catalytic core subunits are replaced with 5 different more catalytically efficient subunits and the 19S regulatory cap is replaced with the 11S cap that lacks requirement for ubiquitination [69-71]. Upon these changes, the proteosome becomes the immunoproteosome and degrades intracellular proteins more much efficiently.

MHC loading and antigen presentation by innate immune cells

MHC class I molecules are expressed on virtually all nucleated cells and are necessary for the presentation of peptides derived from intracellular proteins [72]. While presentation of degraded self-proteins such as old degraded proteins or defective ribosomal translation products is common [73, 74], the classical MHC class I antigen presentation pathway is also responsible for processing and presentation of viral peptides translated by the host ribosomes to CD8 T cells from infected cells. Upon recognition of peptide:MHC, virus-specific CD8 T cells will eliminate the virally infected cells.

MHC class one molecules are heterodimers composed of a 45000 MW type I integral membrane glycoprotein and a 12000 MW soluble β_2-microglobulin (β_2M) protein [75]. The dimeric protein is composed of four domains: three alpha domains (α1, α2, and α3) derived from the glycoprotein and the fourth domain deriving from β_2M. The combination of the α1 and α2 domains form the antigen-binding groove **(Figure 1.1A)**.

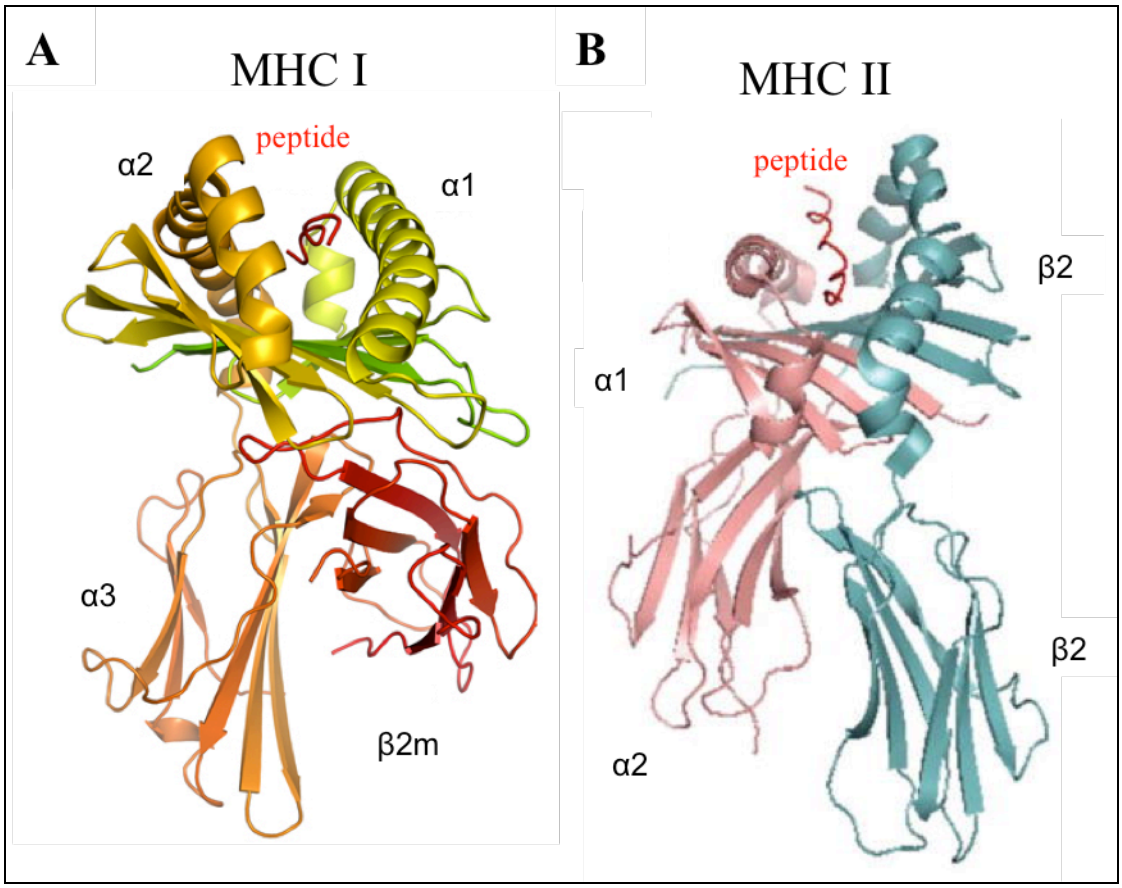

Figure 1.1: Crystal structures of major histocompatibility complex class I and II. **[A]** The crystal structure of the MHC class I molecule illustrates the interaction between the alpha chain composed of three subunits and the interlocked β2m peptide. Alpha helices of the α1 and α2 subunits form a groove for peptide binding. **[B]** The crystal structure of MHC class II molecule illustrating a heterodimer formed of both an alpha and beta chain, each composed of two subunits. Alpha helices on the α1 and β1 subunits fold together, forming the MHC II peptide-binding site. MHC I crystal structure adapted from Zhang *et al.* 2012 [76] and MHC II crystal structure adapted from Painter *et al.* 2011 [77].

This binding site is suitable for displaying peptides of 8-10 amino acids in length and is a required step in the synthesis of the MHC molecules. MHC class I molecules are assembled in the lumen of the rough endoplasmic reticulum (ER), and as loading of peptide is a required step of MHC synthesis, cytosolic peptides must be transported into the lumen of the ER. This peptide transport is carried out by the transporter associated with antigen processing (TAP) complex [78, 79]. In a complex with three other ER resident chaperone proteins, calreticulum, tapasin and ERP57, TAP associates with the nascent MHC class I complex and facilitates the loading of peptide [80, 81]. Following peptide loading, MHC class I complexes (with associated peptide) dissociate from the tap complex and are exported to the Golgi and sent on to integrate into the cell plasma membrane.

MHC class II molecules are more restricted in cell types expressing them (Dendritic cells, macrophages and B cells) [82] and are specifically responsible for the presentation of internalized exogenous antigens to CD4 T cells. The MHC class II molecule is also a dimeric protein complex, however it is composed of two integral membrane chains (α and β). Each chain constitutes two domains ($\alpha1$, $\alpha2$, $\beta1$, and $\beta2$) to the total MHC class II complex with the two distal domains $\alpha1$ and $\beta1$ combining to form the antigen binding groove **(Figure 1.1B)**.

MHC class II assembly again takes place in the ER however, unlike MHC class I assembly, construction of the MHC class II molecule is not dependent upon intra-ER loading of MHC with peptide. Instead, association of α chains with β chains is accomplished with the help of the invariant chain (Ii) [83]. Annealing of Ii to the antigen-

binding groove prevents premature peptide binding and facilitates complete assembly of an MHC class II molecule and allows its exit from the ER. Additionally, Ii also facilitates transport of the MHC class II molecule to an endosome (then referred to as MIIC) at which point Ii will be degraded by endosomal proteases leaving behind a small fragment referred to as CLIP bound to the antigen-binding groove of the MHC II molecule [84-86]. Following Ii degradation, endosomally degraded exogenous proteins will be exchanged for the CLIP molecule to assess the MHC II peptide binding. Three MIIC-resident chaperone proteins HLA-DM, HLA-DR, and HLA-DO mediate CLIP/peptide exchange [87, 88]. Once an MHC II has successfully bound to peptide, the peptide:MHC complex will be delivered to the cell membrane for antigen presentation to CD4 T cells.

The previously mentioned pathway of degradation and presentation of intracellular antigens on MHC class I molecules is the mechanism behind most if not all cellular presentation of viral peptides. More recently discovered however is that the initiation of most immune responses (including extracellular infections) is mainly mediated by dendritic cells [89]. For viral infections, some viruses can infect dendritic cells directly and therefore provide the viral antigens for classical MHC I presentation. However, for extracellular infections such as some bacterial, parasitic, and fugal infections, the dendritic cells will not directly express the antigens. For these infection types, dendritic cells must phago- or pinocytose the antigens from the extracellular environment and then ‘cross present’ the peptides on their MHC I molecules. There are two main models, referred to as the “vacuolar” and “cytosolic” pathways, for how dendritic cells accomplish this cross presentation of extracellular antigens on MHC I

molecules [90]. The cytosolic pathway model proposes that partially degraded phagocytosed antigen is released into the cytosol where it is then processed by the proteosome and enters the classical MHC I pathway **(Figure 1.2A)**, or re-enters the phagosome through TAP and is loaded onto MHC I that have been recycled from the membrane down into the phagosome **(Figure 1.2A)**. The vacuolar pathway, similar to the second cytosolic pathway, involves the recycling of membrane-bound MHC I molecules into endosomes and/or phagosomes where they are then loaded with exogenous antigen **(Figure 1.2B)** [91]. Regardless of the mechanism, cross presentation allows dendritic cells to present extracellular antigen and antigen from intracellular infection of non-dendritic cells on MHC I molecules to stimulate CD8 T cell responses [56].

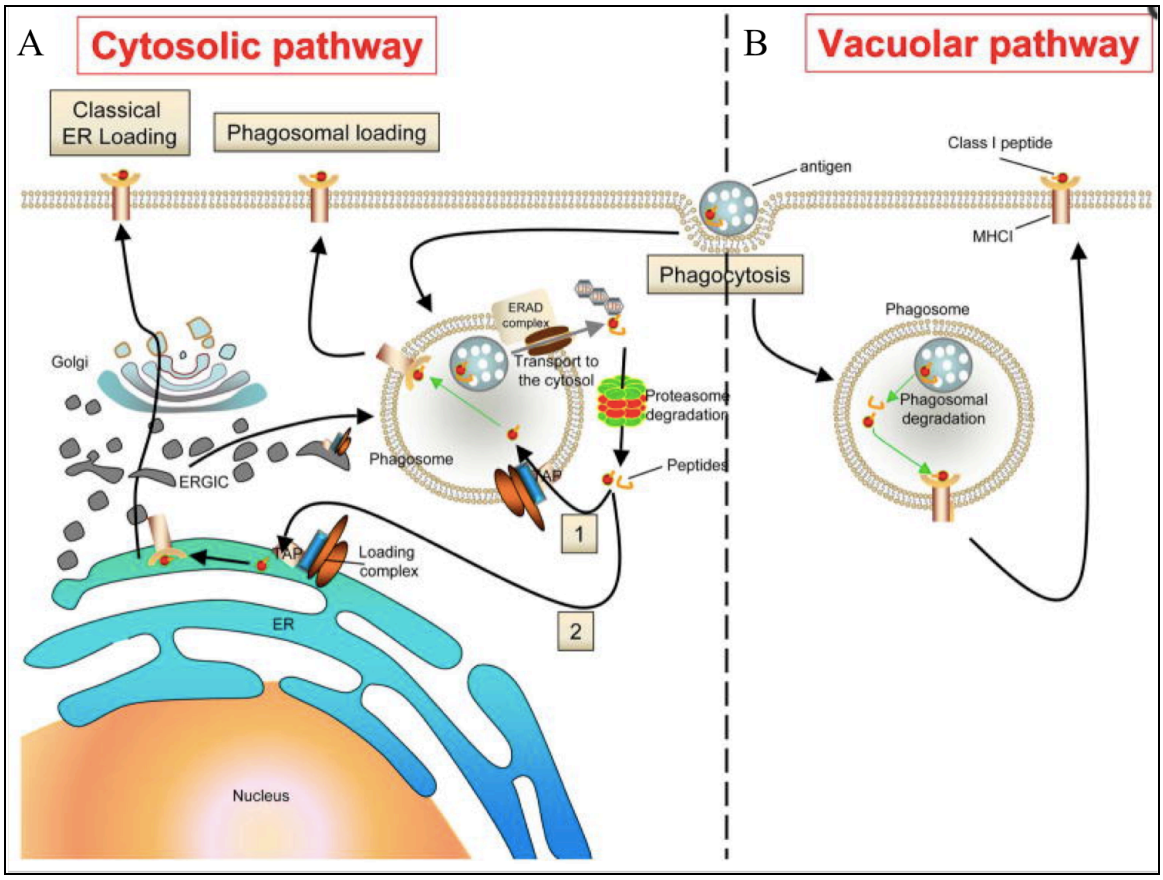

Figure 1.2: The two models of extracellular antigen cross presentation on MHC I. **[A]** The cytosolic pathway model proposes that peptides from phagocytosed antigens are exported to the cytosol where they will be ubiquitinated and targeted to the proteosome for degradation. Following degradation, the peptides will either be (1) translocated back into the phagosome through TAP and loaded onto MHC I that have been recycled from the cell membrane, or (2) fed into the classical MHC I loading pathway in the ER. **[B]** The vacuolar pathway proposes that cross presentation is independent of both the TAP and proteosome. For this pathway, peptides are generated from phagocytosed antigens by cathepsins and low pH in the phagosome, and are subsequently loaded on MHC I in the phagosome that have migrated from the cell membrane. Figure 1.A and 1.B have been adapted from Mantegazza *et al.* 2013 [56].

The adaptive immune system is required for protection

While the term "immunity" refers to an encompassing ability of a host to resist being destroyed by infectious microbes [92], a great dichotomy exists between branches in the vertebrate immune system. Researchers of the innate immune system specialize mainly on how microbes are recognized via innate immune receptors, how innate cells kill invading microbes, and how innate cells alert the system of an infection via cytokine signaling. Adaptive immunologists have spent the last few decades seeking to understand the generation of B and T cell receptor diversity, immune self-tolerance, immunological memory, and the generation of protective antibody and T cell responses.

While the two branches of the immune system perform many different roles, they are both in many ways dependent upon one another. In the late 1960s it was discovered that mononuclear phagocytic cells were required for effective lymphoid cell responses to antigens [93, 94]. Now more thoroughly understood, the adaptive immune system is highly dependent on the innate system for antigen presentation and activation. Even more interesting was the discovery that different "types" of adaptive response can be triggered depending on the mechanism of innate cell activation. Specifically, intracellular pathogens such as viruses require a more robust T helper type 1 (T_H1) response, while some extracellular pathogens such as helminths require a less intense T helper type 2 (T_H2) response [95]. It is now incredibly clear that the innate immune system, while less robust in its protective capability, is an integral part of the adaptive immune system's protection. However, while the innate immune system may be involved with preventing

infection, once virulent infections occur, the adaptive immune system is required to clear them.

There are many differences between the innate and adaptive immune systems, including the means by which they recognize microorganisms. Innate cell antigen receptors are proteins encoded in the germ line that allow them to identify broad classes of compounds usually specific to various types of invading microbes. In contrast to this relatively inflexible system of innate receptors is the nearly infinitely diverse repertoire of adaptive cell antigen receptors. B and T cells acquire they're diverse repertoires of antigen receptors by somatically rearranging large sets of genes to create an estimated pool of more than 10^{16} possible receptors per cell type [96]. These highly diverse repertoires of antigen receptors allow the adaptive immune system to cope with the immense genetic variability of microorganisms, but also lends to the possibility of innocuous antigen reactivity (allergies) or autoreactivity (autoimmunity).

A second key difference between the innate and adaptive systems is the ability of adaptive cells to proliferate and develop into effector and/or memory cells. Clones of B and T cells with receptors triggered by antigen will proliferate, creating a pool of millions of antigen-specific B cells and T cells [97]. The effector subset of these proliferated cells differ in numerous ways from the naïve cells from which they were derived, and the changes allow the effector cells to respond quickly and efficiently when they encounter antigen on target cells in the periphery. Upon clearance of the infection, the majority of the proliferated cell fraction will die off, leaving behind a small subset of antigen-specific cells, termed memory cells. While similar to naïve cells, memory cells are quiescent and

require activation by APCs and co-stimulation in order to become effector cells. However, memory cells do not require as rigorous of an activation step, and will therefore differentiate into effectors much faster than naïve cells. This fast activation ability of memory cells is the crux of some types of vaccinations.

The successful activation of T cells requires multiple productive TCR peptide:MHC interactions

T cell activation is a highly crucial checkpoint in adaptive immunity and it is governed through a number of signaling interactions. In order for T cells to initiate and regulate the adaptive immune response, be it against infection, cancer, during autoimmunity or allergy, productive binding between the TCR and antigenic peptides bound to MHC must first occur [98]. Following activation signals, T cells may proliferate, differentiate, kill target cells (CD8 T cells), secrete cytokines (CD4 T cells), or carry out other effector functions [99].

Numerous experiments have evaluated different panels of TCR:peptide/MHC (TCR:pMHC) combinations to establish the relationship between TCR:pMHC stoichiometry required for T cell activation as measured by downstream cytokine secretion or killing ability [100-102]. Some studies have demonstrated that T cell activation is at an optimum when plotted over the TCR:pMHC dissociation time (τ) [103], which is the reciprocal of the off rate ($k_{off} = 1/\tau$) [99, 104]. More simply, high affinity antigens (antigens that bind with long dissociation times) lead to poor activation of T cells [105]. However, different experiments, utilizing intricate pMHC titrations have

shown that antigens with short dissociation times cant produce the same optimal responses as antigens with longer dissociation times [100, 106]. In these dose-response experiments optimal T cell activation was directly correlated to the TCR:pMHC dissociation constant (K_d), and inversely proportional to the off rate (k_{off}). These differing opinions on the relationship between T cell activation parameters have led to the development of numerous models aimed at explaining TCR:pMHC binding kinetics [106-109]. While continued experimental analysis will be necessary for further explanation of TCR:pMHC binding kinetics vs T cell activation, a kinetic proofreading model described by Lever *et al.* in 2014 [99] is an elegant combination of what we currently understand.

In the analysis by Lever *et al.,* they looked over all the various published models of TCR:pMHC binding kinetics with the aim to predict quantitative T cell responses to antigens over varying antigen:TCR affinity [99]. Their analysis revealed five distinct models: (1) The occupancy/affinity model states that T cell activation is directly proportional to the number of TCR:pMHC interactions, (2) the kinetic proofreading model posits that T cell activation is proportional to the fraction of TCR:pMHC interactions maintained for a sufficient time to achieve full signaling, (3) the kinetic proofreading with limited signaling model proposes that T cell activation requires continuous signaling that necessitates serial binding of TCRs to multiple pMHCs, (4) the kinetic proofreading with sustained signaling model states that there is an optimal dissociation time for all T cell activation at any pMHC concentration, and (5) the kinetic proofreading with negative feedback model predicts that T cell activation will exhibit an

optimum as a function of the pMHC dose [99]. In summary, they concluded that the phenotypic model most compatible with experimental data currently available [106, 110-112]was the kinetic proofreading with limited signaling model. This model allows for reversible binding of TCR:pMHC and therefore serial TCR:pMHC interactions.

CD4 T cells recognize antigenic peptides presented on MHC class II while CD8 T cells recognize peptide presented on MHC class I

The molecular details of antigen presentation by MHC class I and II molecules was described nearly fifteen years ago. While both similar in their antigen presentation function, MHC class I and II molecules typically only present peptide fragments to CD8 or CD4 T cells respectively. A key difference between the two types of MHC antigen presentation is the source of the peptides they present. Peptides derived from intracellular antigens such as viral proteins are typically displayed on MHC class I molecules and thus recognized by CD8 T cells, while peptides derived from extracellular sources are commonly displayed on MHC class II and recognized by CD4 T cells [113]. As mentioned above, additional mechanism of antigen presentation have also been elucidated in which exogenous antigens are presented by MHC I (cross-presentation) [114], or endogenous antigens presented by MHC II (autophagy) [115]. A second key difference between the two types of MHC antigen presentation is the mechanisms of antigen degradation. The MHC I pathway commonly utilizes proteosomal degradation of intracellular peptides, while the classic MHC II pathway relies on phagolysosomal

proteolysis. These intrinsic differences between MHC I and II antigen presentation are fundamental in the different effector functions of CD8 and CD4 T cells.

Most likely owing to the fact that they originate from a common founder gene that has been duplicated [116], MHC I and II share several characteristics. Both classes have a similar structure, they express high levels of polymorphism, and serve the same goal of antigen presentation to the immune system. However, MHC class I and II molecules differ in a wide variety of properties. Beyond the differences discussed above, tissue expression for MHC I is found on nearly all cells while MHC II is reserved mainly for professional antigen presenting cells such as dendritic cells, macrophages and B cells. These intrinsic differences in MHC I and II are essential for the different effector functions of the CD8 and CD4 T cells by which they are recognized. CD8 T cells are our main protectors from viral infections, and as such CD8 T cells must be able to recognize viral peptide presented by virtually any cell. While CD4 T cells are more encompassing in their directing role of the immune response, they play a key role antibody production and are thus necessary for directing the immune response against extracellular pathogens that are mainly presented by the previously mentioned professional antigen presenting cells.

The CD4 and CD8 co-receptors are involved with TCR:pMHC binding and intracellular signaling

In the late 1970s, T cells found in the peripheral blood or secondary lymphoid organs were distinguished using monoclonal antibodies (mAbs) by the exclusive

expression of either CD4 (previously known as T4, L3T4, Ly1, or Leu3) or CD8 (previously known as T8, Leu2, or Ly2/Ly3) cell surface proteins [117]. Initial experiments linked T cell expression of CD4 with the ability to initiate B cell production of antibodies and CD8 expression with cytotoxic function [118, 119], however it was yet unknown whether these surface molecules possessed any inherent effector mediating function themselves. It wasn't until the discovery that CD4 T cells only respond to antigen presented on MHC II molecules [120] and that CD8 T cells only respond to antigen presented on MHC I molecules [33] that studies into how the CD4/CD8 co-receptors mediate TCR:pMHC recognition were undertaken.

A number of studies have identified the roles of the CD4 and CD8 co-receptors on stabilizing TCR:pMHC recognition and downstream signaling [102, 121-123]. Specifically it was shown that at the extremes, high-affinity TCR:pMHC interactions were sufficient for downstream activation without the need for CD4/CD8 co-stimulation, but TCR:pMHC interactions with low affinity (K_d > 3μM) necessitated CD4/CD8 co-receptor recognition [102, 121]. Structural data has demonstrated a ternary complex formed between the TCR:pMHC:co-receptor interaction **(Figure 1.3)** [124, 125].

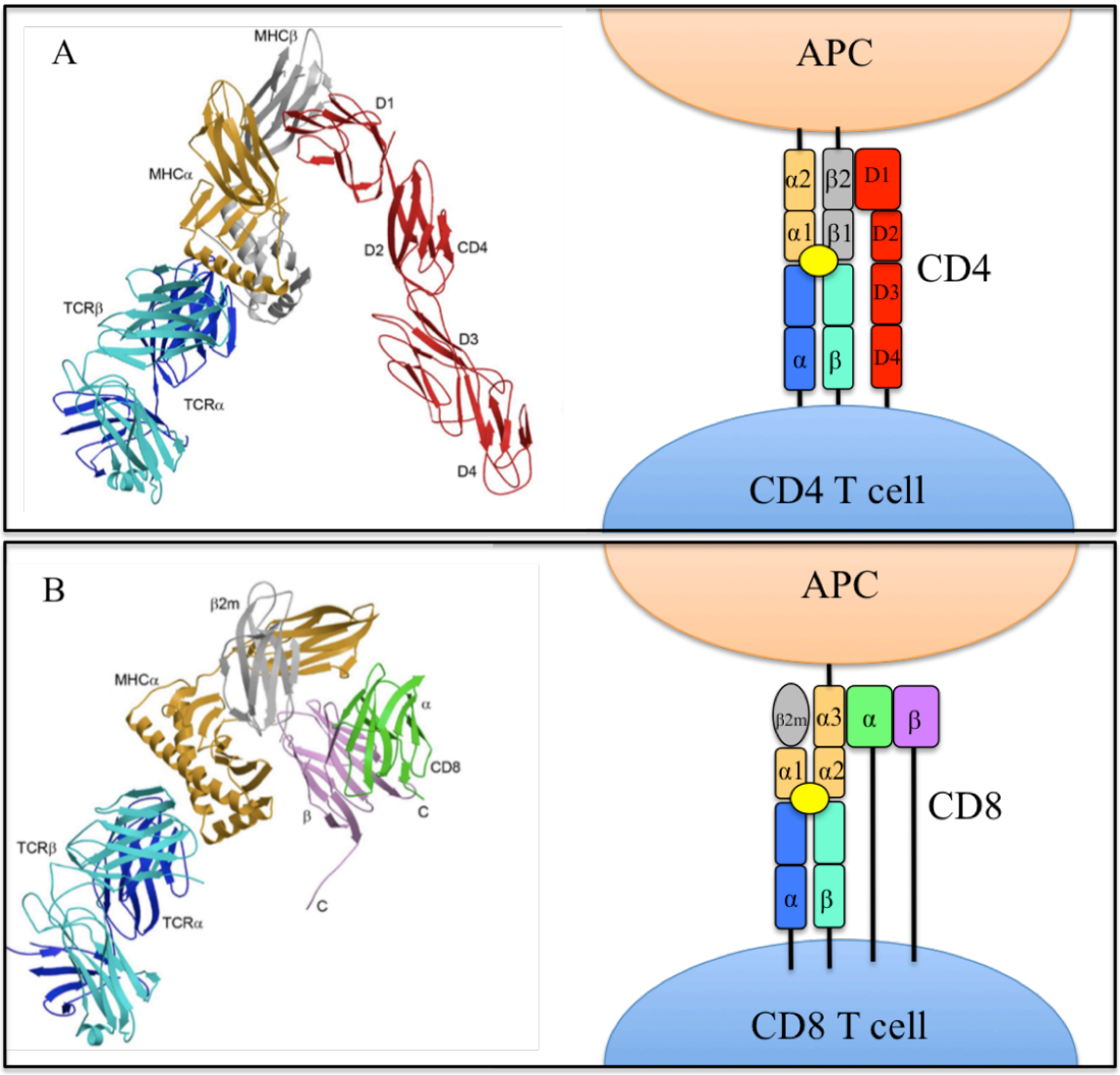

Figure 1.3: Ternary complexes of CD4 and CD8 co-receptors with respective TCR:MHC. **[A]** Crystal structure (left) and corresponding cartoon (right) of TCR:MHC with CD4 co-receptor oriented with CD4 T cell at the bottom and antigen presenting cell (APC) at the top. The domain 1 (D1) of the CD4 co-receptor contacts the β2 subunit of the MHCII complex. **[B]** Crystal structure (left) and corresponding cartoon (right) of TCR:MHC with CD8 co-receptor oriented with CD8 T cell at the bottom and antigen presenting cell (APC) at the top. The α domain (α) of the CD4 co-receptor contacts the α3 subunit of the MHCI complex. Crystal structure images adapted from Li *et al.* 2013 [125].

These ternary complexes are formed by the MHC-dependent recruitment of CD4 or CD8 co-receptors to the immunological synapse formed between the T cell and the antigen-presenting cell [126]. In addition to enhancing TCR:pMHC binding kinetics, the CD4 and CD8 co-receptors are also believed to play a role in the kinetic proofreading model discussed above [99]. Surface plasmon resonance (SPR) studies have shown that the half-lives of co-receptor:MHC interactions to be <35ms for both CD4 and CD8 [123, 127]. In contrast to this brief interaction, the TCR:pMHC interaction can last >10,000ms [128], therefore indicating the possibility of CD4/CD8 co-receptor disengagement 1000X per antigen recognition. Furthermore, it has been demonstrated that while CD8 has been found to stabilize TCR:pMHC binding [129], CD4 does not provide significant stabilization to the TCR with peptide presented on MHC II molecules [130]. It has therefore been proposed that the primary purpose of the CD4 and CD8 co-receptors is to enhance intracellular signaling rather than stabilize the TCR:pMHC interaction [131].

In addition to their role as binders to tighten the key interaction between TCR and pMHC, it has also been demonstrated that the CD4/CD8 co-receptors participate in signal transduction [132-134]. Seminal work by the groups of Schlossman [135] and Bolen [136] revealed that the CD4 and CD8 proteins were both associated with intracellular protein tyrosine kinase (PTK) signaling activity. Specifically, CD4 and CD8 co-receptors augment TCR signaling by increasing the efficiency by which recruitment of the proto-oncogene tyrosine-protein (Src) kinase lymphocyte-specific protein tyrosine kinase (Lck) is brought to the intracellular TCR domains. The likelihood that free Lck will associate

with the TCR signaling complex is very small compared to the corresponding likelihood for Lck bound to the CD4/CD8 co-receptor [131].

The T cell receptor complex includes a dimeric TCR and the CD3 signaling complex

In the early 1980s, monoclonal antibodies were generated against antigen-specific T cell clones in the hopes of identifying the antigen receptor on T cells [19, 137]. These antibodies not only were generated against the dimeric αβ TCR, but also against the nonpolymorphic CD3 proteins [138]. While determination of the antigen-binding function of the αβ TCR was obvious early on, the role of the CD3 proteins was less obvious. Initial studies went into analyzing the stoichiometry between the TCR proteins and the multiple CD3 γ, δ, ε, and ζ proteins. It was discovered that the CD3 proteins exist as a series of hetero- and homo-dimers consisting of γε, δε, and ζζ all in association with a single αβ TCR complex [138].

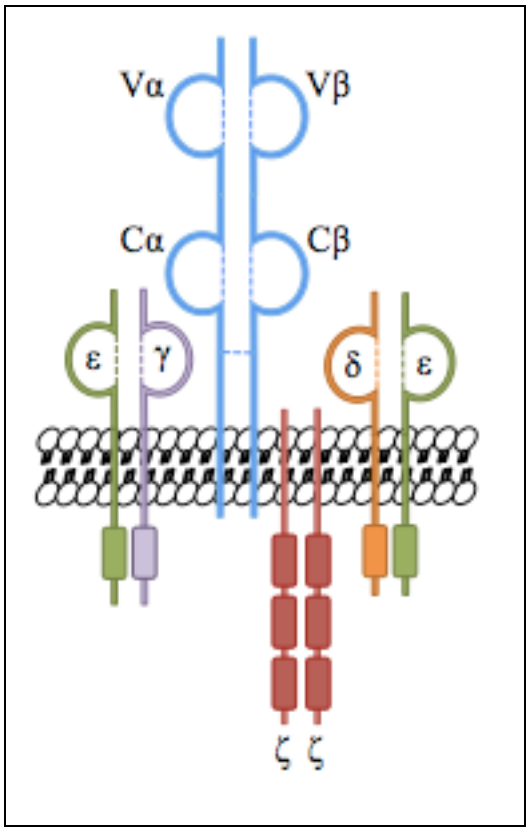

Figure 1.4: The mature αβ T cell receptor complex includes the TCR and the CD3 signaling complex. The αβ T cell receptor is a dimeric structure composed of two type-I glycoprotein chains linked together by disulfide bridges (dashed lines). Each chain is comprised of two extracellular domains (variable and constant), a hydrophobic transmembrane region and a very short intracytoplasmic region. Due to the short nature and lack of any intracellular protein docking sites or signaling motifs, a multimeric complex referred to as CD3 mediates signal transduction following antigen recognition. The three CD3 dimers (εγ, δε and ζζ) contain acidic transmembrane residues that interact with basic residues on the transmembrane portion of the TCR chains. Signal transduction is mediated by intracellular tyrosine activation motifs (ITAMs) contained on the intracellular portions of the CD3 subunits. The γ, δ and ε subunits each carry one ITAM, while the ζ subunits contain three. Once phosphorylated upon TCR activation, the ITAMs will recruit several tyrosine kinases such as ZAP70, triggering a signaling cascade leading to T cell activation and/or effector function. Image adapted from Schoettle *et al*. 2015 [139].

Several lines of evidence pointed towards the CD3 complex of proteins involvement in signal transduction: unlike the αβ TCR proteins that have extremely short intracellular domains the CD3 proteins had long cytoplasmic tails, anti-CD3 antibodies resulted in T cell activation, and experiments using cell lines expressing αβ without CD3 or CD3 without αβ revealed the obligatory co-expression of the two protein complexes [140]. Further studies revealed that engagement of the TCR and/or CD3 led to increases in intracellular Ca^{2+} concentrations both released from intracellular pools released from the endoplasmic reticulum (ER) stores in response to inositol triphosphate (IP_3) and from influx of Ca^{2+} from outside of the cell replenishing the ER stores [141]. Additional studies revealed that hydrolysis of the membrane lipid phosphatidyl inositol 4,5 bisphosphate by the enzyme phospholipase C (PLC) generated both IP_3 and diacylglycerol (DAG), an important regulator of protein kinase C (PKC) [142], leading the researchers to believe that the TCR functioned through regulation of PLC activity [98].

A key finding in the unraveling of TCR signal transduction was the knowledge that T cell activation required PTK function [143], which led researchers to discover TCR recruitment of cytosolic PTKs to activate a secondary messenger system. This finding was brought forth by the discovery that upon stimulation of the TCR, inducible phosphorylation of the ζ chain of the CD3 complex was observed leading to PLCγ function, and the same phosphorylation and PLCγ function could be induced by PTK [144]. Concurrent with these findings, the discovery that cytosolic PTKs lck and fyn

associated with both the TCR and CD4/CD8 co-receptors were being described [136, 145], providing more evidence for the requirement of PTK function in TCR signaling.

The pathway of CD3 signal transduction is initiated by the phosphorylation of tyrosines within the cytoplasmic tails of the CD3 proteins. Designated immunoreceptor tyrosine-based activation motifs (ITAMs), each CD3 ε, δ, and γ tail contains a single tyrosine motif, and each ζ tail contains three [146]. Numerous experiments were conducted creating chimeric cell surface proteins expressing the intracellular ITAM domains that all revealed ITAM tyrosine phosphorylation upon receptor ligation [147], indicating tyrosine phosphorylation of ITAMs as a requirement for TCR signaling and T cell activation. Following phosphorylation, the ITAMs were then discovered to serve as a docking site for the 70 kDa syk kinase family protein ζ-association protein (ZAP-70) [148]. Once recruited, ZAP-70 functions as an active PTK, phosphorylating numerous downstream substrates leading to activation and proliferation of the T cell.

Following TCR:pMHC ligation and ITAM phosphorylation by the PTKs lck and fyn, recruitment of ZAP-70 leads to phosphorylation of two key targets: the transmembrane adapter protein linker for the activation of T cells (LAT), and the 76-kDa cytosolic adapter protein src homology 2 (SH2) domain-containing leukocyte phosphoprotein (SLP-76) [149, 150]. Knockout experiments demonstrated that the loss of either LAT or SLP-76 results in near complete loss of TCR signal transduction [151]. It was determined that these two chaperone proteins form a docking site complex, allowing for correct spatial arrangement and timing of multiple downstream signaling pathways.

Upon engagement with the TCR complex, nine tyrosine residues on LAT will be phosphorylated. These tyrosines act as a docking site for PLCγ. Following PLCγ recruitment, SLP-76 is recruited to LAT via a mutual chaperone protein [152]. SLP-76 contains three functional domains involved with the interaction between LAT and PLCγ and other downstream adapter proteins and kinases [153]. LAT, SLP-76, and the other effector proteins making up this initial signaling complex all work in conjunction to recruit each other to the complex, stabilize the complex upon recruitment, and enhance the guanine nucleotide exchange factor (GEF) Vav1 phosphorylation of SLP-76 [154, 155]. The result of these multiple interactions in forming the proximal signaling complex is believed to be required for PLCγ stabilization and thus optimal activity of PLCγ-dependent pathways including Ca^{2+} and DAG mediated responses [156].

Once activated, PLCγ hydrolyzes the membrane lipid phosphatidylinositol 4,5 bisphosphate (PIP_2), releasing the secondary messengers DAG and IP_3 [157]. The PLCγ production of DAG results in activation of two separate pathways, Ras and protein kinase C theta (PKCΘ). Activation of Ras, a guanine nucleotide-binding protein, activates the serine-threonine kinase Raf-1 that activates the mitogen-associated protein kinases (MAPKs) extracellular signal-regulated kinase 1 (Erk1) and Erk2. Following Erk activation, activation of signal transducer and activator of transcription 3 (STAT3) that promotes T cell survival [158]. Ras is additionally involved with activation of the activator protein-1 (AP-1) c-Jun/c-Fos transcription factor complex leading to upregulation of CD69 involved in proliferation [159]. In addition to the Ras pathway, DAG also mediates PKCθ induced activation of nuclear factor kappa-light-chain-

enhancer of activated B cells (NFκB). Briefly, upon T cell activation, the inhibitor of NFκB (IκB) is phosphorylated, ubiquitinylated, and degraded, which allows NFκB to translocate to the nucleus and up regulate transcription of multiple genes involved in survival and effector function of the T cell [160].

The second PIP_2 cleavage product IP_3 activates Ca^{2+}-permeable ion channel receptors (IP_3R) on the ER membrane, resulting in the release of ER stores of Ca^{2+} into the cytosol. The loss of ER-Ca^{2+} leads to an influx of extracellular Ca^{2+} through plasma membrane Ca^{2+} release-activated Ca^{2+} (CRAC) channels via a process referred to as store-operated Ca^{2+} entry (SOCE) [161]. Ca^{2+} ions are somewhat ubiquitous secondary messengers in most eukaryotic cells, however TCR-induced increases in intracellular Ca^{2+} concentration result in the activation of Ca^{2+} and calmodulin-dependent transcription factors including myocyte-enhancing factor 2 (MEF2), downstream regulatory element antagonist modulator (DREAM), and the phosphatase calcineurin. Following activation, calcineurin dephosphorylates the transcription factor nuclear factor of activated T cells (NFAT) allowing it to translocate to the nucleus and increase transcription of the T cell growth factor interleukin-2 (IL-2)[162].

In addition to the activation of transcription factors resulting in the up regulation of growth, proliferation, and survival factors, T cell activation also results in programmed actin cytoskeletal rearrangements [163]. Upon T cell activation, filamentous actin (f-actin) will accumulate at the T cell:APC interaction surface, reorienting the T cell in a more conducible alignment for continued interaction with the activating APC [164]. Additionally, upon activation the T cell will become polarized, where the microtubular

organizing center (mTOC) will reorient towards the TCR:pMHC interaction, an essential step in creating the organized structure between the T cell and the APC known as the immunological synapse [165]. Finally, activation of T cells results in activation of integrins, αβ heterodimeric receptors that mediate cell-cell interactions. The two key integrins activated upon T cell activation are the leukocyte function-associated antigen-1 (LFA-1) and very late antigen-4 (VLA-4). These two integrins will bind their respective ligands intracellular adhesion molecule (ICAM) and vascular cell adhesion molecule (VCAM) on APCs, infected cells, endothelial cells and extracellular matrix proteins [166].

Co-stimulation of the T cell co-receptor CD28 provides additional signals required for T cell activation

When T cells are activated solely through the TCR:pMHC and CD3 complex, they often enter a state of anergy or non-responsiveness [97]. In order to avoid the anergic state and reach a productive activation, co-ligation of other cell surface receptors provides additional signaling input. While there are many other co-receptors that can enhance signaling, the most robust of these is CD28 [98].

CD28 signaling has been shown to enhance cell proliferation, cytokine production, cell survival, and increased cell metabolism [97]. The key signaling intermediate downstream of T cell CD28 ligation with its respective ligands CD80/CD86 on APCs is phosphoinositide 3-kinase (PI3K). Following ligation, PI3K converts PIP_2 to phosphatidylinositol (3,4,5) triphosphate (PIP_3) at the cell membrane which then serves

as a docking site for the protein kinase Akt [167]. The downstream protein phosphorylation by Akt enhances numerous pathways including NFκB survival signaling [97], NFAT regulation of IL-2 production [168], and increased cell surface express of the insulin transporter Glut1 thus allowing for increased T cell metabolism [169].

T cells are derived from bone marrow resident hematopoietic stem cells but migrate to the thymus for development

In adults, each class of blood cells is derived from bone marrow (BM) resident hematopoietic stem cells (HSCs). The different classes emerge as the result of progressive losses in differentiation potential for the other cell lineages [170]. Modern experimental techniques have made it relatively simple to isolate HSCs and create model systems both *in vitro* and *in vivo* for studying blood cell differentiation leading to cell fate and lineage commitment [171]. The first major bifurcation in the lineage commitment process of HSCs is into either the lymphoid or myeloid lineage. The lymphoid lineage gives rise to the adaptive T and B cells, as well as natural killer (NK) cells, and the myeloid lineage comprises platelet producing megakaryocytes, erythrocytes and innate cell types such as monocytes, macrophages, and granulocytes. Additionally, while also of hematopoietic origin and commonly associated with the myeloid lineage, dendritic cell lymphoid/myeloid lineage affiliation is somewhat controversial [172].

One of the first major changes HSCs undergo during lineage commitment is the loss of self-renewal capability [173] at which point the cells may be classified as multipotent progenitors (MPPs). Following progression to MPP status, the cells will then

diverge into either common lymphoid progenitor (CLP) cells or common myeloid progenitor (CMP) cells, with potential for restriction to all cell types within the respective common lineages [174, 175]. Cell surface expression of different proteins allows for the identification and isolation of either subset of common progenitor types; CLPs can be identified as IL-7Rα^{+}Thy-1.1^{-}Lin^{-}Sca-1^{lo}c-Kit^{lo} [174] and CMPs as CD34^{+}FcγRIII$^{-/lo}$Thy-1.1^{-}IL-7Rα^{-}Lin^{-}Sca-1^{-}c-Kit^{+} [175]. This initial discovery of a common lymphoid progenitor suggested a common origin for all lymphocytes (B and T cells), instead of separate pathways of development directly from the HSC stage.

There have been numerous laboratories elucidating the key signal/expression changes on MPP cells that direct cell differentiation to either the CMP or CLP restriction. A key finding was that MPPs could be separated based on Flt3 expression levels [176], and these MPP subsets could be further subdivided into three main subcategories: Flt3^{lo}VCAM-1^{+}, Flt3^{hi}VCAM-1^{+}, and Flt3^{hi}VCAM-1^{-} [177]. *in vivo* characterization of these three subsets demonstrated that only the most primitive Flt3^{lo}VCAM-1^{+} subset of MPPs could give rise to CMPs (as well as CLPs), while the most developmentally advanced Flt3^{hi}VCAM-1^{-} subset were more restricted and preferentially give rise to CLPs only [178].

Upon becoming CLPs, commitment to the lymphoid lineage has been established as the cells have lost all ability to generate myeloid lineage cells. It was observed that injection of CLPs intrathymically resulted in CLP differentiation into T cells [179], however intravenous injection of CLPs resulted in B cell differentiation [174]. It was hypothesized that upon lineage specification, progenitors would preferentially migrate to

distinct loci either within the BM or the thymus that only support differentiation into B cells or T cells, respectively. Known as microenvironments, these anatomical niches are highly specialized structures where specific cellular components and stem/progenitor cells localize to aid in cell differentiation. HSCs utilize a similar model in the bone marrow where osteoblasts at the trebacular region of the bone cavity express the adhesion molecules N-cadherin and VCAM-1 to retain the HSCs in their niche [180]. T and B cell niches in the thymus and BM have also been investigated. As T cells undergo their consecutive stages of development they will migrate to different regions of the thymus [181] and developing B cells will traverse from niche to niche in the BM as they progress through their developmental stages as well [182]. The main homing signals directing cells to the various niches for each stage of their development are chemokines [183]; SDF1α, CCL19, CCL21 and CCL25 have all been shown to aid in the homing of T cells during their development in the thymus [181].

While BM resident HSCs retain the ability to divide, immature thymocytes have lost self-renewal capability and therefore must be constantly replenished by progenitors from the BM to sustain adequate numbers of developing T cells [184]. The deterministic cue dictating whether bi-potential CLPs will enter T cell or B cell lineage commitment is through the signaling molecule Notch. Inhibition of Notch 1 signaling in the earliest developmental stages of thymocytes results in B cell development [185]. Additionally, constitutive expression of Notch 1 can induce T cell development of BM resident MPPs [186]. These findings suggest that the replenishment of developing T cells in the thymus comes from the migration of bi-potential MPPs from the BM.

The final step before T cell development involves the actual circulatory system migration of BM-derived progenitor cells to the thymus. In addition to the intrinsic signals dictating T cell lineage, lymphoid progenitors must be able to mobilize from their niches in the BM and home to specific regions of the thymus for continued lineage commitment and development. A small population of T cell generating MPPs has been found to express increased levels of CD62L. These cells have enhanced ability for thymic homing [187]. Additionally, increased expression of the chemokine receptor 9 (CCR9) has been shown on thymic homing progenitor cells [188]. CCR9 is the receptor for chemokine ligand 25 (CCL25), also known as thymic expressed chemokine (TECK), which is expressed in the thymus but not the BM. Together with NOTCH expression, CD62L and CCR9 expression of BM derived progenitor cells leads to thymic migration and T cell lineage commitment.

T cell development occurs throughout multiple stages in the thymus

Once progenitor cells reach the thymus, T cell development proceeds in a highly regulated, well-defined, sequential manner. As the developing cells progress through developmental stages, cell differentiation becomes more and more restricted [189].

Premature cells enter the thymus at the cortico-medullary junction and proceed to migrate to the outer cortex [190]. Developing cells at this stage express neither TCRs, CD3 signaling complexes, nor CD4/CD8 co-receptors and as such are commonly referred to as “Double negative” (DN) cells. Two of the first surface proteins to be expressed during this developmental stage are the CD44 molecule and the IL-2α receptor (IL-2Rα).

During this stage, increased transcription of the recombinase activating genes 1 and 2 (RAG1 and RAG2) leads to somatic rearrangement of the TCR genes (discussed below). At the double negative stage, only the TCRβ chain genes are rearranged, and only functional rearrangements result in β chains capable of binding the surrogate α chain pTα, the prerequisite for progression to the next developmental stage [191]. Cells not expressing functional β chains will either attempt another round of β chain gene rearrangement or undergo apoptosis. Cells that are capable of expressing functional β chains will successfully join their β chain with the pTα chain to form the pre-TCR that will be expressed on the cell surface along with the CD3 signaling complex for testing of functionality. It should also be mentioned that during the double negative stage of development T cells would also rearrange the gene sets encoding the γδ TCR. This will be discussed in more depth below. Following a successful β chain rearrangement, the cell will proliferate, producing hundreds of clones and the process will repeat in the proliferated cells by rearranging the α chain genes as they attempt to form a mature TCR. The random rearrangement and selection of TCRα and β genes (along with the sloppy joining of the gene segments) is the main mechanism of TCR diversity development which will again be discussed in more depth below [192].

As the cell progresses in development, it will finish rearranging the TCRα genes and upregulate expression of both CD4 and CD8 co-receptors, creating “Double positive” (DP) T cells. The double positive developmental stage again takes place in the thymic cortex and accounts for over 70% of thymocytes [189]. Continued TCRα gene

rearrangement and eventual expression of the mature TCR defines this stage of T cell development. Upon successful α chain rearrangement, expression of the complete αβ TCR will be followed by a series of selective tests. In order to ensure that mature T cells can recognize self-MHC presentation of antigenic peptide, the interactions of the newly expressed TCR will be tested during "Positive selection". Cortical thymic epithelial cells (cTECs) expressing numerous MHC will be utilized to assess TCR:MHC interaction, failure of which will result in the T cell entering an unresponsive state known as anergy, or the cell will die from neglect [193]. Our current understating of this positive selection process involves both the strength of the TCR:MHC interaction (affinity), as well as the number of actual TCR:MHC interactions occurring (avidity). This selection process ensures self-MHC restriction of T cells and will result in final CD4 vs CD8 lineage commitment. While the deciding factor on whether to become CD4 or CD8 ultimately depends on which MHC class (I or II) is recognized, the process of lineage commitment is much more intricate than simple selective kinetics. Initially, DP cells will down regulate expression of both CD4 and CD8 co-receptors and then re-express CD4 only. If these $CD4^{hi}CD8^{lo}$ cells make a productive interaction with MHC II they will be committed to the CD4 lineage and completely down regulate expression of CD8. However, if no successful interaction is made with MHC II, the cells will down regulate expression of CD4 and up regulate expression of CD8. These $CD4^{lo}CD8^{hi}$ cells will then attempt interaction with MHC I molecules and should an interaction of sufficient affinity occur the cell will commit to the CD8 lineage and fully down regulate CD4 expression

[194]. Should no sufficient interactions occur between either MHC II or MHC I, the cell will die from neglect or enter a state of non-responsive anergy.

As the developing cells progress from the DP stage to the single positive (SP) stage, expressing either the CD4 or CD8 co-receptor, they will undergo a final selective process. “Negative selection” is the process by which self-reactive T cells are deleted from the developing repertoire. It involves the interaction between T cells and self-peptide presented by medullary thymic epithelial cells (mTECs), and thus takes place in the thymic medulla [195]. Uniquely, mTECs are capable of expressing a majority of tissue-specific proteins in the thymic medulla, which allows them to present most of the body’s protein to developing T cells. To accomplish this feat, mTECs express the master transcriptional autoimmune regulator (AIRE) transcription factor. AIRE promotes the expression of tissue-specific proteins in the thymic medulla and thus allows for self-reactive T cells whose TCR recognizes the body’s tissue-specific proteins to be deleted [196]. Depending on the strength of interaction, there are ultimately three potential fates for cells following negative selection. Should strong TCR recognition of self-peptide:MHC occur, the developing T cell may be differentiated into a natural regulatory T cell (T_{reg}) [197]. As these cells express a high affinity for self-peptide, they are excellent at ensuring that peripheral autoimmune responses by lower affinity conventional T cells are suppressed. If a moderate level of self-peptide:MHC recognition occurs the cell will undergo apoptosis and die. Only if relatively weak or a lack of recognition occurs will the developmental T cell survive to maturity and thus be allowed to exit to the periphery [197]. While the process of negative selection is formulated to

ensure a lack of autoreactive T cells in the immune repertoire, many autoreactive T cells will escape this selective process and exit the thymus. Peripheral tolerance mechanisms such as regulatory T cells and suppressive cytokines help to prevent autoimmunity from occurring **(Figure 1.5)** [198].

