The Dissertation Committee for Zachary Paul Frye Certifies that this is the approved version of the following dissertation:

Linking TCR sequence, structure, and function in the DO11.10 mouse model

Committee:

Jennifer Maynard, Supervisor

George Georgiou

Hal Alper

Lauren Ehrlich

Ning Jiang

Linking TCR sequence, structure, and function in the DO11.10 mouse model

by

Zachary Paul Frye, B.S.C.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2015

Dedication

I dedicate this work to my parents for their love and support

and to my friends near and far.

Acknowledgements

First, I would like to thank my adviser Dr. Jennifer Maynard for her optimism, patience, and support with my scientific endeavors. Second, I would also like to thank my committee members for their time, advice, and efforts.

I would also like to thank my current and former colleagues in the Maynard lab who I had the pleasure to work with on my many projects. A special thank you to Dr. Benjamin Roy, Dr. Kevin Entzminger, Dr. Annalee Nguyen, Dr. Jordan Lewandowki, Dr. Tarik Kahn, Ellen Wagner, and Chris Stevens who helped me with various aspects of the work described here and assisted me with this dissertation.

My gratitude and appreciation also goes out to the many colleagues, friends, and staff of the chemical engineering department who have all helped me along the way.

Linking TCR sequence, structure, and function in the DO11.10 mouse model

Zachary Paul Frye, Ph.D.

The University of Texas at Austin, 2015

Supervisor: Jennifer Maynard

Thymus-maturated lymphocytes (T-cells) provide the immune system with the ability to appropriately detect and respond to diseases through a unique receptor. These T-cell receptors (TCRs) differentiate between self and foreign peptide antigens and form by hypervariable gene rearrangement and imprecise recombination, both of which can potentially generate upwards of 10^{18} unique TCRs in humans. Our work focused on characterizing the chicken ovalbumin (OVA) antigen specific parent hybridoma DO11.10 T-cell with (1) high-throughput sequencing, (2) structural analyses, and (3) TCR reconstitution into T-cells to determine the important factors linking TCR sequence, structure, and function. The repertoire of TCR sequences generated in mice immunized with OVA antigen were sequenced using a novel mouse TCR gene primer set. Using high-throughput sequencing, we determined the relative gene and hypervariable sequence frequencies in the repertoire. Based on our results in these experiments, no significant repertoire differences were found between mouse treatment groups. Additionally no sequences with similarity to previously studied OVA-specific T-cell clones were recovered. The antigen binding loops of TCRs have a high degree of structural similarity across diverse receptor sequences. Key residues in the complementarity determining,

antigen-binding regions of TCRs form conserved canonical loop structures. Here we evaluated 249 TCRs and TCR-like antibodies to determine the prevalence of these key residues. Expanding on previous work, we found four new canonical class conformations and validated the key residues for these classes. Finally, we report our work to reconstitute engineered TCRs onto the TCR-deficient mouse $58^{-/-}$ hybridoma line. We modified an existing antibody expression system to display the engineered TCRs in the hybridomas for activation assays. Expression of the TCR was found to be highly sensitive to the N-terminal signal sequence on each receptor chain. Additional work began to examine how these signal sequences alter stable expression and surface display of these TCRs. This work is important for the discovery of new methods and biological agents to target and respond to designed antigens, especially in the context of altering and engineering TCR specificity for therapeutic purposes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Background

1.1: Overview of the Adaptive Immune System

The innate and adaptive immune systems consist of a complex cellular network to identify, signal, and neutralize antigens through a generic primary response and a more specific secondary response. Coordination of these highly evolved systems can provide protection against common pathogens, generating unique cellular and molecular responses to protect against the myriad of encountered antigens from the environment. The innate immune system provides an initial non-specific defense against infection by identifying foreign material, initiating the complement cascade, and activating the adaptive immune system. While this first response can protect against a variety of common pathogens it fails to keep pace with and incredibly diverse and often mutating array of antigens and pathogens present in the environment. Approximately four days after the innate response begins, the recruitment of immune cells and activation of the adaptive immune system initializes the generation of a more specific secondary response to the infection.

The activated adaptive immune system provides a pathogen-specific, long-lasting response generated from naïve and effector lymphocytes (white blood cells). The two main cell types within the specific lymphocyte response involved in adaptive protection are B-cells and T-cells. These two major cell types fill distinct and important immunological roles by using both antibody and cell-mediated responses to neutralize pathogens and target infected cells for removal. Within these lymphocyte cell types are additional cell subsets, which function to monitor, regulate, and remove cells as well as provide long-lasting immunological memory.

After maturation in the bone marrow and thymus respectively, naïve B-cells and T-cells produced during the initial adaptive immune response are clonotypic effector cells, meaning each cell type produces a unique protein construct in the form of an antibody or

membrane-bound T-cell receptor (TCR) to target the pathogen. This complex maturation process generates receptor diversity of approximately $10^{13} – 10^{18}$ for antibodies and TCRs respectively in humans, which is explained more in **Chapter 1.3**. The cells successfully targeting the infectious antigens, and/or pathogens are activated and stimulated further to help clear the infection from the host.

Once the adaptive immune system resolves the infection, the majority of the activated effector cells are cleared by phagocytosis after death by neglect. A small fraction of the effector cells are reserved and recruited as memory cells to guard against future infections from the same pathogenic material. The maintenance of these memory cells allows individuals to build a cellular repertoire which can more rapidly respond and proliferate in the case of repeat exposure. The failure to convert effector cells to memory cells or the loss of memory cells in general can lead to immune disorders.

The critical job of the adaptive immune system is the generation of an enormous diversity of antibodies and T-cell receptors which can theoretically respond against any antigen. Proper immune function requires the coordination of a complex network of genes, chemicals, proteins, receptors, and cells. Deficiencies in the adaptive immune system are characteristic of a variety of serious conditions including autoimmune disease, persistent infections, and severe combined immunodeficiency.

1.2: Lymphocyte Maturation and Development

1.2.1: B-cell Maturation

Progenitor lymphoid cells in the bone marrow committed to becoming B-cells undergo somatic recombination and gene rearrangement to form an antigen binding clonotypic B-cell receptor (BCR). Negative selection deletes any immature B-cells with self-antigen reactivity the population. The remaining B-cells emerge from the bone marrow

and migrate to peripheral lymphoid organs where they wait for an activating signal. Thymus-dependent B-cell maturation requires two signals: first, binding of the BCR by a specific antigen, and second, interaction with an activated helper T-cell which is displaying a peptide from the same antigen. Alternatively, B-cell activation can occur in a T-cell independent manner for common antigens such as bacterial lipopolysaccharides (LPS).

After receiving the activation signal, the mature B-cell undergoes somatic hypermutation and class switching to differentiate the cell further into either plasma or memory B-cells. Somatic hypermutation mutates antibody genes to increase the antigen affinity and specificity. Activated plasma B-cells proliferate and produce large quantities of secreted antibodies containing the antigen-binding domain of the BCR with a different constant domain for altered functionality and processing by effector molecules and cells. Memory B-cells have a prolonged life for future encounters with its cognate antigen. Succeeding antigen interactions reactivate the memory cells causing proliferation and further affinity maturation through somatic hypermutation in germinal centers. The maturation process rapidly responds to clear infections and provide a means to produce more effective immune responses over time with additional interactions with cognate antigens or pathogens.

1.2.2: T-cell Maturation

Similar to B-cells, T-cells begin as progenitor lymphoid cells in the bone marrow which migrate to the thymus through the blood for thymic-specific maturation. This thymic-dependent maturation is how T-cells get their designation as thymic lymphocytes. T-cell maturation involves a series of steps within the thymus (**Figure 1.1**) as a progressive migration within sections of the thymus before moving to peripheral tissues and sites of infection. These maturation steps control the affinity and activity of the clonotypic T-cell

receptor (TCR) responsible for recognizing foreign antigens and delivering activation or kill signals to other cells.

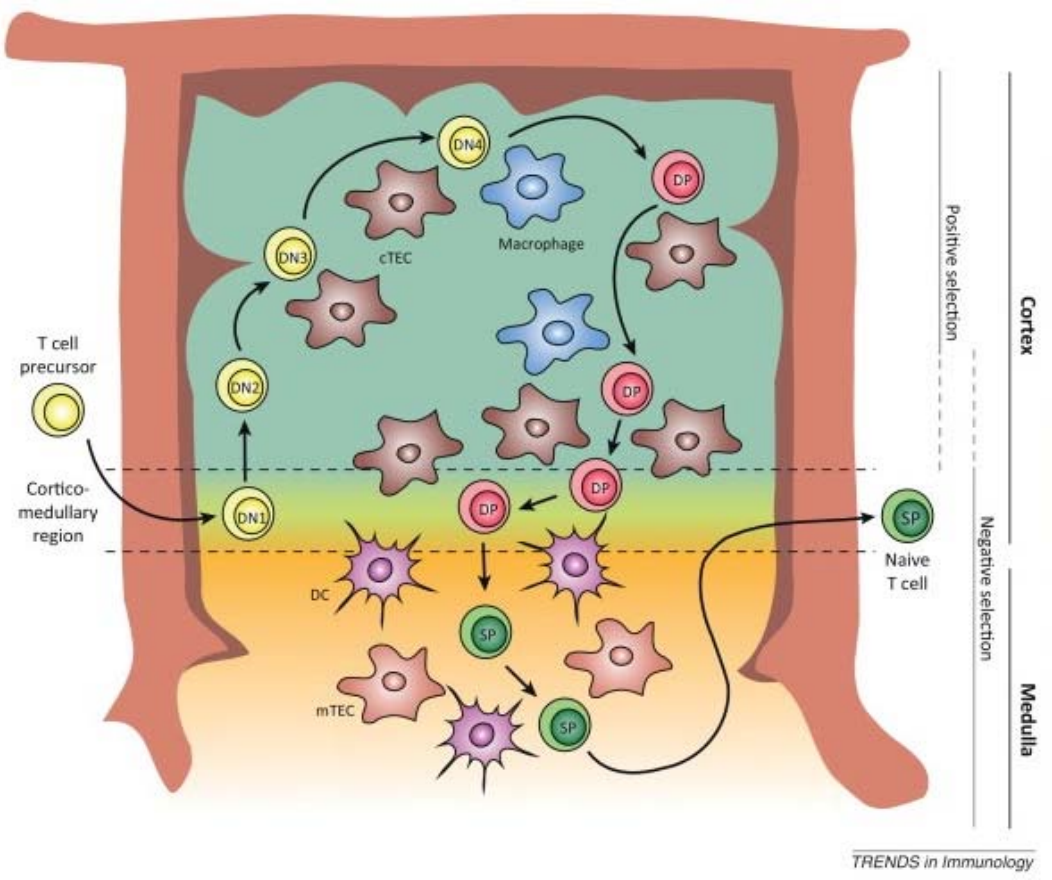


Figure 1.1: Diagram of T-cell development in a thymus capsule.

T-cell precursors enter the thymus at the cortico-medullary region and migrate through the cortex then medulla for positive and negative selection respectively before leaving as naïve T-cells ready for antigen activation in the peripheral lymphoid tissue[1]**.**

Upon arriving to the thymus, T-cell precursors enter a capsule within the thymic cortex to undergo somatic rearrangement to assemble the TCR genes. The imprecise recombinations generated during somatic rearrangement also cause the majority of gene products to fail from out-of-frame recombinations or undesirable stop codons. Cells with productive, in-frame TCRs then move through the cortex space for positive selection.

Positive selection sets the lower threshold affinity so the TCR can bind the MHC strongly enough to interrogate bound peptide. Cells failing to bind cortical epithelial cells during positive selection do not receive survival signals and die by neglect. Positively selected cells then migrate through the cortico-medullary junction into the medulla for negative selection where they interact with dendritic cells (DC) and medullary thymic epithelial cells (mTECs). Negative selection removes strongly binding TCRs providing central tolerance of self-peptides to prevent autoimmune disorders from self-reactive T-cells.