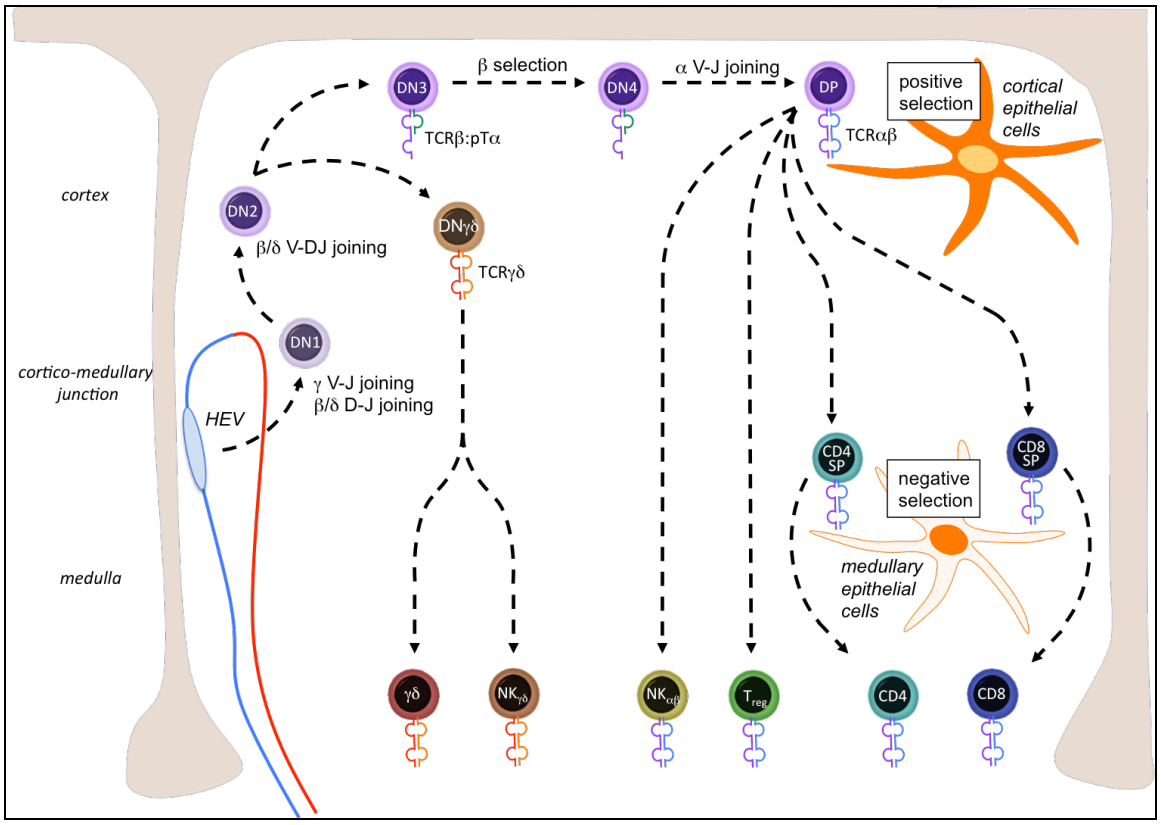

Figure 1.5: T cell development is a highly ordered process of events occurring in distinct stages and locations within the thymus. Common lymphoid progenitor cells travel to the thymus and enter through the high endothelial venule (HEV). Now double negative 1 (DN1) cells, these early progenitors begin β-, γ-, and δ-chain gene rearrangements. Reaching DN2, the cells migrate towards the cortex and complete β-, γ- and δ-chain rearrangements. Should successful γδ rearrangement occur before successful β-chain, the cells will express γδ TCRs and exit the thymus as mature γδ T cells, forgoing any further selective processes. If productive β-chain rearrangement occurs first, the β-chain will be paired with a surrogate α-chain (pTα) as the cell moves into the DN3. If a productive β- chain rearrangement occurred, the cell will enter DN4 stage and proliferate. In the early double positive (DP) stage, the cells will undergo α-chain rearrangement. Upon successful α-chain recombination, the cell will shuttle its αβ TCR to the cell surface, up regulate both CD4 and CD8 co-receptors and undergo positive selection mediated by the cortical epithelial cells. Following positive selection, the cell will become either a single positive (SP) CD8 or CD4 T cell. At the SP stage, the cell migrates back towards the cortico-medullary junction and undergoes the final process of negative selection carried out by medullary epithelial cells. Following negative selection, the newly matured naïve T cell will leave the thymus back through the HEV and exit to the periphery in search of its cognate antigen. Figure adapted from Schoettle *et al.* 2015 [139].

Overall, the positive and negative selective processes are extremely stringent and the vast majority of developing T cells will never reach maturity. It has been estimated that as low as 1-2% of progenitor cells that enter the thymus will ever reach complete T cell maturity [199]. Less dense than the cortex, the thymic medulla acts as a reservoir of newly matured T cells that will remain for roughly two weeks in the thymus before exiting to the peripheral circulation as they undergo approximately six rounds of cell division [200].

Organization of the T cell receptor genetic loci

One of the hallmarks of adaptive immunity is the high degree of specificity of the antigen receptors found on B and T cells. In order to maintain sufficient diversity of receptors allowing for identification of a multitude of pathogens and transformed cells encountered in an organism's lifetime, a repertoire of highly diverse TCRs (and BCRs) must be generated. The vast majority of T cells express heterodimeric antigen receptors composed of both an α and β chain (αβTCR) while a small subclass of T cells express a γ and δ chain (γδTCR). The ability of T cells to generate such a large diversity of antigen receptors is dependent upon recombination of variable (V), diversity (D), and joining (J) segments for TCRβ and δ chains, and V and J segments for TCRα and γ. The most variable parts of the TCR, commonly referred to as the complementarity determining regions (CDRs), make the major contact with the p:MHC complex and thus constitute TCR specificity [201] and will be discussed below.

The human germline TCRβ locus on chromosome 7q34 contains 64 V, two D, 14 J, and two constant gene segments; similarly the TCRα locus on chromosome 14q11 has 44 V, 61 J, and one constant gene segment (relative numbers for the murine TCR gene locus are only slightly lower) [202]. Interestingly, while the loci for the α, β, and γ chain gene segments are located on different chromosomes or discreet from one another on the same chromosome (TCRγ is found discreet from TCRβ on chromosome 7q34), the δ chain locus is interspersed entirely within the α chain locus on chromosome 14q11 **(Figure 1.6)** [202].

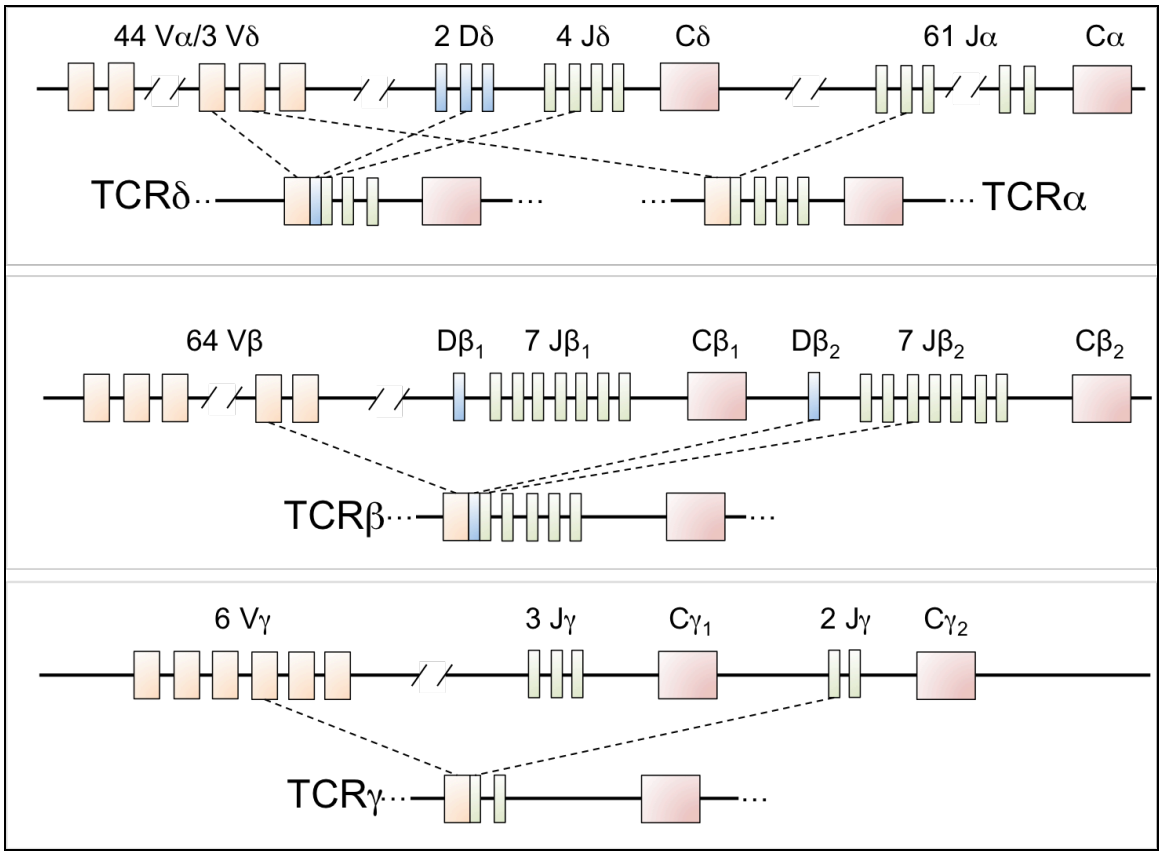

Figure 1.6: Somatic recombination of semi-randomly selected germline encoded TCR gene segments is required to form the variable regions of TCRs. The α- and γ-chains are composed of only V and J segments and therefore require only a single V→J rearrangement. The β- and δ-chain variable regions are composed of V, D and J segments, the first rearrangement joins a D→J, and the second rearrangement joins a V→DJ. While the β (middle) and γ (bottom) genes are located discreetly on chromosome 7, the δ segments are interspersed within the α locus (top) on chromosome 14. The multiplicity of V, D, and J gene segments allows successive rearrangement events to occur if an unproductive rearrangement leads to an inadequate receptor chain. This process may continue until either a productive rearrangement occurs of the supply of gene segments is exhausted. Figure adapted from Schoettle *et al.* 2015 [139].

During the lineage commitment process of T cell development, simultaneous reorganization of the TCRβ, γ, and δ loci will occur prior to TCRα reorganization. Signaling from the productive reorganization the TCRβ locus producing a TCRβ chain paired with the pre-TCRα chaperone protein will halt further reorganization and allelic exclusion of the γ and δ loci, and follow with TCRα gene rearrangement leading to αβ T cell lineage commitment [203]. Conversely, should productive rearrangement and signaling via the γδTCR occur prior to β chain rearrangement, allelic exclusion of the TCRβ and α loci will occur as the cells progress to γδ T cell commitment. The nature of this process may account for the quantitative bias towards commitment of T cells to the αβ lineage rather than γδ lineage as only successful rearrangement of a single loci (TCRβ) needs to be accomplished for αβ commitment whereas both the γ and δ chain loci need to be successfully rearranged for γδ commitment [204], however subsequent positive and negative selective processes αβ T cells are required to pass may negate overall numbers as γδ T cells are not subjected to these restrictions.

The key enzymes involved in the somatic recombination of the TCR gene loci are the recombination activation genes 1 and 2 (RAG1/2). The temporal control of expression of these genes determines the order by which the TCR alleles and loci are rearranged. The RAG1/2 complex is responsible for joining different gene segments together and mediating the splicing and rejoining events that lead to gene recombination [205]. It was recently demonstrated that RAG1/2 recruitment sites are most likely based on differential histone methylation of pairing TCR loci, ensuring the maintenance of tight

control of recombination only between loci on the same chromosome and also possibly for allelic exclusion [206], however it remains unclear what signals might mediate such differential methylation. Following recruitment, RAG1/2 recognition of specific palindromic recombination signal sequences (RSSs) ensures correct gene segment selection within each chromosome. These RSSs will employ either a 12mer or 23mer spacer region (corresponding to one and two turns of the DNA double helix, respectively) and will be flanked by semi-conserved heptamer and nonamer sequences preceding and/or following each V, D, and J gene segment [207]. The 12 and 23 nucleotide (nt) spacers may bind to the RAG1/2 complex differently which allows for pairing of 12mer RSSs with 23mer RSSs. Gene segments flanked by a 12mer RSS will only recombine with segments flanked by a 23mer RSS [208]. More importantly, for TCRα chain recombination this '12/23 rule' ensures that V segments (flanked by 23mer RSSs on the 3' side) recombine to J segments (flanked by 12mer RSSs on the 5' side) and not with other V segments, and for TCRβ chain recombination D segments (flanked by 23mer RSSs on the 3' side) recombine with J segments (flanked by 12mer RSSs on the 5' side) and V segments (flanked by 23mer RSSs on the 3' side) recombine to D segments (flanked by 12mer RSSs on the 5' side). The temporal organization of these recombination events is coordinated to the developmental stage of the precursor T cell [208] and will be discussed in detail below **(Figure 1.7)**.

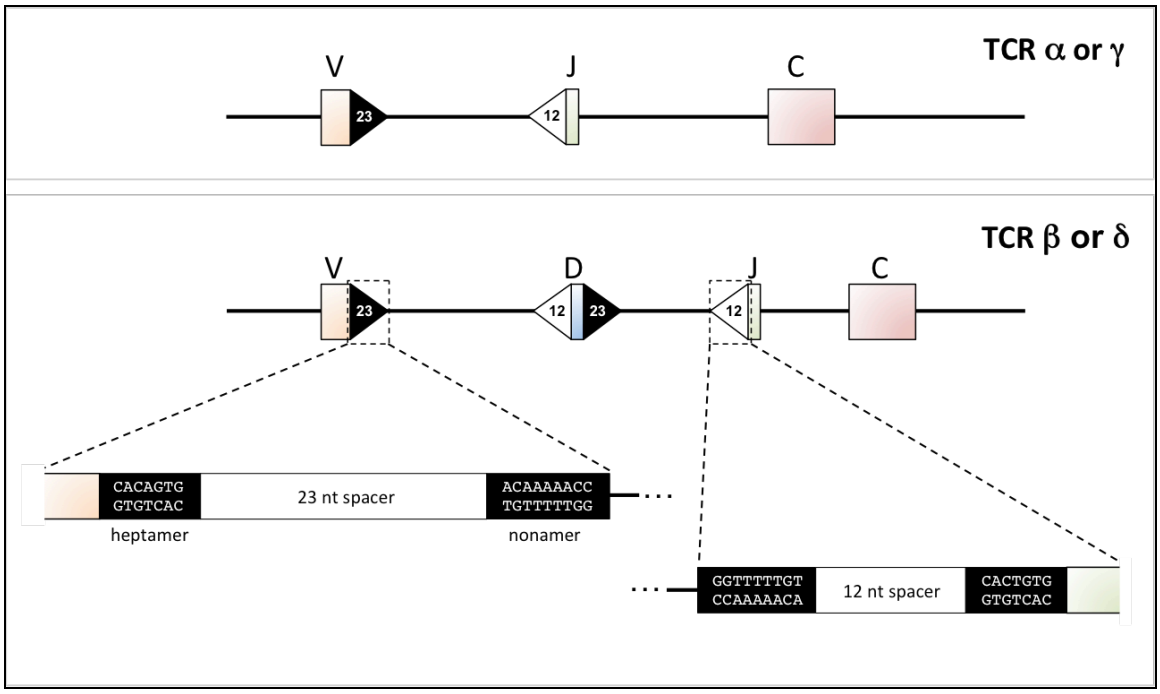

Figure 1.7: Recombination signal sequences (RSSs) ensure correct V(D)J recombination. The recombination of V, D, and J gene segments is directed by sequences referred to as recombination signal sequences (RSS), which flank the 3' side of V segments, both sides of D segments and the 5' side of J segments. There are two types of RSSs, one containing a 12nt spacer and one with a 23nt spacer. Recombination of VDJ genes can only occur between genes flanked by different RSSs ("12/23 rule"). Both RSSs are comprised of their respective 12 or 23nt spacer flanked on either side by a 7nt heptamer and 9nt nonamer. In addition to their function of providing recognition sites for enzymes to cut and rejoin the DNA, RSSs also ensure that correct joining of gene segments occurs: V→J for TCRα/γ (top), and D→J followed by V→DJ for TCRβ/δ (bottom). Figure adapted from Schoettle *et al.* 2015 [139].

One important consequence of this process relies on the inherent diploid organization of the human (and mouse) T cell genome. As diploid cells, T cells possess two alleles for each TCR loci. However, the majority of T cells in an individual express only a single TCRβ chain on each cell. While allelic exclusion of the alternate TCRγ/δ loci is most likely RAG1/2 expression mediated, it is currently unknown whether any additional signaling is required or if it is a stochastic process. Should an unproductive rearrangement at one TCRβ allele occur, there is an equal probability for the second allele to undergo rearrangement or for the first allele to rearrange again [209]. However, upon successful β chain rearrangement, pairing with the surrogate pTα chain, and intracellular signaling, down regulation of RAG1/2 expression will terminate rearrangement of the γ, δ, and other β allele [203]. While the vast majority of mature T cells express a single β chain, current estimates suggest that a significant fraction of T cells (30%) will express two distinct α chains [210, 211]. This is believed to be due to the fact that TCRα rearrangement may continue for some time after the first successful α chain rearrangement. Unlike the events following TCRβ rearrangement, where immediate signaling halts further loci rearrangements, upon successful TCRα rearrangement arrest of successive rearrangements depends on signaling during positive selection (MHC restriction) and CD4 vs CD8 lineage commitment which allows sufficient time for additional TCRα rearrangements on the other allele [210]. However, while many cells may express two distinct TCRα chains, only one of the two is likely to recognize antigen presented by self-MHC when paired with the TCRβ chain (thus

passing positive selection), and therefore even cells expressing two TCRα chains will be specific to a single antigenic epitope although it has been postulated that these cells may be responsible for some cases of allergy or autoimmunity [212]. Continued research into simultaneous quantification of both TCRα and TCRβ from individual cells will be required to rectify these potential discrepancies.

Somatic rearrangement of TCR genes occurs in precise steps corresponding to consecutive stages of T cell development

Following NOTCH signaling from bone marrow stromal cells, precursor T cells will upregulate CCR9 and migrate from the bone marrow to the thymus to complete their development [213]. Entering as double negative (DN) cells, these cells will progress through multiple stages of development (i.e. DN I, DN II, DN III, DN IV, early DP, late DP, single positive) during which a highly organized and regulated series of events will lead to TCR recombination. These events are both spatially and temporally regulated throughout the thymus, ensuring that only T cells with productively rearranged TCR that are not self-reactive are permitted to leave the thymus to the peripheral circulation **(Figure 1.5)**.

The process of somatic recombination occurs in both T cells and B cells creating TCRs and BCRs, respectively. These recombination events occur between semi-randomly selected gene segments (V and J or V, D, and J) encoded in the T or B cell's germline DNA. This highly regulated and intricate process involves the selection, splicing, and recombination of discreet gene segments from discrete sites on the same

chromosome. The variable domains of TCRα (and γ) result from the recombination of V and J segments, while the variable domains of TCRβ (and δ) result from the recombination of V, D and J segments. While only a single recombination event is necessary for alpha chain rearrangement (V→J), two recombination events are required for beta chain rearrangement (D→J proceeded by V→DJ). The recombination events between these gene segments are coordinated to the developmental stage of the precursor T cell. It is hypothesized that the specific sequences of the RSSs juxtaposed to each gene segment influence the order of the recombinations by differential binding to the RAG1/2 complex [208]. Some form of hierarchical RAG1/2 binding preference may help to ensure D→J recombination is always followed by V→DJ for the TCRβ chain, such that no V→J recombination events occur.

Following RAG1/2-mediated synaptic complex formation between two gene segments, the RAG1/2 complex will cut the strands by generating double stranded breaks (DSBs) in the DNA between the RSS and the coding portion of the V/D/J segment [209]. Initial RAG1/2 cleavage occurs between the coding sequence 5' end of the heptamer portion of the RSS and is followed by cleavage on the opposite strand, generating hairpins in the RAG-maintained coding strands and free ends for the noncoding strands containing the excised RSSs. For the recombination of two gene segments in the same 5'→3' orientation, the excised portion of DNA (referred to as an excision circle or signal sequence) will be bound by the Ku70/80 protein complex (also involved in nonhomologous end joining DNA repair mechanisms) and ligated together by DNA Ligase IV. For the joining of two sequences inverted to one another in reading frame, an

inversion of the DNA strand will occur, and the intervening non-coding sequence will be sequestered at a distal site in the genome [209].

Following RAG1/2 strand cleavage and hairpin formation the gene segments to be recombined will be held in close proximity to each other by the RAG complex, and the binding of Ku70/80 will then stabilize the hairpin structures. At this point junctional diversity is generated between the gene segments. Activation of the enzyme Artemis by the protein kinase DNA PKcs will allow Artemis to randomly cleave the coding sequence hairpins held by Ku70/80. This random cleavage creates palindromic (P-) nucleotide tails on the ends of the coding strands [214]. While discussed in further detail below, it should be mentioned that along with the random selection of each of the many possible gene segments to recombine, this represents one of the key steps in the generation of diversity in the TCR repertoire. After the random cleavage of the DNA strands by Artemis, the P-nucleotide tails will be aligned by XRCC4 and the Cernunnos proteins allowing for the recruitment of the next key enzyme terminal deoxynucleotide transferase (TdT) [215]. TdT mediates the next key step in TCR diversity generation by adding non-templated (N-) nucleotides to the 3' ends of the palindromic tails. Short stretches of complementarity between the P- and N-nucleotides most likely facilitates brief pairing of the two strands which allows DNA exonucleases and polymerases (Polλ and/or Polμ) to remove unpaired nucleotides and fill in any remaining gaps in between the DNA strands. Once the junction between the two gene segments is aligned and all regions have been complimentarily paired, recombination will be completed by DNA Ligase IV that seals the final nick between the two strands **(Figure 1.8)** [207].

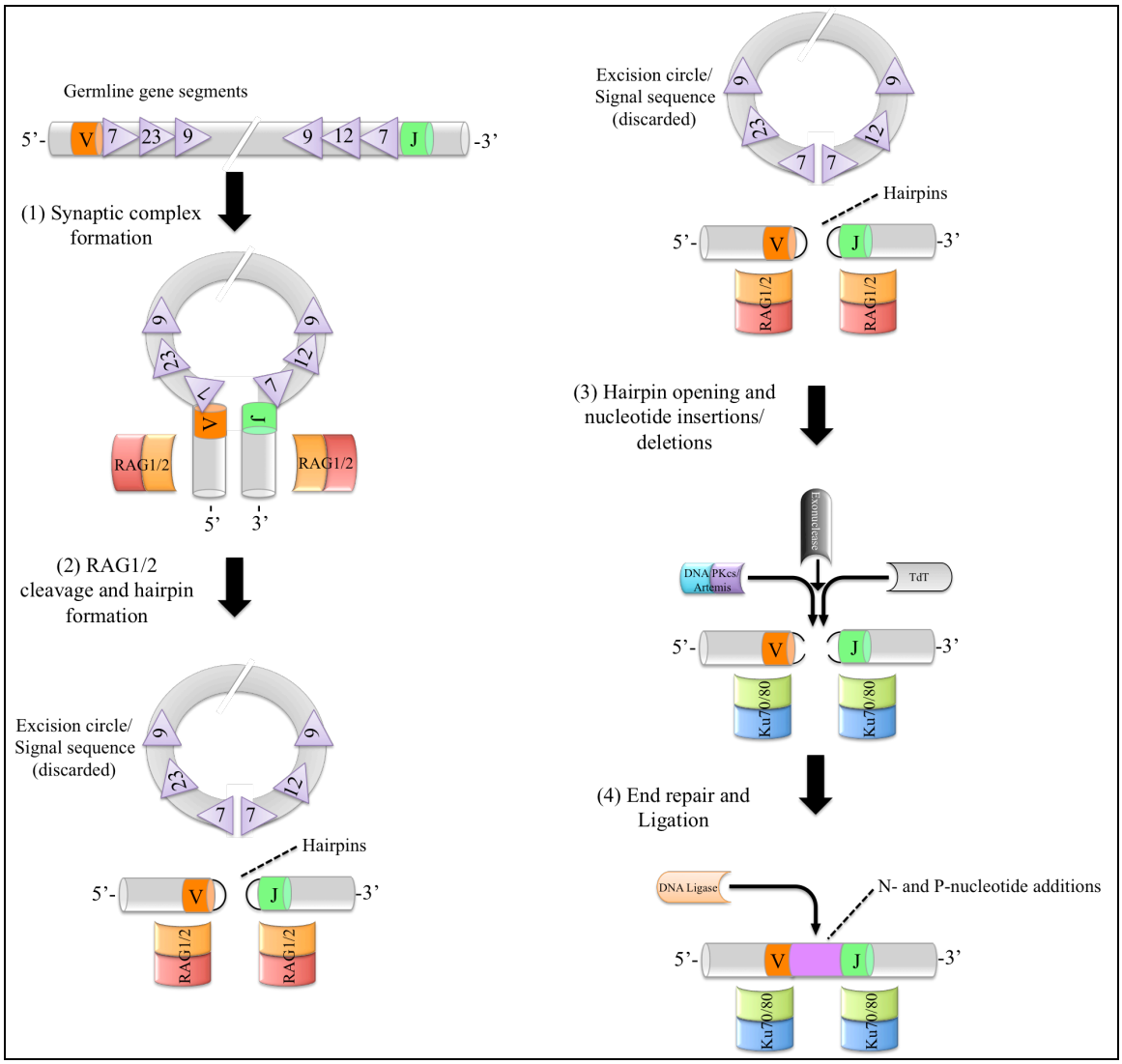

Figure 1.8: Somatic recombination of TCR gene segments. T cell receptor gene rearrangement of variable (V), diversity (D) and joining (J) gene segments generates an enormous repertoire of antigen receptors with different specificities. (1) RAG1/2-mediated synaptic complex formation occurs between semi-randomly selected gene segments (V and J depicted above) encoded in the T cell's germline DNA. (2) Cutting the strands by inducing DSBs in the DNA between the RSS sequences and the coding portion of the V/D/J genes, the RAG1/2 complex then generates hairpin loops in both stands. The excised portion of DNA, referred to as the excision circle or signal sequence, will be bound together by Ku70/80 and ligated by DNA Ligase (not shown). (3) Binding of Ku70/80 will stabilize the hairpin structures as DNA PKcs activates Artemis who will randomly cleave the hairpins creating P-nucleotide tails. TdT and exonucleases will then add/remove N-nucleotides to the 3' ends of the palindromic tails. (4) Following alignment and pairing of the two strands, DNA Ligase IV will seal the final nick.

While the events that occur during somatic recombination are under high levels of enzymatic regulation, the order of the recombination events is also under high levels of regulation throughout T cell development. The entire process leading to expression of a membrane-bound αβ (or γδ) TCR is preceded by the progressive expression of specific enzymes and/or cell-surface proteins that mediate each consecutive genetic rearrangement or cell proliferative event. As previously mentioned, precursor cells entering the thymus do so without expression of either CD4 or CD8 co-receptors. The first portion of this DN stage (DN I) involves the initial expression of the RAG1/2 proteins, resulting in initial Vβ and Vδ D→J rearrangements as well as Vγ V→J rearrangement. As the cells finish their initial TCRβ/γ/δ rearrangements the T cell precursors will advance to the DN II stage where they will complete TCRβ/δ recombination of V→DJ gene segments. During the DN III stage the cells will be analyzed for expression of either a γδ TCR or a pre-TCR consisting of a β chain paired with the surrogate pTα. Expression of a functional β:pTα pre-TCR results in pTα cross-linking between pre-TCR on the same cells that leads to CD3 signaling-mediated down regulation of RAG1/2 expression through the ETS1 transcriptional regulator thus temporarily halting additional recombination events. Once this has occurred, termination of expression of pTα via the ID3 transcriptional regulator will release the inhibition on RAG1/2 expression for downstream recombination events [216]. However, should successful γδ rearrangement precede β chain rearrangement and pre-TCR signaling, the γδ T cells will exit the thymus with no apparent positive or negative selection events [217].

Following successful DN III stage signaling through the β:pTα pre-TCR, the cells will begin to proliferate as they enter the DN IV stage. Sometimes referred to as 'β selection' [218], this event is a critical checkpoint during T cell development that allows for expansion of T cell precursors that have successfully rearranged their TCRβ chain [207]. This creates a pool of mono-clonal TCRβ expressing cells that may independently rearrange their TCRα chains creating unique TCRαβ combinations, giving rise to the final level of TCR diversity; αβ combinatorial diversity.

Following the proliferation of cells at the end of the DN IV stage, re-expression of the RAG1/2 proteins initiates TCRα V→J rearrangements. While enhancer-/promoter-dependent changes in chromatin structure are the mechanisms behind RAG accessibility to RSSs, it is still poorly understood how the delayed 'opening' of the TCRα locus until after TCRβ/γ/δ 'opening' and rearrangement [209]. Accompanying the completion of TCRα rearrangement, expression of both CD4 and CD8 co-receptors advances the precursor cells to the DP stage of T cell development. Upon reaching the early portion of the DP stage, the cells may express a complete αβ TCR and the process of positive selection ensuring self-MHC restriction ensues. Once a threshold TCR:MHC interaction has occurred, the cells will progress to the late DP stage and CD4/CD8 lineage selection will cause down regulation of one of the two co-receptors creating single positive (SP) cells. Upon reaching the SP stage, cells will undergo the final process of negative selection where autoreactive cells will be deleted from the developing repertoire. As mentioned previously, due to the rigorousness of the multiple genetic rearrangements,

positive selection, and negative selection, the vast majority of developing cells will not survive development, with only 1-2% ultimately exiting to the periphery [199].

Complementarity determining regions and the structural basis for T cell antigen recognition

The process of somatic recombination of V, D, and J gene segments produces variable regions on the distal portions of the TCRα and TCRβ chains and thus constitute the repertoire of T cell antigen receptors. These variable regions recognize antigenic peptide fragments presented on MHC molecules by infected, cancerous, or antigen presenting cells. Following somatic recombination, each protein variable region will contain three highly diverse loops, termed complementarity-determining regions (CDRs) 1, 2, and 3 that make direct contact with the peptide and/or MHC molecule [219]. The CDR1 and 2 loops are directly encoded within the randomly selected V gene segment portion of the TCRα or TCRβ chain, however the CDR3 loops are encoded by a region spanning the 3' end of the V gene segment, the D gene segment (β chain only), the 5' end of the J gene segment, and the random P- and N- nucleotide junctions between them [220]. Therefore, compared to the CDR1 and 2 loops, CDR3 loops are significantly more diverse due to the nucleotide additions and losses between the junctions of the V-J (α chain), and V-D-J (β chain) gene segments. Furthermore, in addition to the diversity in amino acid sequence, CDR3 loops may also differ in loop length [221]. While CDR3 loop lengths between paired BCR/antibody heavy and light chains (and γδ TCR chains) are relatively broad and unmatched [219], the loop lengths between paired TCR α and β

chains are narrow and closely matched, suggesting the possibility that functional αβ chain pairing may be determined by CDR3 loop length [219, 222].

Similar to antibodies that recognize protein ligands [223], the site of pMHC antigen recognition on the TCR is relatively flat, which is consistent with similarly flat nature of the pMHC complex [224, 225]. Using x-ray crystallography, it has been determined that most TCRs have a cleft between the α and β chain CDR3s possibly for accommodation of a central upward facing side chain from the antigenic peptide [226]. This property has led to the speculation that the majority of peptide discrimination is contingent on CDR3 interaction, consistent with the higher level of diversity observed in CDR3s than their counterpart CDR1 and 2s. While the majority of antigenic peptide contact occurs via the CDR3 loops, CDR1 and 2 make major contacts with the conserved structural elements of the MHC molecule **(Figure 1.9)** [205].

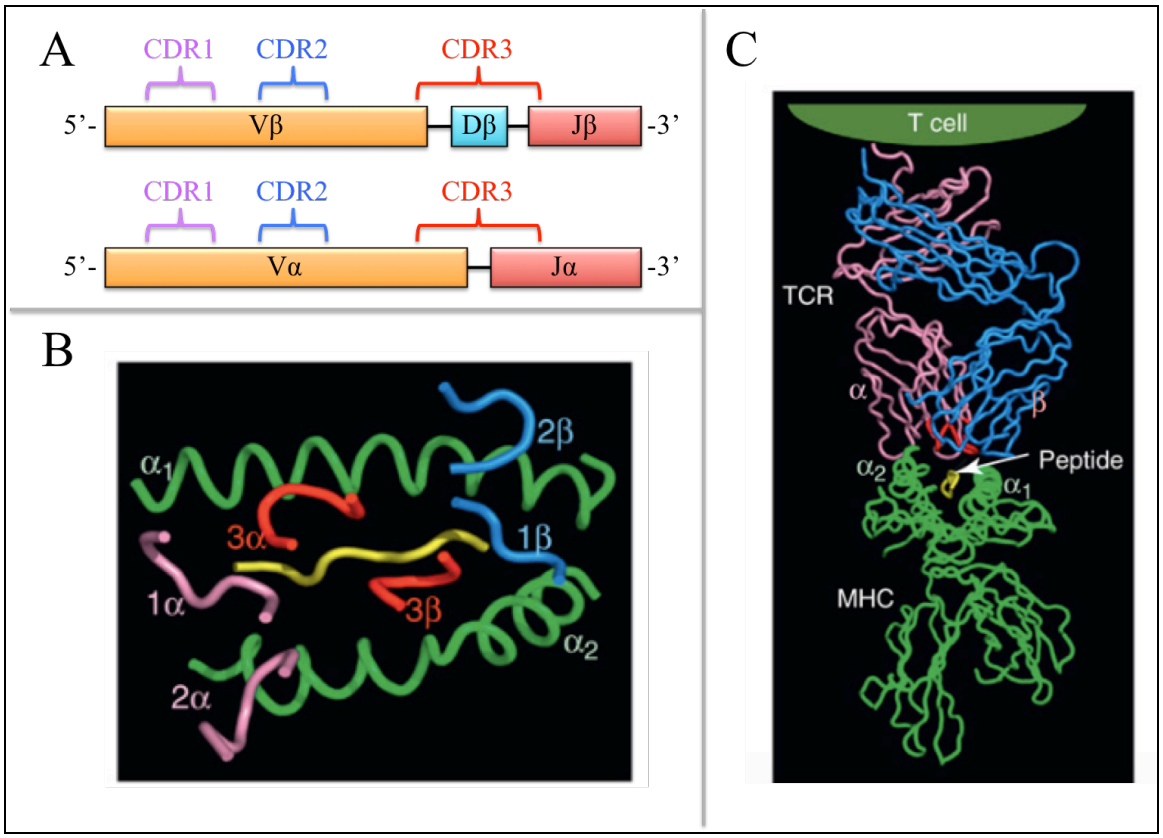

Figure 1.9: Complementarity determining regions (CDRs) 1 and 2 mainly contact the self-MHC protein while the highly diverse CDR3 region makes major antigenic peptide contact. **[A]** The coding sequences for the CDR regions correspond to three hypervariable regions (HVRs) of DNA for each TCR chain. While HVR 1 and 2 (corresponding to CDR1 and 2) are germline encoded within each given V gene, HVR3 (corresponding to CDR3) is encoded by the imprecise junction of the V, D, and J gene segments for TCRβ or V and J gene segments for TCRα. **[B]** Crystal structure solutions of TCR:pMHC complexes have repeatedly shown that although all three CDR can contact both MHC and bound peptide epitopes, CDR1 and 2 (pink and blue, respectively) make contact primarily with the MHC protein (green) while CDR3 (red) is typically centered over the bound peptide (yellow). **[C]** Crystal structure of complete TCR:pMHC complex. Crystal structures adapted from Garcia *et al.* 2012 [227].

An overview of TCR diversity

In order for the immune system to combat the ever increasingly diverse number of pathogens it may encounter in an individual's life time it has developed an immensely large repertoire of T cells with unique TCRs where only a few clones will be specific for any given antigen. Estimations for the number of different possible TCRs encoded in the genome (the potential repertoire) and methods for quantifying the number of TCRs present in an individual (the realized repertoire) have both been attempted numerous times over the past 16 years and as technology improves the values for both estimates continue to rise [192, 228-232].

A functional immune system must possess both vigorous reactivity against pathogens and a lack of overt self-reactivity [228]. These defining features are inherently postulated by the clonal selection theory [233], which states that lymphocytes express a highly diverse, clonally distributed set of antigen-specific receptors, and upon antigen recognition, expansion of antigen-reactive cells provides sufficient protection from the infection. As the immune system cannot predict which pathogens it may encounter in a given lifetime, it relies on maintenance of a highly diverse TCR repertoire. Following the elucidation of the process of somatic recombination [220], the puzzle of how such diversity is generated was discovered.

As discussed above, diversity in reference to the TCR repertoire is mainly confined to the CDRs that make contact with the pMHC. Specifically, the diversity of generated CDR3 regions is due to 4 main contributors: (1) the stochastic selection of one from a number of possible V, D, and J gene segments, (2) semi-random DNA cleavage

by the enzyme Artemis leaving P-nucleotide hairpins, (3) N-nucleotide addition and subtraction by TdT and exonucleases at the junctions of the V, D, and J genes, and (4) the pairing different α and β chains to generate the functional receptors [205]. The seminal paper by Arstilla *et al.* published in 1999 [192] estimated diversity generated by this process to be capable of producing 10^6 unique TCRβ chains per person and similar estimates for TCRα, yielding a possible 10^{12} unique TCRαβ pairs (should completely random pairing of α and β chains be possible). However, more recent investigations have increased this estimate to 10^{16}-10^{20} unique pairs [96, 230]! Interesting to note, adult humans have roughly10^{12} T cells [192] (mice have roughly 10^8 [234]) and it is known that some T cells will express the same TCR [228], so while the theoretical diversity may be as high as 10^{16}-10^{20} possible receptors, a given human will actually express fewer than 10^{12} at any given time.

Estimates of the **actual** αβ T cell diversity expressed in humans were first obtained by extrapolation from small sub-samples of molecular measurements of TCR diversity using the 'specratype' or 'immunoscope' technique [222, 235]. Calculations from such methods produced an estimate of approximately $2.5x10^7$ unique naïve αβ T cells per person; with 100 fold fewer unique antigen-exposed memory T cells. Interestingly, although four orders of magnitude smaller in cell number (10^{12} T cells in humans vs 10^8 T cells in mice), the actualized mouse αβ TCR repertoire was estimated to be near $2x10^6$, only 10x less diverse than estimates for humans [236]. These early human estimates were most likely skewed however, as the methodology relied heavily on exhaustive capillary-based sequencing of rearranged TCR genes expressed in small, well-

defined subsets of the repertoire, and then extrapolation of these subsets was applied to the entire repertoire [96]. The biases inherent in extrapolation severely limited these early estimates and recent advancements in technology have demonstrated that the actual number of unique αβ T cells may be 2-4 times as large [232, 236]. While all these estimates are thought to be somewhat minimalistic, given the limits in the total number of circulating cells they have been widely held to most likely convey realistic values [228].

As previously mentioned, the diversity of the theoretically achievable TCR repertoire may be as high as four orders of magnitude greater than the actual repertoire (due to the limited number of T cells in an individual). This discrepancy may account for the differences in TCR repertoires between any two given individuals. In fact, it was found that even between genetically identical inbred mice, only 20-25% of TCRβ sequences were shared [236]. Additionally, alymphoid animals receiving spleen cells from the same donor were analyzed for TCR sequence homology, however 80% of discovered TCR sequences were unique [237]. Even genetically identical animals infected with the same strains of lymphocytic choriomeningitis virus (LCMV) responded with proliferation of different TCR sequences [238]. Therefore, while relatively high levels of diversity exist across the TCR repertoire, most clones will be found at a very limited copy number, and the constraints in the total number of cells found in a single individual only allows for a fraction of the potential diversity to be expressed at any given time.

Current approaches for analyzing TCR diversity

In the original experiments assessing the diversity of immune receptor repertoires a technique known as spectratyping/immunoscoping was utilized. Spectratype analysis utilizes polymerase chain reaction (PCR) technology to amplify cDNA reverse transcribed (RT) from immune receptor mRNAs [239]. The rearranged mRNA transcripts maintain different CDR3 lengths from specific TCR (or BCR) variable region genes due to the random VDJ gene selection and insertion/deletion of nucleotides at the gene segment junctions. Extrapolation from a small subset of these samples to account for the entire repertoire is then performed to estimate total TCR diversity.

A seminal review by Nikolich-Zugich outlines the four main steps to the spectratyping procedure [228]. First, the CDR3 of a selected Vβ-Jβ rearrangement (i.e. VB1-Jβ1) from a population of T cells is amplified by PCR following RT of purified T cell mRNA. Products from the PCR reaction will be separated by polyacrylamide gel electrophoresis (PAGE) based on CDR3 length (commonly referred to as CDR3 length polymorphism analysis [222, 235]). All the products of a given CDR3 length (N-nucleotides) will migrate as a single band; typically 6-10 bands will be visible for any given Vβ-Jβ rearrangement, distributed in a Gaussian profile [192, 235, 236]. Second, quantification of the relative abundance of each band corresponding to CDR3 length is performed (i.e. a 10 amino acid CDR3 corresponding to a 30 nt band may constitute 11% of all amplified sequences). The band from the chosen Vβ-Jβ rearrangement of specific CDR3 length is excised from the gel for downstream analysis. Third, once the DNA from the excised band has been purified, the PCR products belonging to said band are

exhaustively sequenced until no new sequences are detected. The resulting number of sequences (X) corresponds to the diversity of TCRVβ sequences belonging to the selected Vβ-Jβ combination (i.e. Vβ1-Jβ1) with CDR3 length corresponding to the selected band excised from the gel (i.e. 30 nt). Finally, extrapolation to all TCRVβ sequences belonging to the selected Vβ-Jβ rearrangement (i.e. Vβ1-Jβ1) regardless of CDR3 length (for the previous example, the 30 nt band constituted 11% of all Vβ1-Jβ1 rearrangements: X x 0.11 = % of total combination). The summary of this statistic represents the total β chain diversity. The number of different α chains found to pair to the given β chain (originally estimated to have a lower limit of ≥25 α chains per β chain [192]) is then used to calculate the total diversity of the repertoire. Using this method, it was determined that the human repertoire contained $2.5x10^7$ different TCR clonotypes and the mouse repertoire was an order of magnitude smaller at $2x10^6$ different clonotypes [192, 236].

While highly novel and groundbreaking at the time, these previous assessments of diversity relied on massive extrapolation from extremely limited subsamples applied to the entire repertoire. As previously stated, due to the necessity of such extrapolations, accuracy and precision of this method was severely limited, and researches sought to improve upon methodologies for estimation of actual immune receptor repertoire diversity analysis.

The recent developments in high throughput sequencing technology has provided us with a much more powerful tool than the capillary-based sequencing technology that was afforded researchers when the analysis of TCR diversity began nearly 20 years ago.

Specifically, the illumina-based platforms allow for parallel sequencing of millions of short reads simultaneously in a few hours [240]. This has both increased the accuracy and precision by which diversity estimates can be made.

Harlan Robins and a team of researchers in Seattle have pioneered current modern immune receptor repertoire diversity analysis by adapting standard next generation sequencing technology specifically tailoring it for the analysis of large polyclonal immune receptor samples. Using unique sets of multiplex primer systems, 'ImmunoSEQ' technology allows for the simultaneous sequencing from genomic DNA of the differently rearranged TCRβ CDR3 regions expressed by millions of T cells. What became the standard of immune repertoire analysis, ImmunoSEQ enables the direct sequencing of a much larger (significant) fraction of CDR3 sequences than spectratyping, and due to the massively parallel nature of next gen sequencing, the samples may contain diversity far exceeding the capability of capillary-based DNA sequencing instruments. Finally, the use of high throughput sequencers also allows for estimation of the relative frequency for each CDR3 sequence discovered in a given population [232].

Generation of the library of human TCRβ CDR3 encoding sequencing amplicons using the ImmunoSEQ technology requires the use of multiple comprehensive multiplex primer sets. Equimolar pools of 45 Vβ forward primers (Vβ F primers) and 13 Jβ reverse primers (Jβ R primers), each specific to the known functional human TCRβ V and J gene segments were generated. Additionally, each primer contains at the 5' end the universal forward and reverse primer sequences compatible with the base Illumina sequencing technology. While the entire amplicon generated by this method is roughly 200bp, the

average length of any given CDR3 region is 35 ± 3 base pairs [241], so all sequencing reads progressing from the J gene segment primers should capture the entire CDR3 regions [239]. Using the ImmunoSEQ technology, the number of unique TCRβ sequences in healthy male adults was estimated to be between 3-4x10^6 [232], roughly 4x greater than estimates using the spectratyping method **(Figure 1.10)**.

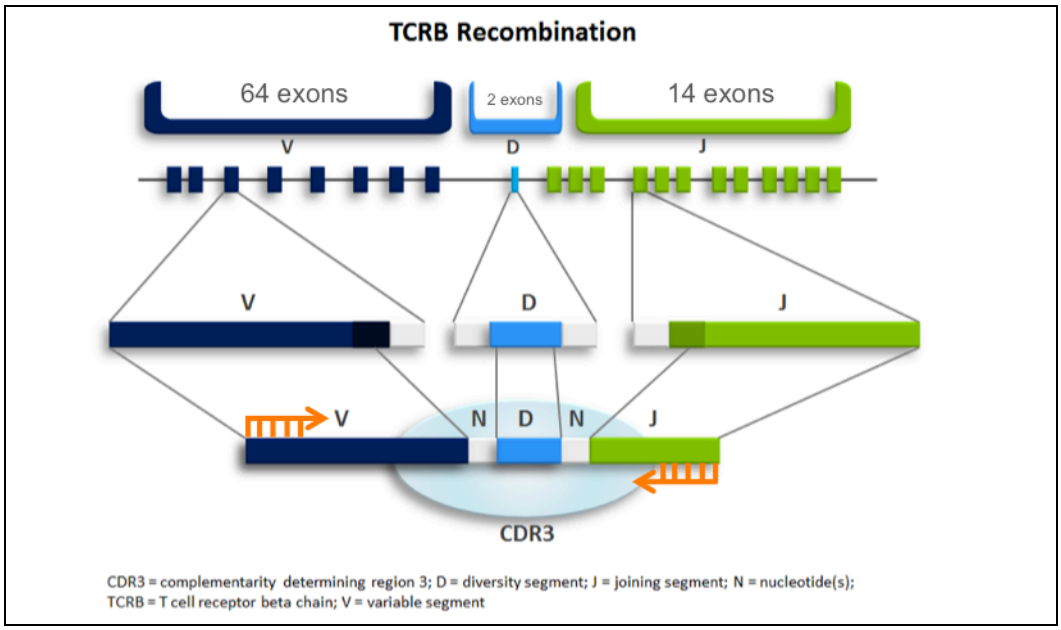

Figure 1.10: ImmunoSEQ strategy for TCRβ chain deep sequencing. Equimolar pools of 45 Vβ forward primers and 13 Jβ reverse primers, each specific to the known functional human (or mouse - not shown) TCRβ V and J gene segments are combined in a dual multiplex PCR reaction to amplify all recovered TCRβ DNA. Proprietary software offered by Adaptive Biotechnologies allows customers to analyze millions of individual TCRβ chain V, D, and J gene frequencies, amino acid sequences, or even CDR3 nucleotide sequences at a massively parallel level. Image adapted from Adaptive Biotechnology's home web site [242].

Chapter 2 outlines initial data from single chain (TCRβ) analyses that led to the finding of a lack of a 'publicly' distributed TCR repertoire.

Limitations of currently available technology for the analysis of immune repertoires

The technological advancements in high throughput massively parallel sequencing have allowed for a much more comprehensive subsampling of the full immune repertoire. However, even with our improved sequencing capability, accurate estimation of the diversity of TCRβ CDR3 sequences in an entire αβ T cell repertoire requires measuring diversity in a finite number of sampled T cells followed by estimation of the number of CDR3 sequences not present in the sample. Statisticians have long since dealt with this issue in numerous fields of science, especially in population ecology. Commonly referred to as the 'unseen species problem' [243], estimation of species diversity in large populations from measurements in random, finite samples can be applied to the estimation of TCRβ CDR3 diversity [232, 244].

TCR diversity can be defined using the Venturi strategy for measuring the diversity of species [245, 246]. Here, a species refers to a unique TCRα/TCRβ pair. The simplest measure of diversity used by ecologists, species richness, equals the total number of species, but ignores the relative abundance of different species. To provide a more accurate assessment of species diversity, we will use the Simpson Diversity Index ($D = 1- \Sigma\Phi f2$) where Φf is the frequency of BCR clones with abundance f in the population [247, 248]. Although diversity indices that encapsulate diversity in terms of a single number can be easily estimated by determining the frequency of sampling the

same species repeatedly, the data collected allows researchers to generate a much more comprehensive measure of species diversity: the frequency spectrum, sometimes called a Preston plot in ecology [249, 250]. The frequency spectrum will give us both the number of unique species (TCRαβ pairs) as well as their relative frequencies in the population. This approach was also used in the pioneering studies by Kourilsky *et al.* (among many others),when they measured the sequence diversity of TCR α and β chains [192, 232, 236].