After negative selection, the non-self-reactive T-cells become naïve T-cells and leave the thymus for peripheral lymphoid organs in search of cognate antigen. The stringent affinity maturation process allows T-cells to screen cell surfaces rapidly to increase the probability of engaging cognate antigen. When a T-cell binds and activates from interactions with an antigen presenting cell (APC), the T-cell ceases lymphoid tissue migration, proliferates by clonal expansion, and differentiates. Activated T-cells differentiate into effector and memory cells capable of regulating other T-cell responses and have a prolonged life to ward off future infections. The scale of the cellular surfaces to monitor through one-to-one interaction with a recognized antigen requires the immune system to generate a diverse repertoire of TCRs and save the effective antigen engaging cells for immune memory.

1.2.3: Antigen presentation

T-cell mediated activation requires proper antigen presentation on the APC within the MHC binding groove. As the antigen loads into MHC, peptide residues interact with hydrophobic pockets in the binding groove (**Figure 1.2**) anchoring the peptide in place and exposing alternate residues for TCR interaction[2]. Modifying the anchoring and exposed residues in the peptide antigen has been shown to change the TCR-pMHC interaction and subsequent T-cell activation[3]. Antigen presentation allows T-cells a method to screen the

contents of cells or extracellular proteins collected from the environment for pathogenic material. The different processes for collecting endogenous and exogenous material also determine the MHC class molecule presenting the peptide and the types of responding T-cell types. Endogenous peptides presented in MHC-I are interrogated by $CD8^+$ T-cells and exogenous peptides in MHC-II by $CD4^+$ T-cells.

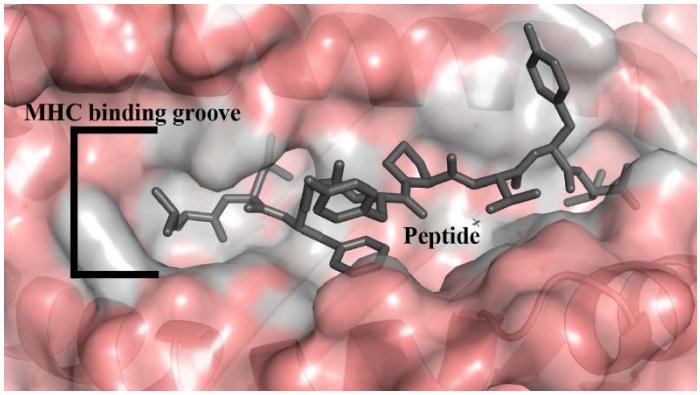


Figure 1.2: MHC binding groove with peptide anchoring pockets.

This image shows the peptide (gray) loaded in the MHC binding groove flanked on either side by the α-helices with anchoring residues protruding into downward into the hydrophobic pockets of the MHC molecule.

Nearly all cells in the body regularly present endogenous peptides regardless of infection status. Intracellular proteins are digested and transported to the endoplasmic reticulum (ER) where the peptide fragments load into MHC-I molecules. The peptide-MHC then moves to the cell surface for recognition by cytotoxic $CD8^+$ lymphocytes (CTLs). If an activated CTL recognizes the presented endogenous antigen it releases a focused blast of cytotoxins to kill only the neighboring infected cell. This controlled method to kill cells stems the spread of infection while preserving the health of surrounding cells and tissues. The cytotoxicity of CTLs is so potent and destructive, these cells require additional activation signals from either helper ($CD4^+$) T-cells or mature dendritic cells.

In order to monitor extracellular material throughout the body, exogenous pathogens and materials are internalized through endocytosis. As the endocytic vesicle travels deeper into the cell it becomes more acidic and the internalized materials become digested by proteolytic activity. The antigen-laden endosomes fuses with other vesicles containing MHC-II molecules. Then with the aid of HLA-DM, the antigen-peptide loads into the MHC-II binding groove. The MHC-II binding groove has open ends, which allows presentation of longer peptide chains. This places fewer restrictions on the alignment or register of peptide within the binding groove. The different MHC registers also change the exposed peptide residues and subsequent T-cell activity[4]. Once the peptide loads into the MHC-II, the complex moves to the cell surface for presentation and interrogation by $CD4^+$ T-cells.

Unlike MHC-I, only a subset of immune cells are equipped with MHC-II molecules for regular antigen presentation. T-cells engaging pMHC-II on APCs are effector $CD4^+$ T-cells because they mediate the activation of both CTLs to kill infected cells and stimulate plasma B-cells to produce antibodies. The cellular coordination of antigen recognition and neutralization between B-cells and T-cells allows the adaptive immune system to respond appropriately to the many disorders hosts face from viruses, bacteria, and other immune diseases.

1.2.4: The immune complex and T-cell activation

T-cells provide the body with constant monitoring of self and foreign peptide antigens encountered on the surface of APC by a clonotypic αβ TCR heterodimer. The central α- and β-chains of the TCR are part of a larger immune complex including three dimers of membrane proteins forming a tetramer of dimers around the αβ-chain core (**Figure 1.3**). Two hetero-dimers comprised as CD3δ-CD3ε and CD3γ-CD3ε associate with the constant and transmembrane (TM) domains of the α- and β-chains respectively.

The third is a CD3ζζ homodimer, which associates with the positively charged arginine on the TM domain of the α-chain.

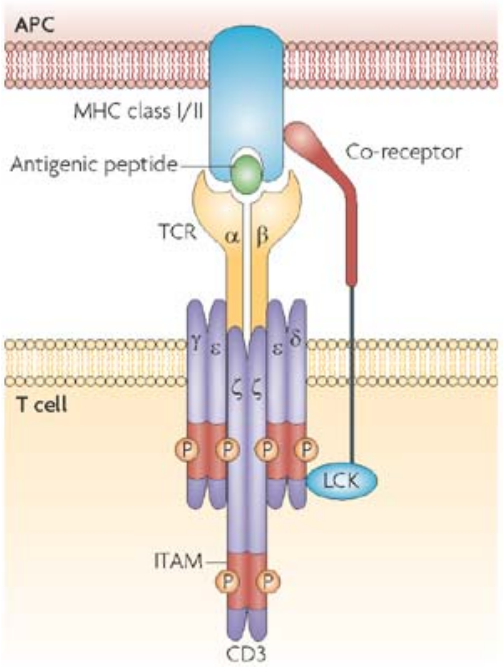


Figure 1.3: Diagram of the T-cell immune complex.