The use of two multiplex primer systems (V gene segment multiplex and J gene segment multiplex) is required for amplification prior to sequencing. Inherent differences in primer-kinetics may potentially generate systematic bias in the inferred relative abundances of CDR3 sequences [232]. Furthermore, errors introduced by the polymerases used in the amplification and sequencing may artificially create 'new' CDR3 sequences. This will produce a Luria-Delbrück distribution of final CDR3 error frequencies, and as these errors will likely occur at random and be present at low copy numbers, the raw counts of rare CDR3 clones may be artificially inflated [251]. Both of these issues can be accounted for using unique algorithms that take into account both individual primer-kinetics and base polymerase fidelity error rates, respectively.

To estimate the magnitude of bias introduced due to primer-kinetics differences, samples of roughly 30,000 unique T cells were amplified for 25 PCR cycles, samples were split in half, and one half was amplified for an additional 15 cycles (40 cycles total). Both samples were then sequenced independently, and compared to one another for discrepancies. A linear correlation was observed after plotting the number of observations

of the 30,000 sequences from each sample. Attributing the mean variation about the line of correlation (1.5 fold) entirely to PCR bias, it was determined that each cycle of PCR amplification potentially induces bias on average of $1.5^{1/15} = 1.027$ [232]. Use of the immunoSEQ strategy, which employs 25 rounds of dual-multiplexed CDR3 amplification before sequencing, would thus potentially introduce biases of average magnitude $1.027^{25} = 1.95$ in the inferred relative abundance of distinct TCRβ CDR3 sequences.

Similarly, if *n* number of CDR3 regions is sequenced, each of length *L*, and the empirically defined per-base error rate is *ε*, one should expect *nLε* sequences to contain errors. The chance of the same error occurring multiple times, *m*, to the same sequence is:

$$\sum_{c} cL\epsilon\left(\frac{\epsilon}{3}\right)^{m-1} f_c$$

where *c* is the true copy number of the CDR3 clone and *fc* is the probability of a clone being present in *c* copies. Thus, the chance of multiple coincident errors decreases exponentially. Correcting for these sequencing errors is analogous to challenges faced during statistical inference from high coverage, pooled, or metagenomic sequencing projects in which the observed data are modeled as a convolution of the true data and the error probabilities [252].

While the above-mentioned problems due to PCR bias and enzyme fidelity can be accounted for both empirically and systematically, there is an additional area of concern that has yet to be addressed. As heterodimeric receptors, αβ TCRs are composed of two chains generated by independent rearrangements of **both** the TCRα loci and the TCRβ loci. The theoretical diversity that can therefore be generated is thus greatly increased due to combinatorial efforts by pairing any one of the numerous TCRα chains with any one of the numerous TCRβ chains. And

while it has been estimated that any given unique TCRβ chain will pair with on average 25 different TCRα chains [192], determining the actualized number of unique TCRαβ pairs in an entire TCR repertoire will require technological advancements that can somehow covalently link the α and β chain CDR3 sequences from individual T cells into a single 'amplicon' suitable for downstream sequencing. It then goes with out saying that pursuit of such innovation has been the gold standard of modern immune receptor repertoire research.

Current technological advancements in single cell sorting and additional approaches for pairing TCR mRNA sequences

Numerous strategies have been attempted to accomplish high-throughput single-cell sequence analysis of polyclonal cell populations. There are many biological questions that require single-cell analysis of gene sequences, including analysis of clonally distributed dimeric immunoreceptors on lymphocytes [192, 211, 253, 254] and the accumulation of driver/accessory mutations in polyclonal tumors [255-258]. Lysis of bulk cell populations results in mixing of gene sequences, making it impossible to know which pairs of gene sequences originated from any particular cell and obfuscating analysis of rare sequences within large populations. Although current single-cell sorting technologies can be used to address some of these questions, such approaches are expensive, require specialized equipment, and lack the necessary high-throughput capacity for comprehensive analysis [259-261]. Water-in-oil emulsion approaches for single cell sorting have been developed [262] but droplet-based single-cell lysis and analysis have proven inefficient and yield high rates of false pairings. Ideally, molecular

approaches for linking gene sequences from individual cells could be coupled with next-generation high-throughput sequencing to overcome these obstacles, but conventional approaches for linking gene sequences, such as by transfection with bridging oligonucleotides, result in activation of cellular nucleases that destroy the template, precluding this strategy as well [263-265]. Finally, combinatorics based methodology utilizing discreet high copy number sequences have been validated [266], but such combinatorics-based approaches are not capable of pairing low copy number sequences that constitute the majority of the TCR repertoire and more importantly may give rise to rare clones with high-anti-tumor/anti-pathogen activity.

Recent advances in the synthesis and fabrication of modular DNA origami nanostructures have resulted in new possibilities for addressing these and many other current and long-standing scientific and technical challenges in biology and medicine [267-270]. One exciting application of DNA nanotechnology in reference to immune repertoire profiling is the capability of the nanostructures with regards to intracellular capture and subsequent sequence analysis of mRNA from individual cells within heterogeneous cell populations [263, 271]. DNA nanostructures can be transfected into individual cells to capture and protect mRNA for specific expressed genes, and incorporation of a set of unique matching barcodes into the origami nanostructure may facilitate pairing and analysis of mRNA from individual cells by high-throughput next-generation sequencing. This approach is highly modular and can be adapted to virtually any two (or possibly more) gene target sequences, and therefore has a wide range of potential applications for analysis of diverse cell populations such as understanding the

relationship between different immune cell populations, development of novel immunotherapeutic antibodies, or improving the diagnosis or treatment for a wide variety of cancers.

Objectives

The following chapters outline a proposal to develop a novel DNA origami nanotechnology approach to analyze T cell receptor genes expressed in single cells within polyclonal populations, without the need for single-cell sorting. Specifically, it will be outlined how to use DNA origami nanostructures to capture and protect both TCRα and β chain mRNA in transfected T cells as a means of both quantifying actual TCR diversity (including that of $\alpha\beta$ chain pairing) and making predictions about the payoff between diversity and protection. It will be shown that DNA origami nanostructures can be transfected into lymphocytes by electroporation with high efficiency, thus avoiding the endosomal/lysosomal degradation pathway, and are able to selectively bind and preserve lymphocyte immunoreceptor mRNA for subsequent analysis. Integral fluorescent labels on the DNA origami facilitate identification of transfected cells, and reisolation of the nanostructures with bound mRNA is achieved using integral biotin labels and avidin column purification. An important design feature of the DNA origami is that the mRNA capture sequences on any given nanostructure are linked to one another by a unique 5'-5' 'bowtie' linkage that includes matching complimentary barcodes; the recovered mRNA are reverse transcribed utilizing the origami capture probes as gene-specific RT primers thereby linking the cDNA to the bowtie-barcodes. Furthermore, as the DNA origami capture probes are required for priming the RT reaction, any unbound mRNAs

will not be amplified, improving selectivity and avoiding false pairing. Thus, the following chapters will be used to outline and validate this approach for obtaining paired TCRα and TCRβ sequences from single cells in large oligo- or polyclonal T cell populations (>10^6 cells/run), without the need for single-cell sorting. This approach will reveal potentially important rare T cell receptor sequences that are not identifiable in conventional approaches and avoid the high costs associated with single-cell sorting.

CHAPTER 2

ANALYSIS OF THE PUBLIC VERSUS PRIVATE NATURE OF THE CD8 T CELL TCRB CHAIN CDR3 REPERTOIRE AT THE V/J GENE SEGMENT AND AMINO ACID LEVEL

ABSTRACT

The complex mechanism of somatic recombination of V, D, and J gene segments gives rise to the TCR and in doing so generates the diverse repertoire of lymphocyte receptors that is necessary for defense against infection. Modern deep-sequencing approaches enabled us to probe the 'public' versus 'private' nature of the TCRβ CDR3 repertoire. We found that naïve TCR repertoires in any given individual are not "flat", but rather display a gradient of TCR diversity with some precursors present at low numbers and others present in relatively high frequency. When comparing V and J gene usage between individuals, repertoires were very consistent and recombination frequencies between V and J gene segments were conserved as well. However, when comparing repertoires from different individuals at the amino acid level, major differences were observed, with only ~11.5% of sequences shared between any two individuals. As would be expected, the shared (public) fraction of CDR3 sequences correlates to the few sequences that are highly expressed in any given individual. These differences in TCR repertoires have many novel implications, such as why individuals display differing competencies in immune protection, for analysis of individuals post vaccination, or for analysis of immune reconstitution (such as after chemoablation).

IMPORTANCE

Thus far, TCR diversity has been described in terms of simple summary statistics, such as the number of distinct TCR (based solely on analysis of TCRβ chain sequences), which is analogous to species richness or the Simpsons diversity index in ecology that provides a single number to describe the relative abundance of different species [245-247]. However, through the use of TCRβ multiplex sequencing, our findings suggests that simple summary statistics are inadequate to describe total diversity as these fail to include the frequency of each individual TCR species in the population, which can vary over 10,000-fold [272, 273]. Furthermore, previous studies (again based solely on TCRβ sequence data) have also roughly described individual TCR in absolute terms as either public (occurring in many individuals) or private (occurring rarely in individuals) [274, 275]. By comparison of TCR repertoires between multiple genetically identical individuals we have determined that the naïve T cell receptor repertoire is very consistent among different individuals at the V and J gene level. Furthermore, when analyzing different subsets of V gene expression, the J gene recombination frequencies are conserved as well. However, analysis of these individual's repertoires at the amino acid level revealed that major differences arise among even genetically identical individuals. Finally, as one might expect, the small shared (public) fraction of TCR gene sequences correlates to the few TCR gene sequences that are highly expressed in most individuals. These vast differences in "public" TCR gene sequences may help to explain the immunological variance in self-protective ability among different individuals.

INTRODUCTION

The adaptive immune system relies on a highly diverse repertoire of antigen receptors (TCRs and BCRs) with the goal enabling the host to mount highly specific immune responses tailored to virtually any encountered pathogen. The immune system's response to pathogenic infection involves activation of highly biased profiles of antigen-specific T cells, selected from a highly diverse naïve cell repertoire [233]. In the vast majority of T cell responses, the specific repertoire of T cells activated during a given immune response will be distinct between any two given individuals [276]. This 'private' T cell response constitutes the distinct T cells activated in immune responses to specific epitopes bearing TCRs that are rarely observed between any two individuals. Contrastingly, on some occasions antigen-specific TCR repertoires may be observed in multiple individuals, constituting a 'public' TCR repertoire. Interestingly, such public immune responses have been reported in a variety of immune responses, not only to pathogenic infections but also in malignancy and autoimmunity **(Table 2.1) [229].**

Table 2.1: Examples of public TCRs in humans. Table adapted from Li *et al.* 2012 [276].

Disease	***Antigen***	***TCRβV***	***TCRβJ***	***TCRαV***	***TCRαJ***	***Ref.***
Infectious Diseases						
EBV	EBNA $3A_{339-347}$	7-6	2-7	26-2	52	[277]
Cytomegalovirus	$IE1_{316-324}$	5-1	1-3	Unknown	Unknown	[278]
Cytomegalovirus	$pp65_{103-114}$	28	2-7	8-6	30	[279]
Parvovirus B19	$NS1_{572-580}$	5-1	2-1	Unknown	Unknown	[280]
Clostridium tetani	Tetanus toxin	5-4	2-3	41	Unknown	[281]
HSV	Virion $P22_{49-57}$	10	2-1	8-1	27	[282]
HIV	$Gag_{162-172}$	19	1-2	5	13	[283]
Malignancy						
Melanoma	Melanin-A_{26-35}	27	2-1	12	34/45	[284]
Cancer (multiple)	NY-$ESO1_{157-165}$	12-3	2-1	17	31	[285]
Autoimmunity						
MS	MBP_{83-99}	6-5	2-7	23	10	[286]
Reactive Arthritis	Unknown	5	2-3	Unknown	Unknown	[287]
Aplastic Anemia	Unknown	9	2-1	Unknown	Unknown	[288]
Psoriasis vulgaris	Unknown	3	2-7	Unknown	Unknown	[289]
Systemic sclerosis	DNA Topo I	30	1-1	Unknown	Unknown	[290]
Sarcoidosis	Unknown	Unknown	Unknown	12-1	15	[291]
Rheumatoid Arthritis	Unknown	27	2-7	22	1	[292]

In some instances, the observed public TCRs were shown to correlate with favorable outcomes, including SIV [293]. Additionally, studies of HIV-infected individuals demonstrating long-term non-progressive disease revealed a repertoire of shared TCRs displaying effective cross-reactivity of epitope variants [290, 294-296]. However, the effects observed from antigen-specific publicly shared TCR repertoires are not always favorable. It has also been observed that public TCR usage in SIV studies facilitated viral immune evasion [297]. The relative cost vs benefit of publicly distributed antigen-specific TCR repertoires will most likely be antigen-specific and require further research [276].

While a publicly distributed T cell response to a given antigen between different individuals may seem obvious due to the fact that any given invading antigen (or self-protein in terms of cancer or autoimmunity) will only provide a limited number of activating epitopes (and thus possible TCRs capable of recognizing and responding), the inherent prerequisite for such public T cell responses is the sharing of discreet TCRs in the naïve T cell repertoire between individuals. Therefore, in order for any public T cell response to occur to a given antigen, there must exist some amount of overlap of the naïve TCR repertoire [298]. The factors contributing to such overlap have been previously hypothesized [299-301].

In order for public T cell responses to occur, there must be mature naïve T cells in different individuals with identical TCRs. In fact, previous studies have demonstrated that on average, any two individuals share 10.5% (±1.8%) of their expressed CDR3 amino acid sequences [302]. These public subsets of T cells could arise from favorable

positive selection during T cell development or be due to a stochastic communal production during recombination events across individuals (or both) [276]. Mechanisms involving biases of thymic selective events during T cell development have been proposed [299, 300], however repertoires in DP thymocytes and mature naïve T cells have been demonstrated to show high levels of similarity [276], indicating thymic selective biases to play a relatively minor role in the public distribution of naïve TCR repertoires. Therefore, the majority of any commonality among naïve TCR repertoires between individuals must rely mainly on selective biases in VDJ recombination.

Pioneering studies by Venturi and colleagues [301, 303-305] outline the phenomenon of 'convergent recombination', and demonstrated that biases during VDJ combinatorial events are the major contributors of public TCR repertoires. Described as:

> *"...the process whereby multiple recombination events converge to produce the same nucleotide sequence, and multiple nucleotide sequences converge to encode the same amino acid sequence, resulting in different TCR sequences generated with differential frequencies during recombination."*

Venturi and others outlined biased V/D/J gene usage and further combinatorial biases (including not only biases between selected V, D, and J genes, but also biases in the number of nucleotide insertions/deletions at the coding ends of the VDJ junctions) [301, 303-309]. With its inherent prediction that different TCR sequences have different expression frequencies, convergent recombination outlines why specific clonotypic frequencies of different TCRs display high levels of public distribution variation [276].

This theory was evaluated by the demonstration that TCRβ sequences with convergent features are present at higher copy numbers within individuals and are also shared between individuals at a statistically higher than average level [310], however it was determined that random convergent recombination processes are an insufficient cause of the relatively significant overlap observed in DP thymocytes, indicating involvement in other mechanisms as well [276].

While Venturi's theory of convergent recombination demonstrates statistically significant predictions about the extent to which naïve TCR repertoires are shared between individuals, it accounts for less than half of the TCRβ sequence overlap observed in actualized TCR repertoires from separate individuals [276]. Furthermore, multiple groups have demonstrated that there are multiple TCR combinatorial sequences that should be preferentially produced due to convergent recombination that are present at lower clonotype frequency in any given individual, and therefore makeup a smaller proportion of the public TCR repertoire than would be predicted [274, 276, 303-305]. To account for this discrepancy, VDJ recombinatorial biases have been proposed as contributors to naïve TCR repertoire overlap that is not accounted for by convergent recombination [309, 311-313].

Extensive preferences have been observed in the usage frequencies and pairing frequencies of different VDJ gene segments during TCR recombination [309, 311]. Skewed usage patterns of individual Vβ, Dβ, and Jβ gene segments have been observed in human lymphocytes [311], as well as well as non-random usage of Jβ gene segments

in human Vβ17 gene segment repertoires [309]. Furthermore, statistically preferential pairing between specific Vβ genes with specific Dβ and Jβ genes has also been shown [208, 312]. Additionally, there are statistically significant differences between the various Vβ and Jβ gene segments in terms of the numbers of 'random' exonuclease-removed nucleotides from the 3' end of V gene segments and the 5' end of J gene segments [314]. Should a given exonuclease cleavage of a gene segment preferentially result in a multiple of 3 nucleotides in its junction it may be expressed (due to functionality) at a higher level than a gene segment resulting in random numbers of remaining nucleotides. Furthermore, the specific base usage frequencies during TdT N-nucleotide addition for given VDJ junctions is also not random [276, 308, 309].

While the aforementioned convergent recombination and recombinatorial biases during TCR repertoire development may account for the ~10.5% TCRβ CDR3 homology between individuals, it should be mentioned that these biases were demonstrated in the pre-selected DP thymocyte population, and therefore losses due to positive and negative selective events may reduce the actualized intrinsic convergent/recombinatorial biases inducing public repertoire formation in the actual circulating repertoire [276, 315]. Overall however, it is clear that the development of TCR repertoires is not as random as once proposed, and the selective pressures due to convergent recombination and recombinatorial bias may contribute to the small level of a 'publicly' distributed naïve TCR repertoire [229, 276].

In chapter two we utilize single chain (TCRβ) sequence analysis of mouse naïve CD8$^+$ T cells to provide key insight into the public vs private nature of somatic recombination of naïve CDR3 sequences. Specifically, we demonstrate homology between individuals at the Vβ and Jβ gene usage level consistent with the previous studies mentioned above. However, at the amino acid level, differences begin to arise among even genetically identical individuals when analyzing the supposedly 'public' distribution of naïve TCRβ CDR3 sequences.

MATERIALS AND METHODS

Mice: 6-8 week old C57BL/6 mice were purchased from Jackson laboratories (Bar Harbor, ME) maintained in our ASU animal facilities. All mice were maintained under specific-pathogen free conditions at The Biodesign Institute and experiments were performed in compliance with institutional guidelines as approved by Institutional Animal Care and Use Committee of Arizona State University.

Cell sorting: CD8$^+$ T cells were purified by positive immunomagnetic cell sorting (>95% CD8$^+$; Miltenyi Biotec) as previously described [316] from spleens of 6-week-old donor C57BL/6 mice.

Cell surface antibody staining: Single cell suspensions were prepared from splenocytes as previously described [36]. Erythrocytes were lysed with ammonium chloride lysis (ACK) buffer purchased from Lonza (Allendale, NJ) and FACS staining

was done as previously described [37] in 96 well plates with flurochrome-labeled monoclonal antibodies: anti-CD8 (clone 53-6.7), anti-CD44 (clone IM7), anti-CD4 (clone GK1.5). Samples were then fixed in 1% paraformaldehyde solution and immediately acquired on a BD LSR II Fortessa flow cytometer (San Jose, CA) and analyzed using FlowJo Software (Tree-Star, Ashland, OR). All monoclonal antibodies were purchased from BD Pharmigen (San Diego, CA) or eBiosciences (San Diego, CA).

TCRβ CDR3 sequencing and bioinformatic analysis: Three samples consisting of between 1.16 x 10^6 – 1.48 x 10^6 $CD8^+$ sorted cells from C57BL/6 mouse spleens were shipped to Adaptive Biotechnologies (Seattle, WA) for standard ImmunoSEQ TCRβ profiling [232]. All data and statistics were generated from the proprietary ImmunoSEQ analyzer software using a previously described VDJ gene-calling algorithm [317].

RESULTS

The naïve TCRβ repertoire is very consistent among different individuals at the V and J gene expression level. By utilizing the ImmunoSEQ high-throughput sequencing approach [232] on genomic DNA extracted from purified $CD8^+$ T cells, we were able to analyze the CDR3 sequence repertoire from >1 x 10^6 T cells from each of the spleens of three C57BL/6 mice. Samples consisted of 1.16 x 10^6, 1.34 x 10^6, and 1.48 x 10^6 $CD8^+$ sorted T cells from each of the respective mice. Each of the three samples yielded roughly 10^5 CDR3 sequence reads **(Table 2.2)** with an average CDR3 length of

around 42 nucleotides for both total reads **(Figure 2.1A)** and productive reads **(Figure 2.1B)**.

Table 2.2: Sample cell counts and respective TCRβ CDR3 sequences and CDR3 lengths obtained from ImmunoSEQ analysis.

Sample	*Total number of cells sampled*	*Productive rearrangements*	*Average CDR3 length*
Spleen 1	1.16×10^6	9.5624×10^4	42.25 nt
Spleen 2	1.34×10^6	1.10998×10^5	42.29 nt
Spleen 3	1.48×10^6	1.27103×10^5	42.36 nt

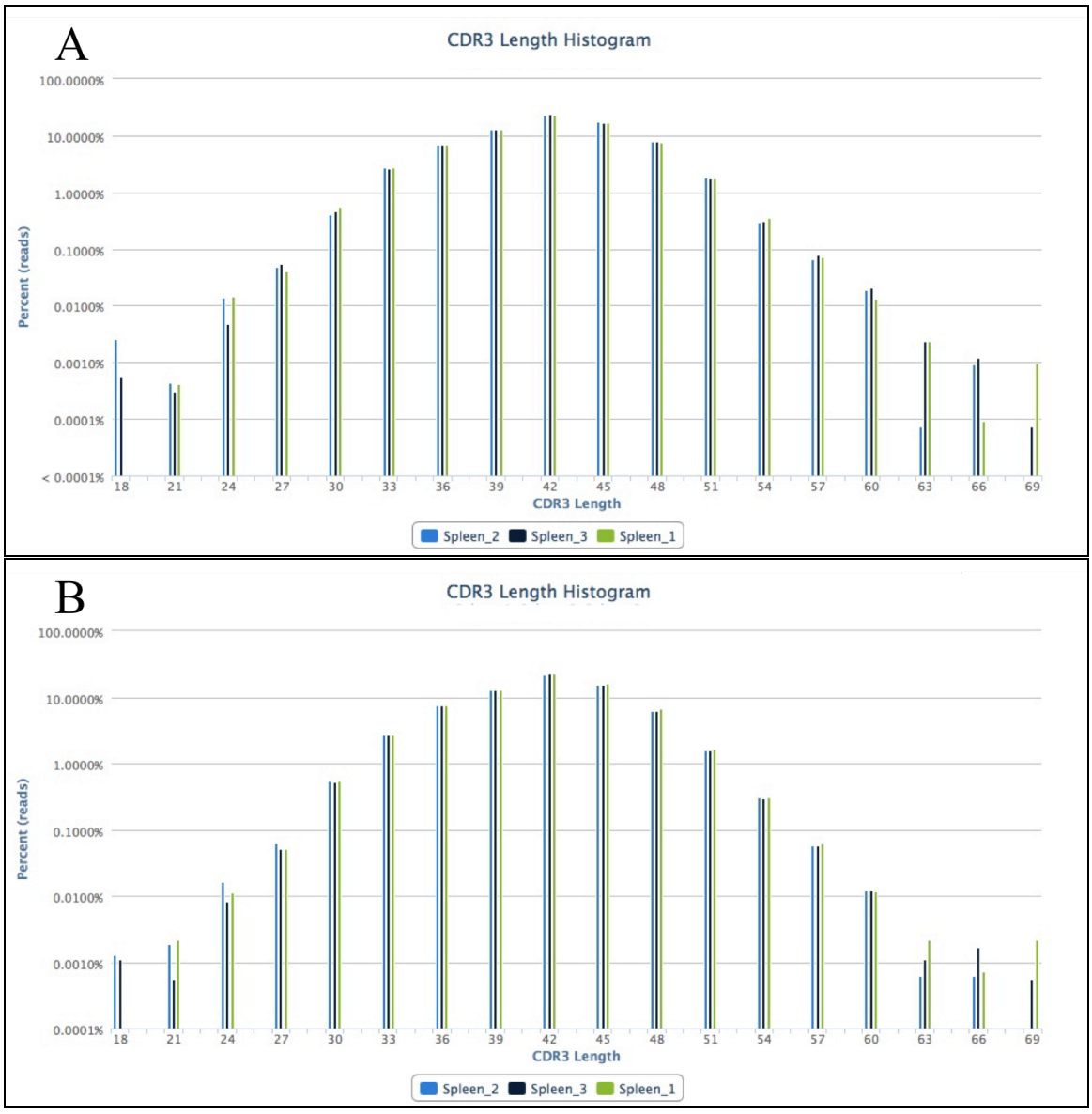

Figure 2.1: TCRβ CDR3 length distributions. CDR3 length distribution patterns (peak profile representation) reflect a typical Gaussian-like profile. An average CDR3 length of 42 nucleotides was observed from $CD8^+$ T cells sorted from splenocytes from all three mice for both the total **[A]** read profile as well as the subset of productive reads only **[B]**.

The mouse TCRβ chain is composed of one each V region, D region, and J region. Previous studies have reported that public TCR repertoires across these regions are due to convergent recombination and selective bias during VDJ recombination [274, 276, 301]. Utilizing the ImmunoSEQ technology, we were able to determine that the frequency with which specific V and J genes were used was highly variable (>10,000 fold in some cases) within an individual, however these V and J gene usage frequencies are conserved between the three individuals analyzed. Analysis of the frequency of expression of each individual V gene shows that there was no significant difference between the three mice **(Figure 2.2A)**. Additionally, all three mice had virtually identical J gene expression frequencies as well **(Figure 2.2B)**. Analysis of the Vβ gene segment utilization frequency demonstrates that ten of the thirty-two Vβ gene segments account for >90% of the sequences collectively observed in the three donors, again providing support for selective biases in VDJ gene selection during somatic recombination.

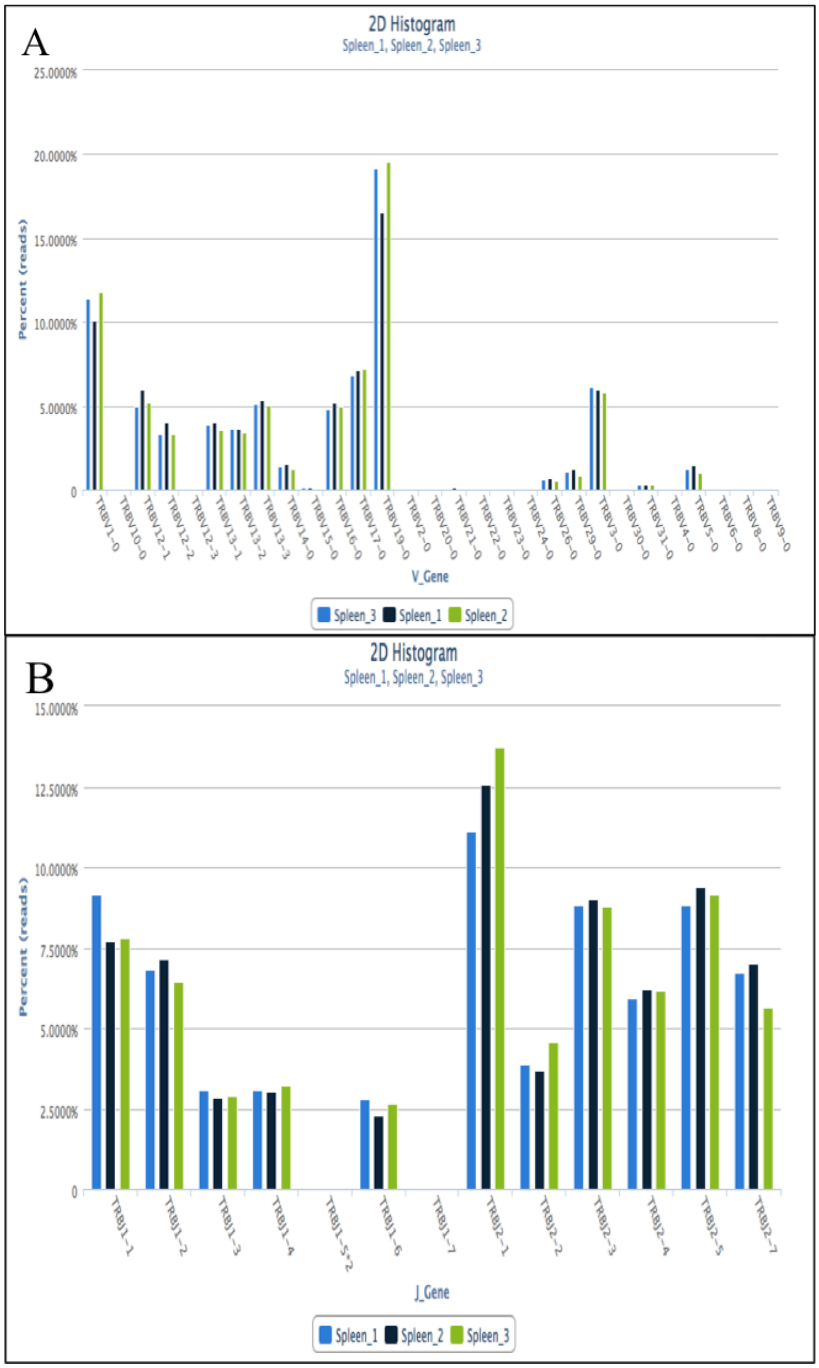

Fig 2.2: Public distribution of TCRβ V gene and J gene expression frequencies between three inbred individuals. **[A]** All three mice display virtually identical V gene expression frequencies for all 31 of the functional mouse V genes. **[B]** Similarly, the 14 J genes were observed to show no significant differences among the three mice analyzed.

CDR3 sequencing reveals V-J gene segment recombination frequencies are conserved between individuals. Consistent with previous data [306], the frequency with which specific Vβ→Jβ recombinations were observed in an individual was highly variable. Remarkably however, while the majority of possible V→J recombinations were observed to be highly variable within an individual, specific Vβ→Jβ recombination frequencies were highly conserved between the three different individuals, especially for the more rare Vβ→Jβ recombinations **(Figure 2.3-2.10)**. Vβ→Jβ recombination frequency graphs were not generated for Vβ genes that were not sampled at least once by all three mice (Vβ6, 7, 8, and 9). It has previously been reported that a small fraction of the TCRβ CDR3 sequences observed from genomic DNA extracted from naïve $CD8^+$ T cells would generate out-of-frame transcripts, not encoding functional TCRβ chains [306]. Consistent with our data from in-frame transcripts, the Vβ→Jβ recombination frequencies of the out-of-frame CDR3 sequences also displayed high levels of non-uniformity. As these out-of-frame transcripts would be selected against during thymic selection, the variability in Vβ→Jβ recombination must attributable, at least in part, to VDJ gene segment selection biases (occurring prior to the thymic selection events).

Analysis of the individual Vβ→Jβ recombination frequencies demonstrates that high levels of non-conformity exist for each individual V gene segment in terms of the relative frequencies to which J gene segments it pairs. Again however, while non-conformity exists in Vβ→Jβ gene pairing within an individual for each specific V gene segment, homology in Vβ→Jβ gene selection is observed between the three individuals.

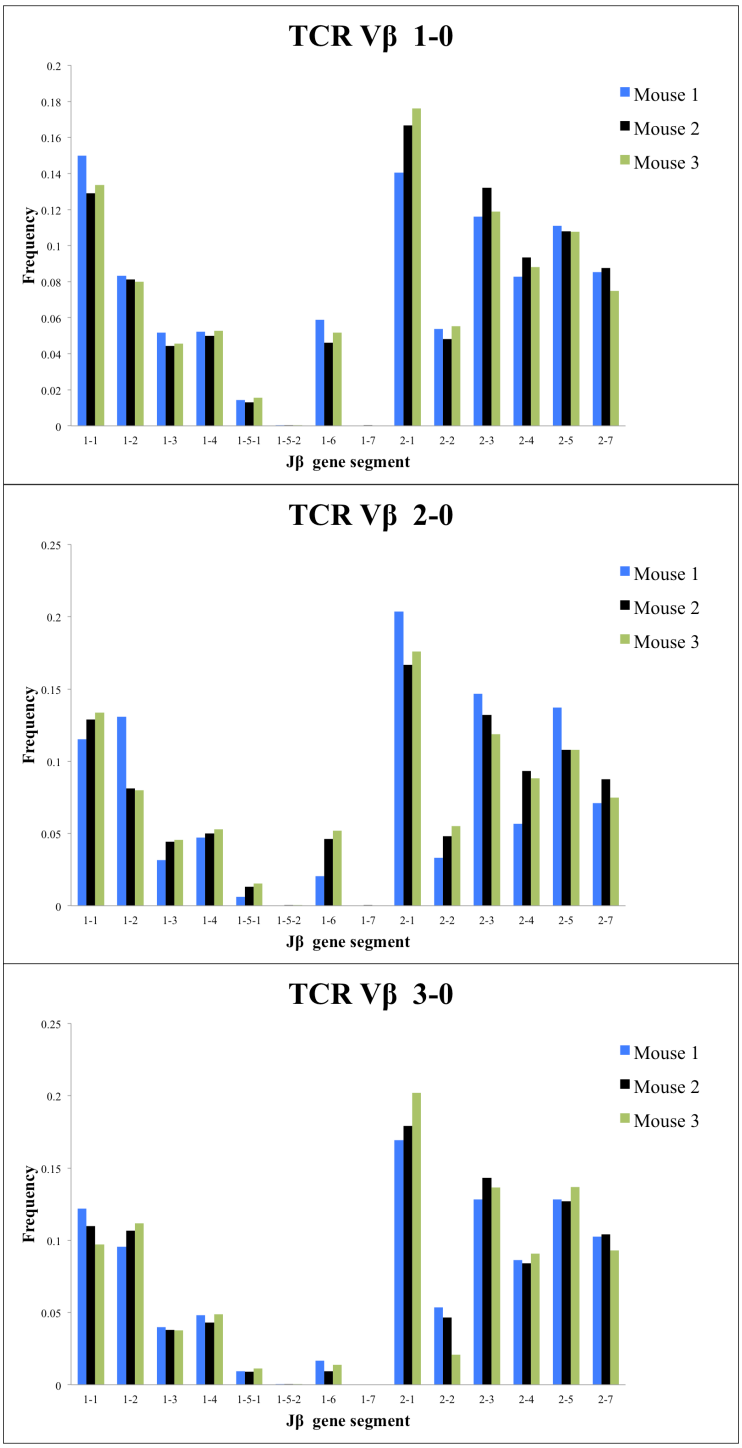

Figure 2.3: TCRβV1-3 gene segments display public recombination frequencies between individuals. Analysis of TCRβV1-3 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

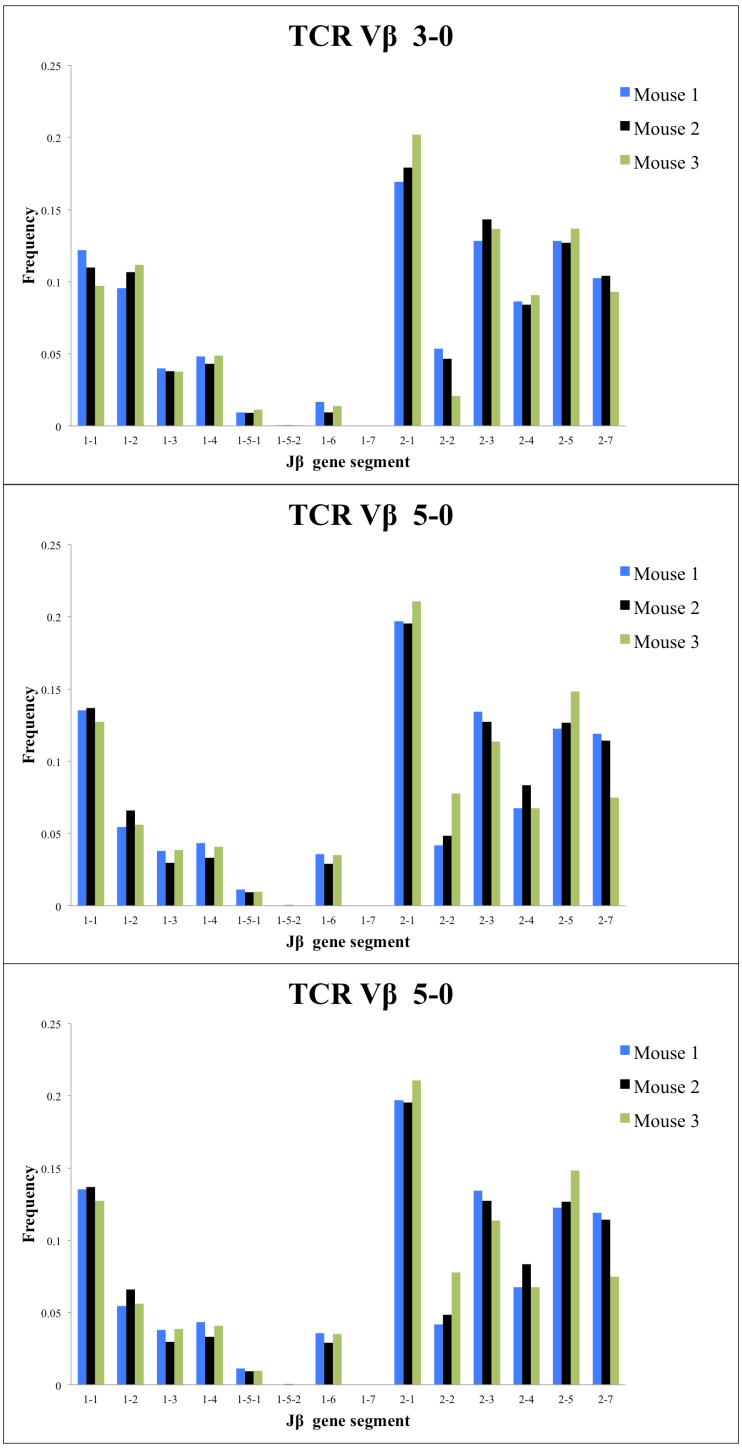

Figure 2.4: TCRβV4, 5, and 10 gene segments display public recombination frequencies between individuals. Analysis of TCRβV4, 5, and 10 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

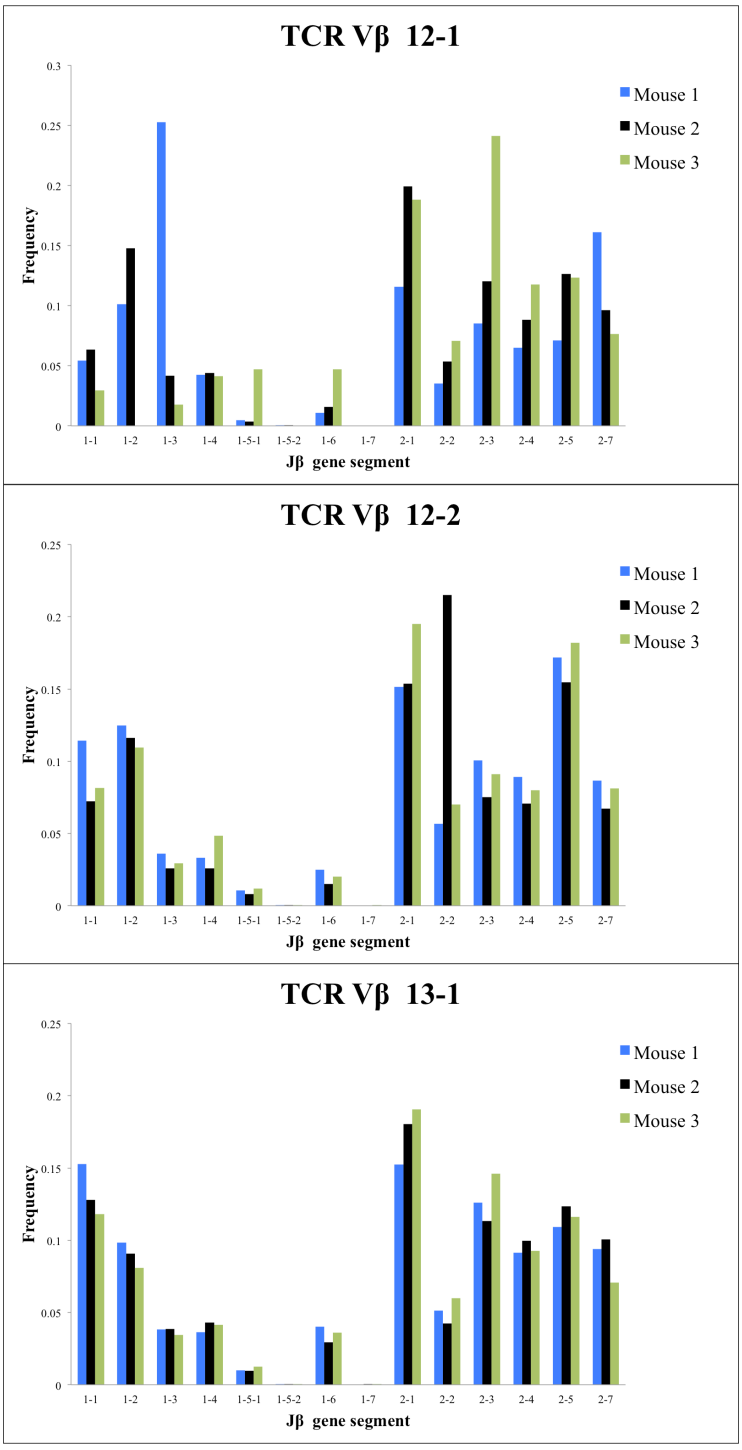

Figure 2.5: TCRβV12-1, 12-2, and 13-1 gene segments display public recombination frequencies between individuals. Analysis of TCRβV12-1, 12-2, and 13-1 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

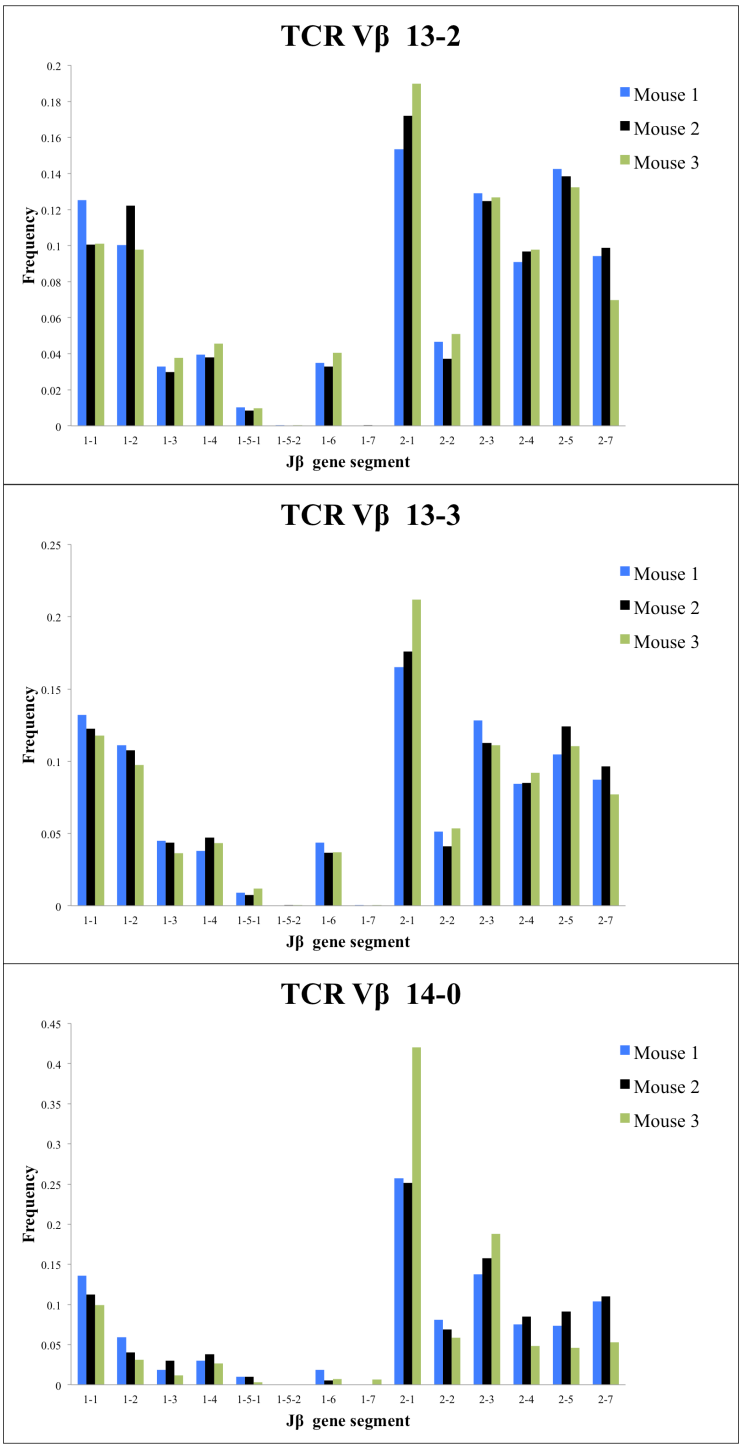

Figure 2.6: TCRβV13-2, 13-3, and 14 gene segments display public recombination frequencies between individuals. Analysis of TCRβV13-2, 13-3, and 14-0 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

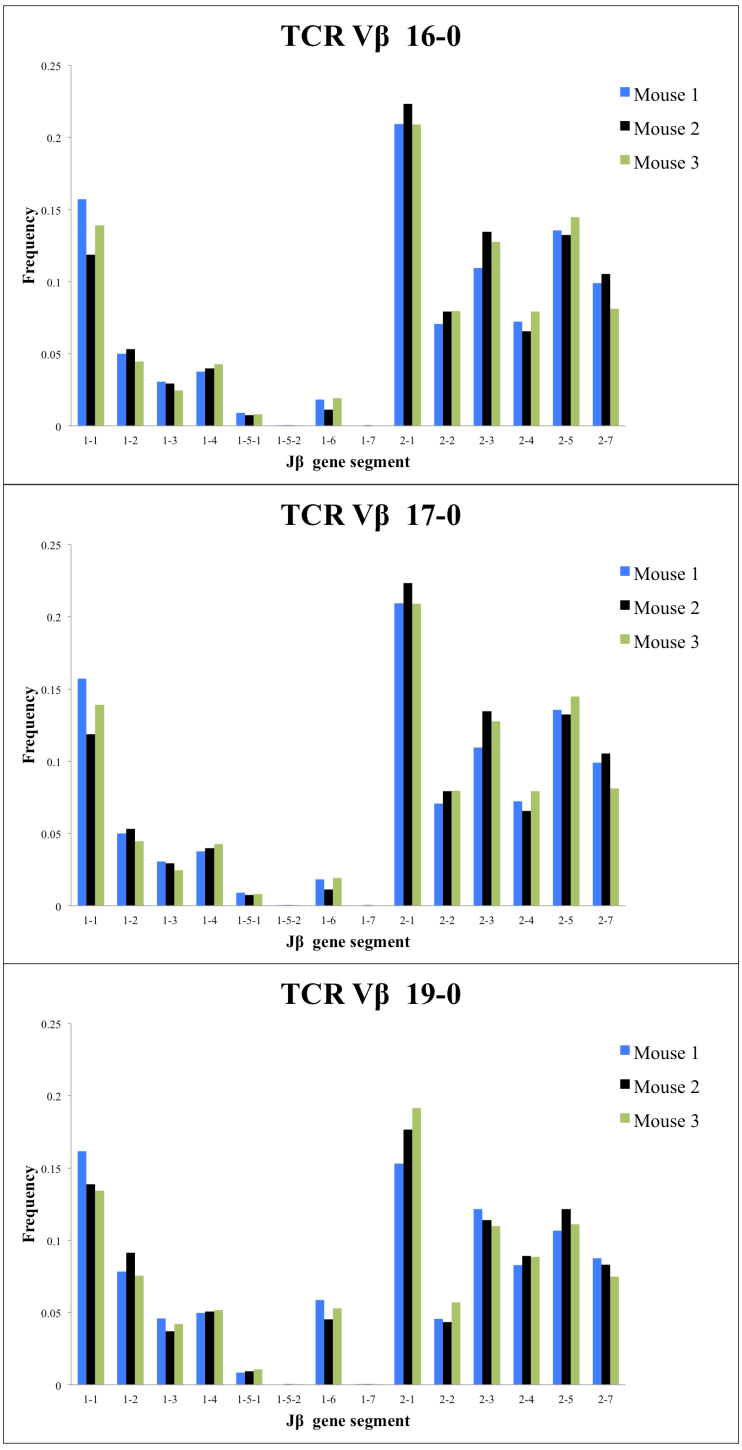

Figure 2.7: TCRβV16, 17, and 19 gene segments display public recombination frequencies between individuals. Analysis of TCRβV16, 17, and 19 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

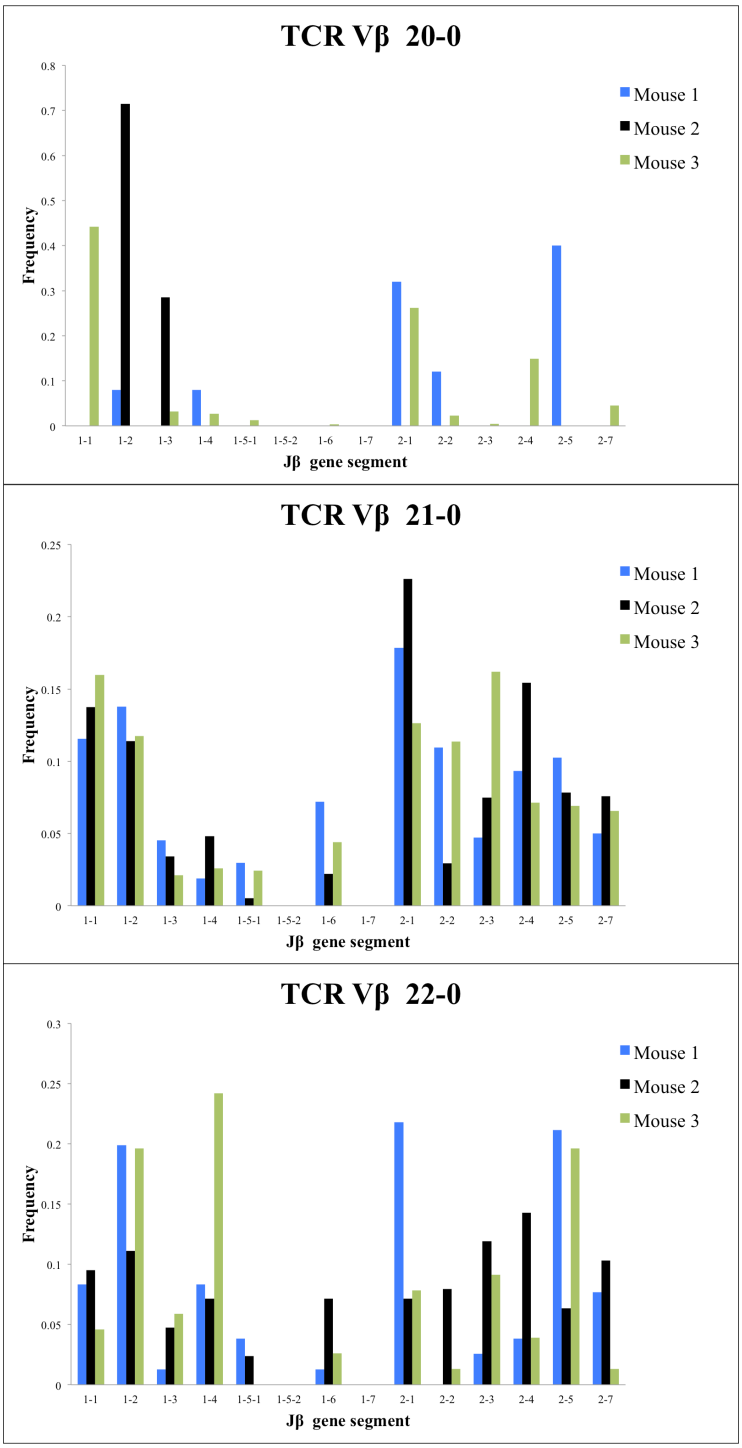

Figure 2.8: TCRβV20-22 gene segments display public recombination frequencies between individuals. Analysis of TCRβV20-22 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

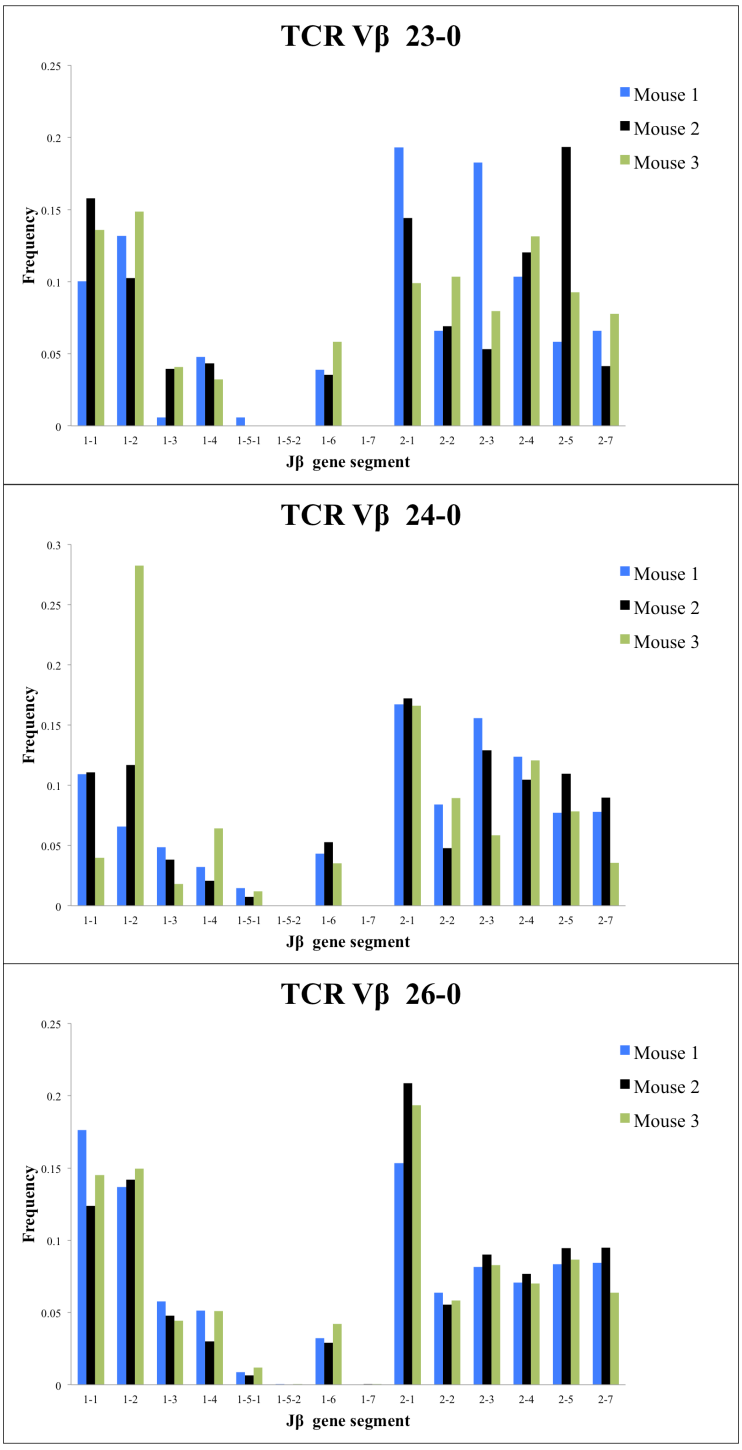

Figure 2.9: TCRβV23, 24, and 26 gene segments display public recombination frequencies between individuals. Analysis of TCRβV23, 24, and 26 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

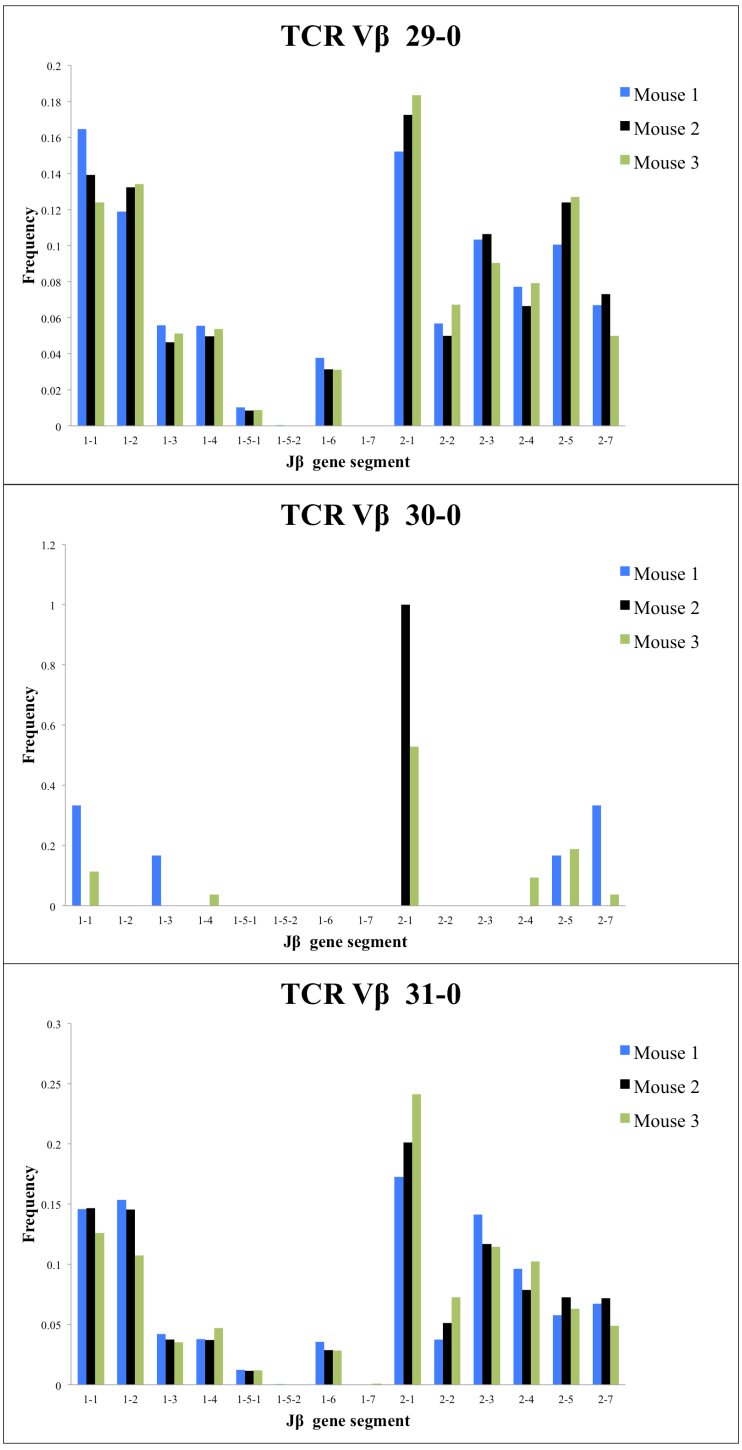

Figure 2.10: TCRβV29-31 gene segments display public recombination frequencies between individuals. Analysis of TCRβV29-31 gene segment recombination frequencies reveals conservation in J gene segment pairing between three individuals.

The majority of TCRβ CDR3 sequences are present at low frequencies while significantly fewer sequences occur at relatively higher frequencies. Previous estimates of diversity have borrowed an approach from ecology, specifically population diversity studies, in which TCR diversity was described by enumerating the number of distinct clones (species) in a given repertoire (population) as well as their respective frequencies. Commonly referred to as the Simpson diversity index [318], these summary approaches compress all the diversity information into a single number. In order to develop a more accurate and comprehensive model to assess a quantitative description of TCR diversity, the frequency distribution of different clone sizes must also be included in the metrics used [230]. By plotting the frequency distribution of TCRβ chain sequences from the mouse naïve $CD8^+$ T cell repertoire we found that a majority of sequences are present at relatively low frequencies, and that a much smaller fraction of sequences occur at relatively higher frequencies **(Figure 2.11)**. It should be mentioned however, that this data only represents TCRβ chain information, and that without paired TCRα information the discrepancy in our estimates of clone size frequencies were most likely even further underestimated.

Specifically, we found that while several clones have very high frequencies (~10%), possibly due to the biased selection of V/D/J gene segments and/or due to convergent selective events, the majority of clones (~90%) are found fewer than 100 times in a given individual.

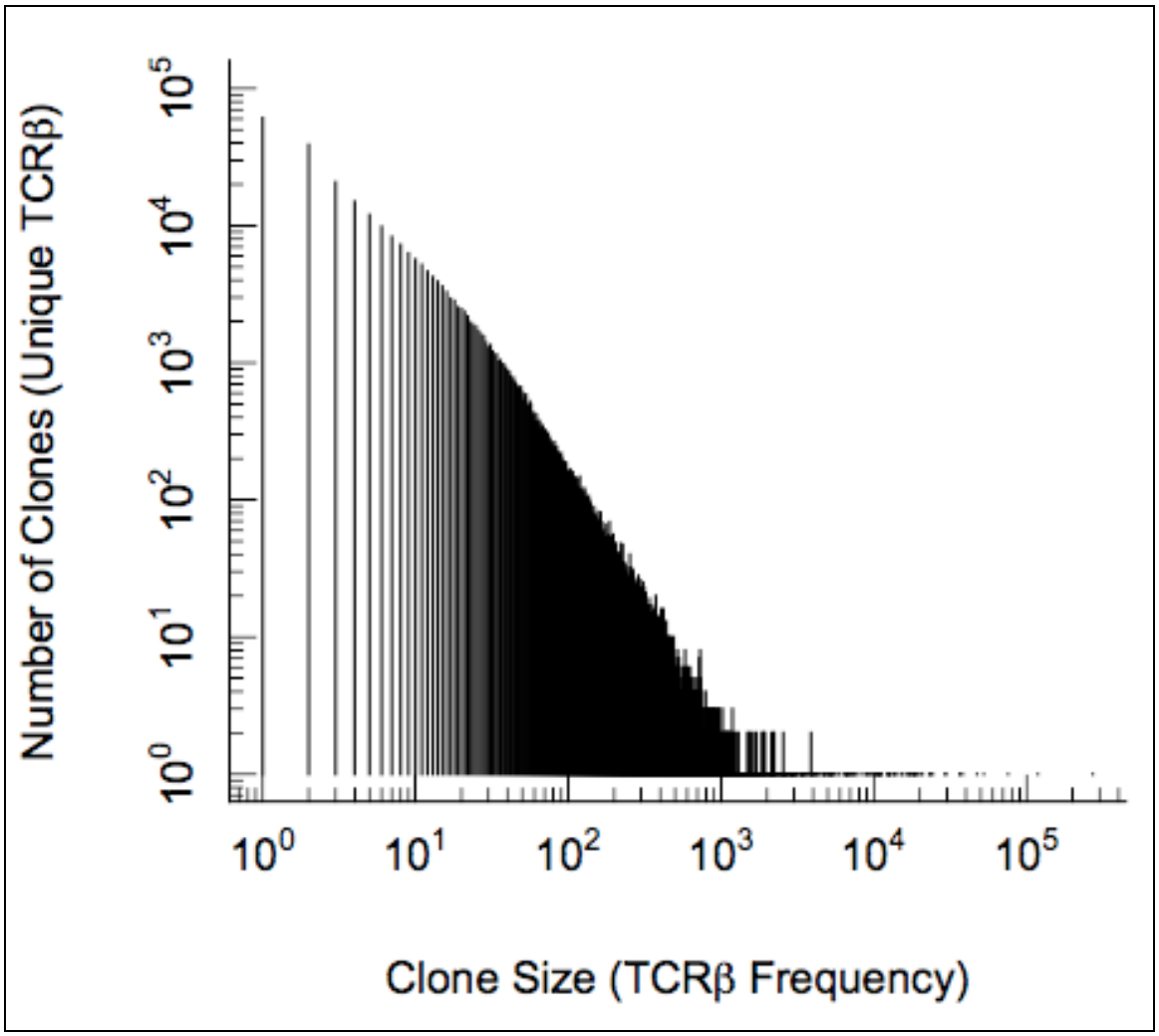

Figure 2.11: Plot of the frequency distribution in the TCRβ CDR3 sequences of naïve CD8$^+$ T cells. CD8$^+$ T cells from C57Bl/6 mice were isolated by magnetic beads and >98% purity confirmed by flow cytometry. Genomic DNA was subjected to TCRβ V-J multiplex DNA sequencing and the distribution of unique in-frame CDR3 sequences is plotted. While few sequences occur at relatively high frequencies, the majority of CDR3 sequences are present at low frequencies (>100 copies). Figure adapted from Schoettle and Blattman's data in Zarnitsyna, 2013 [230].

At the amino acid level and even more pronounced nucleotide level, major differences arise among individuals when analyzing the public vs private TCRβ CDR3 sequence repertoire. While high levels of homology are observed between individuals at the V/J gene segment usage level, this public distribution is drastically diminished when analyzing CDR3 sequences at the amino acid level and even further disparaging at the nucleotide sequence level. Furthermore, while it has been suggested that an adequate sampling of individual TCR repertoires would demonstrate the true prevalence of public TCR sequences [298, 303, 305], our deep sequencing approach to investigate the relative 'publicness' of the TCR repertoire of three genetically identical mice reveal very little true sharing of CDR3 repertoires between individuals.

We found that on average, any two mice in our data set share ~11.5% of the CDR3 amino acid sequences expressed between the two individuals (**Figure 2.12A-C).** Even more revealing, samples between two individuals share only ~2% of nucleotide sequence homology **(Figure 2.13A-C)**. On average, comparison of amino acid sequences between samples 1, 2, and 3 yielded 15,608 shared sequences out of 135,489 total sequences (11.5%) **(Table 2.3)**. From the analysis of nucleotide sequences we observed 12,980 shared sequences from 653,470 total sequences (2.0%) **(Table 2.4)**. Alignment of CDR3 regions was defined by the CASS consensus amino acid sequence. Frame shift and non-productive sequences were excluded. All comparison analysis was performed with ImmunoSEQ data analysis software and standard spreadsheet programs.

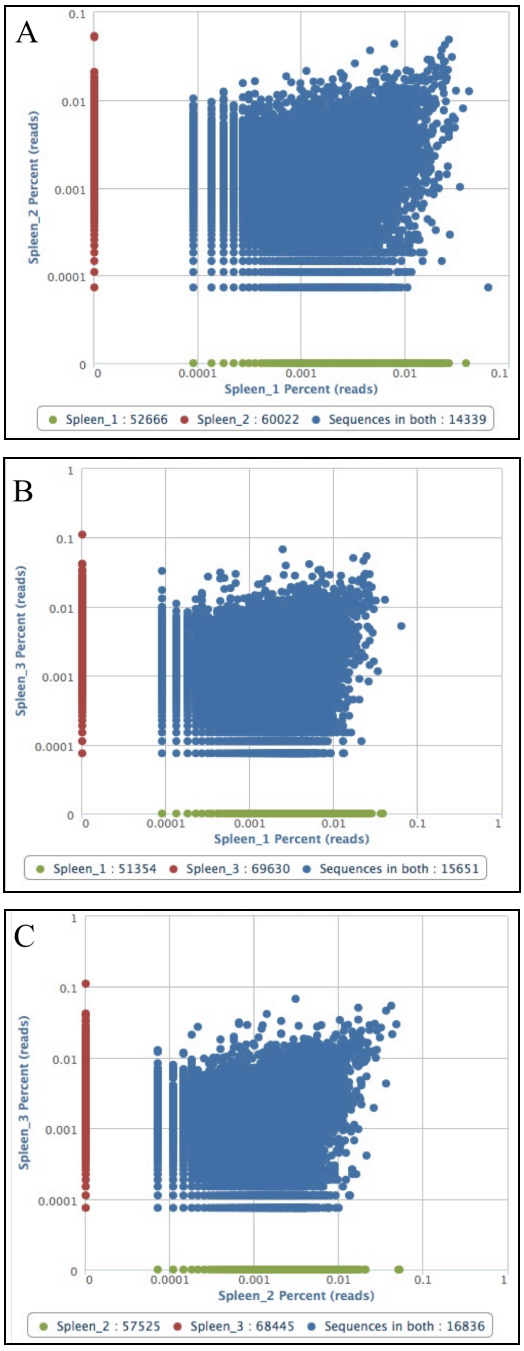

Figure 2.12: A decrease in public TCRβ CDR3 sequence repertoire is observed at the amino acid level. Although relatively conserved at the gene frequency level, only ~11.5% of amino acid sequence homology (blue dots) was observed between three genetically identical C56BL/6 mice. Comparing splenocytes from mouse 1 vs 2 **[A]**, 1 vs 3 **[B]**, and 2 vs 3 **[C],** the majority of discovered CDR3 amino acid sequences (~88.5%) were private to any given mouse (green and red dots).

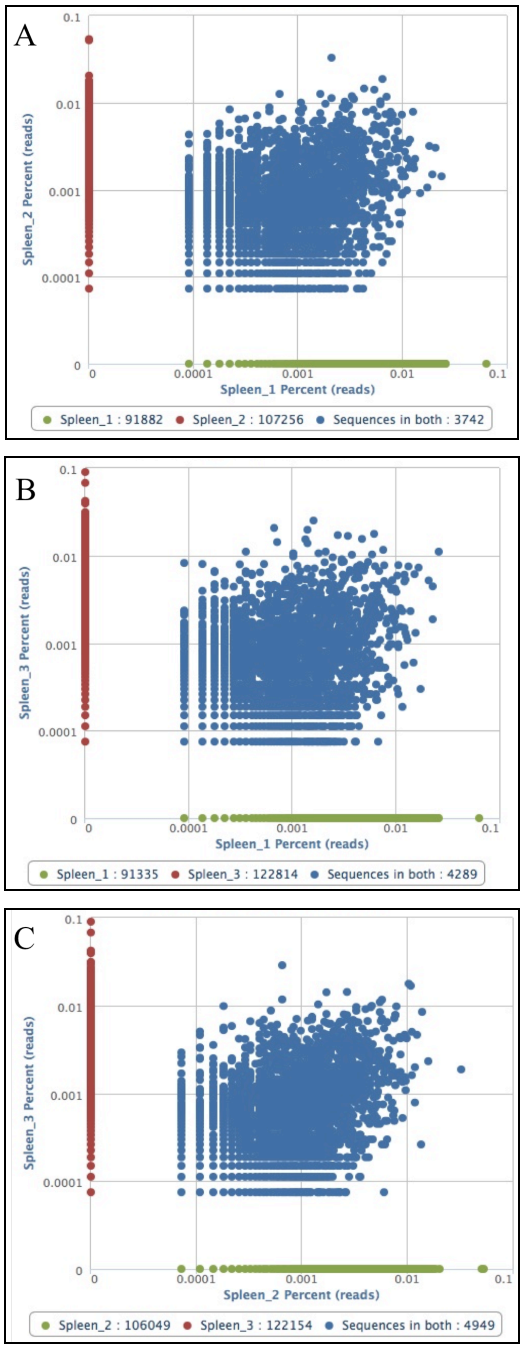

Figure 2.13: A further decrease in public TCRβ CDR3 sequence repertoire is observed at the nucleotide level. While ~11.5% of TCRβ CDR3 amino acid sequence homology was observed between the three individuals, only 2% sequence homology (blue dots) was observed at the nucleotide level. Comparing splenocytes from mouse 1 vs 2 **[A]**, 1 vs 3 **[B]**, and 2 vs 3 **[C],** the majority of discovered CDR3 nucleotide sequences (98%) were private to any given mouse (green and red dots)..

Table 2.3: Comparing public vs private $CD8^+$ T cell CDR3 amino acid sequences between three C57BL/6 mice. Public fraction percentages between compared mice are highlighted in yellow.

	1 vs 2	1 vs 3	2 vs 3
Total # Seq	*127,027*	-	-
1 only	*52,666*	-	-
2 only	*60,022*	-	-
Shared (#)	*14,339*	-	-
Shared (%)	*11.3*	-	-
Total # Seq	-	*136,635*	-
1 only	-	*51,354*	-
3 only	-	*69,630*	-
Shared (#)	-	*15,651*	-
Shared (%)	-	*11.5*	-
Total # Seq	-	-	*142,806*
2 only	-	-	*57,525*
3 only	-	-	*68,445*
Shared (#)	-	-	*16,836*
Shared (%)	-	-	*11.8*

Table 2.4: Comparing public vs private $CD8^+$ T cell CDR3 nucleotide sequences between three C57BL/6 mice. Public fraction percentages between compared mice are highlighted in yellow.

	1 vs 2	1 vs 3	2 vs 3
Total # Seq	*202,880*	-	-
1 only	*91,882*	-	-
2 only	*107,256*	-	-
Shared (#)	*3,742*	-	-
Shared (%)	*1.8*	-	-
Total # Seq	-	*218,438*	-
1 only	-	*91,335*	-
3 only	-	*122,814*	-
Shared (#)	-	*4,289*	-
Shared (%)	-	*2.0*	-
Total # Seq	-	-	*233,152*
2 only	-	-	106,049
3 only	-	-	122,154
Shared (#)	-	-	*4,949*
Shared (%)	-	-	*2.1*

Public TCRβ CDR3 sequences are found in relatively higher frequencies than private CDR3 sequences. Previous studies have reported that public TCRs display a higher level of convergent recombination than private sequences [274, 276, 303, 310]. Additionally it has been shown that public TCR sequences differ from private sequences in their gene segment usage and CDR3 lengths [302]. On average, the more public CDR3 amino acid sequences tend to be roughly one amino acid shorter than those expressed privately in individuals. Additionally, public CDR3 amino acid sequences showed significantly fewer nucleotide insertions and deletions in the V-D and D-J junctions of the TCRβ chain. All of the aforementioned findings indicate that public CDR3 sequences tend to be closer to germ-line DNA configurations [302, 306].

Previous findings have shown that public CDR3 sequences demonstrate a biased and restricted V and J gene segment usage compared to the more inclusive private sequence repertoire, and furthermore that the public V and J gene usage frequencies do not mirror the private gene usage frequencies [302]. To further assess these findings, we next analyzed the frequency of each TCRβ CDR3 amino acid sequence with respect to its degree of sharing between individuals, and we found that public TCR sequences tend to be more abundant than private TCR sequences **(Figure 2.14)**.

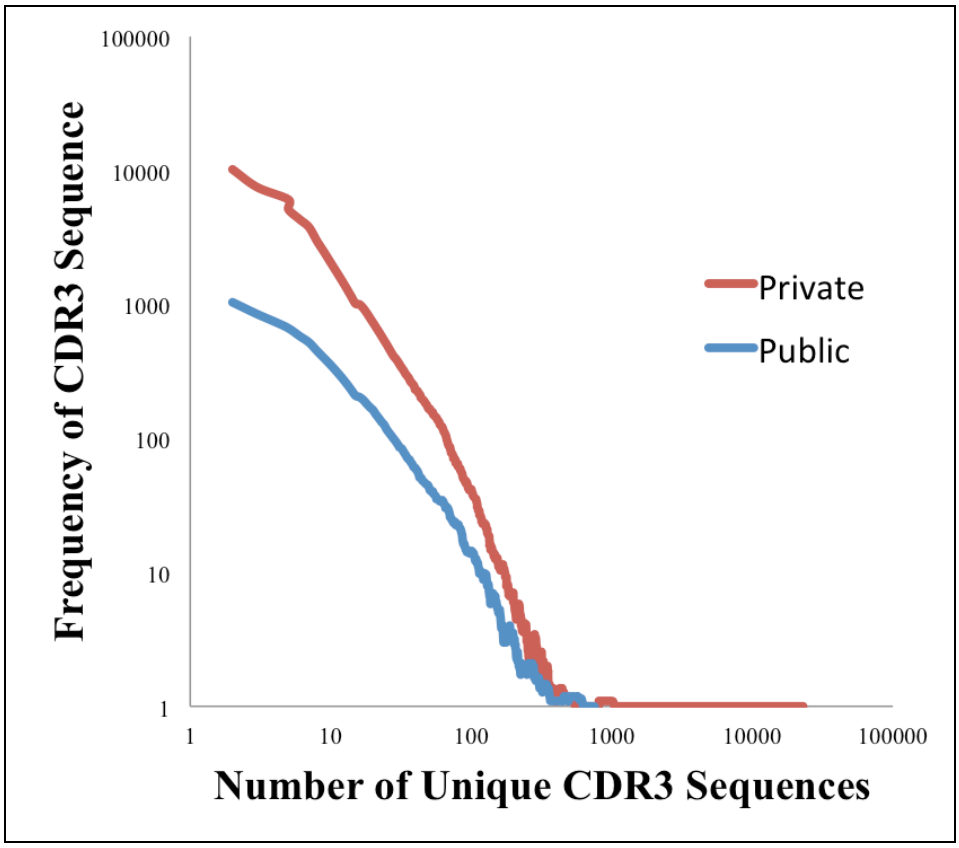

Figure 2.14: The majority of public TCR sequences are found in relatively high frequencies. Comparing frequency (y-axis) to the number (x-axis) of unique TCRβ CDR3 sequences of both public (blue) and private (red) sequences we find that the shared or public fraction of TCR gene sequences correlates with the few TCR gene sequences that are found in high numbers across individuals.

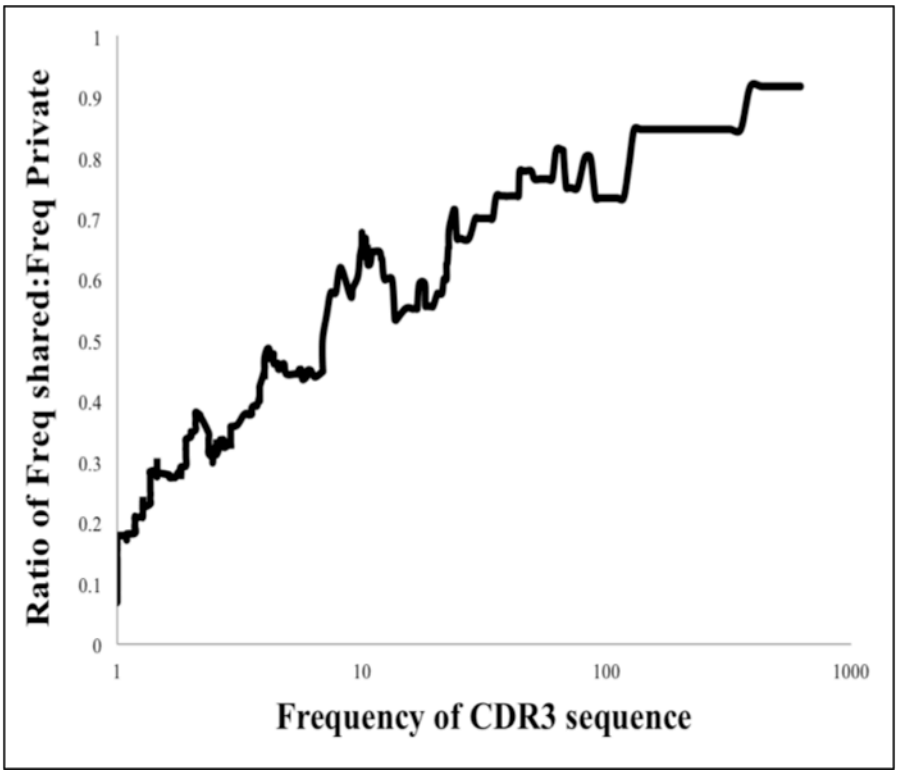

Figure 2.15: Sequences found at higher frequencies tend to have a higher ratio of public:private distribution. As the frequency for specific CDR3 sequences increase (x-axis), the ratio of the number of public:private sequences (y-axis) increases as well. More clearly stated, the more frequently a CDR3 sequence is expressed, the more likely it is to be a public CDR3 sequence shared between individuals. However, CDR3 sequences expressed at relatively low frequencies (<10 copies) tend to be privately expressed by a single individual.

DISCUSSION

Using ImmunoSEQ TCRβ CDR3 sequencing we were able to compare the naïve $CD8^+$ T cell receptor repertoires of three genetically identical mice. We discovered that most sequences are present at very low frequencies (<10 copies/individual) and only a few sequences are present at high frequencies. Furthermore, we compared the relative distributions of V and J gene segment usage between the mice and found that at the whole gene level, V and J expression is conserved between individuals. We also determined that average CDR3 length (12 amino acids) was conserved between the mice both for total reads as well as productive in-frame reads. Additionally, when analyzing V→J recombination frequencies between individuals we also found that the specific V→J recombination frequencies were conserved as well.

However, when analyzing the CDR3 sequences at the amino acid level, divergence between individuals began to emerge. While there was definitely a conserved public fraction of CDR3 sequences shared between individuals (~11.5%), the majority of CDR3 sequences in any individual are private, specific to that individual only. Even more revealing, when analyzing the public CDR3 sequences shared between individuals at the nucleotide level only ~2% of sequence homology was observed. This lack of a true publicly distributed TCR repertoire can most likely be attributed to the fact that the potential repertoire of $>10^{10}$ possible TCRβ CDR3 sequences is much greater than the actual number of cells found in a given individual (mice).

Closer examination of the public vs private fractions of TCRβ CDR3 sequences revealed many additional insights into the attributes of shared CDR3 sequences. The majority of public TCR sequences are found in relatively high frequencies. Comparing frequency to the number of unique TCRβ CDR3 sequences of both public and private sequences we found that the shared or public fraction of TCR gene sequences correlates with the few TCR gene sequences that are found in high numbers across individuals. Furthermore, sequences found at higher frequencies tend to have a higher ratio of public:private distribution. The more frequently a CDR3 sequence is expressed, the more likely it is to be a public CDR3 sequence shared between individuals. However, CDR3 sequences expressed at relatively low frequencies (and thus the majority of CDR3 sequences) tend to be privately expressed by a single individual. These findings bring up additional questions about population dynamics of public vs private immune repertoires.

Our findings, as well as numerous studies to date, highly suggest that the observed sharing of TCR CDR3 sequences between individuals is determined by two key processes: (1) convergent recombination [274, 303], and (2) recombinatorial biases during VDJ gene segment selection [229, 312, 319]. Even though a relatively small fraction of the total repertoire, it is still intriguing that VDJ recombination is not a totally random process, which has the potential to generate a much more diverse repertoire within an individual and also set of repertoires across a population [276]. It has been proposed that the discrepancy between the massively diverse repertoires VDJ recombination is capable of producing and although small, shared fraction of immune repertoire actually found in populations may benefit the species as a whole [276].

Additionally, recombinatorial biases during VDJ selection have also been proposed as a form of natural selection, a co-evolutionary event, to maintain population-wide immunity to the ever evolving smorgasbord of pathogens we are exposed to as a species [320]. Comparing phylogeny of immune repertoire diversity vs pathogen evolution over thousands of years may shed insight into these questions.

Defining the biological utility of public recombinatorial biases may play an additional role in the realm of prophylactics, diagnostics, or even therapeutics. If manipulation of recombinatorial biases were achievable, one could theoretically skew an individual's immunity to specific pathogens. The regulation of VDJ recombination has been analyzed in depth with respect to recombination signal sequences, enhancers, and promoters in the TCR genetic loci [314, 321-323]. Additionally studies have revealed the role of epigenetic factors in the regulation of somatic recombination by altering the chromatin accessibility [324-326]. Additional studies and investigations into how these upstream signals regulate VDJ recombination may potentially provide insight into how TCR repertoires can be manipulated, although the ethics of such manipulation will be a topic of future debate.

As previously mentioned, the analysis of public vs private distributions of TCR repertoires brings up a fundamental question of immunology. Is VDJ recombination somehow pre-determined by genetically programmed processes that are 'immune' to peripheral stressors, or is the development of immune repertoires a responsive evolution to the ever changing immune stressors encountered by the host? As proposed by Li *et al.* [276] and reviewed by many others, could the immune system be utilizing 'adaptive

mutation', the process by which organisms change their genetic information to adapt to encountering different pathogens [327-331]? On one hand, evolution of the immune repertoires over time to specifically combat the pathogens of that era would allow for an evolutionary benefit to populations. However, retention of public fractions of the TCR repertoire would limit diversity, and thus make populations more vulnerable to rare pathogens [276].

It is clear that a large repertoire is required to generate an effective immune response to the highly diverse array of pathogens one may encounter in a lifetime. Studies by our group [230] and others [332, 333] have investigated the relationship between repertoire diversity and protection. To provide reliable protection against a given pathogen there must exist some number of T cell clones present in the naïve repertoire that are specific for the pathogen [230]. The existence of diverse private repertoires across a population plays an important role in defense against new pathogens, but increasing diversity of the private repertoire renders many in the population susceptible to these pathogens. The delicate balance between public vs private immune repertoires ensures that at although many will be susceptible to newly encountered pathogens, at least some individuals in the population will be able to mount an effective immune response and thus survive. Natural selection and the proceeding evolution would then suggest that the once private TCRs would then become part of the public repertoire fraction.

Through the use of TCRβ CDR3 repertoire next generation sequencing, we have demonstrated that there exists VDJ gene usage homology between individuals in the naïve $CD8^+$ TCR repertoire. Furthermore, recombination frequencies between specific V→J genes are also conserved among individuals. However, analysis of the amino acid and nucleotide sequences of TCRβ CDR3 sequences reveals a much smaller public fraction of CDR3 sequences shared between individuals, although this fraction is much larger than would be predicted by stochastic VDJ rearrangement. Finally, analysis of the publicly distributed fraction of sequences reveals these sequences to be present in relatively higher frequency than those found privately. It must be pointed out however, that T cell receptors are heterodimers whose two constituent chains are generated by independent rearrangement events of the TCRα and TCRβ loci. The theoretical diversity of the TCR repertoire is thus significantly increased by the potential pairing of any of the possible α chains with any of the possible β chains. And although it has been estimated that each unique TCRβ chain is on average paired with 25 different TCRα chains [192], this estimate was taken from an extremely small subset of the TCR repertoire and extrapolation of such may include extreme bias. Understanding the balance between immune diversity and immune protection has been limited by a lack of methods for accurately estimating total TCR diversity as this is a function of both sequence diversity within each somatically recombined TCR loci as well as pairing diversity due to combination of different TCRα and TCRβ chains. Single cell sorting and sequencing of both TCR chains for the pre-immune repertoire remains unfeasible while molecular strategies for linking TCR sequences within individual cells have not been adequately

developed. Determining the number of unique TCRαβ combinations in the entire TCR repertoire will require new techniques for linking TCRα and TCRβ sequence information. *In chapter three we outline a template molecule for such an approach.*

CHAPTER 3

DESIGN AND DEVELOPMENT OF DNA NANOSTRUCTURES FOR CAPTURE AND REISOLATION OF TCRA AND TCRB MRNA FROM INDIVIDUAL CELLS

ABSTRACT

DNA origami nanostructures have the potential to become dynamic biomolecular tools. Applications of these nanostructures include use as smart therapeutics, drug delivery systems, and intricate molecular machines. One immediate application is as nanoscale gene chips at the single cell level for interrogating gene expression or sequences. Current tools for single-cell analyses such as single-cell sorting or oil-in-water emulsion droplet sorting are limited by cell throughput, while high-throughput molecular methods for single cell analysis of more than a single gene remain inadequately developed. Here, we demonstrate the use of DNA origami nanostructures as a novel strategy to capture and protect TCRα and β mRNAs from individual cells within large polyclonal cell populations without the need for single-cell sorting. We demonstrate that these nanostructures are highly transfectable, are inherently resistant against intracellular degradation, and due to their modular design can capture mRNA from virtually any gene pair facilitating high-throughput sequence analysis of paired genes from individual cells. While applied here for the capture of TCR mRNAs, this tool has the potential for a wide range of applications including the rapid advancement of immunotherapy for cancer by providing both immune receptor genes from individual T cells, rapid development of monoclonal antibodies, or enhancing personalized approaches to cancer by analysis of progressive mutations to maximize therapeutic efficacy.

IMPORTANCE

Heterogeneity is becoming recognized as a major problem in complex biological systems including immunology and cancer. In particular, advances in single cell analyses have led to the realization that both immune responses and cancer are diverse populations of cells based on either somatic recombination/hypermutation or progressive mutations. Single cell sorting of antigen-reactive lymphocytes can be used for identification of novel immunotherapeutic or experimental antibodies, but this approach remains prohibitively expensive and time-consuming for large cell populations.

One limiting factor in the analysis of cellular gene sequence heterogeneity is that cells must be individually sorted in order to maintain integrity of gene pairing. To our knowledge, the ability to constrain multiple mRNA species from individual cells from large cell populations without first pre-sorting has yet to be demonstrated. We show here the use of DNA origami nanostructures for capture and recovery of mRNA from individual cells without pre-sorting these cell populations. These nanostructures are highly transfectable in primary cell lines using electroporation and are robust against intracellular degradation during the time periods required for mRNA capture. Moreover, the highly modular nature of the DNA origami nanostructures make them easily adaptable to capture virtually any mRNA pairs, and modifications to the nanostructures allows for identification of transfected cells and for purification and reisolation of the nanostructures with bound mRNAs from cell lysate. These nanostructures represent an ideal balance between development of new technological approaches and application of to fundamental biological questions.

INTRODUCTION

The burgeoning field of DNA origami seeks to create molecular structures, tools, or molecular machines out of nucleic acids using predictable Watson-Crick base pairing. Originally, this field was constrained exclusively to the *in vitro* design and construction of these structures. However, recent advancements and successes in the nanotechnology field have led to an expansion in the potential applications for DNA origami nanostructures both *in vivo* and *ex vivo* [267-270].

One major area of interest for application of DNA nanotechnology is for sensing the intracellular environment [334-336] and/or capturing and sequencing mRNAs from individual cells [263, 271]. Such investigations are essential for assessing clonally distributed T cell or B cell dimeric receptor genes in highly diverse immune repertoires or for defining the diversification of cancer cells during disease progression. Previous studies by our group and others have shown the stability of DNA nanostructures in cell lysates [337] as well as serum [338], and Mao *et al*. showed that origami structures were conducive to cellular uptake [269]. The ease of engineering and adaptability of DNA nanostructures, combined with their intracellular stability provides a unique approach for capture, protection and reisolation of multiple mRNA species from individual cells from large heterogeneous cell populations without the need for single cell sorting. This approach has several advantages over current methods, such as single cell sorting followed by deep sequence analysis, for identification and/or analysis of multiple gene sequences from individual cells [259-261] including reduced cost, higher throughput, and the lack of need for specialized equipment. In contrast, conventional molecular

approaches for linking gene sequences, such as by transfection with bridging oligonucleotides, typically result in activation of cellular nucleases that destroy target mRNA and preclude downstream analysis [339]. More recently developed combinatorics-based approaches are useful for high copy number sequences from high frequency clones but provide little to no information on rare sequence variations [340].

DNA origami nanostructures offer an ideal and somewhat hybrid solution to these problems as they can be transfected into individual cells to capture and protect mRNA for specific genes, and using a novel system of barcoding the origami mRNA capture probes the mRNA bound to the DNA nanostructures can be bioinformatically linked following next-generation sequencing. The main innovation of our technology is the development of a highly novel strategy for capture and downstream analysis of multiple gene sequences from single cells within heterogeneous cell populations of massive sample size without the need for single-cell sorting. While this discussion focuses on applying the approach to the rapid identification of murine CD8 T cell receptor sequences, as mentioned above there are potentially a tremendous number of additional applications for this technology. Due to the highly modular nature of the origami molecules, the developed techniques can be (and currently are being) adapted by relatively simple modifications to the DNA origami nanostructure sequences, for the analysis of cells expressing other immunoreceptors (e.g. B cell receptors and novel antibodies, γδ T cell receptors), immune cells from other species, or specific mutations in cancer cells or cancer stem cells. In chapter 3, we describe the development of a robust method for the capture of sequence information from multiple genes from single cells without the need

for single-cell sorting or specialized equipment that is not only cost-effective and has the capacity for analysis of large cell populations, but is also easily adaptable for probing sequence diversity of multiple genes from other heterogeneous cell populations.

MATERIALS AND METHODS

DNA origami design: DNA is an excellent nano-construction material. The development of "scaffolded DNA origami" represented a breakthrough in the field of DNA nanotechnology [341]. In this technique, single-stranded, M13mp18 genomic DNA (~7 kilobases) acts as a scaffold strand that is folded into a target shape (~60x90 nm in scale) as directed by 200+ short, ssDNA "staple" strands. A one-pot nanomolar-scale synthesis produces over 10^{14} origami nanostructures with a folding yield approaching 100%. DNA origami nanostructures are fully addressable molecular pegboards, with more than 200 six-nanometer pixel positions resulting from the unique sequences of each of the staple strands, providing capability for addressable functionalization. Specific staple strands can be substituted to include extended single-stranded "probe" sequences that protrude from the DNA origami structure and are complementary to the selected conserved regions of either the TCRα or TCRβ mRNA coding sequences.

For initial development and validation of this approach, we used the P14 TCR transgenic mouse line that expresses a TCR specific for the glycoprotein 33 (GP33) peptide from the lymphocytic choriomeningitis virus (LCMV) (discussed below). Thus, selected staple strands were extended with complementary sequences to the CDR3-

proximal conserved regions of TCRα and TCRβ. Site-directed attachment of fluorescent tags (FITC) facilitated gel analysis and detection of successfully transfected cells by flow cytometry. Biotin tags were included on some staple strands for subsequent isolation and purification of transfected DNA origami nanostructures from cell lysates with bound TCR mRNA. It is important to note that more than one hundred trillion probe tiles can be fabricated in a one-pot annealing step (within 12 hours), which renders this method suitable for high-throughput applications such as next generation sequencing. Additionally, our group previously demonstrated that DNA origami structures retain their structure and are resistant to nuclease digestion in cell lysate [337]. Thus, it is possible to use these structures for mRNA capture and analysis due to their inherent intracellular stability.

DNA origami synthesis: The single-stranded M13mp18 bacteriophage genome (7.4 kb) was purchased from a commercial vendor (Affymetrix). All oligonucleotide staples were synthesized by and procured from a commercial vendor (IDT) and mixed in equimolar amounts resulting in a master pool. While standard staples are mixed in excess, the TCRα/β mRNA capture probe strands and biotin-tagged staples were PAGE-purified and quantified (by measuring absorbance at 260 nm) prior to being included in the origami folding reaction. Scaffold M13 ssDNA, staple DNA, and functionalized staples were then mixed in a fixed 1:5:3 ratio (50 nM scaffold, 250 nM staples, 150 nM mRNA probes/biotinylated tags/fluorophore tags) in aqueous buffer (1x TAE with 12 mM Mg^{2+}), followed by thermal denaturation (90°C) and gradual annealing (to 20°C) over 12 hours. Folded DNA origami was purified from non-folded products and excess

staples by centrifugation through 100K nominal molecular weight limit (NMWL) Amicon filters. Such purification typically results in a solution of 20-50 nM of target DNA origami nanostructures; the final concentration can be measured by A_{260}/A_{280} absorbance and standardized to either 20 or 50nM depending on the downstream application. The DNA origami nanostructures were also visualized by atomic force microscopy (AFM) to verify proper folding of the designed structures.

Mice: 6-8 week old C57BL/6 mice were purchased from Jackson laboratories (Bar Harbor, ME) and bred in our ASU animal facilities. P14 transgenic mice, in which CD8 T cells express TCR specific for the D^bGP33- 41 epitope of LCMV, were obtained initially from Dr. Rafi Ahmed and bred in our animal facilities. All mice were maintained under specific-pathogen free conditions at The Biodesign Institute and experiments were performed in compliance with institutional guidelines as approved by the Institutional Animal Care and Use Committee of Arizona State University.

***in* vitro Transcription of P14 TCRα and β mRNA:** DNA origami nanostructures were first optimized for binding to TCRα and β mRNA sequences using *in vitro* transcribed mRNA. Spleens from 4-6 week old P14 TCR-transgenic mice were prepared by mechanical disruption followed by red blood cell lysis in 0.83% NH_4Cl buffer. $CD8^+$ T cells were separated from total splenocyte cell suspensions by magnetic cell separation (MACS, Miltenyi Biotec) and >95% purity of sorted populations confirmed by flow cytometry. Total RNA was isolated using standard endotoxin free RNA extraction kits (Qiagen) and quality assessed by measuring A_{260}/A_{280} absorbance

ratios (optimally between 1.8-2.0). To generate plasmids for production of large amounts of *in vitro* transcribed TCR mRNA, P14 TCRα and β cDNA was prepared using a Qiagen OneStep RT-PCR kit with P14 TCRα or TCRβ specific primers. RT-PCR products were sequenced and confirmed to be P14 TCR cDNA and then further amplified with AmpliTaq Gold PCR kits to create 3'-A overhangs necessary for cloning into pCR4-TOPO plasmids (Invitrogen). Resulting plasmids were sequenced to confirm correct insertion of the PCR product. TCRα and TCRβ mRNA was then synthesized from EcoRI linearized plasmids by *in vitro* transcription using the T7 phage polymerase priming site and the Invitrogen MEGAscript T7 kit. Transcribed mRNA was purified with Invitrogen MEGAclear kits followed by assessment of concentration and purity (A_{260}/A_{280}=1.8-2.0).

DNA origami binding to immunoreceptor mRNA: P14 TCRα and TCRβ mRNA (0.5 μg/μL) was denatured for 5 minutes at 65°C and incubated with purified origami (20nM) in 1X TAE-Mg^{2+} so hybridization could occur (37°C, 1 hour). Negative controls included: 1) DNA origami ***without*** TCR-specific staples mixed with *in vitro* transcribed P14 TCRα and TCRβ mRNA and 2) DNA origami ***with*** TCR-specific staples mixed with anti-sense P14 TCRα and TCRβ mRNA generated using T3 polymerase from the opposite orientation on the pCR4 plasmid. To directly visualize mRNA binding, hybridized product was analyzed by AFM. Furthermore, analysis of gel migration shifts due to mRNA binding was assessed by agarose gel electrophoresis of incubation products.