The immune complex on the surface of T-cells includes the αβ chains, CD3γδεζ –chains with ITAMs labeled in red, and a CD4 or CD8 co-receptor with Lck as they associate on the membrane and engage either MHC-I/MHC-II[5].

These six associated proteins translate the TCR-pMHC interactions into chemical signals through activation motifs. Intracellular immunereceptor tyrosine-based activation motifs (ITAMs) on these proteins initiate activation signaling though the biophysical activity of the αβ TCR chains[6]. Proteins CD3γ, -δ, and -ε each have one ITAM and CD3ζ has three for ten total ITAMs in the immune complex. For stable surface expression of the immune complex most of the CD3 proteins must be present in the ER and able to associate with the TCR αβ chains through coordinated residue charges in TM regions[7]. Failure to properly express or associate all the CD3 chains can critically diminish TCR surface

expression and cripple T-cell activation due to immune complex instability and decreased ITAMs[8].

The immune complex has two functionally distinct roles split between the extracellular and intracellular domains. The extracellular peptide-antigen recognition translates into intracellular cell activation signals through the biomechanical interactions between the proteins in the immune complex. Until TCR complexation occurs, the sensitive intracellular activation motifs remain occluded from the kinase activity of activation enzymes[6,9]. A current theory on this signal translation involves a two-step, serial triggering of TCR-pMHC engagement which coordinates the spatial orientation and structural deformations occurring during interaction between these surface molecules[10].

The first step involves the alignment of the TCR over the MHC α-helices in a diagonal orientation using MHC residues (α-65, α-69, α-155 for MHC-I and α57, α-61, and β-50 for MHC-II) (**Figure 1.4**) as guides[11]. The MHC interaction with CD4 or CD8 also restricts TCR alignment. The second step is the CDR3 peptide interrogation which either pushes or pulls the central strands of the TCR dimer tugging on the F-G loop in the β-chain constant domain. The deformation from the binding region through to the constant and TM domains reorients the inner-membrane domains of the surrounding accessory proteins in the immune complex exposing the CD3 ITAMs[12]. The piston-like TM translational movement of the TCR αβ chains also exposes the CD3ζ ITAMs from their association with the inner membrane. The lymphocyte –specific protein tyrosine kinase (Lck) co-receptor phosphorylates exposed tyrosine residues in the ITAMs.

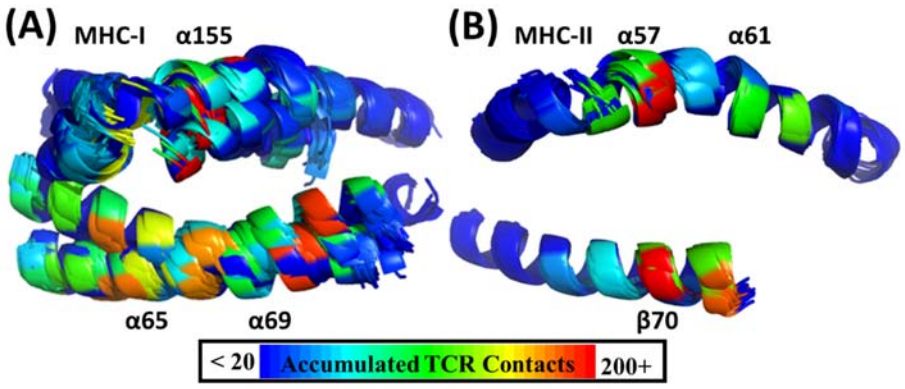


Figure 1.4: TCR-PMHC contact analysis.

Alignments of MHC molecules colored based on the number of TCR contacts made using rainbow spectrum where blue is minimal contacts and red is the maximum number of contacts made for (**A**) MHC-I and (**B**) MHC-II molecules.

The low probability of any given T-cell finding its cognate pMHC within the body requires TCRs to scan APC surfaces rapidly. The binding rate kinetics of TCR-pMHC interactions allow for rapid pMHC scanning due to a slow k_{on} and rapid k_{off}[13]. These binding kinetics give TCRs weaker but more specific interactions with the pMHC when chaperoned by co-accessory proteins in the immune complex. This also allows TCRs to quickly recognize antigen, activate, and move aside so another TCR on the same T-cell can interrogate the pMHC[10]. In this way a single pMHC can serially trigger a large number of immune complexes on the T-cell, amplifying the overall signal activation.

During pMHC engagement and immune synapse formation, CD4 or CD8 co-receptor is recruited futher activating the immune complex. This co-receptor serves to both stabilize and orient TCR binding by interacting with the invariant portion of the MHC and recruiting Lck to the synapse[11]. Lck associates with the endodomains of the CD4/CD8 receptor and phosphorylates exposed CD3 ITAMS. Once ITAM phosphorylation reaches high enough levels, another protein molecule, ZAP-70 (CD3ζ-chain-associated protein) associates with phosphorylated CD3ζ, which can then act as a kinase to trigger a linker of

activated T-cells (LAT) scaffold protein causing further cascades of activation signals. Throughout this process, co-stimulatory or inhibitory signals from molecules and receptors regulate these T-cell activation molecules. Current technologies are investigating ways to utilize these regulatory signals for therapeutic purposes[14].

The functions of activated T-cells depend on their lineages as $CD4^+$ or $CD8^+$ cells. Each type is responsible for controlling different aspects of the immune system. Activated $CD4^+$ cells release small soluble proteins called cytokines, such as interleukin-2 (IL-2) or interferon-γ (INF-γ), which may promote effector cell activation, induce MHC expression, or inhibit cell growth. Activated $CD8^+$ cells or CTLs release cytotoxic proteins including perforin or granzyme which induce apoptosis in target cells. T-cells remain activated until the loss of activating signals through effector molecules or antigen engagement. Additionally without memory cell commitment these activated cells survive for only a few days which is long enough to resolve most infections[15].

1.3: LYMPHOCYTE REPERTOIRE GENERATION

The adaptive immune system utilizes a diverse antigen binding repertoire with rigorously selected binding affinities for robust and specific antibody targeting and T-cell activation. This diversity in antigen binders is imperative for protecting against the nearly limitless variety of pathogens in the environment. Production of this diversity in B-cells and T-cells employs several mechanisms including combinatorial gene rearrangement, imprecise junctions between those recombined genes, and somatic hypermutation for antibody affinity maturation in B-cells. Combined these processes can potentially generate over 10^{13} and 10^{18} different antibody and TCR sequences respectively in humans[16].

The first level of genetic diversity occurs from gene rearrangements to combine different gene regions which generates millions of different gene combinations[17]. Antibodies and TCRs use three segments of genes to produce functional proteins consisting

of Variable-, Diversity-, and Joining-region (VDJ) genes. These gene segments recombine in two configurations as VJ in light or α-chains and VDJ in heavy- or β-chains in antibodies and TCRs respectively. Clusters of these genes are arrayed on the chromosome locus as shown in (**Figure 1.5A**).

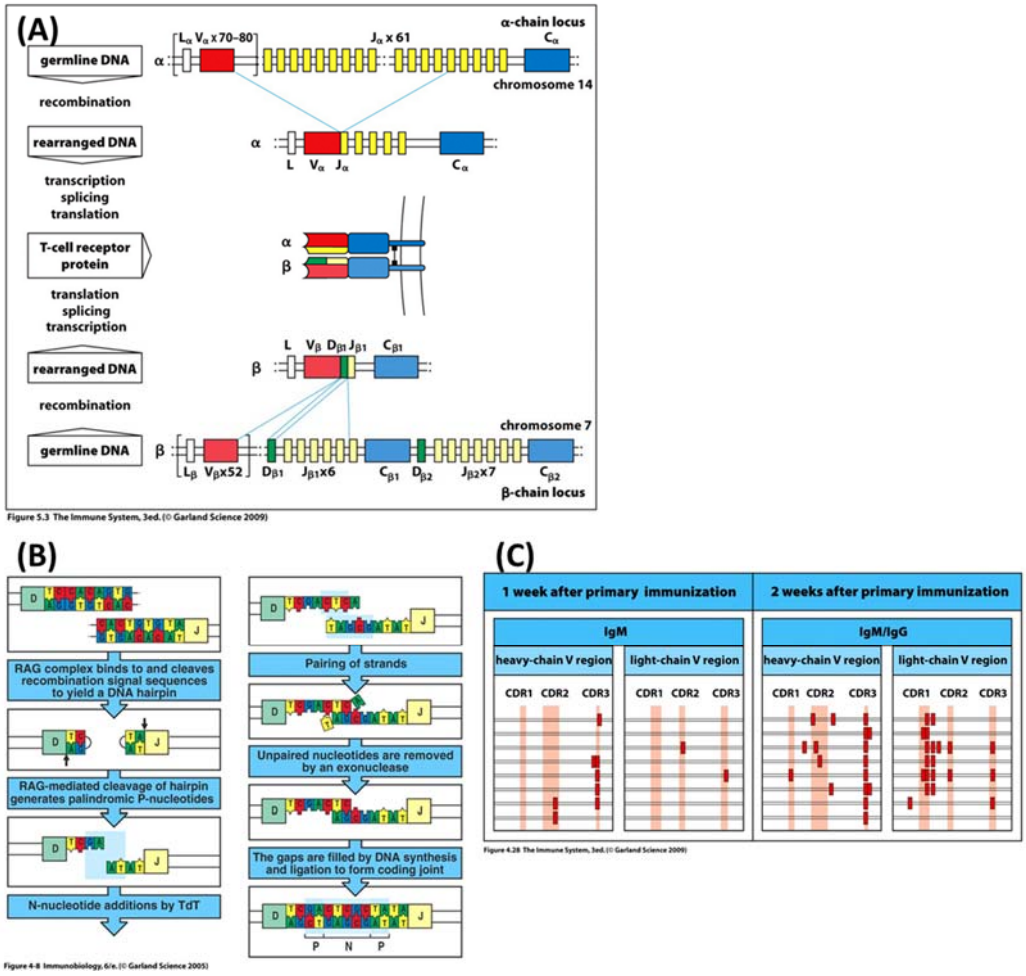


Figure 1.5: Diversity generation in the adaptive immune system.

The immune system generates diversity in three main ways to engage foreign antigens with a limited amount of gene information though (**A**) V-, D-, and J-region gene recombinations from the various gene segments on the chromosome, (**B**) the imprecise junctions with formation of palindromic hairpins and N-nucleotide addition in the hypervariable CDR3, and (**C**) somatic hypermutation in the antibody genes to mutate and selectively maturate the antigen binding portions of the antibody (Source: The Immune System 3rd and 6th editions, Garland Science 2014)[18].

The second level of diversity occurs in the imprecise hypervariable junctions formed between these recombined genes through addition and subtraction of nucleotides during rearrangement (**Figure 1.5B**). Junctional diversification occurs in all recombined chains except the antibody light chain. The random addition and subtraction of nucleotides at these junctions produces non-productive proteins with out-of-frame mutations or premature stop codons in the majority of these hypervariable junctions[19]. Cells with unproductive rearrangements will continue gene rearrangement with the remaining VDJ segments until either a functional antibody or TCR chain is produced or the cell dies.

A third level of diversity in these immune receptors comes from the pairing between light- and heavy-chains or α- and β-chains in antibodies and TCRs respectively. These pairings can control structural features in the receptor including the binding region orientation angles. Theoretically every VJ combination can partner with any VDJ combination, but based on sequencing data there is some amount of pairing bias occurring possibly due to incompatible interfaces between the two combinations of variable domains[20].