DNA origami transfection into primary T cells: Previous studies have utilized receptor-mediated endocytosis of DNA nanostructures via the caveolin-dependent pathway leading to microtubule transport to lysosomes and therefore breakdown of the nanostructure. The obvious problem with this strategy is that for mRNA capture and subsequent reisolation, our nanostructures must bypass the endocytosis pathway, and instead cross the membrane directly to the cytoplasm. Electroporation is the simplest method for cytoplasmic entry and thus avoidance of the deleterious effects of endosomal degradation. However high cell mortality rates have commonly been associated with this methodology. Recent advancements in microporation technology (the use of μL volume electroporations) have shown high transfection efficiencies (>93%) as well as high cell viabilities (>86%). For our T cell transfection experiments we used P14 TCRα/β transgenic mice. Splenocytes from 4- to 6-week-old P14 TCRα/β transgenic mice were prepared by mechanical disruption and red blood cell lysis in 0.83% NH_4Cl. $CD8^+$ T cells were purified by magnetic cell sorting and >95% purity of sorted populations were confirmed by flow cytometry. Cells were pelleted by centrifugation (1,200 rpm, 5 minutes, 4°C), washed with MACS media (Miltenyi Biotec), and resuspended in MACS media at $1x10^7$ cells/mL. For electroporation, a both the BTX Cuvette electroporation system (Harvard Apparatus) and the Neon syringe transfection system (Thermo Scientific) were used. Settings for the BTX cuvette system consisted of 2 mm gap cuvettes, 300 V, 10 ms, and one pulse. Settings for the Neon system consisted of 100 μL syringe tips, 2,000 V, 10 ms, 1 pulse. Samples consisted of 100μL cell suspensions and either 25 μL (50 nM) DNA origami suspension (in 1X TAE-Mg^{2+}) or a mock transfection

control of 25 μL 1X TAE-Mg^{2+} buffer. Immediately following electroporation, cells were transferred to 100 μL fresh RPMI-1640 culture medium with 10% fetal calf serum and incubated at 37°C for 2-24 hours in individual wells of a 96 well plate. To assess transfection efficiency, cells were visualized on a LSR Fortessa flow cytometer after successive incubation periods; the fluorescein isothiocyanate (FITC; 488 nm excitation, 518 nm emission) tag incorporated into each nanostructure allowed for successfully transfected cells to be identified by flow cytometry. To verify that the DNA origami nanostructures had actually entered transfected cells, rather than binding non-specifically to the cell surface, we also included DNase pre-treatment (Turbo DNase, Ambion, Life Technologies) of transfected cells followed by FACS analysis. Furthermore, we verified that DNase resulted in destruction of DNA origami by incubating refolded origami in 1X Turbo DNase (37˚C, 30 min) followed by electrophoresis on a 1% agarose gel. It should be noted that Turbo DNase is a proprietarily engineered version of DNaseI and has a markedly higher affinity for DNA than wild type DNaseI, which we found to be much less effective at degradation of origami nanostructures.

Intracellular TCR mRNA binding, cell lysis, and origami purification: DNA origami with extended mRNA capture probes specific for TCRα and TCRβ conserved regions was purified from transfected P14 transgenic T cells by utilization of the four biotinylated origami staple strands. At 12 hours post-transfection, cells were lysed in 100μL 1% NP-40 buffer including 2μL Ribolock RNase inhibitor for 30 minutes on ice. DNA origami with bound TCR mRNA was then isolated from cellular debris and any unbound mRNA by filtration through a streptavidin resin column (Pierce Streptavidin

UltraLink Resin, Sigma Prep Columns, 7-20μM pore size). Unbound cellular debris and mRNA was washed through the column using 3x350μL 1X TAE-Mg2+ rinses followed by centrifugation at 2,000 rpm for 30s. Specific capture of TCRα and TCRβ mRNA by DNA origami with probe staples was confirmed by Sanger sequencing of RT-PCR products and comparison to known P14 CDR3 sequences (discussed below).

Reverse transcription and P14-specific TCR chain CDR3 amplicon generation: Once mRNA-bound origami was purified from cell lysate in the streptavidin columns, reverse transcription (RT) was performed directly in the columns. Using the 5'→3' TCRα and TCRβ mRNA capture probes as RT primers, elongation was achieved by simply adding 40μL of the RT mastermix (Omniscript, Qiagen) directly into the purification column, and incubating for 1 hr in a 37°C heat block. The RNA was then removed by addition of an RNaseH cocktail (NEB), and TCRα/β cDNA was then eluted by heating the column to 95°C via heat block for 5 min, and centrifuging at 10,000 rpm for 5min. Finally, a standard polymerase chain reaction (PCR) using a single primer for each of the conserved priming sequences on either end of the mRNA capture probes, and additional single primers for the known P14 TCRα and β V-families (P14 TCR-Vα2 and TCR-Vβ8.1 genes) was performed to generate a pool of amplicons with corresponding P14 CDR3 regions within the resulting PCR product. Confirmation of the correct P14 TCRα and TCRβ genes was confirmed by standard dye termination sequencing analysis of gel-isolated products.

RESULTS

AFM imaging confirms successful synthesis of DNA origami nanostructures. The nanostructure template chosen for our studies was the well-studied 2D rectangle first published by Rothemund in 2006 and was assembled and purified following previously described methods [341]. As the first application of our technology and what is presented here is for immune repertoire analysis, we designed our structures for capture of clonally distributed T cell receptor α and β chain mRNAs. For mRNA capture, each DNA origami molecule was refolded with selected staple strands extended to include a 3' region complimentary to the constant region of either TCRα or TCRβ mRNA allowing for capture of up to 6 copies of each mRNA **(Figure 3.1A)**. Additionally, four staple strands were conjugated to biotin for downstream purification and one/two strands were conjugated with fluorescein isothiocyanate (FITC) and/or 5-carboxytetramethylrhodamine (TAMRA) for identification of successfully transfected cells **(Table 3.1)**. Successful folding of DNA origami was verified by atomic force microscopy **(Figure 3.1B)**.

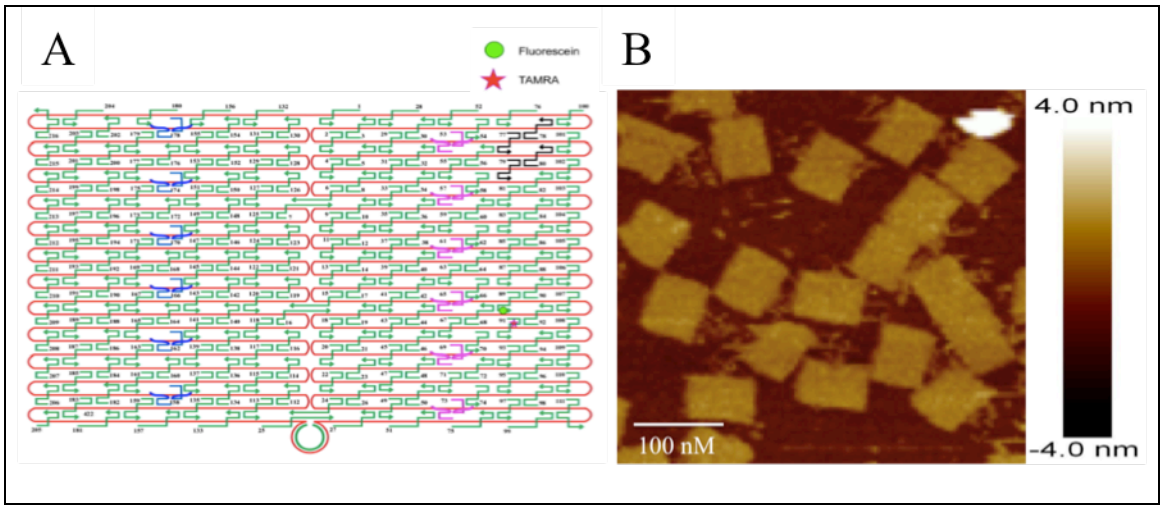

Figure 3.1: DNA origami design and synthesis. **[A]** Basic DNA origami design showing circular M13 ssDNA (red) folded into a rectangular shape by annealing with "staple" primers (green), complementary to the M13 ssDNA. Selected staples are extended with complementary sequences to TCRα and TCRβ conserved regions (pink and blue, respectively) for binding mRNA. Other staples are biotinylated (black) for purification, or have fluorescent probes (green dot, red star) attached for transfection analysis. **[B]** Validation of properly folded DNA origami molecules was visualized by AFM, demonstrating the anticipated rectangular shape of the nanostructure. Each nanostructure is roughly 60 x 90 nM is size.

Table 3.1: Origami nanostructure staple modifications. Original sequences from [341], modifications include additions of gene-specific capture probes (A or B), biotin tags (/3Bio/), and fluorescent markers (/3Fitc/ or /3Tamra/).

Staple ID	Modification	Original Sequence	Modified Sequence
73	Alpha probe	GCCACGCTATACGTGGCA CAGACAACGCTCAT	A'-73-1 GCCACGCTATACGTGGTTTGAAGATATCTTG A'-73-2 GGTGGCGTTGGTCTCCACAGACAACGCTCAT
69	Alpha probe	GCGCAGAGATATCAAAAT TATTTGACATTATC	A'-69-1 GCGCAGAGATATCAAATTTGAAGATATCTTG A'-69-2 GGTGGCGTTGGTCTCATTATTTGACATTATC
65	Alpha probe	CATATTTAGAAATACCGA CCGTGTTACCTTTT	A'-65-1 CATATTTAGAAATACCTTTGAAGATATCTTG A'-65-2 GGTGGCGTTGGTCTCGACCGTGTTACCTTTT
61	Alpha probe	TTTTGTTTAAGCCTTAAAT CAAGAATCGAGAA	A'-61-1 TTTTGTTTAAGCCTTATTTGAAGATATCTTG A'-61-2 GGTGGCGTTGGTCTCAATCAAGAATCGAGAA
57	Alpha probe	AATCACCAAATAGAAAAT TCATATATAACGGA	A'-57-1 AATCACCAAATAGAAATTTGAAGATATCTTG A'-57-2 GGTGGCGTTGGTCTCATTCATATATAACGGA
53	Alpha probe	CCTCAAGAATACATGGCT TTTGATAGAACCAC	A'-53-1 CCTCAAGAATACATGGTTTGAAGATATCTTG A'-53-2 GGTGGCGTTGGTCTCCTTTTGATAGAACCAC
158	Beta probe	AGTTTGGAGCCCTTCACC GCCTGGTTGCGCTC	B'-158-1 AGTTTGGAGCCCTTCAGTGTGACAGGTTTGG B'-158-2 CTGCACTGATGTTCTCCGCCTGGTTGCGCTC
162	Beta probe	CAGCTGGCGGACGACGAC AGTATCGTAGCCAG	B'-162-1 CAGCTGGCGGACGACGGTGTGACAGGTTTGG B'-162-2 CTGCACTGATGTTCTACAGTATCGTAGCCAG
166	Beta probe	GGTAGCTAGGATAAAAAT TTTTAGTTAACATC	B'-166-1 GGTAGCTAGGATAAAAGTGTGACAGGTTTGG B'-166-2 CTGCACTGATGTTCTATTTTTAGTTAACATC
170	Beta probe	TACCTTTAAGGTCTTTACC CTGACAAAGAAGT	B'-170-1 TACCTTTAAGGTCTTTGTGTGACAGGTTTGG B'-170-2 CTGCACTGATGTTCTACCCTGACAAAGAAGT
174	Beta probe	TTTCAACTATAGGCTGGC TGACCTTGTATCAT	B'-174-1 TTTCAACTATAGGCTGGTGTGACAGGTTTGG B'-174-2 CTGCACTGATGTTCTGCTGACCTTGTATCAT
178	Beta probe	ATATATTCTTTTTTCACGT TGAAAATAGTTAG	B'-178-1 ATATATTCTTTTTTCAGTGTGACAGGTTTGG B'-178-2 CTGCACTGATGTTCTCGTTGAAAATAGTTAG
77	Biotinylation	TGCTCAGTCAGTCTCTGA ATTTACCAGGAGGT	TGCTCAGTCAGTCTCTGAATTTACCAGGAGGTTTTTT/3 Bio/
78	Biotinylation	GGAAAGCGACCAGGCGG ATAAGTGAATAGGTG	GGAAAGCGACCAGGCGGATAAGTGAATAGGTGTTTTT /3Bio/
79	Biotinylation	TGAGGCAGGCGTCAGACT GTAGCGTAGCAAGG	TGAGGCAGGCGTCAGACTGTAGCGTAGCAAGGTTTTT/ 3Bio/
80	Biotinylation	TGCCTTTAGTCAGACGAT TGGCCTGCCAGAAT	TGCCTTTAGTCAGACGATTGGCCTGCCAGAATTTTTT/3 Bio/
89	FITC	AGAGGCATAATTTCATCT TCTGACTATAACTA	AGAGGCATAATTTCATCTTCTGACTATAACTA/3Fitc/
91	TAMRA	TATGTAAACCTTTTTTAAT GGAAAAATTACCT	TATGTAAACCTTTTTTAATGGAAAAATTACCT/3Tamra/

Successful cloning and *in vitro* transcription of the TCR transgenic P14 TCRα and β mRNA was confirmed by sequencing. The P14 transgenic mouse generates $CD8^+$ T cells that express only TCR specific for the DbGP33-41 co-dominant epitope of LCMV. The P14 animal is widely used and gene sequencing of its TCRα and β chains has been well documented [342]. Specifically, the P14 animal is a TCRα knockout that is homozygous for a transgene encoding a Vα2/Vβ8.1 TCR specific for the LCMV peptide mentioned above. As such, it is deficient in the TCRα gene and therefore does not develop endogenous mature αβ T cells. Virtually all of the peripheral T cells are $CD8^+$ and will express the transgenic TCR **(Figure 3.2A)**. The mice are commonly used for *in vitro* studies of $CD8^+$ T cell activation and differentiation, as well as a source for adoptive transfer of LCMV-specific T cells into wild type C57BL/6 mice. For our studies, the use of this model is especially valuable in order for proof-of-principle experiments to be conducted without the additional challenges of heterogeneity in captured sequences. Once plasmid vectors were constructed and *in vitro* transcription of large pools of both TCRα and TCRβ mRNA was conducted and confirmed by standard Sanger sequencing for homology to the known Vα2/Vβ8.1 P14 TCR sequences **(Figure 3.2B)**.

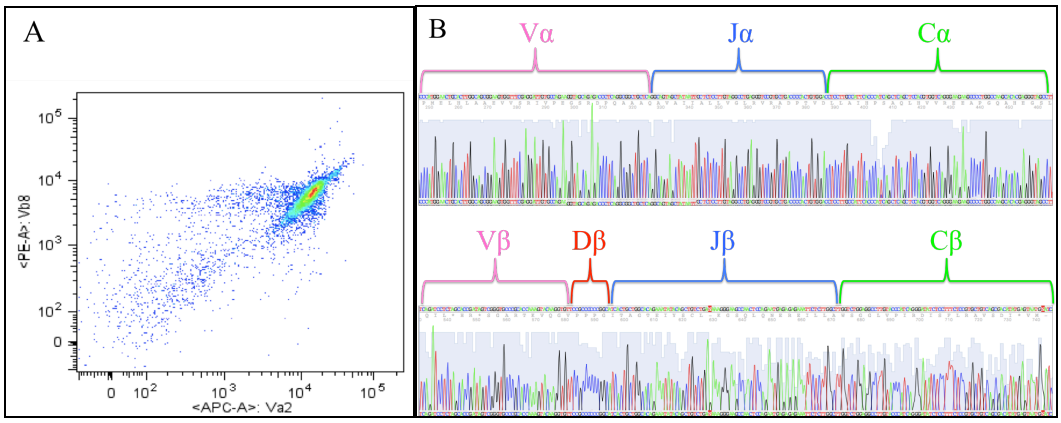

Figure 3.2: Sequencing confirms successful *in vitro* transcription of transgenic P14 TCR mRNAs. **[A]** Flow cytometry of sorted $CD8^+$ T cells from spleens from P14 mice demonstrate >99% of $CD8^+$ T cells to express the transgenic Vα2/Vβ8.1 TCR. **[B]** Sequencing traces of both TCRα (top) and TCRβ (bottom) CDR3 regions from RT-PCR products *in vitro* transcribed P14 mRNAs. 100% homology was observed between reported P14 sequences and traces generated from *in vitro* transcribed mRNA RT-PCR products.

DNA origami with TCR mRNA-specific probes can specifically hybridize with *in vitro* transcribed TCR mRNA. In order to optimize origami:mRNA binding, and for *in vitro* validation, P14 TCR transgenes encoding either TCRα or TCRβ genes were cloned into expression vectors and mRNA was *in vitro* transcribed. Secondary structures of the conserved C-domain of both TCRα and TCRβ mRNA were predicted using available software (mfold) to identify accessible regions within the first C-region exon for origami-probe annealing. Origami-probe sequences were designed to be complimentary to these sites located 3' to the CDR3 sequence of interest **(Figure 3.3A-B and Table 3.1)**. Specific binding of both TCRα and TCRβ mRNA by origami constructed with both TCR mRNA-specific probes was observed by atomic force microscopy **(Figure 3.4A)**. No binding of mRNA was observed for DNA origami refolded without TCR-specific capture staples mixed with TCRα and TCRβ mRNA **(Figure 3.4B)** nor was binding observed for DNA origami with TCR specific staples mixed with "reverse sequence" TCRα + TCRβ mRNA. Furthermore, incubation of either TCRα or TCRβ mRNA alone with origami constructed with both TCRα and β probes yielded origami with binding of mRNA to only one side of the nanostructure, demonstrating mRNA-specific binding of both probes **(Figure 3.4C-D)**. Additionally, under these saturating conditions, we observed specific binding of origami nanostructure to TCR mRNA by changes in mobility during electrophoresis of 20μL of each sample on a 1% non-denaturing agarose gel **(Figure 3.5A)**.

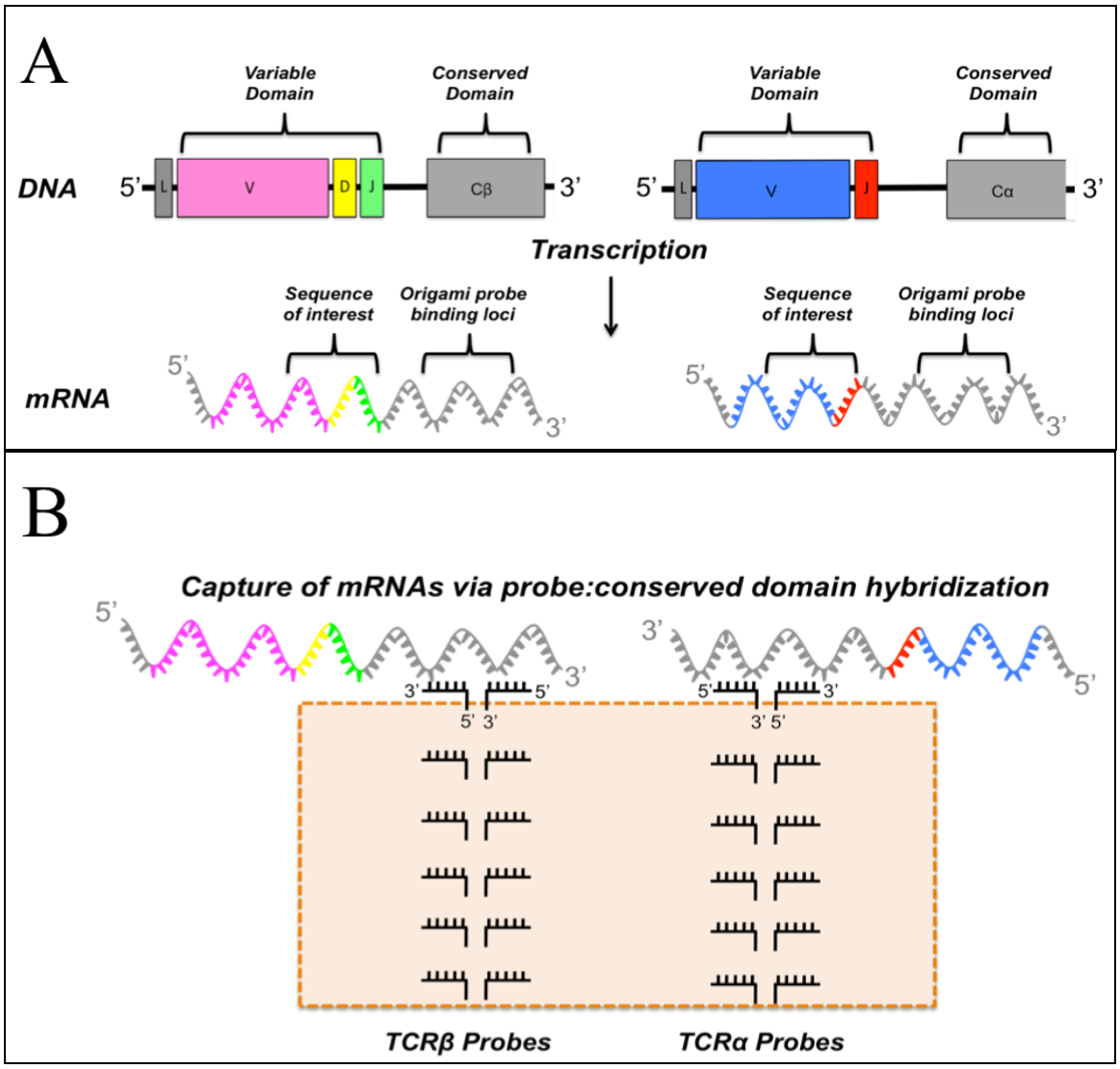

Figure 3.3: Extended mRNA capture probes on DNA origami scaffold are designed to hybridize to conserved regions of either TCRα or TCRβ mRNA. **[A]** Organization of TCRβ (left) and TCRα (right) genes including a 5' variable region containing the heterogeneous CDR3 sequences of interest followed by a 3' conserved region serving as the origami probe hybridization loci. **[B]** Graphic representation of both TCRα and TCRβ mRNA transcripts hybridized to origami probes. 5'→3' orientation of both mRNA transcripts as well as both sets of origami probes is indicated. As mentioned above, each DNA origami molecule was designed to include 6 sets of mRNA capture probes on each side for each gene allowing for capture of up to 6 mRNAs of each species, however to simplify the illustration the origami molecule is depicted binding to a single TCRα and TCRβ mRNA.

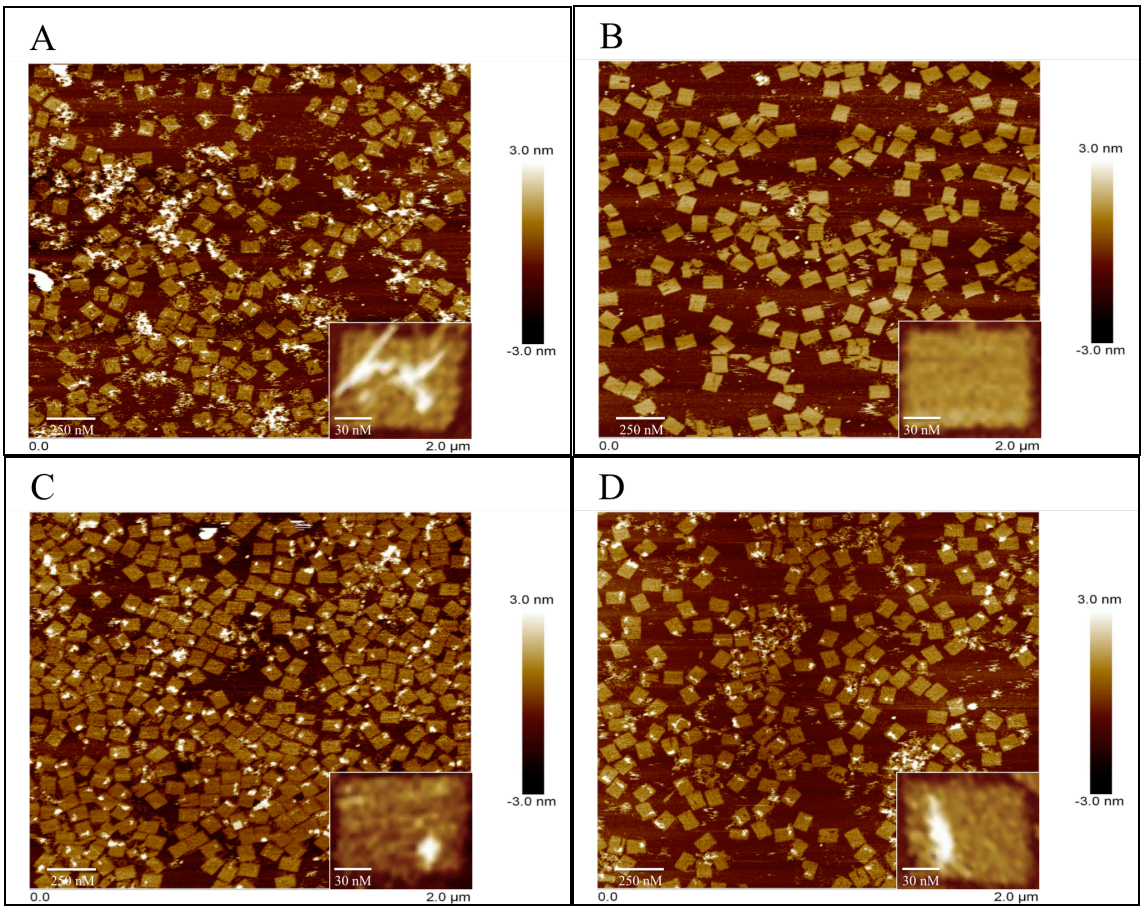

Figure 3.4: DNA origami binding to TCRα/β mRNA visualized by atomic force microscopy. TCRα and β mRNAs were in vitro transcribed and incubated with DNA origami constructed with or without TCR probes. **[A]** Wide view AFM image of origami constructed with both TCRα/β probes incubated with both TCRα and β mRNAs. Visual binding of RNAs to both sides of numerous origami nanostructures demonstrates the ability of the nanostructures to bind both TCRα and TCRβ mRNAs simultaneously. **[B]** Wide view AFM image of origami constructed *without* mRNA probes incubated with both TCRα and β mRNAs illustrates no detectable non-specific binding of mRNA by the origami nanostructures. **[C]** Wide view AFM image of origami constructed with both TCRα and TCRβ probes incubated with TCRα mRNA only. **[D]** Wide view AFM image of origami constructed with both TCRα and TCRβ probes incubated with TCRβ mRNA only. Binding of mRNA to only one side of each nanostructure in C-D indicates that the alpha and beta probes are specific for binding to only alpha or beta mRNA respectively.

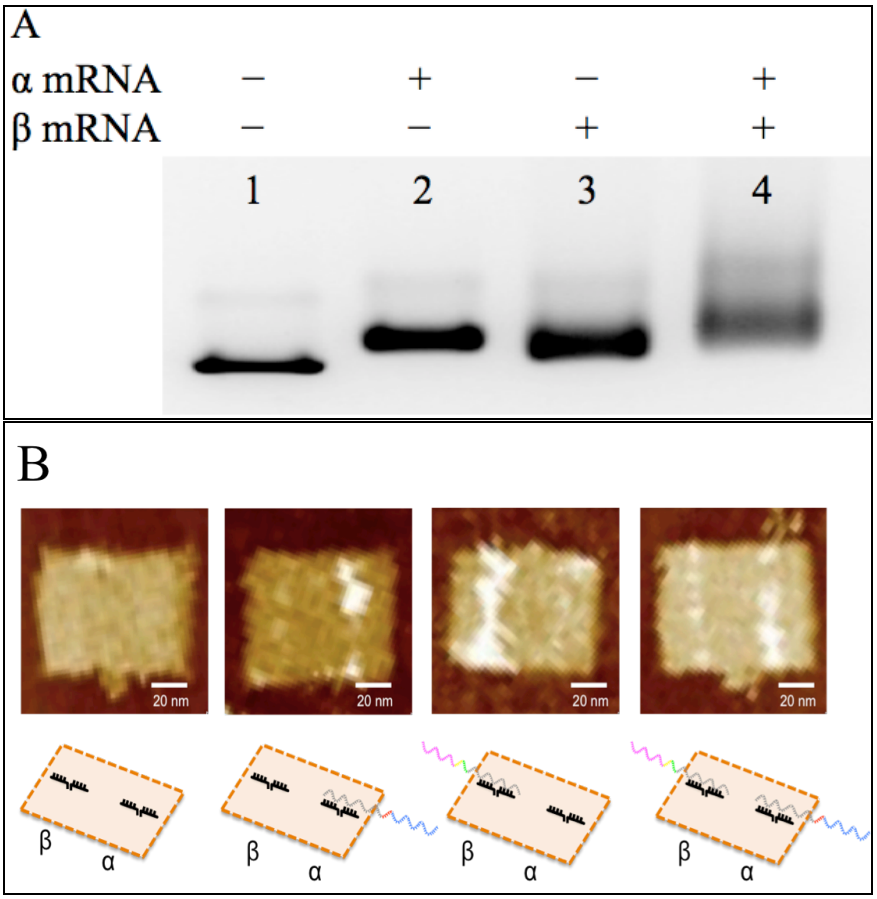

Figure 3.5: Gel electrophoresis of products from origami incubated with various *in vitro* transcribed mRNA combinations. TCRα and β mRNAs were *in vitro* transcribed and incubated with origami at 37°C. **[A]** Products of various mRNA:origami combinations were analyzed by agarose gel electrophoresis; Lane 1 - Origami only, Lane 2 - Origami with alpha mRNA only, Lane 3 - Origami with beta mRNA only, Lane 4 - Origami with both alpha and beta mRNA. **[B]** Selected AFM images were captured of incubation products from origami alone (left), binding to individual alpha or beta mRNA (center two, respectively), or both mRNAs (right) with corresponding graphic illustrations below.

Electroporation of DNA origami nanostructures routinely yields high transfection efficiencies. While the majority of the previously mentioned DNA nanostructure studies relied on simple cellular uptake of the nanostructures either with or without the aid of lipid-based transfection reagents, we found that electroporation of cell suspensions with DNA origami nanostructures yields robust transfection efficiencies while maintaining high cell viability. To assess transfection efficiency, cells were stained with anti-CD8-APC antibody (1:200 dilution, BD Biosciences) and after washing twice in FACS buffer, immediately acquired on an LSR Fortessa flow cytometer utilizing the origami FITC tag as a readout for successfully transfected cells. After establishing optimal voltages and pulse lengths for electroporations to maintain cell viabilities, routine transfection efficiency of >85% was observed for both systems after 20-24 hr incubation times **(Figure 3.6A-B)**.

Additionally, to verify that the DNA origami nanostructures had actually entered transfected cells, as opposed to binding non-specifically to the cell surface and giving false positive signal, transfection samples were incubated with a concentrated DNase pretreatment (Turbo DNase, Ambion, Life Technologies) and then followed by flow cytometry analysis. No loss in transfection was observed following DNase treatment compared to samples without DNase treatment, indicating the nanostructures were in fact entering the cell membranes **(Figure 3.7A)**. To ensure the origami nanostructures were sensitive to the DNase treatment resulting in their destruction should they be unprotected on the outside of the cell membranes 10 μL of origami (50 nM) incubated in 1 μL DNase

(30 min, 37°C) on a 1% agarose gel revealed no detectable nanostructure present following DNase treatment **(Figure 3.7B)**.

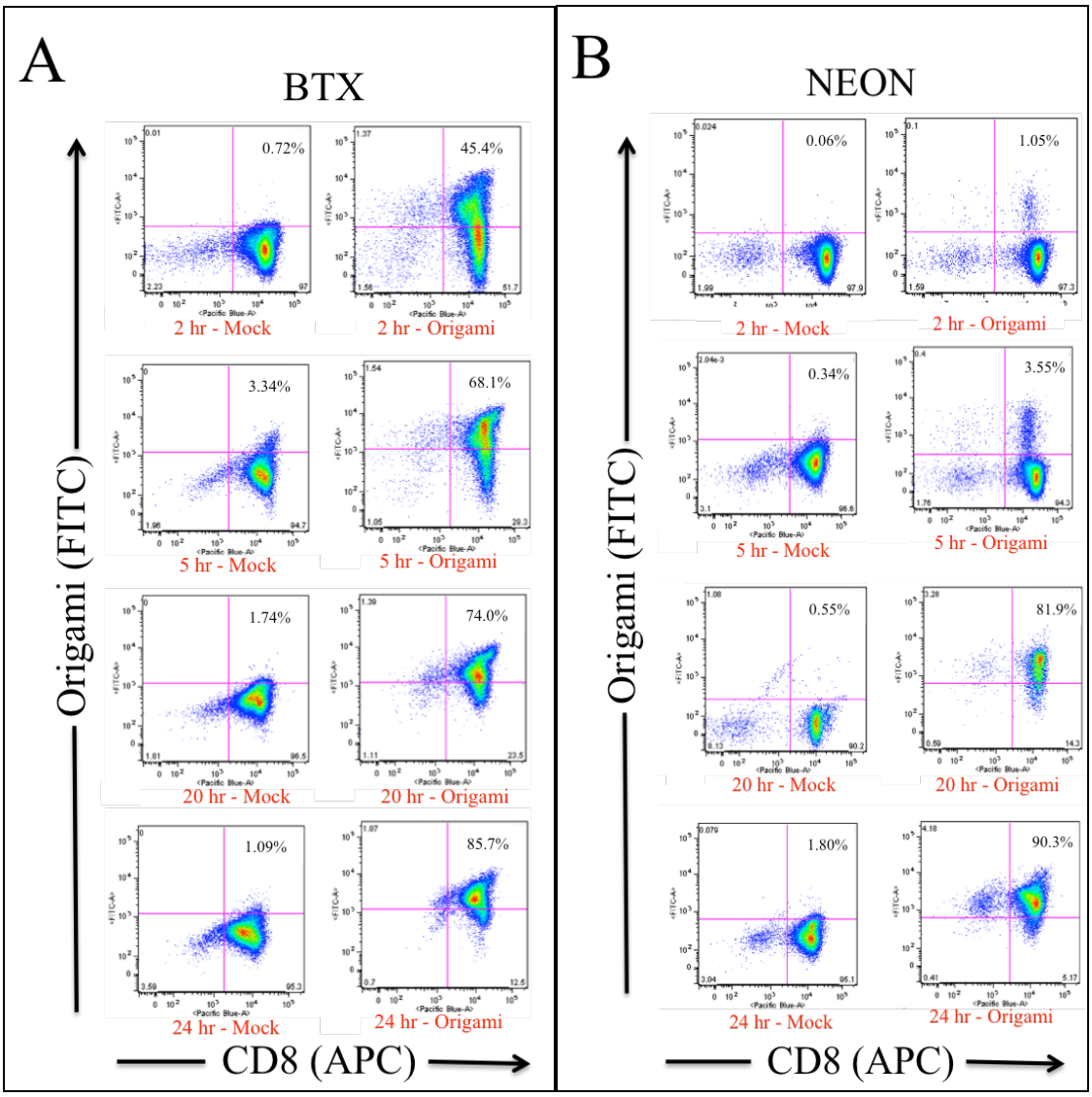

Figure 3.6: High transfection efficiencies achieved by both BTX cuvette and NEON syringe transfection systems. **[A]** Flow cytometry analysis of electroporation samples of sorted $CD8^+$ T cells with DNA origami (right) using the BTX cuvette system demonstrates transection efficiency of >85% after 24hr incubation (bottom) when compared to mock transfection controls (left). **[B]** Flow cytometry analysis of electroporation samples of sorted $CD8^+$ T cells with DNA origami (right) using the Neon Syringe electroporation system demonstrates transfection efficiency of >90% after 24 hr incubation (bottom) when compared to mock transfection controls (left). Samples consisted of $1x10^6$ cells in 75 µL MACS buffer and 25 µL DNA Origami (50nM) in 1X TAE-Mg2+ and were compared to negative control mock transfections of $1x10^6$ cells in 75µL MACS buffer and 25 µL 1X TAE-Mg^{2+}.

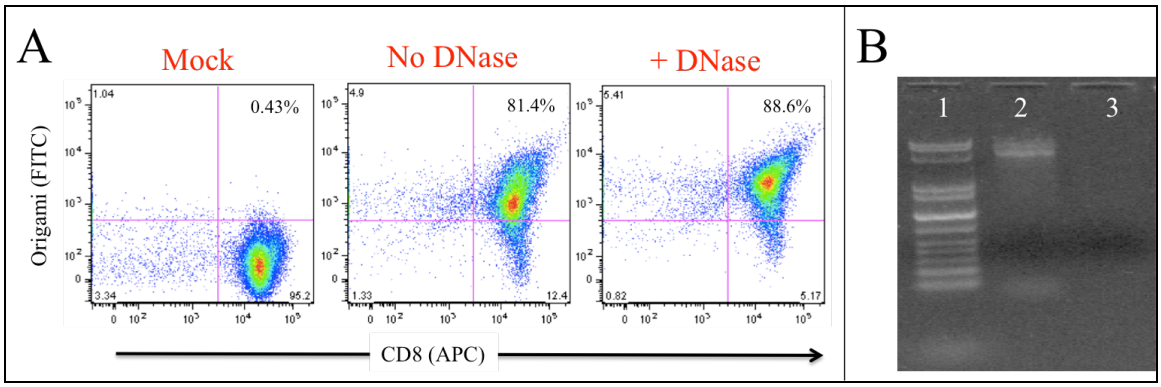

Figure 3.7: DNase treatment of transfected cells reveals DNA enters cell membranes rather than binding non-specifically to cell surfaces. [A] To ensure that origami structures were entering cell membranes, rather than binding to the cell surface, cells were treated with a DNase cocktail after transfection prior to flow cytometry analysis. No significant loss in transfection efficiency was observed from samples treated with DNase (right) when compared to samples not treated with DNase (middle) relative to mock-transfected controls (left). **[B]** DNAse digestion destroys DNA origami nanostructures: Lane 1- MW marker, Lane 2: DNA origami (upper band), Lane 3- DNA origami incubated with DNAse.

Biotinylated origami nanostructures can be successfully purified from unbound mRNA using avidin column filtration. The incorporation of four biotinylated staples into the DNA origami nanostructure design allowed for relatively simple isolation of nanostructures from unbound *in vitro* transcribed mRNA via streptavidin column filtration **(Figure 3.8A)**. After one hour of incubation with *in vitro* transcribed mRNA, samples containing origami with (or without) bound mRNA was then isolated from unbound mRNA by filtering the lysate through a primed streptavidin resin column. Purified origami was then eluted from the streptavidin columns by a solution of excess (2mM) biotin. Successful reisolation of mRNA bound origami was observed by gel electrophoresis **(Figure 3.8B)**. To determine relative efficiency of origami elution and unbound mRNA purification successive elutions were analyzed by gel electrophoresis and compared to unpurified origami incubated with in vitro transcribed mRNA. While difficult to directly quantify, it was determined that pooling of 3 successive elutions may be necessary for high efficiency reisolation of the nanostructures from the purification columns. Additionally, the absence of a band corresponding to unbound mRNA corroborates RT-PCR data (discussed below) demonstrating undetectable levels of unbound mRNA in origami elution fractions.

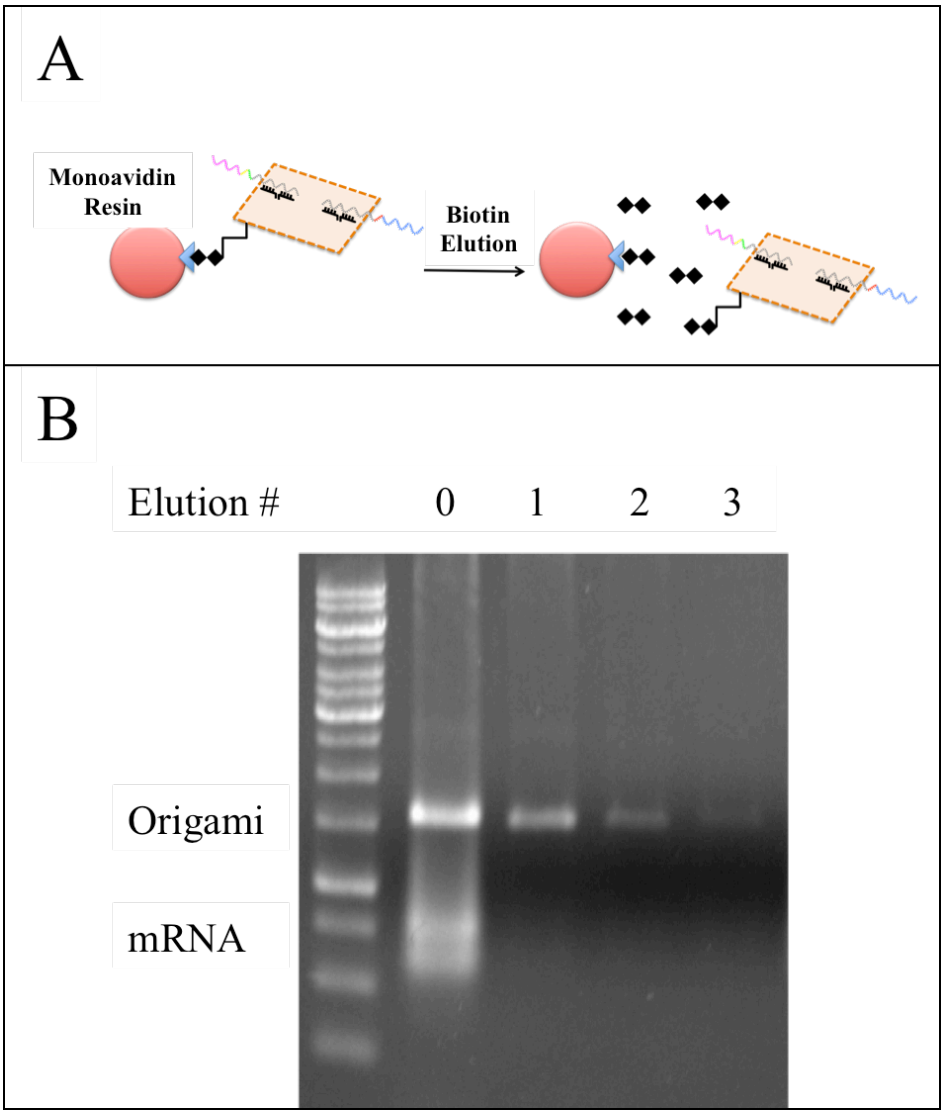

Figure 3.8: Purification of DNA origami nanostructures from *in vitro* transcribed mRNA and cell lysate with bound mRNA using avidin filtration columns. [A] Following transfections or incubations with *in* vitro transcribed mRNA, nanostructures with bound mRNAs can be reisolated using avidin resin column purification. Following 1 hr incubation of cell lysate in avidin columns, three rinses with 350 µL 1X TAE-Mg^{2+} can be performed to remove any unbound mRNA and cellular debris. Origami with bound mRNA can then be eluted from the avidin columns with 100 µL 2mM biotin. **[B]** *in vitro* Transcribed TCRα/β mRNA was incubated with origami nanostructures and samples were purified via avidin column filtration and visualized by agarose gel electrophoresis; Lane 1: ladder, Lane 2: unpurified origami incubated with mRNA, Lane 3: 1st biotin elution of purified origami product, Lane 4: 2nd biotin elution of purified origami product, Lane 5: 3rd biotin elution of purified origami product. Subsequent rounds of biotin elutions can be pooled to increase efficiency of nanostructure reisolation.

DNA origami mRNA-capture probes function as gene specific primers in reverse transcription reactions of captured mRNAs. To covalently link captured sequences to the DNA origami, the mRNA capture probes were used as primers for reverse transcription. Optimization reactions of reverse transcription were conducted on the surface of the DNA origami scaffold using in vitro transcribed TCRα and TCRβ mRNA. Briefly, origami with TCR mRNA-specific probes were incubated with in vitro transcribed TCR mRNA at 37°C for 20 minutes. The reaction mixture was purified via streptavidin columns to isolate origami with bound mRNA from any unbound mRNA. Origami with bound mRNA was then either eluted from the columns with 2mM biotin in 1X TAE-Mg2+ (for AFM post-analysis) or retained in the column to perform the reverse transcription reactions directly in the purification column.

Four basic steps were identified and analyzed for sequence capture by the origami nanostructures (Fig. 4a): 1) The origami nanostructures were synthesized with gene-specific mRNA capture probes 2) mRNAs were then successfully captured by the complimentary probes and thus annealed to the origami templates, 3) successful reverse transcription of the mRNA validated as elongation of the probes during a 1 hr 37°C incubation resulting in the sequence of interest being tethered into the origami molecule as an elongated cDNA strand and 4) an RNase cocktail was administered to the sample resulting in the degradation of any bound mRNA. Aliquots of product for validation of each step were analyzed by gel electrophoresis **(Figure 3-9A)** and steps 1, 2, and 4 by AFM **(Figure 3-9B)**.

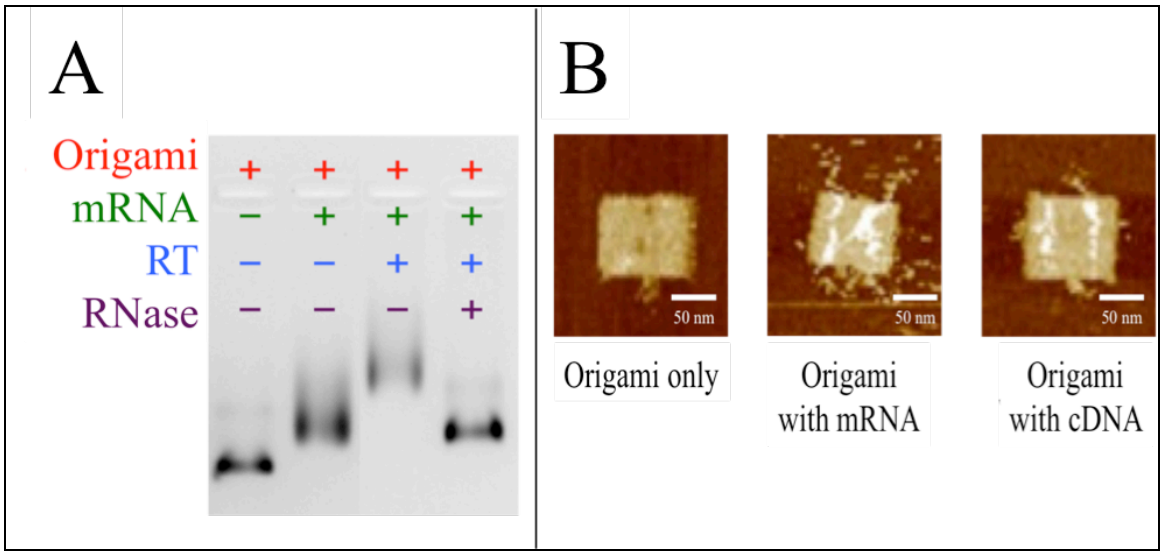

Figure 3.9: DNA origami capture probes successfully function as mRNA reverse transcription primers to create cDNA extended from the origami molecules. **[A]** Agarose gel electrophoresis of products from the four stages of the amplicon generation process show migration shifts due to binding of mRNA and/or cDNA. Lane 1: DNA origami only. Lane 2: DNA origami incubated with *in* vitro transcribed mRNA. Lane 3: DNA origami incubated with *in vitro* transcribed mRNA after a reverse transcription reaction (origami has both cDNA elongation with mRNA still bound). Lane 4: DNA origami incubated with *in vitro* transcribed mRNA after a reverse transcription reaction and incubated in an RNase cocktail to remove the bound mRNA. **[B]** Selected AFM images demonstrate products from stages 1 (origami only), 2 (origami with bound mRNA), and 4 (origami with extended cDNA only) of the cDNA synthesis (stage 3 not imaged).

Sequencing confirms successful transfection of DNA origami nanostructures, intracellular capture of mRNA, reisolation of purified nanostructures with bound mRNA, use of origami probes as RT primers, and amplification of P14 $CD8^+$ T cell TCRα and TCRβ genes. One major advantage of electroporation is that the delivery of DNA nanostructures is relatively quick; the time-dependent factor relies on the incubation length necessary for origami to come into contact and thus bind cytoplasmic mRNA. The stability of the rectangular DNA origami nanostructure used for our experiments in cell lysate has been previously evaluated by our group and found to be highly stable at 12-24 hr post-incubation [337]. Compared to our previous studies, the number of cells used for the transfections described here was increased by 3 orders of magnitude ($1x10^6$ cells/transfection), therefore sufficient purification of cellular debris and biotin from the elution buffer required for clear AFM visualization was not attainable. However, we were able to confirm specific mRNA capture and isolation after 12-24 hr post-transfection incubations by gene-specific RT-PCR amplicon generation and sequencing.

To confirm capture of TCRα/β mRNA, P14 TCRα and β PCR reactions were performed using the RT product from purified origami with bound mRNA from transfected cells as input material. The probes on the origami nanostructures acted as suitable reverse transcription primers eliminating the need for [primer:template] optimization. Furthermore, as binding of mRNA to origami is required for reverse transcription, any unbound mRNA remaining in the purification columns were not transcribed, as the mRNAs would lack a primer for the RT reaction. Positive controls for

the RT-PCR reactions consisted of PCR reactions performed using the eluted RT product from origami incubated with *in vitro* transcribed mRNA for both TCRα and TCRβ genes.