Beyond the productive rearrangement of antibody genes, activated B-cells can undergo somatic hypermutation wherein the antigen-binding genes randomly mutate to select for higher affinity variants (**Figure 1.5C**). This process occurs in the germinal center of peripheral lymphoid organs and allows antibody producing B-cells to evolve more effective antibodies. With each succeeding B-cell reactivation, the antibody genes can undergo further somatic hypermutation. Unlike antibodies in B-cells, T-cells cannot further maturate TCRs after leaving the thymus.

1.4: Antibody and TCR Structures

TCRs are multi-domain heterodimers with the antigen binding region positioned at the N-terminus of the protein chains followed by invariant constant domains (**Figure**

1.6A). Each domain in these receptors consists of two parallel β-sheets linked by a disulfide bond between highly conserved cysteine residues to form an approximate β-barrel shape. The variable domain of each chain has three pMHC-binding loops known as the complementarity determining regions (CDRs) (**Figure 1.6B**). The third CDR of each chain comprises the hypervariable region formed during the rearrangement of the Variable-, Diversity-, and Joining-region genes. The antigen binding region of the receptor comprises the dimer formed from the α- and β-chains which locates CDR1 and CDR2 peripheral to the hyper-diverse CDR3s in a rotationally symmetric orientation. The variable domain is linked to the invariant constant and transmembrane domains, which closely associate with their complementary chain to stabilize the TCR dimer.

Antibodies are large (~150 kDa) tetramer structures forming a dimer of light and heavy chain dimers. The N-terminus of each chain comprises a variable antigen-binding domain and the other domains are invariant constant domains. The variable domains form the antigen-binding region with six CDR loops directed towards the antigen interface. The CDR3 loops in antibody heavy and light chains are larger and more flexible hypervariable junctions, which often provide the majority of the interactions with the antigen. The light chain has one constant domain while the heavy chain has three. Heavy chain constant domains 2 and 3 form di-sulfide bonds between them giving the antibody bivalency for binding regions.

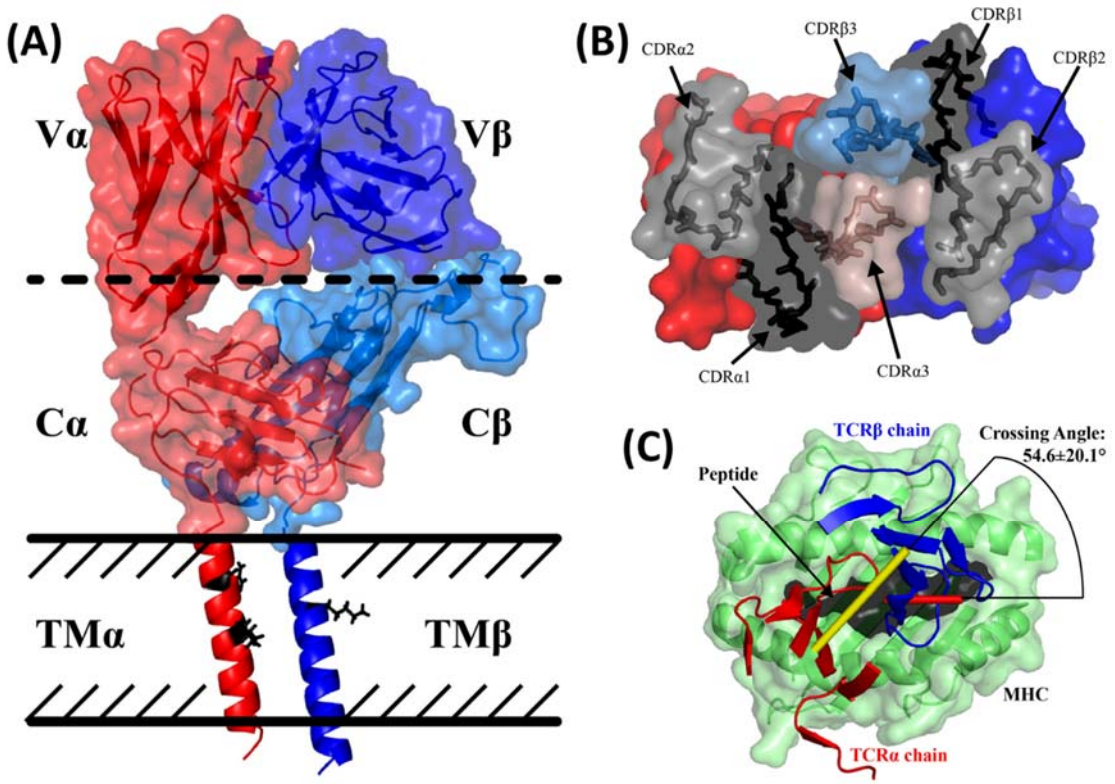


Figure 1.6: TCR structure and complex orientation.

TCRs have a conserved structure including (A) the variable, constant, and transmembrane domains with the charged transmembrane residues depicted as black sticks. (PDB: 1BD2), (B) the arrangement of the antigen binding CDRs with CDR3 central to the peptide binding core and CDR1 and CDR2 on the peripheral interacting with MHC, and (C) the semi-conserved binding orientation of the TCR-pMHC complex (PDB: 1QSF).

The peripheral arrangement of germline-encoded CDR1 and CDR2 loops restricts interactions to the MHC binding groove α-helices which stabilizes and orients the TCR over the pMHC[21]. The majority of the peptide recognition comes from the CDR3 loop, in particular the CDRβ3 due to its longer and more diverse structure. Unlike other types of protein-protein interactions, the membrane-bound TCRs bind in a conserved diagonal orientation with an average angle of 54.6±20.1° (**Figure 1.6C**). This measurement represents the angle between the linear approximation of the MHC binding groove and line through the center of mass of the TCR variable domains. A diagonal orientation allows a

mixed interaction between each of the αβ CDR3 to interrogate the bound antigen peptide generating the extra specificity and binding affinity[22].