Elution of RT product from transfection of origami constructed with TCRα/β probes yielded identical results as *in vitro* transcribed mRNA positive controls **(Figure 3.10A)**. In contrast, negative control samples consisting of elution of RT product from transfection of origami *without* TCRα/β probes did not yield a PCR product **(Figure 3.10A)**, indicating undetectable levels of nonspecific mRNA binding or amplification of unbound mRNA. Sequence confirmation of purified PCR products from transfected cells was obtained by comparison to known P14 TCRα/β sequences [202] confirming the ability of DNA origami nanostructures to capture, protect, and reisolate cytoplasmic mRNAs following post-transfection purification, and be used in downstream amplification and sequencing analysis **(Figure 3.10B)**.

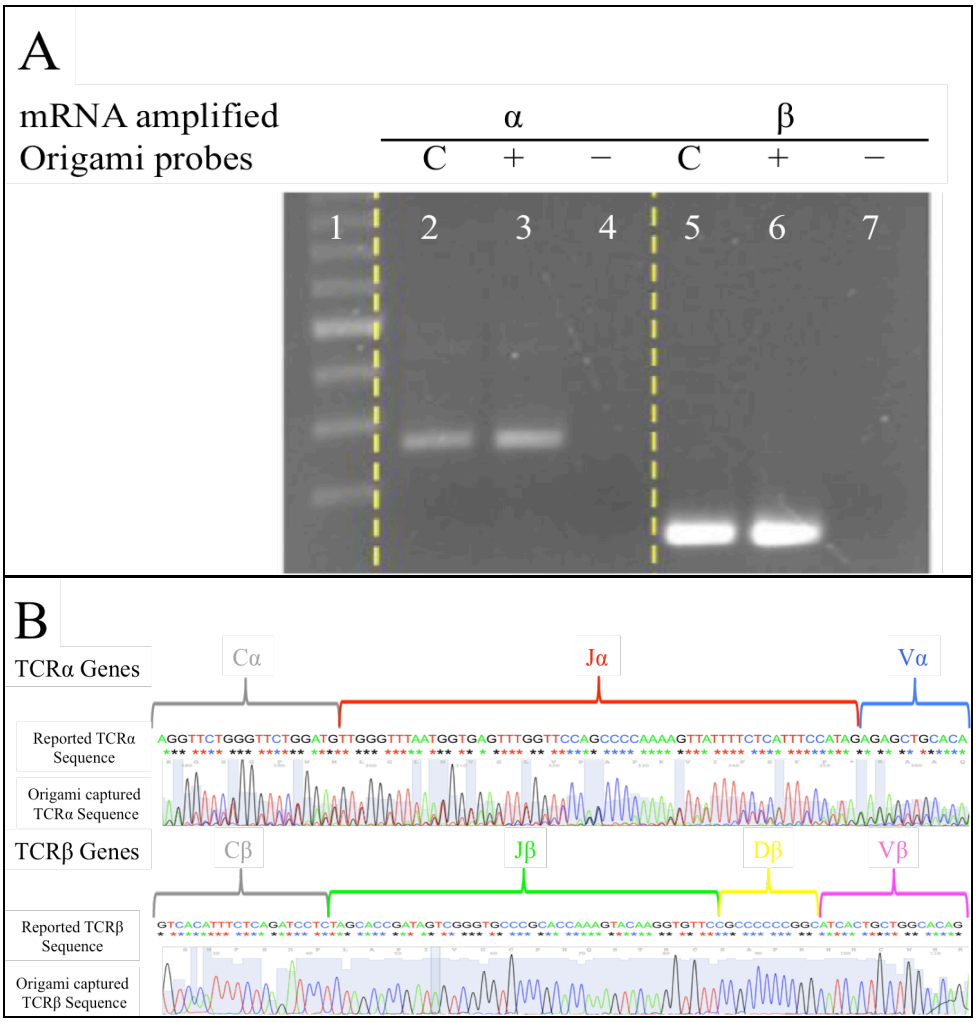

Figure 3.10: RT-PCR and sequencing confirm ability of DNA origami nanostructures to be transfected and capture both TCRα and TCRβ mRNA from individual P14 CD8$^+$ T cells. **[A]** Agarose gel electrophoresis of TCRα and TCRβ RT-PCR products amplified from mRNA captured by origami purified from post-transfected cell lysate. Using *in vitro* transcribed TCRα/β mRNA as positive controls, origami *with* TCRα/β probes transfected into P14 CD8$^+$ T cells were able to bind, protect, and be reisolated with TCRα/β mRNA that was then reverse transcribed and amplified yielding identical bands as controls (lanes 2 vs 3 and lanes 5 vs 6). However, origami synthesized *without* TCRα/β probes did not yield RT-PCR products (lanes 4 and 7), indicating a lack of detectable non-specific mRNA capture by the nanostructures. **[B]** Sequencing traces of both TCRα (top) and TCRβ (bottom) from RT-PCR products from transgenic P14 mice with reported TCRα and TCRβ gene sequences listed above. 100% homology was observed between reported P14 sequences and traces generated from post-transfected RT-PCR products.

DISCUSSION

In order to recognize and combat a diverse array of pathogens, T cells express a large repertoire of clonotypic αβ dimeric T cell receptors, resulting in an enormous number of specificities at the population level [228, 343]. TCR diversity is due to two processes. First, somatic recombination of V, D (for the TCRβ chain), and J gene segments, together with junctional diversity, results in diverse TCRa and TCRb sequences [205]. The CDR3 diversity of TCRα and β chains generated by recombination of gene segments has been directly estimated for each chain separately to be ~10^6 unique sequences in humans and only a slightly lower estimate of ~$8x10^5$ unique sequences in mice [192, 236]. Second, pairing between different TCRα and TCRβ chains results in potentially a one million-fold increase in TCR diversity: completely non-random pairing of each TCRα with a single TCRβ would result in a total diversity of ~10^6 unique TCR while completely random pairing of any TCRα with any TCRβ would yield a maximum combinatorial diversity of ~10^{12} (or greater) unique TCR. This diversity due to pairing of TCRα and TCRβ chains has not been systematically examined. The brute force method of sequencing both TCR chains at the single cell level is financially unfeasible for large naïve T cell populations; each sequencing reaction costs ~$2 so for a single naïve mouse with ~10^7 total T cells this would be a $20,000,000 experiment [260]! Molecular strategies for linking TCRα and TCRβ mRNA have not been adequately developed to generate suitable input material for standard multiplex deep sequencing of TCR CDR3 regions that would provide information on both TCR chains from a single cell. One major

limitation to such approaches is that hybrid structures, generated by transfection with oligonucleotides complementary to the constant regions of TCRα or TCRβ mRNA results in activation of nucleases that destroy the template and therefore, preclude TCR sequence analysis [344, 345]. Our novel strategy to determine the contribution of pairing of different TCRα and TCRβ chains to total T cell receptor repertoire diversity utilizes DNA origami nanostructures to specifically bind the constant region of TCRα and TCRβ mRNA and protect these hybrid structures from destruction.

The fundamental properties of DNA origami nanostructures make them highly programmable, low-cost molecules that many labs can synthesize. These qualities have pushed researchers to identify the capabilities of using DNA nanostructures as therapeutic delivery systems, intracellular machines, and a wide bevy of technologies in the molecular biology toolbox. The use of DNA origami described here is the first assessment of how these ever evolving nanostructures can be manipulated and used as a "single-cell gene chip" for analysis of multiple mRNAs from individual cells without the need for single cell sorting. Additionally, the infancy of the field of DNA origami suggests that increasingly more complex structures and devices may soon be established, adding to the capability of projects such as ours in that new structures may lend to the ability to capture even more copies of mRNA and/or withstand intracellular degradation for longer periods of time which may be required for capture of rare mRNA species.

A central problem in immunology is that the immune system must balance diversity in immune populations with maintenance of sufficient precursor cells specific

for any given pathogen in order to mount effective responses. Quantitatively defining this balance between diversity and protection has been problematic, in large part due to the lack of methods for quantitating ***total*** TCR diversity; despite the quantitation of ***sequence*** diversity independently for TCRα and TCRβ chains in the naïve T cell repertoire for both humans and mice, a systematic accurate measurement of TCR combinatorial diversity, due to ***pairing*** of different TCRα and β chains, has not been attempted [192, 236]. Single cell sequencing of both TCR chains remains unfeasible while molecular strategies for linking TCR sequences have not been adequately developed [260]. We have developed novel DNA origami nanostructures that can be transfected into T cells to capture and protect both TCRα and TCRβ mRNA, and after re-isolation, can be used as a template for reverse transcription reactions, and following with PCR amplification, can provide input material for sequencing to obtain CDR3 sequence information for both TCR chains from individual cells within polyclonal T cell populations.

Due to the highly adaptable nature of the origami molecules, simple changes to the sequence of the mRNA capture probes allow this technology to be used in a wide variety of cell types in a wide variety of species (including humans). In summary, by providing a novel method for obtaining genetic information from individual cells within mixed cell populations without the need for single cell sorting, DNA origami nanostructures have created a biochemical tool that may be applied by immunologists and many other scientists, creating a unique collaboration between synthetic chemists and biologists that has been all too rare in the past.

In chapter 4, together with Illumina paired end deep sequencing technology, discussion of an approach to link sequences captured by individual origami molecules will be discussed.

CHAPTER 4

DEVELOPMENT OF MOLECULAR BOWTIE BARCODES FOR LINKING ORIGAMI-CAPTURED TCRA AND TCRB SEQUENCE INFORMATION FROM INDIVIDUAL CELLS

ABSTRACT

The immune system must be able to recognize virtually any pathogen (providing immunological diversity) while maintaining enough cells specific for each pathogen in order to mount an effective response (ensuring sufficient protection). T cells generate diversity by imprecise joining of gene segments to generate alpha/beta heterodimeric T cell receptors (TCR) during a process called somatic recombination. Current methods do not allow for the simultaneous quantitation of both TCRα **and** TCRβ chain diversity from the same cells, which may result in virtually no sharing of clonotypes between individuals. Linking sequence information for TCRα and TCRβ chains from individual cells has been problematic due to the cost of single cell sorting and inadequate molecular approaches for linking the mRNAs encoding these proteins. To address this problem, we have developed novel DNA origami nanostructures to capture and protect both TCRα and TCRβ mRNAs from individual cells, which can then be physically and bioinformatically linked via a unique "bowtie-barcode" capture strand, sufficient to generate individual amplicons containing sequence information from both TCR from individual cells that can then be paired bioinformatically via next generation sequencing. This approach is directly amenable to single cell analysis of virtually any heterogeneous cell population for which sequence information on any two genes is required.

IMPORTANCE

The antigen receptors of mature $CD8^+$ αβ T cells are heterodimers composed of two constituent chains generated by the process of somatic recombination of V, D, and J genes. The rearrangements for the α and β chains are generally thought to be independent processes [192, 236], culminating in the pairing of a single α chain to its complimentary β chain. While vast diversity can be generated for each chain by the somatic reorganization of TCR genes (including the random N- and P-nucleotide additions at the junctions), the theoretical diversity of an individual's complete TCR repertoire is significantly increased by combinatorial diversity; the potential of constructing the heterodimeric receptor. Completely random pairing of any one of the numerous possible α chains with any of the possible β chains could yield $>10^6$ fold increase in theoretical diversity, while on the other hand completely non-random pairing where every α chain can only pair with a single β chain would mean there is no increase in diversity due to combinatorics.

Original estimates concluded that each unique TCRβ chain expressed in the naïve $CD8^+$ TCR repertoire is on average paired with 25 different TCRα chains (however in the effector fraction of the repertoire each TCRβ chain was estimated to pair with only a single TCRα chain) [192]. However these estimates were extrapolated from an extremely small subset of the repertoire, and may not be universally distributed across the many possible TCRβ chain rearrangements. Determining an accurate estimate for the number of unique TCRαβ combinations in the entire TCR repertoire requires the development of

a technique capable of covalently/bioinformatically linking the α and β chain mRNAs from individual cells for subsequent sequencing.

INTRODUCTION

One of the key tenants of the adaptive immune system is specificity. This specificity is driven by a repertoire of hypervariable antigen receptors (as well as antibodies) on both B cells and T cells. The process utilized to create the vast repertoire of antigen receptors involves genomic recombination events, enabling the cells and/or antibodies to recognize a multitude of potential pathogens that the host may encounter in a given lifetime. While the receptors on B cells (BCRs) are necessary to bind antigens and produce effective humoral immune responses, the crux of this study focuses on the critical nature of the T cell antigen receptor (TCR) for peptide:MHC recognition, and cellular immunity.

While two classes of T cells exist in animals (γδ and αβ), 90% of circulating T cells in humans express an αβ TCR [205] and was the focus of this study. It should be mentioned however, that relatively simple modifications to the capture probes (discussed below) would allow for analysis of γδ TCR genes, or even BCR/antibody heavy-light chain genes. Diversity within the TCR repertoire is generated in a number of ways, and similar to the BCR genetic loci, the TCR genetic loci contains many different variable (V), diversity (D), and joining (J) gene segments. These gene segments are subjected to a rearrangement process for both the α and β TCR chains during the early lymphoid differentiation of T cells in the thymus [205, 220]. In addition to the stochastic selection

and recombination of the V, (D), and J gene segments, addition and deletion of non-templated nucleotides at the junctions between the V(D)J genes further increases the diversity of the TCR repertoire [222, 346]. With respect to the TCRα locus in humans, 47 V genes, 57 J genes, and a single constant (C) gene allows for the 105 genes to rearrange into 2679 unique α chain VJC gene combinations. The TCRβ chain contains 54 V genes, 2 D genes, 13 J genes, and 2 C genes, allowing these 71 gene segments to form a possible 2808 unique β chain VDJC rearrangements. Combinatorial diversity (due to the pairing of α chains with β chains) creates an astonishing 7,522,632 possible unique αβ chain combinations [347]. What's even more astounding is the fact that this value does not take into account the random insertions and deletions of non-templated nucleotides at the gene segment junctions, which result in predicted theoretical repertoires of $10^{15\text{-}20}$ different TCR combinations [229, 230].

Described by some as "infinite" [347], the theoretical diversity of the TCR repertoire far exceeds the actual size of the repertoire, as restriction by deletion of both over- and under-reactive cells occurs during negative and positive thymic selective processes, respectively. Furthermore, every given individual's repertoire is dynamic in the sense that the clones (and their relative frequencies) are constantly being molded by clonal expansion of cells responding to antigen during said individual's lifetime [228, 348]. Original estimates for the human adult TCR repertoire predicted roughly 2.5×10^7 unique αβ T cell clones [192]. However, advancements in technology have allowed for improvement upon estimates such as that by Arstilla *et. al.* to be considerably conservative, with the actual upper bound of TCR diversity in an individual being

potentially 4 orders of magnitude higher (10^{11}) [343]. It should be mentioned that the B cell antigen receptor (BCR) diversity might be even more pronounced as the process of somatic hypermutation (a process of stepwise single nucleotide substitutions into the V gene segments and assessment for enhanced antigen binding) increases the diversity of expressed V gene segments [307, 349]. For all of the adaptive immune system antigen receptors (TCRs, BCRs, and antibodies), the majority of diversity arises in the third complementarity-determining region (CDR3) that makes major contact with the antigenic peptide [350], and as such, the diversity of the TCR repertoire is directly proportional to (and can mainly be attributed to) the diversity of the CDR3 repertoire [347].

Over the last twenty years many different strategies have been employed to attempt to investigate the actual size and diversity of the TCR repertoire. Fluorescence activated cell sorting (FACS) analysis of subsets of T cells has been commonplace in standard immunology labs, however inherent restrictions from this approach lend to inaccuracies of predicted evaluations of the repertoire. Specifically, the availability of monoclonal antibodies for each of the TCR Vα and/or TCR Vβ gene segments restricts the number of analyzed clones per sample. Additionally, no sequence information can be garnered from such approaches [351]. Once the disadvantages of FACS were realized, methods utilizing TCR gene-specific PCR such as multiplex V-gene segment PCR and rapid amplification of cDNA ends (RACE) PCR began to emerge. Additional problems however, were immediately identified as not all of the V gene segments had been sequenced and thus primer sets did not include an encompassing repertoire of V-gene complimentary primers. Furthermore, the introduction of large sets of multiplex primer

systems may induce amplification biases in PCR products, and cross-reactivity between V-gene subfamilies have required multiplex primer system optimization experiments to ensure precise V-gene specific amplification [352] that is still undergoing. The immunoscope or CDR3 spectratyping technique to analyze CDR3 polymorphisms was the next advent to immune repertoire analysis. The drawbacks to this method however, are inherent in the necessity to extrapolate large portions of the repertoire from a very small subsample of actual sequence data [353]. As mentioned, spectratyping does not give an actual statistically significant portion of sequence data from a full repertoire sample. More importantly, the extrapolation necessary for analysis does not allow for a quantitative comparison of the clonal aberrations across the repertoire from different CDR3 arrangements. Finally, the low input of sequences used for extrapolation may induce false positives for sequences [347]. While introduction of Sanger sequencing into TCR repertoire analysis allowed for resolution of sequences at the single-cell level, the laborious and limiting nature of Sanger sequencing only allows for hundreds to possibly thousands of sequences to be assessed for any given investigation [354].

The recent advancements in next generation sequencing (NGS) technologies allow for the possible sequencing of tens of millions of short DNA sequence reads (i.e. TCR receptor CDR3 region sequences) to be both sequenced and analyzed using novel pipeline algorithms for both data conversion and read calling in routinely basic experiments [240]. Numerous groups have been evaluating methods (using multiplex primer systems, RACE amplification, or both) to harness the power of NGS capability for the use of immune repertoire sequencing at a massively high-throughput level, however

each approach relies on utilization of one of three commercially available sequencing platforms, and each comes with both advantageous attributes as well as inherent limitations **[Table 4.1]**.

Table 4.1: Advantages and disadvantages of NGS methods for analysis of immune repertoire diversity. Data adapted from Bolotin *et al*. [355] and Hou *et al.* [347].

Method	Advantages	Disadvantages
Roche 454 Sequencing	Longest read lengths	Lowest number of reads, bottlenecks in read calling, and frame shift errors
Illumina Sequencing	Greatest read numbers	Highest error rate, shortest read length
Ion Torrent Sequencing		Frame shift errors, moderately short read lengths, largest amplification biases

While advancements in NGS technology have undoubtedly sent immune repertoire sequencing into a field all its own, most of the methodologies yield data from only one of the two receptor chains (i.e. either TCRα chain or TCRβ chain CDR3 sequence data) and therefore are incapable of providing sequence information from a complete dimeric receptor pair of mRNAs [356]. The "holy grail" of immune receptor repertoire analysis is the capture, sequencing, and analysis of paired TCR (or BCR/Antibody) chain sequences to identify complete antigen receptor antigen binding sequences that can then be evaluated for introduction into genetically engineered chimeric antigen receptor (CAR) T cells, or for the production of broadly neutralizing antibody production. To this end, recently developed systems seeking to acquire paired immune receptor gene sequences have been developed. A high throughput VH-VL gene sequence oil-emulsion pairing technique has been proposed by De Kosky *et al*. [261] for the evaluation of BCR/Antibody heavy and light chain pairings, single-celled sorting and subsequent sequencing methods have been adapted for both TCRs and BCRs [357], and combinatorics based approaches allow for the pairing of high copy number sequencing via bioinformatics [266].

While advancements in sequencing technology have allowed researchers to develop platforms for paired immune receptor sequencing from individual cells, each comes with its own crippling limitations. Any single-cell sorting-based approach requires isolation of single cells into either 96- or 384-well plates, limiting sample sizes necessary for complete repertoire analysis [357]. Oil-emulsion systems, such as that proposed by De Kosky and colleagues have proven inefficient and result in error-prone pairings due to

multiple cells per oil-emulsion droplet, and finally pairSEQ technology, developed by Adaptive Biotechnologies can only resolve clones present at ≥1/200,00 frequency, meaning that rare clones (which may include the rare, highly effective tumor infiltrating lymphocytes) may not be included in final sample analysis [266].

In summary, advanced high throughput sequencing technologies (specifically those tailored for the pairing of immune receptor gene sequences) are powerful tools for analysis of immune repertoires at an extremely high resolution. The inherent limitations of currently available techniques however, has prompted our investigation into the development of a novel system single-cell paired gene sequencing, with the hopes of eliminating both the need for single-cell or oil-emulsion-based sequencing, and any combinatorics-based resolution techniques.

To this end, in chapter 4 we propose a novel strategy for paired analysis of TCR sequences from single cells within heterogeneous cell populations of massive sample size without the need for single cell sorting. A key design feature different from what was outlined in chapter 3, is that the DNA origami utilized for pairing of immunoreceptor genes contain mRNA capture probes on any given nanostructure that are bioinformatically linked to one another by a unique 5'-5' bowtie barcoded capture strand that includes 'matching' (complimentary) barcodes; the recovered mRNAs are reverse transcribed utilizing said capture probes as gene-specific RT primers thereby allowing for the bioinformatical linking of the TCRα and TCRβ CDR3 cDNAs back to one another during sequencing.

MATERIALS AND METHODS

DNA Origami Design for both mRNA capture and pairing: As mentioned in chapter 3, DNA is an excellent nano-construction material. Our design for the DNA scaffold used to not only capture multiple mRNAs from individual cells but to also pair said captured mRNAs is very similar to the design discussed in chapter 3 with one major exception. As the backbone for our nanostructures we again used the single-stranded, M13mp18 genomic ssDNA (~7 kilobases) and fold this strand into the target rectangular shape (~60 x 90 nm in scale) using the same cohort of 200+ short, ssDNA "staple" strands. While the biotinylated and/or fluorescent tag conjugated staples remain consistent with the previously described nanostructure from chapter 3, the mRNA capture probes from the nanostructure have been replaced with their basic staple strand counterparts. Our design for this nanostructure for capture *and pairing* of mRNAs includes only a single mRNA capture strand integrated into the M13mp18 ssDNA backbone with an internal 5'-5' "bowtie" linkage, allowing for 5'→3' strand directionality on both ends that can be used for dual mRNA capture. Additionally, the two ends of each "bowtie strand" synthesized (discussed below) has a unique but complimentary set of barcodes (12 nt-mers) that allow for captured mRNA to be bioinformatically linked back to one another following sequencing **(Figure 4.1)**.

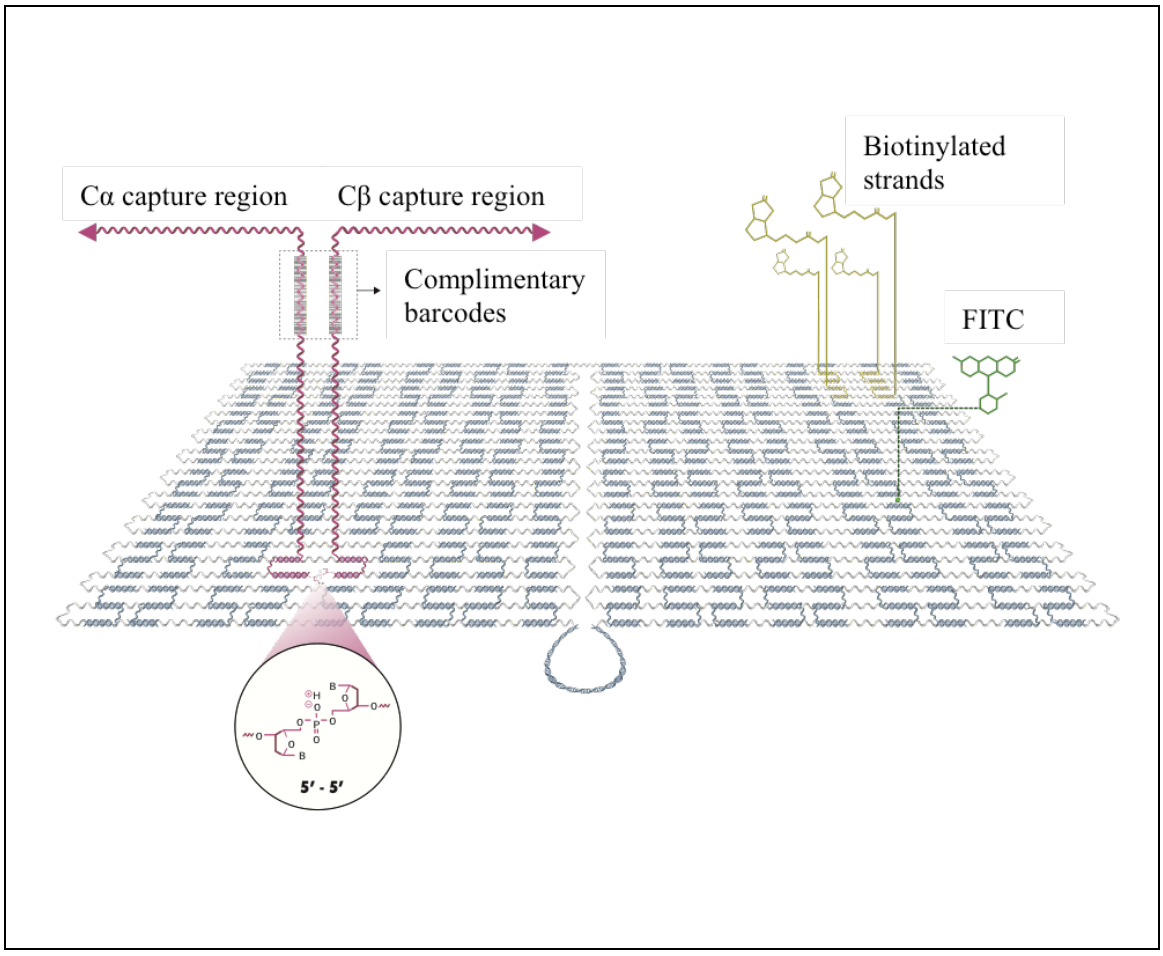

Figure 4.1: DNA origami design for bowtie strand barcoding of captured mRNA sequences. Organization of 5'-5' bowtie mRNA capture probe includes a region complimentary to the M13mp18 phage ssDNA that includes a 5'-5' phosphodiester or "bowtie" linkage, resulting in both ends of the bowtie strand to have 5'→3' directionality. These two ends are designed to be complimentary to the constant regions of either the TCRα or TCRβ mRNAs and therefore serve as our Cα and Cβ capture probes and downstream RT primers. Additionally, each capture probe includes a 12-mer "barcode" upstream from the capture sequence that retains sequence complementarity to one another. These barcodes can be used to bioinformatically link captured TCRα and TCRβ gene information from the same cell during downstream sequencing. The locations of the four biotinylated strands used for purification and FITC conjugated strand used for FACS assessment of transfection are included as well.

Design of 5'-5' "bowtie barcode" mRNA capture strand: The Illumina 2x250 paired-end sequencing platform is ideal for analysis of DNA origami-captured gene sequences because of its ability to generate short sequence reads in massively parallel reactions. To utilize this technology, we developed and optimized a novel bowtie-barcode pairing strategy for linking two complimentary TCR sequences from individual T cells during downstream sequencing. This system is highly modular, and can be easily adapted to link the TCRα and TCRβ mRNA from T cell receptor genes or virtually any two mRNAs from individual cells by relatively simple changes to the mRNA-capture sequences in the bowtie barcode strands. Our design requires that a long ssDNA be constructed containing a central 5'-5' bowtie linkage (pre-synthesized and ordered from commercial vendor IDT), allowing for both ends to run 5'→3' in direction. One side of the bowtie linkage contains a specific sequence complimentary to the M13mp18 phage ssDNA origami backbone, and both sides contain conserved PCR primer sites for downstream amplification, unique barcodes complimentary to one another, and mRNA capture sites complimentary to the conserved region of the genes of interest **(Figure 4.2A)**. This long bowtie mRNA capture strand is then included in the DNA origami mastermix and the portion of the bowtie barcode strand complementary to the M13mp18 sequence allows for self-assembly with the origami nanostructure with the complementary mRNA capture portion of the bowtie barcode strands extending off the nanostructure template **(Figure 4.2B)**.

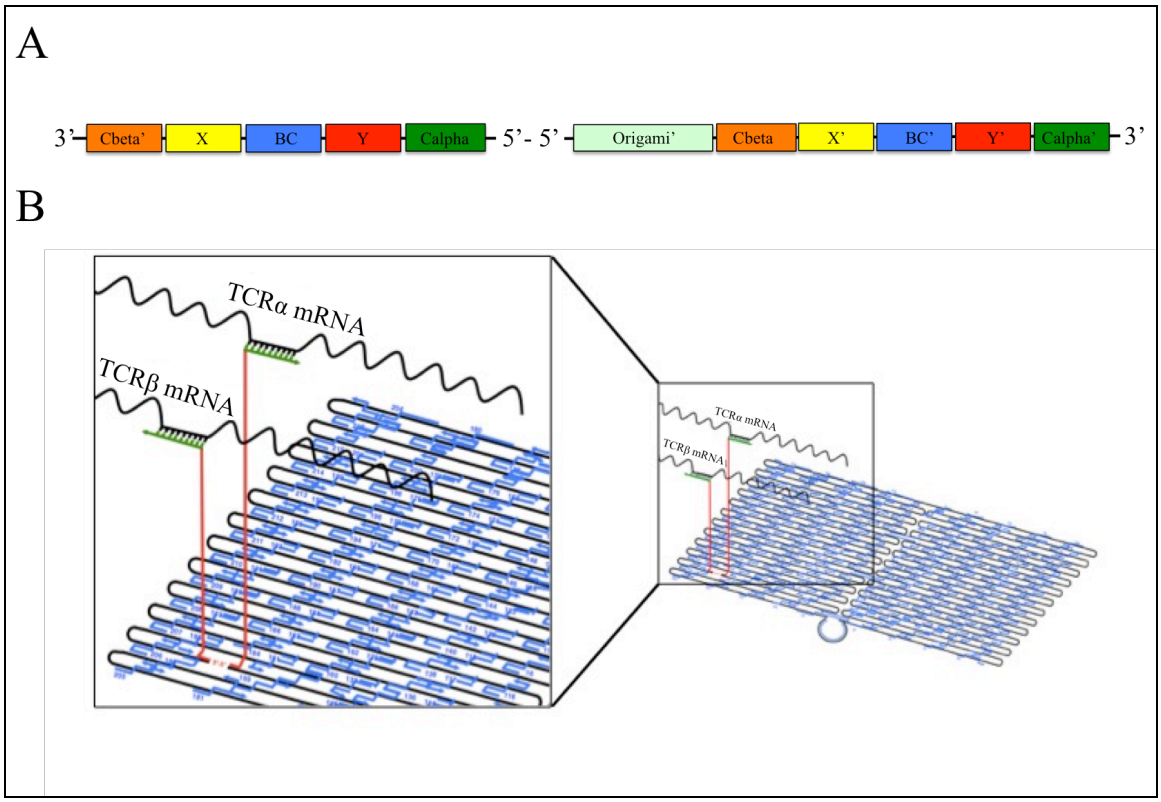

Figure 4.2: Bowtie barcode design and integration into DNA origami nanostructure. **[A]** Organization of 5'-5' bowtie barcode mRNA capture probe. Various regions of the bowtie barcode strand allow for incorporation into DNA origami nanostructures (Origami'), conserved priming sites for downstream PCR amplification (Y and X'), barcode pairing (BC and BC'), and TCRα/TCRβ mRNA capture (Cbeta' and Calpha'). **[B]** Schematic visualization of 5'-5' bowtie barcode mRNA capture strand extending from the surface of an origami nanostructure while annealing with TCRα and TCRβ mRNA (note: the length of the 5'-5' bowtie barcode strand has been exaggerated to allow for visualization of the structure).

Oil-emulsion set-up for synthesis of bowtie barcode mRNA capture probes:

Water-in-oil emulsion droplets are utilized to ensure that complimentary barcodes are incorporated onto each side of individual bowtie strands and that a different set of unique barcodes is incorporated into each different bowtie strand synthesized. A large aliquot of "ssDNA bowtie backbone" strand (including an M13mp18 complimentary sequence and 2 conserved annealing sites) is procured from a commercial vendor (IDT) **(Figure 4.3)**. Additionally, unique barcode-labeled strands consisting of in order 5'→3': 1) a sequence complementary to one side of the 5'-5' backbone strand whose complement will serve to encode the complementary mRNA capture probe for TCRβ (Cbeta), 2) a sequence to encode a conserved PCR priming site (x'), 3) a random 12-mer (4^{12} = 16,777,216 unique barcode strands) nucleotide barcode (BC'), 4) a second sequence to encode a conserved PCR priming site (Y'), and 5) a second sequence complementary to the other side of the 5'-5' backbone strand whose complement will serve to encode the complementary mRNA capture probe for TCRα (Calpha') **(Figure 4.3)**. The ssDNA barcode strands were converted into dsDNA by a first-strand-synthesis molecular reaction with the *E. coli* DNA Polymerase I Large (Klenow) Fragment [358] as per manufacturer's recommendations (New England Bioscience) resulting in each of the dsDNA strands obtaining complimentary barcodes to one another (BC/BC') **(Step 1)**. The ssDNA bowtie backbone and dsDNA barcode strands were then incorporated at a 1:1:1 molecule-to-molecule-to-droplet ratio in an oil-water emulsion droplet system using the alternative protocol from Williams *et al.* [359] using ABIL EM90 surfactant (Evonik Industries) centrifuging the emulsions **(Step 2)**.

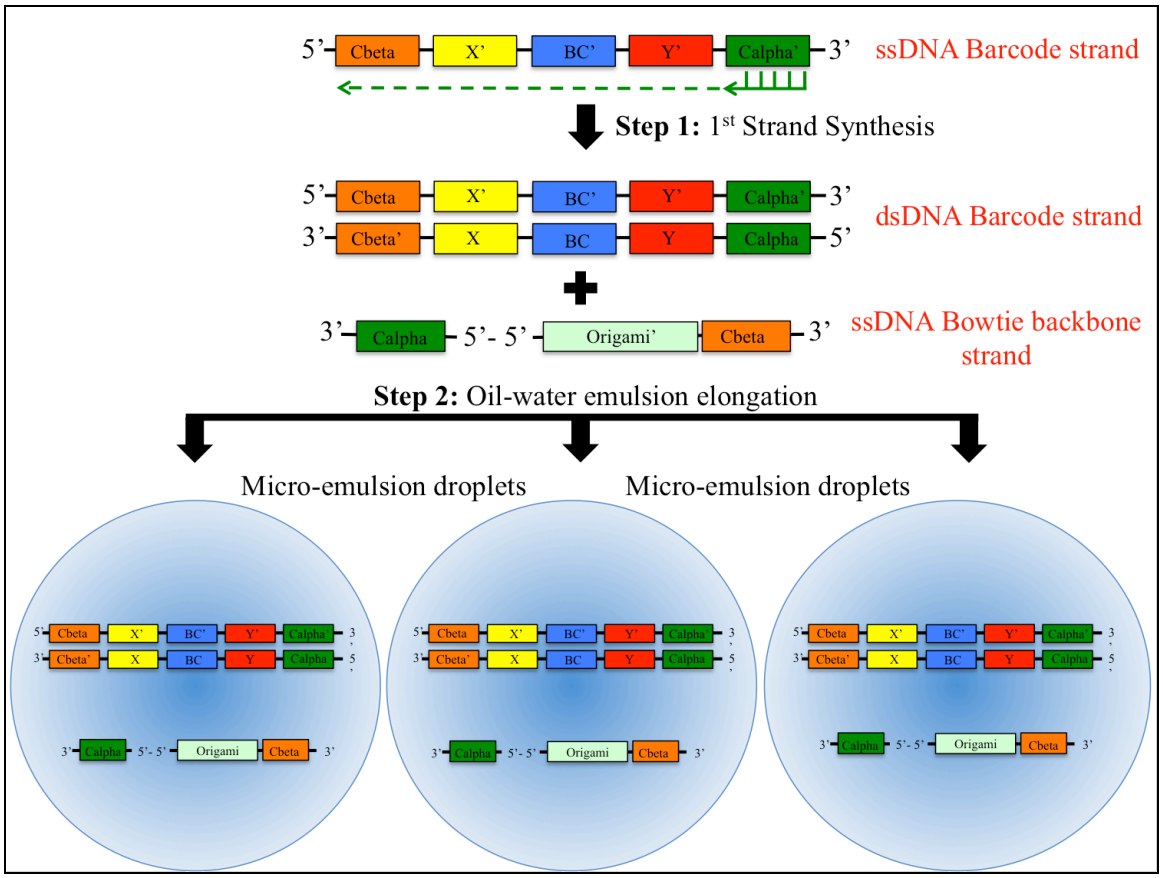

Figure 4.3: Set-up of oil-water emulsion elongation system. Construction of the 5'-5' bowtie barcode mRNA capture probes containing complimentary barcode sequences requires input of two commercially acquired ssDNA oligos. Step 1) First strands synthesis of ssDNA barcode strands using a primer complimentary to the Calpha' sequence on the 3' end of the ssDNA barcode strand creates dsDNA barcode strands. Step 2) Using a 1:1 stoichiometric ratio, the dsDNA barcode strands and ssDNA bowtie backbone strands are incorporated into individual droplets ensuring that both sides of each bowtie backbone strand will incorporate complimentary barcodes and that each bowtie barcode strand synthesized will have a unique set of barcodes.

Overlap extension elongation and purification of bowtie barcode strands: As the aqueous phase of the emulsion droplets will comprise the necessary enzyme, nucleotides, buffer, and other reagents necessary for common PCR, standard denaturation (90°C-30s), annealing (55°C-30s), and elongation (72°C-30s) steps were carried out in each droplet **(Figure 4.4)**. Following dissociation, the hybridization between individual strands of the dsDNA barcode with their complementary annealing sequences on either end of the ssDNA bowtie strand **(Step 1)** acts as a primer and template system for overlap extension elongation, incorporating the conserved priming PCR priming sites, complimentary 12-mer barcodes, and TCRα/β capture sites on either end of the 5'-5' bowtie linkage **(Step 2)**. While this system does not serve as an exponential PCR amplification, multiple (20) cycles were utilized to ensure that both ends of each 5'-5' backbone strand were elongated, minimizing single-sided elongation products. Overlap extension products were then extracted from the oil-water emulsion system using a standard organic solvent ether/ethyl acetate extraction protocol [359], and ssDNA bowtie barcode mRNA capture probes were then purified from remaining nucleic acids by standard denaturing PAGE gels. Final products at this point contained (in order from 5'-5' central bowtie linkage extending towards both 3' ends) the following: an (one side only) M13mp18 complimentary sequence (Origami'), the annealing sites for oil-water emulsion hybridizations (Calpha/Cbeta), internal conserved PCR priming sites (Y/X'), complementary 12-mer barcodes (BC/BC'), external conserved PCR priming sites (X/Y'), and mRNA capture sequences complementary to conserved regions of TCRα and TCRβ mRNAs (Cbeta'/Calpha') **(Step 3)**.

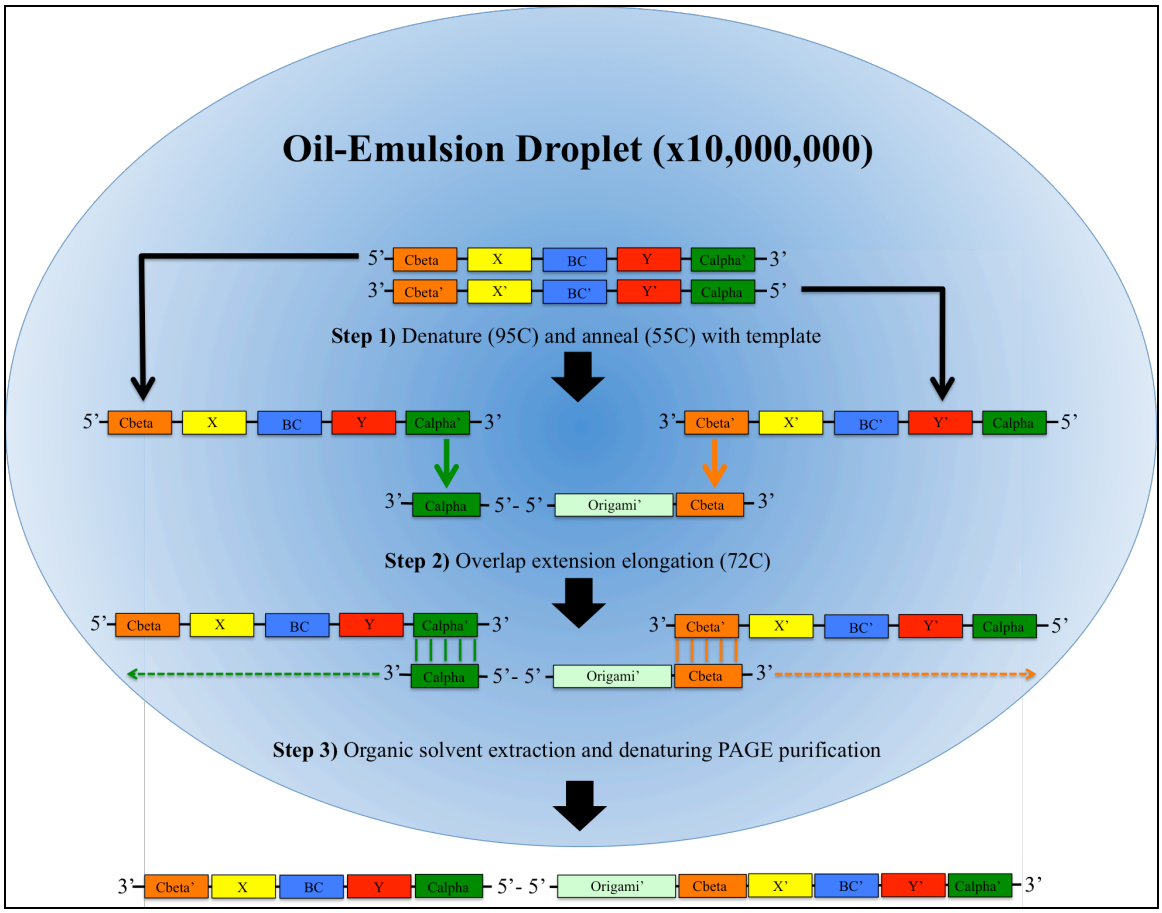

Figure 4.4: Overlap extension and purification of complete bowtie barcode strands. Each droplet formed during the oil-emulsion contains a single dsDNA barcode strand and a single ssDNA bowtie strand. **Step1)** Following 95°C denaturation, each of the dsDNA barcode strands will hybridize (55°C) with their respective complementary sequences on either side of the ssDNA bowtie strand (Calpha'/Calpha and Cbeta'/Cbeta). **Step 2)** Once hybridized, overlap extension elongation is carried out (72°C) on both ends of the ssDNA bowtie strand to incorporate internal conserved PCR priming sites (Y' and X), complementary barcodes (BC and BC'), external conserved PCR priming sites (X' and Y), and TCRβ and TCRα complementary mRNA capture sequences (Cbeta' and Calpha'). **Step 3)** Following 20 cycles of the denature/anneal/elongate protocol, standard organic solvent extraction and denaturing PAGE purification yields purified bowtie barcode strands suitable for incorporation into DNA origami nanostructures.

Synthesis of DNA origami with bowtie barcode mRNA capture strand:

Synthesis of the DNA origami nanostructures follows the same protocol as that outlined in chapter 3. While standard staples are mixed in excess, the bowtie barcode mRNA capture strands and biotin-tagged staples were PAGE-purified and quantified (by measuring absorbance at 260 nm) prior to being included in the origami folding reaction. Scaffold M13mp18 ssDNA, staple DNA, and functionalized staples (including the bowtie barcode strand) were then mixed in a fixed 1:5:3 ratio (50 nM scaffold, 250 nM staples, 150 nM bowtie barcode/biotinylated tags/fluorophore tags) in aqueous buffer (1x TAE with 12 mM Mg^{2+}), followed by thermal denaturation (90°C) and gradual annealing (to 20°C) over 12 hours. Final concentration was measured by A_{260}/A_{280} absorbance and standardized to either 20 or 50 nM depending on the downstream application.

Mice: 6-8 week old female C57Bl/6J mice were purchased from Jackson laboratories (Bar Harbor, Maine). P14 transgenic mice, in which CD8 T cells express TCR specific for the DbGP33-41 epitope of LCMV, were obtained from Dr. Rafi Ahmed and bred in our animal facilities. All studies were conducted according to animal protocol 12-1229R under the approval and guidance of the Arizona State University Institute for Animal Care and Use Committee.

DNA origami transfection into primary T cells: Cell sorting and transfections were carried out following the protocol outlined in chapter 3.

Intracellular TCR mRNA binding, cell lysis, and origami purification: Cell lysis and column purification were performed as outlined in chapter 3 with one alteration.

As the nanostructures purified from cell lysate and unbound mRNA was not eluted from the purification columns for imaging as it as directly used for reverse transcription of captured TCRα/β mRNA, streptavidin conjugated resin was used for purification rather than monomeric avidin. Monomeric avidin is mainly designed for use in simple affinity chromatography purifications of proteins, antibodies and other molecules with a biotin tag. The advantage of monomeric avidin over native avidin, a tetrameric molecule, and streptavidin (also tetrameric) is that monomeric avidin has a much lower biotin binding affinity, $Kd=10^{-7}$ as opposed to native avidin/streptavidin $Kd=10^{-15}$ [360]. The decreased binding affinity of monomeric avidin allows more efficient elution of molecules with mild elution buffers (2mM Biotin in 1X PBS), as opposed to the strong denaturing buffers (8M Guanidine-HCl, pH 1.5) required for the higher affinity avidin/streptavidin column elution.

Reverse transcription and TCR chain CDR3 amplicon generation: Reverse transcription reactions were again performed directly in the streptavidin purification columns using the protocol outlined in chapter 3. The TCRα/β capture sequences on the ends of the bowtie barcode strands perform identically to the mRNA capture probes outlined in chapter 3 in terms of acting as suitable reverse transcription (RT) primers, again eliminating the need for [primer:mRNA] optimization. Once mRNA-bound origami was purified from cell lysate in the streptavidin columns, reverse transcription (RT) was performed directly in the columns. Using the 5'→3' TCRα and TCRβ capture regions on the bowtie barcode strands as RT primers, elongation was achieved by simply adding 40 μL of the RT mastermix (Omniscript, Qiagen) directly into the purification column, and

incubating for 1 hr in a 37°C heat block. The RNA was then removed by addition of 2 μL of an RNaseH cocktail (NEB), and barcoded TCRα/TCRβ CDR3 cDNA can be dissociated from the origami nanostructures by heating the columns to 95°C via heat block for 10 min (dissociating the nanostructure DNA strands), and centrifuging at 10,000 rpm for 1 min to collect the bowtie barcode strand containing TCRα and TCRβ CDR3 sequence information.

To validate this approach, initial transfections were conducted using TCR-transgenic P14 $CD8^+$ T cells. As these cells express only a single TCRα and TCRβ chain, comparing captured sequences to the known P14 sequences allowed for confirmation of successful mRNA capture, RT, amplification, and sequencing. Amplification products were achieved using standard PCR (Phire kit, Thermo Scientific) protocols with two primer sets, one for amplification of P14 TCRα CDR3 sequence information and the other for P14 TCRβ CDR3 sequence information (see below).

RESULTS:

The use of an Oil-emulsion droplet PCR technique created individualized PCR reaction vessels. One of the most novel parts of the bowtie barcode synthesis system is the utilization of oil emulsion droplets as individual vesicles for bowtie elongations. Controlling the stoichiometry between the number of droplets and input of bowtie strands vs dsBC strands ensures that both sides of each bowtie strand will be elongated by only a single dsBC strand (ensuring complementary barcodes on each bowtie strand). We first evaluated time and temperature of oil emulsion set ups to try and

establish the most uniform droplet distribution we could containing sufficient numbers of droplets to meet the synthesis needs. It was determined that a speed of 2000 rpm for 5 min at 4°C yields repeatable emulsions with droplets of fairly uniform size at a high enough count to satisfy our production needs [**Figure 4.5].** Slower spinning and different aqueous:oil volumetric ratios have previously been reported to generate ~10^9 PCR-competent compartments per milliliter of emulsion [359], we estimate that our system can generate an even greater number of droplets >10^{10} droplets/mL (based on droplet size comparisons), which is sufficient for our input of 1.2 x 10^8 backbone bowtie molecules per emulsion.

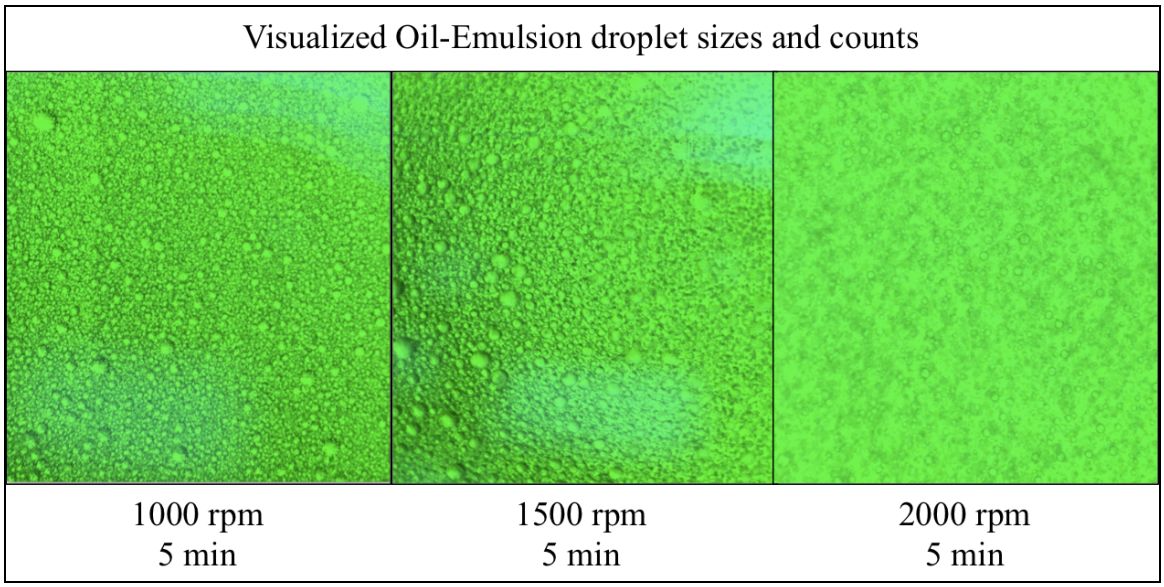

Figure 4.5: Oil-emulsion droplet size and count depends on centrifugation speed and duration of spin. Uniform droplet sizes at sufficient quantity were obtained by centrifugations at 2000 rpm, for 5 min, at 4°C (right). Slower centrifugation speeds yielded emulsions with droplet sizes varying wildly in diameter (left and center). Furthermore, oil-emulsions established at RT did not produce repeatable results (not shown).