Comparisons of TCR CDR structures and variable region sequences finds conserved loop conformations called canonical classes. These structural conformations form with the presence of key, shape-defining residues which interact internally or with neighboring CDRs[23–25]. In 2000, Chothia *et al* predicted the locations and amino acids for these key residues in TCR variable regions with a limited amount of aligned sequencing data and seven solved TCR structures[23]. The majority of these key residue and canonical class predictions remain accurate with the addition of more TCR structures. Analyzing the diverse TCR structures, sequence repertoires, binding kinetics, and functionalities reveals how TCRs bind a variety of disease antigens in MHC molecules with low affinity and high specificity.

1.5: PROTEIN ENGINEERING

Antibodies and TCRs have evolved drastically different binding kinetics due to their distinct roles in fighting disease. Soluble antibodies must bind a wide variety of antigens with relatively high specificity and affinity ($K_D << 1$ μM). These antigens consist of small chemicals, organic molecules, proteins, or whole pathogens requiring a significant amount of diversity in the antibody binding regions. Alternatively membrane-bound TCRs bind antigens in a highly controlled and stereotyped manner in the context of MHC with the coordination of other accessory molecules. TCRs also require high specificity for appropriate activation, but the rapid scanning of cell surfaces requires a much lower affinity. Biochemically, this results in a slow k_{on} and fast k_{off} with affinities often 200-fold reduced relative to antibodies.

Studying these low affinity ($K_D \approx 1 – 100$ μM) TCRs and pMHC interactions requires a variety of highly sensitive assays able to detect rare and unstable binding events.

These assays incorporate methods to increase the signal strength of the TCR-pMHC binding by saturating solutions or making multi-mers of proteins. Some techniques for measuring TCR binding kinetics include enzyme-linked immunosorbent assay (ELISA), phage display, surface plasmon resonance (SPR), flow cytometry, IL-2 release, and proliferation assays. Characterizing the binding, structure, and stability of TCRs can show us how the proteins function during after receiving activation signals from cognate pMHC.

Many characterization assays use soluble proteins with variants of the antigen binding region fused to different constructs and protein domains. Expressing truncated immune-receptors like fragment antibodies (Fab) or single-chain fragments (scTCR) reduces the protein complexity and can increase soluble expression levels[26]. In antibody Fabs the heavy chain constant region domains 2 and 3 and the antibody bivalency are removed, decreasing the complexity of the protein. Bacteria can express large quantities of Fabs and scTCRs, but these soluble variants often require stability mutations[27]. Additionally protein fusions with Fabs provide other formats for characterizing protein-protein interactions.

Engineering antibodies and TCRs to modify antigen-binding domains has shown particular efficacy in generating novel specificities to antigens[28]. Full-length monoclonal antibodies and other soluble protein constructs can serve as therapeutics for protecting immunocompetent hosts. Currently, the FDA has approved over 40 monoclonal antibodies for therapeutic use amounting to over $100 billion worldwide market sales. Antibodies, TCR, and other co-receptor proteins leveraged from the adaptive immune system have advanced the capabilities of biotechnology and medicine. For example, chimeric antigen receptors (CARs) link antigen binding domain of an antibody with functional signaling domains from other immune receptors to stimulate cellular activity.

1.5.1 Directed evolution of TCR affinity and specificity

The process of engineering proteins function involves generating a library of mutated DNA templates for the protein, expressing the protein product of each library member such that the DNA sequence of each variant is specifically recoverable, and screening each library member for altered function. Because of the large amount of degeneracy typically required to find significantly altered mutants, high-throughput screening methods are often required. In this manner directed evolution and library screening can select for higher binding affinity variants[28], map binding epitopes[29], assay enhanced enzymatic activity[30].

For example, scTCR C-terminal fusions to bacteriophage surface coat proteins provide a method to rapidly screen and select billions of receptor variants[31]. While phage libraries can rapidly produce and test libraries in the billions of variants the stability of the displayed proteins often varies dramatically from protein to protein and can bias results to more stable, lower affinity variants. Additionally the resulting library selection and protein assays can vary depending on which phage coat protein is fused to the scTCR[31].

Depending on the complexity of the protein to test, the overall size of the library, and the parameters of the binding assay several systems are available for producing and screening libraries including (in order of increasing complexity) selection screens, ribosomal display, bacteriophage display[32], spheroplast display[33], yeast display[34], or mammalian cell surface display[35]. Library screening should involve a compartmentalized system enabling the testing and selection every protein variant generated in a timely manner. Selection of the preferred trait or activity is also imperative for the screening process. For example with displayed proteins one can use a simple binding assay to select higher binding proteins and recover the genetic material linked to that phenotype.

1.6: CHICKEN OVALBUMIN REACTIVE DO11.10 T-CELLS

In the 1980s, many studies attempted to identify the various lymphocyte lineages in the immune system and determine the peptide-antigen recognition mechanisms involved with adaptive immunity. The simplest form of these studies involves dosing mice with antigen peptides then characterizing the activated T-cell populations with cognate pMHC[36]. In this way, the chicken ovalbumin (OVA) antigen became one of the more studied antigen models. In particular the DO11.10 and OT-II T-cell $CD4^+$ hybridoma clones recognizing peptide $OVA_{323-339}$ when presented by $I\text{-}A^d$ and $I\text{-}A^b$ MHC molecules respectively were well characterized[37].

While the DO11.10 and $OVA_{323-339}$ antigen model provides robust protein binding characterization and T-cell activation data its functionality depends on the register of buried and solvent exposed OVA residues along the MHC binding groove[38,39]. While there is some overlap of TCR recognition between different register presentations, these alterations are mostly deleterious to the overall T-cell activity. Of the 4 verified register presentations of $OVA_{323-339}$: $I\text{-}A^d$ in the DO11.10 system register 2 appears to be the active register for optimal T-cell activity[38].