Optimization of [bowtie backbone strand]:[dsBC strand] yields high efficiency of double-sided elongation products. Water-in-oil emulsion droplets are utilized to incorporate a different set of complimentary barcodes on each bowtie linker strand synthesized. To obtain dsDNA barcode strands for input into the emulsions, first strand synthesis reactions were performed and optimized to produce individual dsDNA barcode strands containing complementary 12mer barcodes on each strand. Synthesis reactions were performed with the Large (KLENOW) fragment of the *E. coli* DNA polymerase I enzyme (NEB) per manufacturer's protocol with slight changes to incubation times and input concentrations. Incubation times of 30 minutes were determined sufficient for optimal dsDNA barcode synthesis **[Figure 4.6A]**, and a 1:1 ratio of ssDNA barcode strand to primer at 5 μM each **[Figure 4.6B]**.

Following synthesis of the dsDNA barcode strands, oil-emulsion strand elongation of the ssDNA bowtie backbone was performed and optimized revealing a 10:1 ssDNA backbone bowtie strand:dsDNA barcode strand (5.0 μM:0.5 μM) per droplet ratio [**Figure 4.7]**. While this system is not an exponential PCR amplification, multiple cycles (up to 30) were evaluated to determine the minimum number of cycles optimal for ensuring that both ends of each backbone bowtie strand were elongated. It was determined that 20 cycles are necessary for highly efficient double-sided elongation as determined by a lack of non-elongated product remaining in the sample **[Figure 4.8].**

Finally, to evaluate whether or not different dsDNA barcode strands were being elongated on either side of a single ssDNA backbone bowtie strand (thus yielding non-

complementary barcodes on a given bowtie strand), we designed a system including two dsDNA barcode strands of different lengths (short = 70 bp and long = 95 bp), included equal concentrations (0.25M) of both strands in the oil-emulsion system. Products from systems including only the short dsDNA barcode strand yielded 150 bp elongation products, while products from systems including only the long dsDNA barcode yielded only 200 bp products. When emulsions were carried out using both short and long dsDNA barcode input, no detectable hybrid structures (>150 bp < 200 bp) were observed **[Figure 4.9]**, indicating that the vast majority of synthesized bowtie barcode strands were elongation products from the same dsDNA barcode strand, and therefore the barcodes contained on either end of a compete bowtie barcode strand should be complementary to one another for bioinformatically linking captured TCRα and TCRβ CDR3 information from the same cell back to one another.

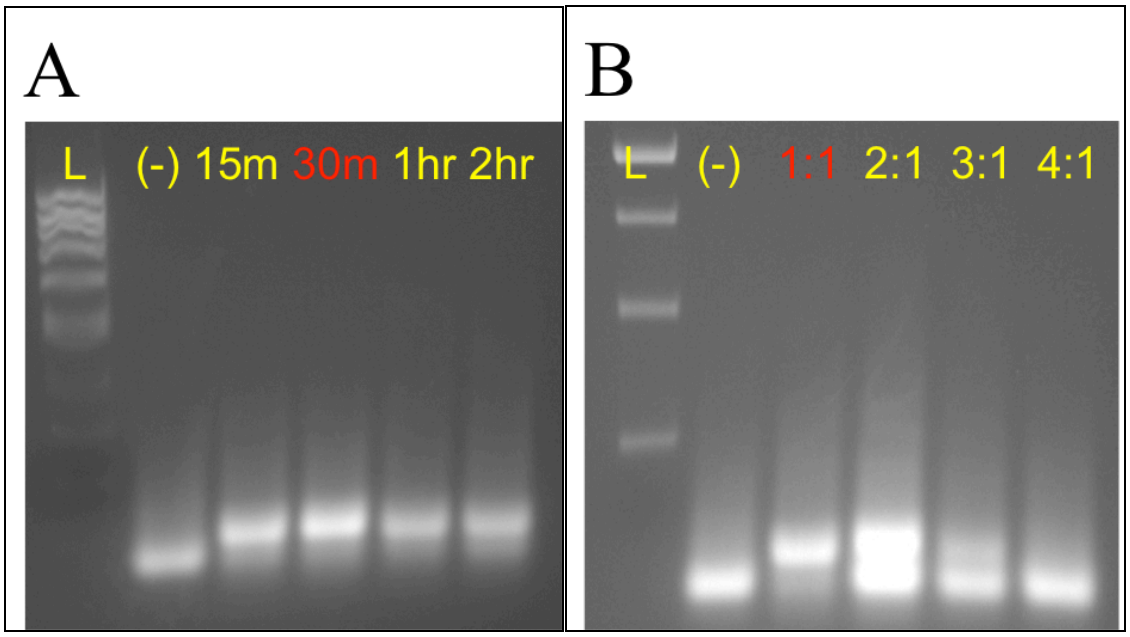

Figure 4.6: First strand synthesis of dsDNA barcode strand. **[A]** Time trial of *E. coli* large (KLENOW) fragment elongation reaction demonstrates efficient synthesis of dsDNA after 30 min elongation at 37°C. **[B]** A titration of input ssDNA barcode strand vs primer was conducted to determine optimal [ssDNA]:[primer] ratio for synthesis of dsDNA. It was determined that a 1:1 ratio of input material (5μM each) yielded the most efficient production of dsDNA.

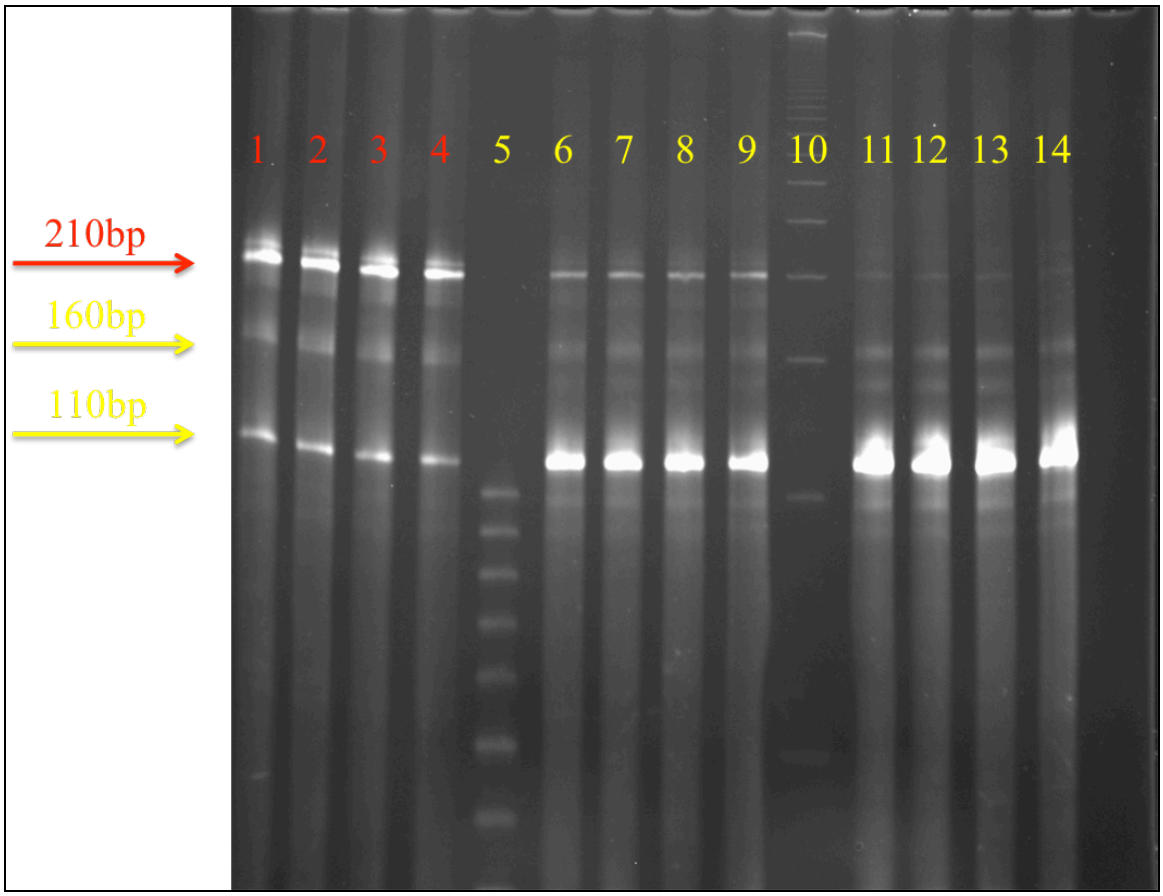

Figure 4.7: Evaluation of DNA input concentrations for optimization of bowtie barcode mRNA capture strand synthesis. Titrations of ssDNA bowtie strand and dsDNA barcode strand were assessed to evaluate the optimal ratio of input material. Complete (double-sided) elongation products are 210bp in length, while single-sided elongation products are 160bp. The dsDNA barcode strand is 110bp. Lanes 1-4: A 10:1 ratio (5.0 μM:0.5 μM) of ssDNA bowtie strand vs dsDNA barcode strand yielded the most efficient production of complete (double-sided) elongation product. Lanes 6-9: Increasing the dilution to 100:1 (5.0 μM: 0.05μM) of ssDNA bowtie strand vs dsDNA barcode strand greatly reduced the synthesis efficiency of complete (double-sided) elongation product. Lanes 11-14: Further dilution to 1000:1 (5.0 μM: 0.005 μM) yielded virtually no complete (double-sided) elongation product.

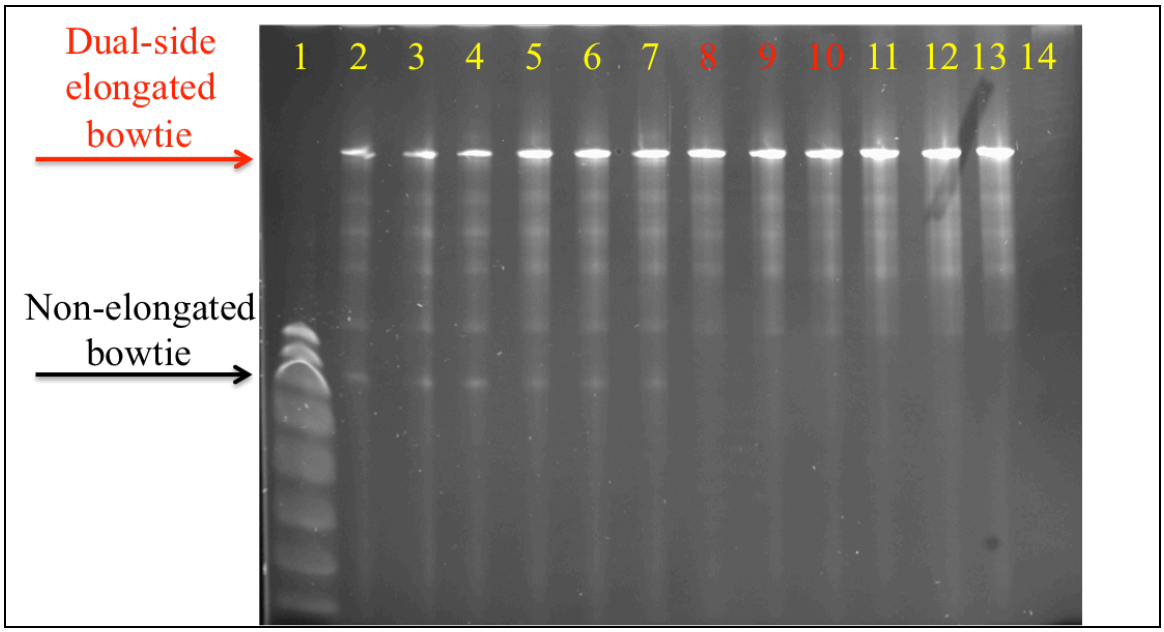

Figure 4.8: Optimization of strand dissociation, annealing, and elongation cycle number. To assess the minimum number of cycles necessary for efficient (double-side) elongation of ssDNA bowtie backbone strands, aliquots of emulsion were run for 5 (lanes 2-4), 10 (lanes 4-7), 20 (lanes 8-10), and 30 (lanes 11-13) cycles and products analyzed by denaturing PAGE. It was determined that 20 cycles (lanes 8-10) are sufficient for dual-sided elongation (a lack of band corresponding to un-elongated ssDNA bowtie at 80 bp).

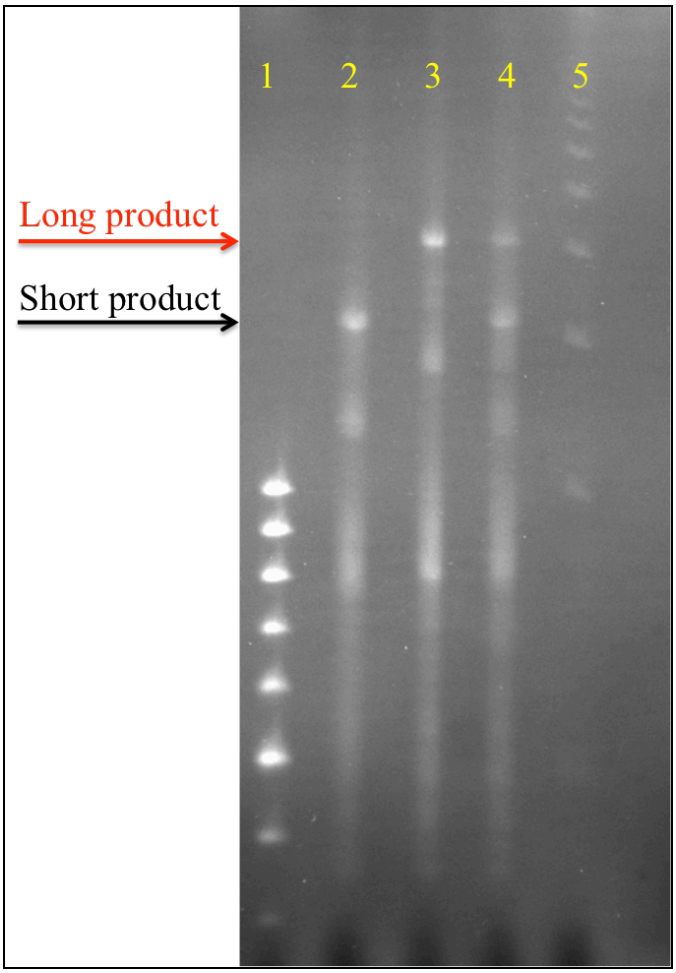

Figure 4.9: Oil-emulsions and limiting dsDNA barcode concentrations ensure complementary barcodes on both ends of bowtie strands. To assess whether ssDNA barcode strands were being elongated by different dsDNA barcode strands (and thus resulting in non-complementary barcodes being incorporated), two dsDNA barcodes of different lengths were included in the same oil-emulsion elongation reaction. Lane 2 demonstrates the elongation product from the short dsDNA barcode strand only (160bp). Lane 3 demonstrates the elongation product from the long dsDNA barcode strand only (200bp). Lane 4 represents the emulsion containing both the short and long dsDNA barcode strands. While products identical to those from the short and long dsDNA barcode only emulsions are present, no hybrid length products (>160bp<200bp) are readily visible, indicating no detectable cross annealing between different dsDNA barcode strands on the same bowtie backbone strand.

DNA origami nanostructures incorporate bowtie barcode mRNA capture strands: Synthesis of the DNA origami nanostructures followed the same protocol as that outlined in chapter 3 with Scaffold M13mp18 ssDNA, staple DNA, and functionalized staples (including the bowtie barcode strand) mixed in a fixed 1:5:3 ratio (50 nM scaffold, 250 nM staples, 150 nM bowtie barcode strand/biotinylated staples/fluorophore conjugated staples) in aqueous buffer (1x TAE with 12 mM Mg^{2+}), followed by thermal denaturation (90°C) and gradual annealing (to 20°C) over 12 hours. Final concentration was measured by A_{260}/A_{280} absorbance and standardized to either 20 or 50 nM depending on the downstream application. Successfully annealed origami nanostructure products were visualized by AFM to confirm incorporation of the bowtie barcode mRNA capture strand.

Sequencing confirms that purification of DNA origami nanostructures with bound TCR mRNA from transfected P14 $CD8^+$ T cells provides suitable input material for RT-PCR, and yields barcoded TCRα and TCRβ CDR3 sequence information. As a proof of principle experiment to validate the bowtie barcode system was capable of capturing, protecting, and being reisolated with both TCRα and TCRβ mRNA, transfections of sorted $CD8^+$ T cells from splenocyte populations from the transgenic TCR P14 mouse were performed. Since the P14 mouse expresses a single known TCRα and TCRβ chain, simple RT-PCR and Sanger sequencing was used to validate sequence capture. Following transfections, nanostructures with bound mRNAs were purified from cell lysate using streptavidin primed purification columns. After

rinsing the columns to remove unbound mRNA and cellular debris, RT reactions were performed directly in the columns using 40 μL of RT mastermix (Omniscript Kit – Qiagen) per manufacturer's protocol with 1 hr incubations in a 37°C heat block. Following RT, the elongated bowtie barcode strands were dissociated from the nanostructures by heating to 95°C for 10 min in a heating block and centrifuging at 10k rpm for 1 min to collect the elongated bowtie barcodes. Standard PCR reactions (Phire kit – Thermo Scientific) were performed using 2 μL of the eluted bowtie barcode. Primers for amplification of the TCRα CDR3 region included a forward primer containing a sequence of the P14 TCRαV14 gene and a reverse primer with the conserved "X" sequence. Amplification of the TCRβ CDR3 region utilized a forward primer for the TCRβV13 gene and a reverse primer with the conserved "Y" sequence. Positive controls consisted of standard gene-specific RT-PCR using extracted P14 $CD8^+$ T cell mRNA and Cα and Cβ RT primers with the same sequence as the bowtie barcode mRNA capture sequences, followed by PCR with the same Vα and Vβ primers described above, but using the Cα and Cβ primers from the RT step (as there was no bowtie barcode in these reactions the use of the X' and Y primers would not have been possible). Additional positive controls using origami with bowtie as RT primers followed by X/Vα14 and Y/Vβ13 PCR were also performed. Negative controls consisted of mock transfections using origami synthesized *without* bowtie barcodes followed by X/Vα14 and Y/Vβ13 PCR. Positive control samples yielded bands of appropriate length based on gel electrophoresis and were confirmed to be correct based on comparison to known P14

TCRα and TCRβ sequences by standard Sanger sequencing **[Figure 4.10A, Lanes 4-5 and 7-8]**. Both amplifications (TCRα and TCRβ) from RT product eluted from purified origami synthesized *with* bowtie barcodes yielded products comparable to positive controls as determined by gel electrophoresis **[Figure 4.10A, Lanes 6 and 9]** and were further confirmed to be correct by comparison to known P14 TCRα and TCRβ sequences as evaluated by standard Sanger sequencing **[Figure 4.10B-C].** Negative controls yielded no detectable RT-PCR product as determined by gel electrophoresis **[Figure 4.10A, Lanes 1-2]**, indicating that no detectable non-specific amplifications were occurring.

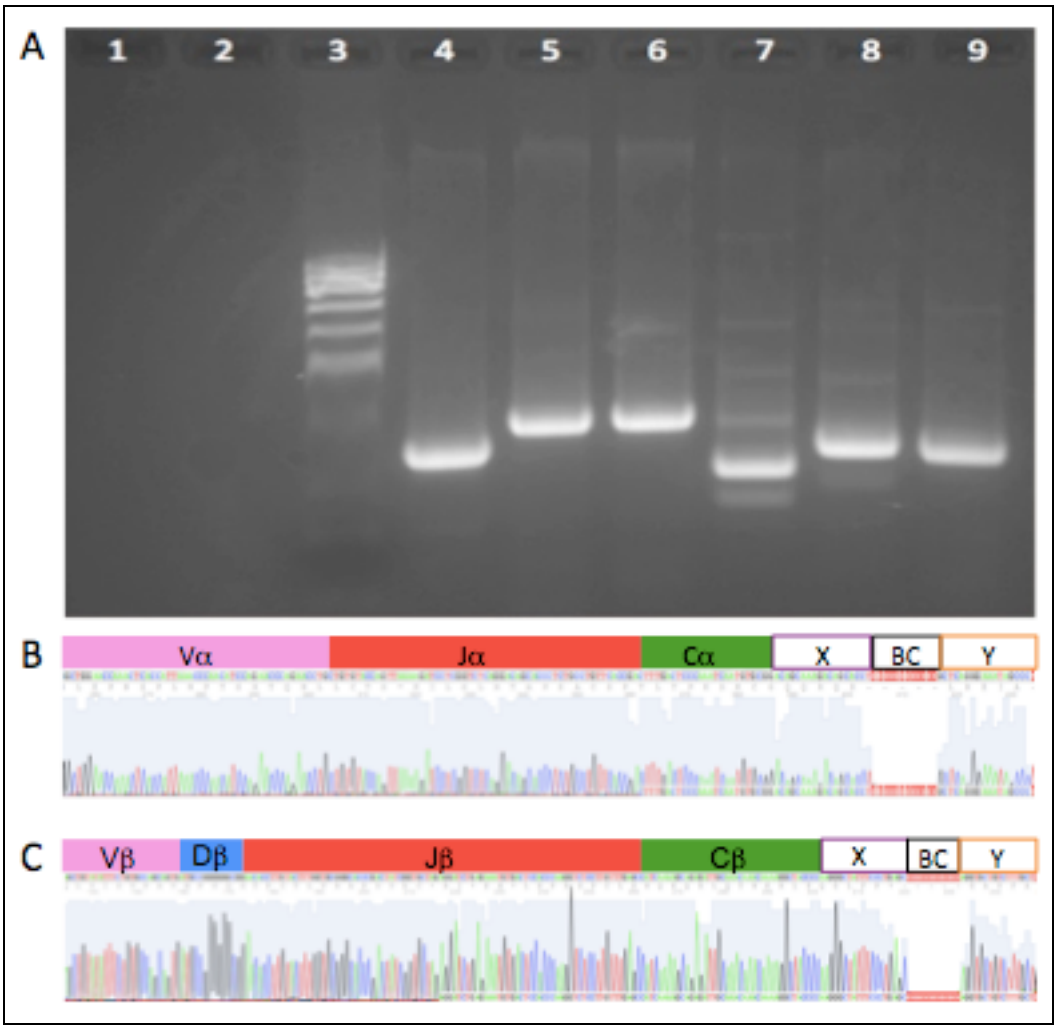

Figure 4.10: RT-PCR and sequencing confirm capture, amplification, and barcoding of TCRα and TCRβ mRNA suitable for downstream genetic analysis. **[A]** Agarose gel electrophoresis of RT-PCR products from origami-transfected P14 CD8 T cells. Lanes 1-2: Negative control RT-PCRs (TCRα and TCRβ respectively) from mock-transfected cells. Lane 3: Ladder. Lanes 4/7: Positive control P14 RNA RT-PCRs using Cα/Cβ RT primers followed by Cα-Vα14/Cβ-Vβ13 PCR primers, respectively. Lanes 5/8: Positive control P14 RNA RT-PCRs using origami as RT primer followed by X-Vα14/Y-Vβ13 PCR primers, respectively. Lanes 6/9: Origami transfected samples with X-Vα14/Y-Vβ13 PCR primers, respectively. **[B]** Sequencing trace of excised band from Lane 6 with corresponding TCRα gene/bowtie sequence aligned above. **[C]** Sequencing trace of excised band from Lane 9 with corresponding TCRβ gene/bowtie sequence aligned above.

DISCUSSION

While the aim of this thesis was to demonstrate the ability of bowtie incorporated DNA origami nanostructures to capture, barcode, and thus link TCR mRNAs from individual cells, future experiments will seek to evaluate the use of the DNA origami nanostructure bowtie barcode approach to pair thousands or even millions of TCR gene sequences from wild type (wt) individuals in a single experiment. Following transfections of wt C57BL/6 $CD8^+$ T cells, cell lysis, and RT, a multiplex PCR will be performed using a single primer for both of the conserved priming sequences on either end of the mRNA capture probes (Y and X) and the well-established multiplex primer system [202] for the 23 functional murine TCRα V-families and the 19 functional TCRβ V-families **(Table 4.2, Figure 4.11)** using manufacturer's protocols of a commercially available multiplex PCR kit (Multiplex Kit – Qiagen) to generate a pool of barcoded amplicons with corresponding TCRα and TCRβ CDR3 sequences within the resulting PCR products. Obtained amplicons should be roughly 350 bp in length and will be immediately suitable for use in standard Illumina paired-end sequencing (to be performed under standard Illumina protocols [240]). Reads will first be subjected to a series of quality control steps for quantifying biases at any given base, and will then be parsed into independent FASTQ files for alignment using BWA-MEM for accurate split-read alignment of the unique CDR3 sequences. Using the well-practiced standard for immune receptor gene calling strategy, each sequence will be required to have a 12 nucleotide match to one of the Vα or Vβ gene segments, corresponding to the CASS consensus amino acid sequence from the second conserved cysteine at the 3' end of the V segment,

as well as a 6 nucleotide match to the J segment corresponding to the conserved phenylalanine [192, 232, 236, 246, 266, 272]. The total number of nucleotides between these codons determines the length and therefore the reading frame of the CDR3 region. Processed sequence data will then deposited in the ASU secure relational database management system, which allows a WebApp front end through JasperSoft Server as well as a secure MongoDB instance allowing Ad Hoc querying. Pairing of TCRα and TCRβ sequences form individual cells will be conducted by a simple "if-then" algorithm, searching for complementary base pairing at the 12-mer-barcode-sequence stretch of each aligned read.

Table 4.2: Multiplex PCR system for amplification of C57BL/6 mouse TCR genes. TRAV AND TRBV primer sequences were obtained from the ImmunoGenetics TRAV/TRBV primer database [202].

Primer	Sequence	Primer	Sequence
TRAV1	CTCCACATTCCTGAGCC	TRBV1	GTATCCCTGGATGAGCTG
TRAV2	ACTCTGAGCCTGCCCT	TRBV2	GGACAATCAGACTGCCTC
TRAV3	GCCCTCCTCACCTGAG	TRBV3	GATATGGGGCAGATGGTG
TRAV4	AGGAACAAAGGAGAAT	TRBV4	CAGGTGGGAAATGAAGTG
TRAV5	GGAGAAGGTCCACAGCTC	TRBV5	GCCAGAGCTCATGTTTCTC
TRAV6A	GGAGAAGGTCCACAGCTC	TRBV12	CCAGCAGATTCTCAGTCC
TRAV6B	CAACTGCCAACAACAAGG	TRBV13	GTACTGGTATCGGCAGGAC
TRAV6C	GTTCTGGTATGTGCAGTATCC	TRBV14	GGTATCAGCAGCCCAGAG
TRAV6D	TCCTTCCACTTGCAGAAAG	TRBV15	GTGTGAGCCAGTTTCAGG
TRAV7	CAGCAGAGCCCAGAATC	TRBV16	GAAGCAACTCTGTGGTGTG
TRAV8	AGAGCCACCCTTGACAC	TRBV17	GAACAGGGAAGCTGACAC
TRAV9	CCAGTGGTTCAAGGAGTG	TRBV19	GGTACCGACAGGATTCAG
TRAV10	CTACACTGAGTGTTCGAGAGG	TRBV20	GCTTGGTATCGTCAATCG
TRAV11	AACAGGACACAGGCAAAG	TRBV23	GCCAGGAAGCAGAGATG
TRAV12	TGACCCAGACAGAAGGC	TRBV24	GCACACTGCCTTTTACTGG
TRAV13	TCCTTGGTTCTGCAGG	TRBV26	GAGGTGTATCCCTGAAAAGG
TRAV14	CTCTGACAGTCTGGGAAGG	TRBV29	GTACTGGTATCGACAAGACCC
TRAV15	TTAGTGGAGAGATGG	TRBV30	GGACATCTGTCAAAGTGGC
TRAV16	ATTATTCTCTGAACTTTCAGAAGC	TRBV31	CTGTTGGCCAGGTAGAGTC
TRAV17	CAGTCCGTGGACCAGC	Y	GGACAGCAAAGACAGCACCT
TRAV18	CAAGATTTCACCGCACG		
TRAV19	GCTGACTGTTCAAGAGGGA		
TRAV21	AATAGTATGGCTTTCCTGGC		
X’	CAAGGGCTATTTCCCTGAGC		

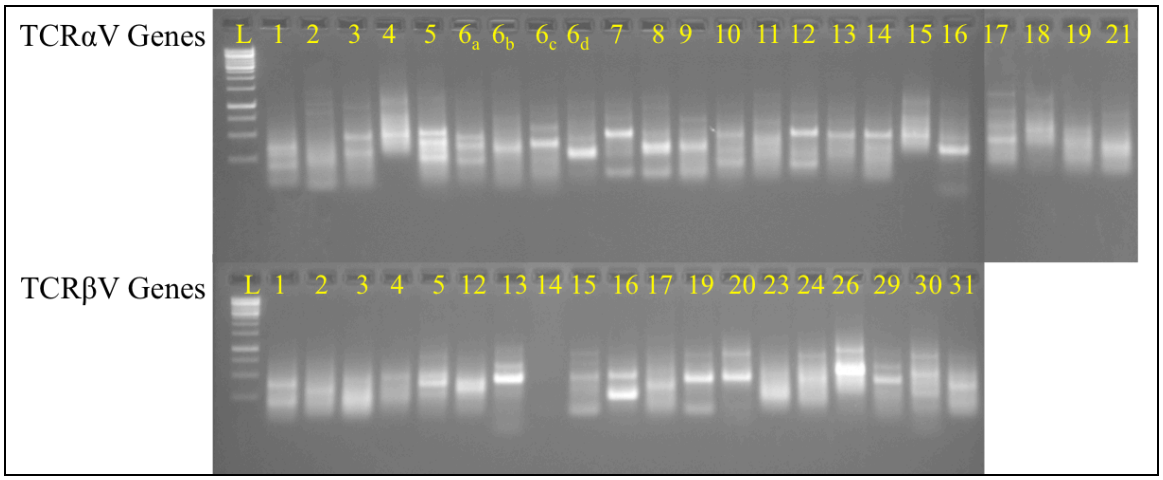

Figure 4.11: Confirmation of C57BL/6 TCRα/β multiplex primer system. Each individual forward primer from the TCRα multiplex pool (23 primers) and TCRβ multiplex pool (19 primers) was subjected to PCR testing using RT products from total cellular $CD8^+$ mRNA RT reactions. (Top) The forward TCRα primers were paired with a single Cα-specific reverse primer and (Bottom) the forward TCRβ primers were paired with a single Cβ-specific reverse primer. Multiple products were observed for many of the primer pairs as wild type mice will express numerous reorganizations for each Vα/Vβ gene selected.

The recent advancements in NGS technologies have profoundly increased the ability of researchers to study and evaluate immune repertoires. The sheer amount of raw data alone of which NGS-based approaches allow for acquisition would have been unheard of merely ten years ago. With improvements on the efficiency and streamlining of methodology of not only our protocol, but all future immune repertoire analysis approaches, the rate limiting step will eventually fall on data management and analysis rather than the actual wet-lab bench work. As proposed by many researchers in the field of immune repertoire analysis, the establishment of communal databases, improvements in computational algorithms and software for analyzing massive repertoire data sets will be paramount in the continued advancement of the study of immunology [340, 347, 361].

Up to this point most robust immune repertoire sequencing approaches have been focused on high-throughput single-chain analysis (i.e. TCRβ) or low-throughput single-cell analysis of both chains (TCRα/β or IgH/L). Single chain approaches do not take into account the cognate pairing of alpha chains and beta chains (or heavy and light chains), and therefore do not provide the true identity of the antigen receptor as they only shed light on 'half the story'. Single-cell sorting approaches only allow for the analysis of hundreds of cells from a given experiment and therefore do not provide a representative sample of any meaningful data set. In order to identify TCRs (or BCRs/antibodies) for therapeutic use, functional analysis, or vaccine efficacy, sequence information from **both** TCRα and TCRβ chains (or heavy and light chains) must be identified as a pair. Single-celled approaches for pairing TCR chains up to this point have relied on numerous different approaches. Some rely on isolating individual lymphocytes and physically

linking the chains by bridge amplification PCR before sequencing. Alternatively, some groups have investigated approaches for uniquely barcoding the chains from individual cells. These approaches however have proven limited both by cell throughput and efficiency and furthermore require fabrication and operation of highly complicated microfluidic devices [262, 357]. Advancements in oil-emulsion PCR systems led other groups to investigate the use of cell encapsulation, lysis, and RT-PCR all within individual oil emulsion droplets, linking the TCR chains by overlap extension PCR. The published work however only demonstrated this approach for a single TCRβV gene subset of cells (TCRβV7) and extremely poor efficiency (~700 TCRα/β pairs from $8x10^6$ input cells) [262]. Improvements on the oil-emulsion approach led another group to investigate the use of bead capture of individual B cell mRNA followed by linkage PCR in emulsion droplets containing one bead each, again however modest paired sequence yields and low efficiency was observed [261].

Here we present a novel approach at high-throughput paired-chain analysis that requires neither initial stage single-cell sorting, nor the need for single-cell/bead oil emulsion droplet capture, the limiting factors in the experiments described above. Our overall goal was to develop a robust method for sequence analysis of linked genes from individual cells without the need for single-cell sorting or specialized equipment that is not only cost-effective and has the capacity for analysis of large cell populations, but was also easily adaptable for probing sequence diversity of paired genes from other heterogeneous cell populations. To such end we have developed novel DNA origami nanostructures that are able to capture and protect mRNA from transfected cells and by

inclusion of novel complementary bowtie barcodes, link cDNA sequences obtained via sequencing. We have developed the DNA origami nanostructures and optimized processes for transfection, mRNA capture and recovery of TCRα and TCRβ mRNA from antigen-specific T cells. We analyzed the sequences captured by this method and demonstrated the ability to barcode captured TCR sequences from transfected cells. We validated our nanotechnology approach to obtain linked sequence information for both TCR chains from individual cells by utilizing transgenic mice expressing a single known TCR. Finally, we have begun the process of adapting this technology to identify heterogeneous TCR expressed by wild type mice.

The proposed DNA origami nanostructure system for linking paired genes from single cells can be applied to a myriad of other biological questions, including current applications for the assessment of heterodimeric B cell receptors, characterization of the sequence mutation history of cancer cells or identification of rare cancer stem cells, and has the potential to revolutionize our understanding of the evolution of both immunity and disease.

REFERENCES

1. Thucydides, R. Crawley, and J. Gavorse, *The complete writings of Thucydides. The Peloponnesian war*. The modern library of the world's best books. 1934, New York,: The Modern Library. xxviii, 516 p.

2. Masopust, D., et al., *A brief history of CD8 T cells.* Eur J Immunol, 2007. **37 Suppl 1**: p. S103-10.

3. Murphy, J.B., *Factors of Resistance to Heteroplastic Tissue-Grafting : Studies in Tissue Specificity. Iii.* J Exp Med, 1914. **19**(5): p. 513-22.

4. Murphy, J.B., *Studies in Tissue Specificity : Ii. The Ultimate Fate of Mammalian Tissue Implanted in the Chick Embryo.* J Exp Med, 1914. **19**(2): p. 181-6.

5. Gowans, J.L. and E.J. Knight, *The Route of Re-Circulation of Lymphocytes in the Rat.* Proc R Soc Lond B Biol Sci, 1964. **159**: p. 257-82.

6. Gowans, J.L. and J.W. Uhr, *The carriage of immunological memory by small lymphocytes in the rat.* J Exp Med, 1966. **124**(5): p. 1017-30.

7. Howard, J.C. and J.L. Gowans, *The role of lymphocytes in antibody formation. 3. The origin from small lymphocytes of cells forming direct and indirect haemolytic plaques to sheep erythrocytes in the rat.* Proc R Soc Lond B Biol Sci, 1972. **182**(1067): p. 193-209.

8. Ellis, S.T. and J.L. Gowans, *The role of lymphocytes in antibody formation. V. Transfer of immunological memory to tetanus toxoid: the origin of plasma cells from small lymphocytes, stimulation of memory cells in vitro and the persistence of memory after cell-transfer.* Proc R Soc Lond B Biol Sci, 1973. **183**(1071): p. 125-39.

9. Miller, J.F., *Immunological function of the thymus.* Lancet, 1961. **2**(7205): p. 748-9.

10. Miller, J.F. and G.F. Mitchell, *The thymus and the precursors of antigen reactive cells.* Nature, 1967. **216**(5116): p. 659-63.

11. Mitchell, G.F. and J.F. Miller, *Immunological activity of thymus and thoracic-duct lymphocytes.* Proc Natl Acad Sci U S A, 1968. **59**(1): p. 296-303.

12. Miller, J.F., *Discovering the origins of immunological competence.* Annu Rev Immunol, 1999. **17**: p. 1-17.

13. Cerottini, J.C., A.A. Nordin, and K.T. Brunner, *Specific in vitro cytotoxicity of thymus-derived lymphocytes sensitized to alloantigens.* Nature, 1970. **228**(5278): p. 1308-9.

14. Golstein, P., et al., *Cells mediating specific in vitro cytotoxicity. I. Detection of receptor-bearing lymphocytes.* J Exp Med, 1971. **134**(6): p. 1385-402.

15. Brondz, B.D., *Complex specificity of immune lymphocytes in allogeneic cell cultures.* Folia Biol (Praha), 1968. **14**(2): p. 115-31.

16. Brondz, B.D. and N.E. Golberg, *Further in vitro evidence for polyvalent specificity of immune lymphocytes.* Folia Biol (Praha), 1970. **16**(1): p. 20-8.

17. Zinkernagel, R.M. and P.C. Doherty, *Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system.* Nature, 1974. **248**(5450): p. 701-2.

18. Zinkernagel, R.M. and P.C. Doherty, *Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytic choriomeningitis.* Nature, 1974. **251**(5475): p. 547-8.

19. Allison, J.P., B.W. McIntyre, and D. Bloch, *Tumor-specific antigen of murine T-lymphoma defined with monoclonal antibody.* J Immunol, 1982. **129**(5): p. 2293-300.

20. Meuer, S.C., et al., *Clonotypic structures involved in antigen-specific human T cell function. Relationship to the T3 molecular complex.* J Exp Med, 1983. **157**(2): p. 705-19.

21. Shimonkevitz, R., et al., *Antigen recognition by H-2-restricted T cells. I. Cell-free antigen processing.* J Exp Med, 1983. **158**(2): p. 303-16.

22. Townsend, A.R., F.M. Gotch, and J. Davey, *Cytotoxic T cells recognize fragments of the influenza nucleoprotein.* Cell, 1985. **42**(2): p. 457-67.

23. Yewdell, J.W., et al., *Influenza A virus nucleoprotein is a major target antigen for cross-reactive anti-influenza A virus cytotoxic T lymphocytes.* Proc Natl Acad Sci U S A, 1985. **82**(6): p. 1785-9.

24. Kappler, J., et al., *The major histocompatibility complex-restricted antigen receptor on T cells in mouse and man: identification of constant and variable peptides.* Cell, 1983. **35**(1): p. 295-302.

25. Kappler, J., et al., *The mouse T cell receptor: comparison of MHC-restricted receptors on two T cell hybridomas.* Cell, 1983. **34**(3): p. 727-37.

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

27. Hedrick, S.M., et al., *Isolation of cDNA clones encoding T cell-specific membrane-associated proteins.* Nature, 1984. **308**(5955): p. 149-53.

28. Hedrick, S.M., et al., *Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins.* Nature, 1984. **308**(5955): p. 153-8.

29. Yanagi, Y., et al., *A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains.* Nature, 1984. **308**(5955): p. 145-9.

30. Bjorkman, P.J., et al., *Structure of the human class I histocompatibility antigen, HLA-A2.* Nature, 1987. **329**(6139): p. 506-12.

31. Bjorkman, P.J., et al., *The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens.* Nature, 1987. **329**(6139): p. 512-8.

32. Blue, M.L., et al., *Evidence for specific association between class I major histocompatibility antigens and the CD8 molecules of human suppressor/cytotoxic cells.* Cell, 1988. **54**(3): p. 413-21.

33. Norment, A.M., et al., *Cell-cell adhesion mediated by CD8 and MHC class I molecules.* Nature, 1988. **336**(6194): p. 79-81.

34. Rosenstein, Y., et al., *Direct evidence for binding of CD8 to HLA class I antigens.* J Exp Med, 1989. **169**(1): p. 149-60.

35. Garboczi, D.N., et al., *Structure of the complex between human T-cell receptor, viral peptide and HLA-A2.* Nature, 1996. **384**(6605): p. 134-41.

36. Murali-Krishna, K., et al., *Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection.* Immunity, 1998. **8**(2): p. 177-87.

37. Altman, J.D., et al., *Phenotypic analysis of antigen-specific T lymphocytes.* Science, 1996. **274**(5284): p. 94-6.

38. Sallusto, F., J. Geginat, and A. Lanzavecchia, *Central memory and effector memory T cell subsets: function, generation, and maintenance.* Annu Rev Immunol, 2004. **22**: p. 745-63.

39. Williams, M.A. and M.J. Bevan, *Effector and memory CTL differentiation.* Annu Rev Immunol, 2007. **25**: p. 171-92.

40. Shin, H. and E.J. Wherry, *CD8 T cell dysfunction during chronic viral infection.* Curr Opin Immunol, 2007. **19**(4): p. 408-15.

41. Masopust, D., et al., *The role of programming in memory T-cell development.* Curr Opin Immunol, 2004. **16**(2): p. 217-25.

42. Beutler, B., *Innate immunity: an overview.* Mol Immunol, 2004. **40**(12): p. 845-59.

43. Fornet, W. and W. Heubner, *Tests on the development of sepsin.* Archiv Fur Experimentelle Pathologie Und Pharmakologie, 1908: p. 176-180.

44. Fornet, W. and W. Heubner, *Experiments on the formation of sepsin. Second report.* Archiv Fur Experimentelle Pathologie Und Pharmakologie, 1911. **65**(5/6): p. 428-453.

45. Nikaido, H., *Studies on the biosynthesis of cell wall polysaccharide in mutant strains of Salmonella. II.* Proc Natl Acad Sci U S A, 1962. **48**: p. 1542-8.

46. Osborn, M.J., *Studies on the Gram-Negative Cell Wall. I. Evidence for the Role of 2-Keto- 3-Deoxyoctonate in the Lipopolysaccharide of Salmonella Typhimurium.* Proc Natl Acad Sci U S A, 1963. **50**: p. 499-506.

47. Kimbrell, D.A. and B. Beutler, *The evolution and genetics of innate immunity.* Nat Rev Genet, 2001. **2**(4): p. 256-67.

48. Poltorak, A., et al., *Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene.* Science, 1998. **282**(5396): p. 2085-8.

49. Chuang, T.H. and R.J. Ulevitch, *Cloning and characterization of a sub-family of human toll-like receptors: hTLR7, hTLR8 and hTLR9.* Eur Cytokine Netw, 2000. **11**(3): p. 372-8.

50. Du, X., et al., *Three novel mammalian toll-like receptors: gene structure, expression, and evolution.* Eur Cytokine Netw, 2000. **11**(3): p. 362-71.

51. Chuang, T. and R.J. Ulevitch, *Identification of hTLR10: a novel human Toll-like receptor preferentially expressed in immune cells.* Biochim Biophys Acta, 2001. **1518**(1-2): p. 157-61.

52. Campos, M.A., et al., *Activation of Toll-like receptor-2 by glycosylphosphatidylinositol anchors from a protozoan parasite.* J Immunol, 2001. **167**(1): p. 416-23.

53. Hoebe, K., et al., *Upregulation of costimulatory molecules induced by lipopolysaccharide and double-stranded RNA occurs by Trif-dependent and Trif-independent pathways.* Nat Immunol, 2003. **4**(12): p. 1223-9.

54. Brinkmann, V., et al., *Neutrophil extracellular traps kill bacteria.* Science, 2004. **303**(5663): p. 1532-5.

55. Trowsdale, J., *HLA genomics in the third millennium.* Curr Opin Immunol, 2005. **17**(5): p. 498-504.

56. Mantegazza, A.R., et al., *Presentation of phagocytosed antigens by MHC class I and II.* Traffic, 2013. **14**(2): p. 135-52.

57. Hildner, K., et al., *Batf3 deficiency reveals a critical role for CD8alpha+ dendritic cells in cytotoxic T cell immunity.* Science, 2008. **322**(5904): p. 1097-100.

58. Savina, A., et al., *The small GTPase Rac2 controls phagosomal alkalinization and antigen crosspresentation selectively in CD8(+) dendritic cells.* Immunity, 2009. **30**(4): p. 544-55.

59. Gordon, S., *Elie Metchnikoff: father of natural immunity.* Eur J Immunol, 2008. **38**(12): p. 3257-64.

60. Flannagan, R.S., V. Jaumouille, and S. Grinstein, *The cell biology of phagocytosis.* Annu Rev Pathol, 2012. **7**: p. 61-98.

61. Swanson, J.A., *Shaping cups into phagosomes and macropinosomes.* Nat Rev Mol Cell Biol, 2008. **9**(8): p. 639-49.

62. Savina, A. and S. Amigorena, *Phagocytosis and antigen presentation in dendritic cells.* Immunol Rev, 2007. **219**: p. 143-56.

63. Ramachandra, L., D. Simmons, and C.V. Harding, *MHC molecules and microbial antigen processing in phagosomes.* Curr Opin Immunol, 2009. **21**(1): p. 98-104.

64. Kimura, H., et al., *New Insights into the Function of the Immunoproteasome in Immune and Nonimmune Cells.* J Immunol Res, 2015. **2015**: p. 541984.

65. Kloetzel, P.M., *Antigen processing by the proteasome.* Nat Rev Mol Cell Biol, 2001. **2**(3): p. 179-87.

66. Shachar, I. and N. Karin, *The dual roles of inflammatory cytokines and chemokines in the regulation of autoimmune diseases and their clinical implications.* J Leukoc Biol, 2013. **93**(1): p. 51-61.

67. Unno, M., et al., *The structure of the mammalian 20S proteasome at 2.75 A resolution.* Structure, 2002. **10**(5): p. 609-18.

68. Ortiz-Navarrete, V., et al., *Subunit of the '20S' proteasome (multicatalytic proteinase) encoded by the major histocompatibility complex.* Nature, 1991. **353**(6345): p. 662-4.

69. Glynne, R., et al., *A proteasome-related gene between the two ABC transporter loci in the class II region of the human MHC.* Nature, 1991. **353**(6342): p. 357-60.

70. Kelly, A., et al., *Second proteasome-related gene in the human MHC class II region.* Nature, 1991. **353**(6345): p. 667-8.

71. Realini, C., et al., *Molecular cloning and expression of a gamma-interferon-inducible activator of the multicatalytic protease.* J Biol Chem, 1994. **269**(32): p. 20727-32.

72. Hewitt, E.W., *The MHC class I antigen presentation pathway: strategies for viral immune evasion.* Immunology, 2003. **110**(2): p. 163-9.

73. Schubert, U., et al., *Rapid degradation of a large fraction of newly synthesized proteins by proteasomes.* Nature, 2000. **404**(6779): p. 770-4.

74. Princiotta, M.F., et al., *Quantitating protein synthesis, degradation, and endogenous antigen processing.* Immunity, 2003. **18**(3): p. 343-54.

75. Jones, E.Y., *MHC class I and class II structures.* Curr Opin Immunol, 1997. **9**(1): p. 75-9.

76. Zhang, J.H., et al., *Narrow Groove and Restricted Anchors of MHC Class I Molecule BF2*0401 Plus Peptide Transporter Restriction Can Explain Disease Susceptibility of B4 Chickens.* Journal of Immunology, 2012. **189**(9): p. 4478-4487.

77. Painter, C.A. and L.J. Stern, *Structural Insights Into HLA-DM Mediated MHC II Peptide Exchange.* Curr Top Biochem Res, 2011. **13**(2): p. 39-55.

78. Abele, R. and R. Tampe, *Function of the transport complex TAP in cellular immune recognition.* Biochim Biophys Acta, 1999. **1461**(2): p. 405-19.

79. van Endert, P.M., et al., *Powering the peptide pump: TAP crosstalk with energetic nucleotides.* Trends Biochem Sci, 2002. **27**(9): p. 454-61.

80. Pamer, E. and P. Cresswell, *Mechanisms of MHC class I--restricted antigen processing.* Annu Rev Immunol, 1998. **16**: p. 323-58.

81. Antoniou, A.N., S.J. Powis, and T. Elliott, *Assembly and export of MHC class I peptide ligands.* Curr Opin Immunol, 2003. **15**(1): p. 75-81.

82. Watts, C., *Capture and processing of exogenous antigens for presentation on MHC molecules.* Annu Rev Immunol, 1997. **15**: p. 821-50.

83. Cresswell, P., *Invariant chain structure and MHC class II function.* Cell, 1996. **84**(4): p. 505-7.

84. Roche, P.A., et al., *Cell surface HLA-DR-invariant chain complexes are targeted to endosomes by rapid internalization.* Proc Natl Acad Sci U S A, 1993. **90**(18): p. 8581-5.

85. Peters, P.J., et al., *Segregation of MHC class II molecules from MHC class I molecules in the Golgi complex for transport to lysosomal compartments.* Nature, 1991. **349**(6311): p. 669-76.

86. Riberdy, J.M., et al., *HLA-DR molecules from an antigen-processing mutant cell line are associated with invariant chain peptides.* Nature, 1992. **360**(6403): p. 474-7.

87. Denzin, L.K. and P. Cresswell, *HLA-DM induces CLIP dissociation from MHC class II alpha beta dimers and facilitates peptide loading.* Cell, 1995. **82**(1): p. 155-65.

88. Kropshofer, H., et al., *Editing of the HLA-DR-peptide repertoire by HLA-DM.* EMBO J, 1996. **15**(22): p. 6144-54.

89. Steinman, R.M., *Dendritic cells: understanding immunogenicity.* Eur J Immunol, 2007. **37 Suppl 1**: p. S53-60.

90. Segura, E. and J.A. Villadangos, *A modular and combinatorial view of the antigen cross-presentation pathway in dendritic cells.* Traffic, 2011. **12**(12): p. 1677-85.

91. Rodriguez, A., et al., *Selective transport of internalized antigens to the cytosol for MHC class I presentation in dendritic cells.* Nat Cell Biol, 1999. **1**(6): p. 362-8.

92. Hoebe, K., E. Janssen, and B. Beutler, *The interface between innate and adaptive immunity.* Nature Immunology, 2004. **5**(10): p. 971-974.

93. Dutton, R.W., *In vitro studies of immunological responses of lymphoid cells.* Adv Immunol, 1967. **6**: p. 253-336.

94. Unanue, E.R. and B.A. Askonas, *The immune response of mice to antigen in macrophages.* Immunology, 1968. **15**(2): p. 287-96.

95. Romagnani, S., *Human TH1 and TH2 subsets: doubt no more.* Immunol Today, 1991. **12**(8): p. 256-7.

96. Robins, H.S., et al., *Comprehensive assessment of T-cell receptor beta-chain diversity in alpha beta T cells.* Blood, 2009. **114**(19): p. 4099-4107.

97. Acuto, O. and F. Michel, *CD28-mediated co-stimulation: a quantitative support for TCR signalling.* Nat Rev Immunol, 2003. **3**(12): p. 939-51.

98. Smith-Garvin, J.E., G.A. Koretzky, and M.S. Jordan, *T cell activation.* Annu Rev Immunol, 2009. **27**: p. 591-619.

99. Lever, M., et al., *Phenotypic models of T cell activation.* Nat Rev Immunol, 2014. **14**(9): p. 619-29.

100. Andersen, P.S., et al., *Role of the T cell receptor ligand affinity in T cell activation by bacterial superantigens.* J Biol Chem, 2001. **276**(36): p. 33452-7.

101. Krogsgaard, M., et al., *Evidence that structural rearrangements and/or flexibility during TCR binding can contribute to T cell activation.* Mol Cell, 2003. **12**(6): p. 1367-78.

102. Holler, P.D. and D.M. Kranz, *Quantitative analysis of the contribution of TCR/pepMHC affinity and CD8 to T cell activation.* Immunity, 2003. **18**(2): p. 255-64.

103. Kalergis, A.M., et al., *Efficient T cell activation requires an optimal dwell-time of interaction between the TCR and the pMHC complex.* Nat Immunol, 2001. **2**(3): p. 229-34.

104. Irving, M., et al., *Interplay between T cell receptor binding kinetics and the level of cognate peptide presented by major histocompatibility complexes governs CD8+ T cell responsiveness.* J Biol Chem, 2012. **287**(27): p. 23068-78.

105. Corse, E., et al., *Attenuated T cell responses to a high-potency ligand in vivo.* PLoS Biol, 2010. **8**(9).

106. Dushek, O., et al., *Antigen potency and maximal efficacy reveal a mechanism of efficient T cell activation.* Sci Signal, 2011. **4**(176): p. ra39.

107. Coombs, D., et al., *Activated TCRs remain marked for internalization after dissociation from pMHC.* Nat Immunol, 2002. **3**(10): p. 926-31.

108. Gonzalez, P.A., et al., *T cell receptor binding kinetics required for T cell activation depend on the density of cognate ligand on the antigen-presenting cell.* Proc Natl Acad Sci U S A, 2005. **102**(13): p. 4824-9.

109. McKeithan, T.W., *Kinetic proofreading in T-cell receptor signal transduction.* Proc Natl Acad Sci U S A, 1995. **92**(11): p. 5042-6.

110. Valitutti, S. and A. Lanzavecchia, *Serial triggering of TCRs: a basis for the sensitivity and specificity of antigen recognition.* Immunol Today, 1997. **18**(6): p. 299-304.

111. Wofsy, C., D. Coombs, and B. Goldstein, *Calculations show substantial serial engagement of T cell receptors.* Biophys J, 2001. **80**(2): p. 606-12.

112. Dushek, O. and D. Coombs, *Analysis of serial engagement and peptide-MHC transport in T cell receptor microclusters.* Biophys J, 2008. **94**(9): p. 3447-60.

113. Vyas, J.M., A.G. Van der Veen, and H.L. Ploegh, *The known unknowns of antigen processing and presentation.* Nat Rev Immunol, 2008. **8**(8): p. 607-18.

114. Kurts, C., B.W. Robinson, and P.A. Knolle, *Cross-priming in health and disease.* Nat Rev Immunol, 2010. **10**(6): p. 403-14.

115. Crotzer, V.L. and J.S. Blum, *Autophagy and adaptive immunity.* Immunology, 2010. **131**(1): p. 9-17.

116. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nat Rev Immunol, 2011. **11**(12): p. 823-36.

117. Bach, F.H., M.L. Bach, and P.M. Sondel, *Differential function of major histocompatibility complex antigens in T-lymphocyte activation.* Nature, 1976. **259**(5541): p. 273-81.

118. Cantor, H. and E.A. Boyse, *Regulation of cellular and humoral immune responses by T-cell subclasses.* Cold Spring Harb Symp Quant Biol, 1977. **41 Pt 1**: p. 23-32.

119. Cantor, H. and E.A. Boyse, *Functional subclasses of T lymphocytes bearing different Ly antigens. II. Cooperation between subclasses of Ly+ cells in the generation of killer activity.* J Exp Med, 1975. **141**(6): p. 1390-9.

120. Doyle, C. and J.L. Strominger, *Interaction between CD4 and class II MHC molecules mediates cell adhesion.* Nature, 1987. **330**(6145): p. 256-9.

121. Chervin, A.S., et al., *Cutting edge: inhibitory effects of CD4 and CD8 on T cell activation induced by high-affinity noncognate ligands.* J Immunol, 2009. **183**(12): p. 7639-43.

122. Wyer, J.R., et al., *T cell receptor and coreceptor CD8 alphaalpha bind peptide-MHC independently and with distinct kinetics.* Immunity, 1999. **10**(2): p. 219-25.

123. Xiong, Y., et al., *T Cell Receptor Binding to a pMHCII Ligand Is Kinetically Distinct from and Independent of CD4.* J Biol Chem, 2001. **276**(8): p. 5659-67.

124. Yin, Y., X.X. Wang, and R.A. Mariuzza, *Crystal structure of a complete ternary complex of T-cell receptor, peptide-MHC, and CD4.* Proc Natl Acad Sci U S A, 2012. **109**(14): p. 5405-10.

125. Li, Y.L., Y.Y. Yin, and R.A. Mariuzza, *Structural and biophysical insights into the role of CD4 and CD8 in T cell activation.* Frontiers in Immunology, 2013. **4**.

126. Yachi, P.P., et al., *T cell activation enhancement by endogenous pMHC acts for both weak and strong agonists but varies with differentiation state.* J Exp Med, 2007. **204**(11): p. 2747-57.

127. Gao, G.F., Z. Rao, and J.I. Bell, *Molecular coordination of alphabeta T-cell receptors and coreceptors CD8 and CD4 in their recognition of peptide-MHC ligands.* Trends Immunol, 2002. **23**(8): p. 408-13.

128. Stone, J.D., A.S. Chervin, and D.M. Kranz, *T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity.* Immunology, 2009. **126**(2): p. 165-76.

129. Luescher, I.F., et al., *CD8 modulation of T-cell antigen receptor-ligand interactions on living cytotoxic T lymphocytes.* Nature, 1995. **373**(6512): p. 353-6.

130. Huppa, J.B., et al., *TCR-peptide-MHC interactions in situ show accelerated kinetics and increased affinity.* Nature, 2010. **463**(7283): p. 963-7.

131. Artyomov, M.N., et al., *CD4 and CD8 binding to MHC molecules primarily acts to enhance Lck delivery.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16916-21.

132. Bank, I. and L. Chess, *Perturbation of the T4 molecule transmits a negative signal to T cells.* J Exp Med, 1985. **162**(4): p. 1294-303.

133. Rosoff, P.M., S.J. Burakoff, and J.L. Greenstein, *The role of the L3T4 molecule in mitogen and antigen-activated signal transduction.* Cell, 1987. **49**(6): p. 845-53.

134. Ledbetter, J.A., et al., *Signal transduction through CD4 receptors: stimulatory vs. inhibitory activity is regulated by CD4 proximity to the CD3/T cell receptor.* Eur J Immunol, 1988. **18**(4): p. 525-32.

135. Samelson, L.E., et al., *A 20-kDa protein associated with the murine T-cell antigen receptor is phosphorylated in response to activation by antigen or concanavalin A.* Proc Natl Acad Sci U S A, 1985. **82**(7): p. 1969-73.

136. Veillette, A., et al., *The CD4 and CD8 T cell surface antigens are associated with the internal membrane tyrosine-protein kinase p56lck.* Cell, 1988. **55**(2): p. 301-8.

137. Haskins, K., et al., *The major histocompatibility complex-restricted antigen receptor on T cells. I. Isolation with a monoclonal antibody.* J Exp Med, 1983. **157**(4): p. 1149-69.

138. Borst, J., et al., *The delta- and epsilon-chains of the human T3/T-cell receptor complex are distinct polypeptides.* Nature, 1984. **312**(5993): p. 455-8.

139. Schoettle, L.N., Poindexter, M. E., and Blattman, J. N., *T-cell Receptors*, in *eLS*. 2015, John Wiley & Sons Ltd,: Chichester.

140. Weiss, A. and J.D. Stobo, *Requirement for the coexpression of T3 and the T cell antigen receptor on a malignant human T cell line.* J Exp Med, 1984. **160**(5): p. 1284-99.

141. Manger, B., et al., *The role of protein kinase C in transmembrane signaling by the T cell antigen receptor complex. Effects of stimulation with soluble or immobilized CD3 antibodies.* J Immunol, 1987. **139**(8): p. 2755-60.

142. Weiss, A. and J.B. Imboden, *Cell surface molecules and early events involved in human T lymphocyte activation.* Adv Immunol, 1987. **41**: p. 1-38.

143. Weiss, R.H. and R. Nuccitelli, *Inhibition of tyrosine phosphorylation prevents thrombin-induced mitogenesis, but not intracellular free calcium release, in vascular smooth muscle cells.* J Biol Chem, 1992. **267**(8): p. 5608-13.

144. Samelson, L.E., et al., *Antigen activation of murine T cells induces tyrosine phosphorylation of a polypeptide associated with the T cell antigen receptor.* Cell, 1986. **46**(7): p. 1083-90.

145. Barber, E.K., et al., *The CD4 and CD8 antigens are coupled to a protein-tyrosine kinase (p56lck) that phosphorylates the CD3 complex.* Proc Natl Acad Sci U S A, 1989. **86**(9): p. 3277-81.

146. Reth, M., *Antigen receptor tail clue.* Nature, 1989. **338**(6214): p. 383-4.

147. Irving, B.A. and A. Weiss, *The cytoplasmic domain of the T cell receptor zeta chain is sufficient to couple to receptor-associated signal transduction pathways.* Cell, 1991. **64**(5): p. 891-901.

148. Chan, A.C., et al., *ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain.* Cell, 1992. **71**(4): p. 649-62.

149. Zhang, W., et al., *LAT: the ZAP-70 tyrosine kinase substrate that links T cell receptor to cellular activation.* Cell, 1998. **92**(1): p. 83-92.

150. Bubeck Wardenburg, J., et al., *Phosphorylation of SLP-76 by the ZAP-70 protein-tyrosine kinase is required for T-cell receptor function.* J Biol Chem, 1996. **271**(33): p. 19641-4.

151. Zhang, W., et al., *Essential role of LAT in T cell development.* Immunity, 1999. **10**(3): p. 323-32.

152. Liu, S.K., et al., *The adaptor protein Gads (Grb2-related adaptor downstream of Shc) is implicated in coupling hemopoietic progenitor kinase-1 to the activated TCR.* J Immunol, 2000. **165**(3): p. 1417-26.

153. Koretzky, G.A., F. Abtahian, and M.A. Silverman, *SLP76 and SLP65: complex regulation of signalling in lymphocytes and beyond.* Nat Rev Immunol, 2006. **6**(1): p. 67-78.

154. Reynolds, L.F., et al., *Vav1 transduces T cell receptor signals to the activation of the Ras/ERK pathway via LAT, Sos, and RasGRP1.* J Biol Chem, 2004. **279**(18): p. 18239-46.

155. Reynolds, L.F., et al., *Vav1 transduces T cell receptor signals to the activation of phospholipase C-gamma1 via phosphoinositide 3-kinase-dependent and -independent pathways.* J Exp Med, 2002. **195**(9): p. 1103-14.

156. Beach, D., et al., *Dual role of SLP-76 in mediating T cell receptor-induced activation of phospholipase C-gamma1.* J Biol Chem, 2007. **282**(5): p. 2937-46.

157. Berg, L.J., et al., *Tec family kinases in T lymphocyte development and function.* Annu Rev Immunol, 2005. **23**: p. 549-600.

158. Genot, E. and D.A. Cantrell, *Ras regulation and function in lymphocytes.* Curr Opin Immunol, 2000. **12**(3): p. 289-94.

159. D'Ambrosio, D., et al., *Involvement of p21ras activation in T cell CD69 expression.* Eur J Immunol, 1994. **24**(3): p. 616-20.

160. Schulze-Luehrmann, J. and S. Ghosh, *Antigen-receptor signaling to nuclear factor kappa B.* Immunity, 2006. **25**(5): p. 701-15.

161. Oh-hora, M. and A. Rao, *Calcium signaling in lymphocytes.* Curr Opin Immunol, 2008. **20**(3): p. 250-8.

162. Savignac, M., B. Mellstrom, and J.R. Naranjo, *Calcium-dependent transcription of cytokine genes in T lymphocytes.* Pflugers Arch, 2007. **454**(4): p. 523-33.

163. Burkhardt, J.K., E. Carrizosa, and M.H. Shaffer, *The actin cytoskeleton in T cell activation.* Annu Rev Immunol, 2008. **26**: p. 233-59.

164. Faure, S., et al., *ERM proteins regulate cytoskeleton relaxation promoting T cell-APC conjugation.* Nat Immunol, 2004. **5**(3): p. 272-9.

165. Monks, C.R., et al., *Three-dimensional segregation of supramolecular activation clusters in T cells.* Nature, 1998. **395**(6697): p. 82-6.

166. Menasche, G., et al., *Regulation of T-cell antigen receptor-mediated inside-out signaling by cytosolic adapter proteins and Rap1 effector molecules.* Immunol Rev, 2007. **218**: p. 82-91.

167. Pages, F., et al., *Binding of phosphatidylinositol-3-OH kinase to CD28 is required for T-cell signalling.* Nature, 1994. **369**(6478): p. 327-9.

168. Beals, C.R., et al., *Nuclear export of NF-ATc enhanced by glycogen synthase kinase-3.* Science, 1997. **275**(5308): p. 1930-4.

169. Frauwirth, K.A., et al., *The CD28 signaling pathway regulates glucose metabolism.* Immunity, 2002. **16**(6): p. 769-77.

170. Lai, A.Y. and M. Kondo, *T and B lymphocyte differentiation from hematopoietic stem cell.* Semin Immunol, 2008. **20**(4): p. 207-12.

171. Weissman, I.L., *Translating stem and progenitor cell biology to the clinic: barriers and opportunities.* Science, 2000. **287**(5457): p. 1442-6.

172. Wu, L., S. Vandenabeele, and K. Georgopoulos, *Derivation of dendritic cells from myeloid and lymphoid precursors.* Int Rev Immunol, 2001. **20**(1): p. 117-35.

173. Morrison, S.J., et al., *Identification of a lineage of multipotent hematopoietic progenitors.* Development, 1997. **124**(10): p. 1929-39.

174. Kondo, M., I.L. Weissman, and K. Akashi, *Identification of clonogenic common lymphoid progenitors in mouse bone marrow.* Cell, 1997. **91**(5): p. 661-72.

175. Akashi, K., et al., *A clonogenic common myeloid progenitor that gives rise to all myeloid lineages.* Nature, 2000. **404**(6774): p. 193-7.

176. Adolfsson, J., et al., *Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment.* Cell, 2005. **121**(2): p. 295-306.

177. Lai, A.Y., S.M. Lin, and M. Kondo, *Heterogeneity of Flt3-expressing multipotent progenitors in mouse bone marrow.* J Immunol, 2005. **175**(8): p. 5016-23.

178. Lai, A.Y. and M. Kondo, *Asymmetrical lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors.* J Exp Med, 2006. **203**(8): p. 1867-73.

179. Krueger, A., A.I. Garbe, and H. von Boehmer, *Phenotypic plasticity of T cell progenitors upon exposure to Notch ligands.* J Exp Med, 2006. **203**(8): p. 1977-84.

180. Zhang, J., et al., *Identification of the haematopoietic stem cell niche and control of the niche size.* Nature, 2003. **425**(6960): p. 836-41.

181. Takahama, Y., *Journey through the thymus: stromal guides for T-cell development and selection.* Nat Rev Immunol, 2006. **6**(2): p. 127-35.

182. Tokoyoda, K., et al., *Cellular niches controlling B lymphocyte behavior within bone marrow during development.* Immunity, 2004. **20**(6): p. 707-18.

183. Suda, T., F. Arai, and A. Hirao, *Hematopoietic stem cells and their niche.* Trends Immunol, 2005. **26**(8): p. 426-33.

184. Schwarz, B.A. and A. Bhandoola, *Trafficking from the bone marrow to the thymus: a prerequisite for thymopoiesis.* Immunol Rev, 2006. **209**: p. 47-57.

185. Wilson, A., H.R. MacDonald, and F. Radtke, *Notch 1-deficient common lymphoid precursors adopt a B cell fate in the thymus.* J Exp Med, 2001. **194**(7): p. 1003-12.

186. Pui, J.C., et al., *Notch1 expression in early lymphopoiesis influences B versus T lineage determination.* Immunity, 1999. **11**(3): p. 299-308.

187. Perry, S.S., et al., *L-selectin defines a bone marrow analog to the thymic early T-lineage progenitor.* Blood, 2004. **103**(8): p. 2990-6.

188. Schwarz, B.A., et al., *Selective thymus settling regulated by cytokine and chemokine receptors.* J Immunol, 2007. **178**(4): p. 2008-17.

189. Robertson, P. and M.C. Poznansky, *T-lymphocyte development and models of thymopoietic reconstitution.* Transpl Infect Dis, 2003. **5**(1): p. 38-42.

190. Lind, E.F., et al., *Mapping precursor movement through the postnatal thymus reveals specific microenvironments supporting defined stages of early lymphoid development.* J Exp Med, 2001. **194**(2): p. 127-34.

191. Shortman, K., et al., *The generation and fate of thymocytes.* Semin Immunol, 1990. **2**(1): p. 3-12.

192. Arstila, T.P., et al., *A direct estimate of the human alphabeta T cell receptor diversity.* Science, 1999. **286**(5441): p. 958-61.

193. Rubin, R.L. and T.M. Hermanson, *Plasticity in the positive selection of T cells: affinity of the selecting antigen and IL-7 affect T cell responsiveness.* Int Immunol, 2005. **17**(7): p. 959-71.

194. Germain, R.N., *T-cell development and the CD4-CD8 lineage decision.* Nat Rev Immunol, 2002. **2**(5): p. 309-22.

195. Klein, L., et al., *Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see).* Nature Reviews Immunology, 2014. **14**(6): p. 377-391.

196. Mathis, D. and C. Benoist, *Aire.* Annual Review of Immunology, 2009. **27**: p. 287-312.

197. Tellier, J., J.P. van Meerwijk, and P. Romagnoli, *An MHC-linked locus modulates thymic differentiation of CD4+CD25+Foxp3+ regulatory T lymphocytes.* Int Immunol, 2006. **18**(11): p. 1509-19.

198. Smith, T.R.F., et al., *Contribution of TCR Signaling Strength to CD8(+) T Cell Peripheral Tolerance Mechanisms.* Journal of Immunology, 2014. **193**(7): p. 3409-3416.

199. Sprent, J. and H. Kishimoto, *The thymus and central tolerance.* Transplantation, 2001. **72**(8 Suppl): p. S25-8.

200. Hare, K.J., et al., *Identification of a developmentally regulated phase of postselection expansion driven by thymic epithelium.* J Immunol, 1998. **160**(8): p. 3666-72.

201. Danska, J.S., et al., *The presumptive CDR3 regions of both T cell receptor alpha and beta chains determine T cell specificity for myoglobin peptides.* J Exp Med, 1990. **172**(1): p. 27-33.

202. Giudicelli, V., D. Chaume, and M.P. Lefranc, *IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.* Nucleic Acids Res, 2005. **33**(Database issue): p. D256-61.

203. von Boehmer, H. and H.J. Fehling, *Structure and function of the pre-T cell receptor.* Annu Rev Immunol, 1997. **15**: p. 433-52.

204. Bonneville, M., R.L. O'Brien, and W.K. Born, *Gammadelta T cell effector functions: a blend of innate programming and acquired plasticity.* Nat Rev Immunol, 2010. **10**(7): p. 467-78.

205. Davis, M.M. and P.J. Bjorkman, *T-cell antigen receptor genes and T-cell recognition.* Nature, 1988. **334**(6181): p. 395-402.

206. Jaeger, S., B. Fernandez, and P. Ferrier, *Epigenetic aspects of lymphocyte antigen receptor gene rearrangement or 'when stochasticity completes randomness'.* Immunology, 2013. **139**(2): p. 141-50.

207. Dudley, E.C., et al., *T cell receptor beta chain gene rearrangement and selection during thymocyte development in adult mice.* Immunity, 1994. **1**(2): p. 83-93.

208. Olaru, A., H.T. Petrie, and F. Livak, *Beyond the 12/23 rule of VDJ recombination independent of the Rag proteins.* J Immunol, 2005. **174**(10): p. 6220-6.

209. Del Blanco, B., et al., *Control of V(D)J Recombination through Transcriptional Elongation and Changes in Locus Chromatin Structure and Nuclear Organization.* Genet Res Int, 2011. **2011**: p. 970968.

210. Marche, P.N. and T.J. Kindt, *Two distinct T-cell receptor alpha-chain transcripts in a rabbit T-cell line: implications for allelic exclusion in T cells.* Proc Natl Acad Sci U S A, 1986. **83**(7): p. 2190-4.

211. Corthay, A., K.S. Nandakumar, and R. Holmdahl, *Evaluation of the percentage of peripheral T cells with two different T cell receptor alpha-chains and of their potential role in autoimmunity.* J Autoimmun, 2001. **16**(4): p. 423-9.

212. Padovan, E., et al., *Expression of two T cell receptor alpha chains: dual receptor T cells.* Science, 1993. **262**(5132): p. 422-4.

213. Radtke, F., H.R. MacDonald, and F. Tacchini-Cottier, *Regulation of innate and adaptive immunity by Notch.* Nat Rev Immunol, 2013. **13**(6): p. 427-37.

214. Lu, H., K. Schwarz, and M.R. Lieber, *Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination.* Nucleic Acids Res, 2007. **35**(20): p. 6917-23.

215. Lieber, M.R., *Mechanisms of human lymphoid chromosomal translocations.* Nat Rev Cancer, 2016. **16**(6): p. 387-98.

216. Rothenberg, E.V., J.E. Moore, and M.A. Yui, *Launching the T-cell-lineage developmental programme.* Nat Rev Immunol, 2008. **8**(1): p. 9-21.

217. Kreslavsky, T., et al., *alphabeta versus gammadelta fate choice: counting the T-cell lineages at the branch point.* Immunol Rev, 2010. **238**(1): p. 169-81.

218. Hager-Theodorides, A.L., et al., *Beta-selection: abundance of TCRbeta-/gammadelta- CD44- CD25- (DN4) cells in the foetal thymus.* Eur J Immunol, 2007. **37**(2): p. 487-500.

219. Rock, E.P., et al., *CDR3 length in antigen-specific immune receptors.* J Exp Med, 1994. **179**(1): p. 323-8.

220. Tonegawa, S., *Somatic generation of immune diversity.* Biosci Rep, 1988. **8**(1): p. 3-26.

221. Hughes, M.M., et al., *T cell receptor CDR3 loop length repertoire is determined primarily by features of the V(D)J recombination reaction.* Eur J Immunol, 2003. **33**(6): p. 1568-75.

222. Pannetier, C., et al., *The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments.* Proc Natl Acad Sci U S A, 1993. **90**(9): p. 4319-23.

223. Davies, D.R. and G.H. Cohen, *Interactions of protein antigens with antibodies.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 7-12.

224. Wilson, I.A. and K.C. Garcia, *T-cell receptor structure and TCR complexes.* Curr Opin Struct Biol, 1997. **7**(6): p. 839-48.

225. Madden, D.R., *The three-dimensional structure of peptide-MHC complexes.* Annu Rev Immunol, 1995. **13**: p. 587-622.

226. Garcia, K.C., L. Teyton, and I.A. Wilson, *Structural basis of T cell recognition.* Annu Rev Immunol, 1999. **17**: p. 369-97.

227. Garcia, K.C., *Reconciling views on T cell receptor germline bias for MHC.* Trends Immunol, 2012. **33**(9): p. 429-36.

228. Nikolich-Zugich, J., M.K. Slifka, and I. Messaoudi, *The many important facets of T-cell repertoire diversity.* Nat Rev Immunol, 2004. **4**(2): p. 123-32.

229. Miles, J.J., D.C. Douek, and D.A. Price, *Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination.* Immunol Cell Biol, 2011. **89**(3): p. 375-87.

230. Zarnitsyna, V.I., et al., *Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire.* Front Immunol, 2013. **4**: p. 485.

231. Wing, J.B. and S. Sakaguchi, *TCR diversity and Treg cells, sometimes more is more.* Eur J Immunol, 2011. **41**(11): p. 3097-100.

232. Robins, H.S., et al., *Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells.* Blood, 2009. **114**(19): p. 4099-107.

233. Burnet, F.M., *The cellular basis of immunology.* Nihon Saikingaku Zasshi, 1961. **16**: p. 620-9.

234. Doherty, P.C., J.M. Riberdy, and G.T. Belz, *Quantitative analysis of the CD8+ T-cell response to readily eliminated and persistent viruses.* Philos Trans R Soc Lond B Biol Sci, 2000. **355**(1400): p. 1093-101.

235. Even, J., et al., *T-cell repertoires in healthy and diseased human tissues analysed by T-cell receptor beta-chain CDR3 size determination: evidence for oligoclonal expansions in tumours and inflammatory diseases.* Res Immunol, 1995. **146**(2): p. 65-80.

236. Casrouge, A., et al., *Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes.* J Immunol, 2000. **164**(11): p. 5782-7.

237. Maryanski, J.L., et al., *Individuality of Ag-selected and preimmune TCR repertoires.* Immunol Res, 2001. **23**(1): p. 75-84.

238. Lin, M.Y. and R.M. Welsh, *Stability and diversity of T cell receptor repertoire usage during lymphocytic choriomeningitis virus infection of mice.* J Exp Med, 1998. **188**(11): p. 1993-2005.

239. Currier, J.R. and M.A. Robinson, *Spectratype/immunoscope analysis of the expressed TCR repertoire.* Curr Protoc Immunol, 2001. **Chapter 10**: p. Unit 10 28.

240. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

241. Giudicelli, V., et al., *IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences.* Nucleic Acids Res, 2006. **34**(Database issue): p. D781-4.

242. Biotechnologies, A. *The most robust, quantitative way to sequence millions of T- and B-cell receptors, period. immunoSEQ*. 2017 [cited 2017 February 16, 2017]; Available from: http://www.adaptivebiotech.com.

243. Fisher, R.A., A.S. Corbet, and C.B. Williams, *The relation between the number of species and the number of individuals in a random sample of an animal population.* Journal of Animal Ecology, 1943. **12**: p. 42-58.

244. Efron, B. and R. Thisted, *Estimating Number of Unseen Species - How Many Words Did Shakespeare Know.* Biometrika, 1976. **63**(3): p. 435-447.

245. Venturi, V., et al., *Method for assessing the similarity between subsets of the T cell receptor repertoire.* Journal of Immunological Methods, 2008. **329**(1-2): p. 67-80.

246. Venturi, V., et al., *Methods for comparing the diversity of samples of the T cell receptor repertoire.* Journal of Immunological Methods, 2007. **321**(1-2): p. 182-195.

247. Hurlbert, S.H., *Nonconcept of Species Diversity - Critique and Alternative Parameters.* Ecology, 1971. **52**(4): p. 577-&.

248. Preston, F.W., *The Commonness, and Rarity, of Species.* Ecology, 1948. **29**(3): p. 254-283.

249. Futschik, A. and C. Schlotterer, *The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples.* Genetics, 2010. **186**(1): p. 207-218.

250. Johnson, P.L.F. and M. Slatkin, *Inference of population genetic parameters in metagenomics: A clean look at messy data.* Genome Research, 2006. **16**(10): p. 1320-1327.

251. Lynch, M., *Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects.* Molecular Biology and Evolution, 2008. **25**(11): p. 2409-2419.

252. Lynch, M., *Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects.* Mol Biol Evol, 2008. **25**(11): p. 2409-19.

253. Bartok, I., et al., *T cell receptor CDR3 loops influence alphabeta pairing.* Mol Immunol, 2010. **47**(7-8): p. 1613-8.

254. Rani, L., et al., *Immunoglobulin heavy chain variable region gene repertoire and B-cell receptor stereotypes in Indian patients with chronic lymphocytic leukemia.* Leuk Lymphoma, 2016: p. 1-12.

255. Li, X., *Dynamic changes of driver genes' mutations across clinical stages in nine cancer types.* Cancer Med, 2016.

256. *New Driver Mutations Detected in NSCLC.* Cancer Discov, 2016.

257. Anoosha, P., et al., *Discrimination of driver and passenger mutations in epidermal growth factor receptor in cancer.* Mutat Res, 2015. **780**: p. 24-34.

258. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18545-50.

259. Busse, C.E., et al., *Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes.* Eur J Immunol, 2014. **44**(2): p. 597-603.

260. Dash, P., et al., *Paired analysis of TCRalpha and TCRbeta chains at the single-cell level in mice.* J Clin Invest, 2011. **121**(1): p. 288-95.

261. DeKosky, B.J., et al., *High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire.* Nat Biotechnol, 2013. **31**(2): p. 166-9.

262. Turchaninova, M.A., et al., *Pairing of T-cell receptor chains via emulsion PCR.* European Journal of Immunology, 2013. **43**(9): p. 2507-2515.

263. Chen, A.K., et al., *Ratiometric bimolecular beacons for the sensitive detection of RNA in single living cells.* Nucleic Acids Res, 2010. **38**(14): p. e148.

264. Tyagi, S. and F.R. Kramer, *Molecular beacons: probes that fluoresce upon hybridization.* Nat Biotechnol, 1996. **14**(3): p. 303-8.

265. Mhlanga, M.M., et al., *tRNA-linked molecular beacons for imaging mRNAs in the cytoplasm of living cells.* Nucleic Acids Res, 2005. **33**(6): p. 1902-12.

266. Howie, B., et al., *High-throughput pairing of T cell receptor alpha and beta sequences (vol 7, pg 309er7, 2015).* Science Translational Medicine, 2015. **7**(309).

267. Dittmer, W.U., A. Reuter, and F.C. Simmel, *A DNA-based machine that can cyclically bind and release thrombin.* Angew Chem Int Ed Engl, 2004. **43**(27): p. 3550-3.

268. Benenson, Y., et al., *An autonomous molecular computer for logical control of gene expression.* Nature, 2004. **429**(6990): p. 423-9.

269. Ko, S., et al., *DNA nanotubes as combinatorial vehicles for cellular delivery.* Biomacromolecules, 2008. **9**(11): p. 3039-43.

270. Chen, Y.J., et al., *DNA nanotechnology from the test tube to the cell.* Nat Nanotechnol, 2015. **10**(9): p. 748-60.

271. Zhang, X., et al., *Quantitative assessment of ratiometric bimolecular beacons as a tool for imaging single engineered RNA transcripts and measuring gene expression in living cells.* Nucleic Acids Res, 2013. **41**(15): p. e152.

272. Emerson, R., et al., *Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data.* J Immunol Methods, 2013. **391**(1-2): p. 14-21.

273. Robins, H., et al., *Ultra-sensitive detection of rare T cell clones.* J Immunol Methods, 2012. **375**(1-2): p. 14-9.

274. Quigley, M.F., et al., *Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire.* Proc Natl Acad Sci U S A, 2010. **107**(45): p. 19414-9.

275. Price, D.A., et al., *Public clonotype usage identifies protective Gag-specific CD8+ T cell responses in SIV infection.* J Exp Med, 2009. **206**(4): p. 923-36.

276. Li, H., et al., *Determinants of public T cell responses.* Cell Res, 2012. **22**(1): p. 33-42.

277. Argaet, V.P., et al., *Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus.* J Exp Med, 1994. **180**(6): p. 2335-40.

278. Khan, N., et al., *Comparative analysis of CD8+ T cell responses against human cytomegalovirus proteins pp65 and immediate early 1 shows similarities in precursor frequency, oligoclonality, and phenotype.* J Infect Dis, 2002. **185**(8): p. 1025-34.

279. Wynn, K.K., et al., *Impact of clonal competition for peptide-MHC complexes on the CD8+ T-cell repertoire selection in a persistent viral infection.* Blood, 2008. **111**(8): p. 4283-92.

280. Kasprowicz, V., et al., *A highly restricted T-cell receptor dominates the CD8+ T-cell response to parvovirus B19 infection in HLA-A*2402-positive individuals.* J Virol, 2006. **80**(13): p. 6697-701.

281. Godthelp, B.C., et al., *Longitudinal analysis of T cells responding to tetanus toxoid in healthy subjects as well as in pediatric patients after bone marrow transplantation: the identification of identical TCR-CDR3 regions in time suggests long-term stability of at least part of the antigen-specific TCR repertoire.* Int Immunol, 2001. **13**(4): p. 507-18.

282. Dong, L.C., et al., *Public TCR Use by Herpes Simplex Virus-2-Specific Human CD8 CTLs.* Journal of Immunology, 2010. **184**(6): p. 3063-3071.

283. Gillespie, G.M., et al., *Strong TCR conservation and altered T cell cross-reactivity characterize a B*57-restricted immune response in HIV-1 infection.* J Immunol, 2006. **177**(6): p. 3893-902.

284. Cole, D.K., et al., *Modification of MHC Anchor Residues Generates Heteroclitic Peptides That Alter TCR Binding and T Cell Recognition.* Journal of Immunology, 2010. **185**(4): p. 2600-2610.

285. Le Gal, F.A., et al., *Distinct structural TCR repertoires in naturally occurring versus vaccine-induced CD8(+) T-Cell responses to the tumor-specific antigen NY-ESO-1.* Journal of Immunotherapy, 2005. **28**(3): p. 252-257.

286. Hong, J., et al., *A common TCR V-D-J sequence in V beta 13.1 T cells recognizing an immunodominant peptide of myelin basic protein in multiple sclerosis.* Journal of Immunology, 1999. **163**(6): p. 3530-3538.

287. May, E., et al., *Conserved TCR beta chain usage in reactive arthritis; evidence for selection by a putative HLA-B27-associated autoantigen.* Tissue Antigens, 2002. **60**(4): p. 299-308.

288. Zeng, W.H., et al., *Limited heterogeneity of T cell receptor BV usage in aplastic anemia.* Journal of Clinical Investigation, 2001. **108**(5): p. 765-773.

289. Prinz, J.C., et al., *Selection of conserved TCR VDJ rearrangements in chronic psoriatic plaques indicates a common antigen in psoriasis vulgaris.* European Journal of Immunology, 1999. **29**(10): p. 3360-3368.

290. Kuwana, M., T.A. Medsger, and T.M. Wright, *Highly restricted TCR-alpha beta usage by autoreactive human T cell clones specific for DNA topoisomerase I - Recognition of an immunodominant epitope.* Journal of Immunology, 1997. **158**(1): p. 485-491.

291. Grunewald, J., et al., *Restricted Usage of T-Cell Receptor V-Alpha/J-Alpha Gene Segments with Different Nucleotide but Identical Amino-Acid-Sequences in Hla-Dr3(+) Sarcoidosis Patients.* Molecular Medicine, 1995. **1**(3): p. 287-296.

292. Sun, W., et al., *Skewed T-cell receptor BV14 and BV16 expression, and shared complementarity-determining region 3 sequence and common sequence motifs in synovial T cells of rheumatoid arthritis.* Arthritis Research & Therapy, 2003. **5**: p. S39-S39.

293. Price, D.A., et al., *Public clonotype usage identifies protective Gag-specific CD8(+) T cell responses in SIV infection.* Journal of Experimental Medicine, 2009. **206**(4): p. 923-936.

294. Yu, X.G., et al., *Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes.* Journal of Virology, 2007. **81**(4): p. 1619-1631.

295. Dong, T., et al., *HIV-specific cytotoxic T cells from long-term survivors select a unique T cell receptor.* Journal of Experimental Medicine, 2004. **200**(12): p. 1547-1557.

296. van Bockel, D.J., et al., *Persistent Survival of Prevalent Clonotypes within an Immunodominant HIV Gag-Specific CD8(+) T Cell Response.* Journal of Immunology, 2011. **186**(1): p. 359-371.

297. Price, D.A., et al., *T cell receptor recognition motifs govern immune escape patterns in acute SIV infection.* Immunity, 2004. **21**(6): p. 793-803.

298. Bousso, P., et al., *Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires.* Immunity, 1998. **9**(2): p. 169-78.

299. Gras, S., et al., *T-cell receptor bias and immunity.* Curr Opin Immunol, 2008. **20**(1): p. 119-25.

300. Turner, S.J., et al., *Structural determinants of T-cell receptor bias in immunity.* Nat Rev Immunol, 2006. **6**(12): p. 883-94.

301. Venturi, V., et al., *The molecular basis for public T-cell responses?* Nat Rev Immunol, 2008. **8**(3): p. 231-8.

302. Madi, A., et al., *T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity.* Genome Res, 2014. **24**(10): p. 1603-12.

303. Venturi, V., et al., *Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination.* Proc Natl Acad Sci U S A, 2006. **103**(49): p. 18691-6.

304. Venturi, V., et al., *The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8+ TCRs between macaques.* J Immunol, 2008. **181**(4): p. 2597-609.

305. Venturi, V., et al., *TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV.* J Immunol, 2008. **181**(11): p. 7853-62.

306. Robins, H.S., et al., *Overlap and effective size of the human CD8+ T cell receptor repertoire.* Sci Transl Med, 2010. **2**(47): p. 47ra64.

307. Weinstein, J.A., et al., *High-throughput sequencing of the zebrafish antibody repertoire.* Science, 2009. **324**(5928): p. 807-10.

308. Gauss, G.H. and M.R. Lieber, *Mechanistic constraints on diversity in human V(D)J recombination.* Mol Cell Biol, 1996. **16**(1): p. 258-69.

309. Candeias, S., et al., *The V beta 17+ T cell repertoire: skewed J beta usage after thymic selection; dissimilar CDR3s in CD4+ versus CD8+ cells.* J Exp Med, 1991. **174**(5): p. 989-1000.

310. Venturi, V., et al., *A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing.* J Immunol, 2011. **186**(7): p. 4285-94.

311. Quiros Roldan, E., et al., *Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells.* Immunogenetics, 1995. **41**(2-3): p. 91-100.

312. Wallace, M.E., et al., *Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1.* Immunity, 2000. **12**(5): p. 547-56.

313. Livak, F., et al., *Genetic modulation of T cell receptor gene segment usage during somatic recombination.* J Exp Med, 2000. **192**(8): p. 1191-6.

314. Melenhorst, J.J., et al., *Contribution of TCR-beta locus and HLA to the shape of the mature human Vbeta repertoire.* J Immunol, 2008. **180**(10): p. 6484-9.

315. Fuschiotti, P., et al., *Analysis of the TCR alpha-chain rearrangement profile in human T lymphocytes.* Mol Immunol, 2007. **44**(13): p. 3380-8.

316. Mora, J.R., et al., *Selective imprinting of gut-homing T cells by Peyer's patch dendritic cells.* Nature, 2003. **424**(6944): p. 88-93.

317. Monod, M.Y., et al., *IMGT/Junction Analysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs.* Bioinformatics, 2004. **20**: p. 379-385.

318. Venturi, V., et al., *Methods for comparing the diversity of samples of the T cell receptor repertoire.* J Immunol Methods, 2007. **321**(1-2): p. 182-95.

319. Fazilleau, N., et al., *Valpha and Vbeta public repertoires are highly conserved in terminal deoxynucleotidyl transferase-deficient mice.* J Immunol, 2005. **174**(1): p. 345-55.

320. Rangarajan, S. and R.A. Mariuzza, *T cell receptor bias for MHC: co-evolution or co-receptors?* Cell Mol Life Sci, 2014. **71**(16): p. 3059-68.

321. Chen, F., et al., *Differential transcriptional regulation of individual TCR V beta segments before gene rearrangement.* J Immunol, 2001. **166**(3): p. 1771-80.

322. Xu, C.R., et al., *Reciprocal patterns of methylation of H3K36 and H3K27 on proximal vs. distal IgVH genes are modulated by IL-7 and Pax5.* Proc Natl Acad Sci U S A, 2008. **105**(25): p. 8685-90.

323. Zhang, Z., et al., *Transcription factor Pax5 (BSAP) transactivates the RAG-mediated V(H)-to-DJ(H) rearrangement of immunoglobulin genes.* Nat Immunol, 2006. **7**(6): p. 616-24.

324. Espinoza, C.R. and A.J. Feeney, *The extent of histone acetylation correlates with the differential rearrangement frequency of individual VH genes in pro-B cells.* J Immunol, 2005. **175**(10): p. 6668-75.

325. Shimazaki, N., A.G. Tsai, and M.R. Lieber, *H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations.* Mol Cell, 2009. **34**(5): p. 535-44.

326. Osipovich, O., et al., *Targeted inhibition of V(D)J recombination by a histone methyltransferase.* Nat Immunol, 2004. **5**(3): p. 309-16.

327. McClintock, B., *The significance of responses of the genome to challenge.* Science, 1984. **226**(4676): p. 792-801.

328. Lucht, J.M., et al., *Pathogen stress increases somatic recombination frequency in Arabidopsis.* Nat Genet, 2002. **30**(3): p. 311-4.

329. Roth, J.R., et al., *Origin of mutations under selection: the adaptive mutation controversy.* Annu Rev Microbiol, 2006. **60**: p. 477-501.

330. Kovalchuk, I., et al., *Pathogen-induced systemic plant signal triggers DNA rearrangements.* Nature, 2003. **423**(6941): p. 760-2.

331. Rosenberg, S.M., *Evolving responsively: adaptive mutation.* Nat Rev Genet, 2001. **2**(7): p. 504-15.

332. De Boer, R.J. and A.S. Perelson, *How diverse should the immune system be?* Proc Biol Sci, 1993. **252**(1335): p. 171-5.

333. Wiegel, F.W. and A.S. Perelson, *Some scaling principles for the immune system.* Immunol Cell Biol, 2004. **82**(2): p. 127-31.

334. Pei, H., et al., *Reconfigurable three-dimensional DNA nanostructures for the construction of intracellular logic sensors.* Angew Chem Int Ed Engl, 2012. **51**(36): p. 9020-4.

335. Modi, S., et al., *A DNA nanomachine that maps spatial and temporal pH changes inside living cells.* Nat Nanotechnol, 2009. **4**(5): p. 325-30.

336. Modi, S., et al., *Two DNA nanomachines map pH changes along intersecting endocytic pathways inside the same cell.* Nat Nanotechnol, 2013. **8**(6): p. 459-67.

337. Mei, Q., et al., *Stability of DNA origami nanoarrays in cell lysate.* Nano Lett, 2011. **11**(4): p. 1477-82.

338. Conway, J.W., et al., *DNA nanostructure serum stability: greater than the sum of its parts.* Chem Commun (Camb), 2013. **49**(12): p. 1172-4.

339. Barry, M.E., et al., *Role of endogenous endonucleases and tissue site in transfection and CpG-mediated immune activation after naked DNA injection.* Hum Gene Ther, 1999. **10**(15): p. 2461-80.

340. Howie, B., et al., *High-throughput pairing of T cell receptor alpha and beta sequences.* Sci Transl Med, 2015. **7**(301): p. 301ra131.

341. Rothemund, P.W., *Folding DNA to create nanoscale shapes and patterns.* Nature, 2006. **440**(7082): p. 297-302.

342. Pircher, H., et al., *Tolerance induction in double specific T-cell receptor transgenic mice varies with antigen.* Nature, 1989. **342**(6249): p. 559-61.

343. Arstila, T.P., et al., *Diversity of human alpha beta T cell receptors.* Science, 2000. **288**(5469): p. 1135.

344. Asirvatham, A.J., et al., *MicroRNA targets in immune genes and the Dicer/Argonaute and ARE machinery components.* Mol Immunol, 2008. **45**(7): p. 1995-2006.

345. Zhang, N. and M.J. Bevan, *Dicer controls CD8+ T-cell activation, migration, and survival.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21629-34.

346. Cabaniols, J.P., et al., *Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase.* J Exp Med, 2001. **194**(9): p. 1385-90.

347. Hou, X.L., et al., *Current status and recent advances of next generation sequencing techniques in immunological repertoire.* Genes Immun, 2016. **17**(3): p. 153-64.

348. Harty, J.T. and V.P. Badovinac, *Shaping and reshaping CD8+ T-cell memory.* Nat Rev Immunol, 2008. **8**(2): p. 107-19.

349. Batrak, V., A. Blagodatski, and J.M. Buerstedde, *Understanding the immunoglobulin locus specificity of hypermutation.* Methods Mol Biol, 2011. **745**: p. 311-26.

350. Woodsworth, D.J., M. Castellarin, and R.A. Holt, *Sequence analysis of T-cell repertoires in health and disease.* Genome Med, 2013. **5**(10): p. 98.

351. Maciejewski, J.P., et al., *Immune-mediated bone marrow failure syndromes of progenitor and stem cells: molecular analysis of cytotoxic T cell clones.* Folia Histochem Cytobiol, 2007. **45**(1): p. 5-14.

352. Diu, A., et al., *Fine specificity of monoclonal antibodies directed at human T cell receptor variable regions: comparison with oligonucleotide-driven amplification for evaluation of V beta expression.* Eur J Immunol, 1993. **23**(7): p. 1422-9.

353. Klarenbeek, P.L., et al., *Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease.* Ann Rheum Dis, 2012. **71**(6): p. 1088-93.

354. Li, S., et al., *IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling.* Nat Commun, 2013. **4**: p. 2333.

355. Bolotin, D.A., et al., *Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms.* Eur J Immunol, 2012. **42**(11): p. 3073-83.

356. Wilson, P.C. and S.F. Andrews, *Tools to therapeutically harness the human antibody response.* Nat Rev Immunol, 2012. **12**(10): p. 709-19.

357. Furutani, S., et al., *Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method.* Analyst, 2012. **137**(13): p. 2951-7.

358. Jacobsen, H., H. Klenow, and K. Overgaard-Hansen, *The N-terminal amino-acid sequences of DNA polymerase I from Escherichia coli and of the large and the small fragments obtained by a limited proteolysis.* Eur J Biochem, 1974. **45**(2): p. 623-7.

359. Williams, R., et al., *Amplification of complex gene libraries by emulsion PCR.* Nat Methods, 2006. **3**(7): p. 545-50.

360. Chaiet, L. and F.J. Wolf, *The Properties of Streptavidin, a Biotin-Binding Protein Produced by Streptomycetes.* Arch Biochem Biophys, 1964. **106**: p. 1-5.

361. Bashford-Rogers, R.J., et al., *Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations.* Genome Res, 2013. **23**(11): p. 1874-84.