The majority of the work described throughout this text specifically involves the analysis and characterization of the DO11.10 TCR. This includes high-throughput sequencing studies, structural comparison, and thymocyte reconstitution of affinity maturated TCR variants.

Chapter 2: High-throughput sequencing of TCR repertoires

2.1: CHAPTER SUMMARY

In recent years, high-throughput sequencing (HTS) technologies have accelerated our ability to analyze heterogeneous DNA samples with many millions of sequence reads per reaction. A few applications of HTS include quantifying the diversity of a prokaryotic microbiome[40], detecting proliferating leukemia cells[41], and repertoire focusing from chronic infections[42]. HTS studies have shown the relationships between TCR repertoire diversity and the relative amount of protection provided against antigens and diseases[43]. By linking specific generation of certain TCRs in terms of VDJ-gene utilization or CDR3 sequence variations, we can better understand of how repertoires shift and expand for inclusive antigen recognition and protection. With the availability of this technology, we sought to analyze the TCR repertoire for antigen-specific sequences resulting from immunization with the well-studied antigen chicken ovalbumin (OVA).

In this experiment, TCR gene repertoires were sequenced from mice immunized with OVA. This antigen challenge is similar to the an initial Kappler and Marrack *et al* study which produced the well-studied DO11 and OT-II T-cell lines[44,45]. The goal of the experiment was to determine if the OVA responding T-cells that were the focus of the original Kappler and Marrack study are highly represented across individual mice exposed to OVA[37,39,45]. We focused on the $CD4^+$ T-cell response more than the $CD8^+$ T-cell response because OVA antigen presentation should occur on MHC class-II molecules through exogenous means described in **Chapter 1.2.3**.

Harvested T-cells were sorted into $CD4^+$ and $CD8^+$ pools then TCR mRNA was reverse transcribed and amplified using a set of designed oligo primers for mouse T-cell Receptor α Variable (TRAV) and T-cell Receptor β Variable (TRBV) genes. These primer sets amplified cDNA spanning the TCR gene signal sequence through the constant domain.

The amplified TCR genes were sequenced using 454 pyrosequencing to produce over 460,796 sequence reads with 276,508 TCR sequences identified by IMGT/HighV-Quest alignment processing.

From these sequence data, we determined there was not a strong repertoire OVA-specific response signal. The probability of producing and detecting over background sequences shared (public) and unique (private) TCR repertoire from a single immunization is low. This expectation is due to the total predicted number of antigen-specific T-cells produced in a single immunization[46]. Even analyzing the sequences for TCR gene and CDR3 sequence similarities to other well-studied OVA-specific T-cells did not find any strong candidates within these current data.

2.2: INTRODUCTION

Technological progress in both HTS and big data analytics has allowed the deep sequencing of diverse polyclonal genes within immune repertoires. HTS can elucidate how the modern immune system has evolved whether through co-evolution with common diseases and viruses or through other mechanisms. Previously, TCR binding affinity was thought to moderate T-cell commitment to regulatory[47] or memory cells[48]. Some deep sequencing studies have found identical TCRs across T-cell subsets meaning TCR affinity and development may rely on more nuanced regulation[49,50]. Early work on immune repertoires in mouse found an increased utilization of TRBV13 with many autoimmune responses[51]. Deep sequencing studies of T-cell repertoires can further characterize chronic illnesses which cause skewed clonal populations with memory inflation as found in some CMV infected individuals[52,53].

Prior to this technology, repertoire analysis could not be performed to any high capacity or repertoire sequence depth. Studies instead used less qualitative methods of PCR and spectratyping to analyze polyclonal T-cell samples[54]. These methods can detect a level

of clonal population bias, variable region gene presence, and CDR3 length distribution, but cannot link variable-region utilization to CDR3 sequence and their frequency in the reperotire[55].

With human T-cell repertoires estimated to number 10^8 different TRBV sequences alone, these less qualitative methods lack the ability to thoroughly discriminate between the various subsets of clonal populations and truly unique sequences[56]. Deep sequencing of antibody and TCR repertoires gives a glimpse at the frequency and distribution of the clonal populations in an individual suggestive of the overall health of the adaptive immune system. The total number of sequences required to accurately measure the repertoire depends on the cell samples and biological question which some studies put at approximately 10^6 sequences[57].

By using the well-studied DO11.10 model we can determine the peptide-specific TCR sequences by removing sequences shared among both the healthy and OVA challenged mice[45]. In this way, we can see if there is a subset of shared sequences either similar in sequence to DO11.10 or a subset not found in the control repertoire. Exposing mice to particular antigens and finding TCR repertoire changes could find the antigen-specific sequences and degree of TCR cross reactivity generated during an adaptive immune response.

2.2.1: High-Throughput Sequencing technologies

Traditional Sanger sequencing with the multi-step process of random nucleotide incorporation into elongating strands combined with gel filtration to separate DNA by length has been the predominant method for sequencing. This technique requires homogeneous purified samples and considerable amount of DNA which often requires additional cloning for gene detection or screens. While this method is still predominantly

used for general sequencing applications, it is unable to efficiently analyze DNA libraries or gene expression patterns in heterogeneous samples.

In the 1990s, a new application combining pyrophosphate chemistry and microfluidics generated a new sequencing technology called pyrosequencing which could sequence large quantities of heterogeneous DNA samples. In 2000, 454 Life Sciences licensed this technology and was later purchased by Roche in 2007. 454 pyrosequencing involves the real-time elongation and detection with pyrophosphate substrate linked nucleotides[58]. Incorporation of these nucleotides results in the oxidation of luciferin by luciferase to release discrete photons (**Figure 2.1A**). Set up of these reactions involves oil-water emulsions with beads containing the sequencing DNA and a massive array of detectors to analyze the photons produced by each of the individual bead reactors able to read greater than 10^7 bases per hour[59]. These pyrosequencing experiments can also process DNA sequencing read lengths of up 450 bases which is long enough to accurately read a TCR variable region through to the constant region. In 2013, Roche announced the discontinuation of the 454 pyrosequencing technology as much of the HTS experiments moved to newer technologies licensed by Illumina.

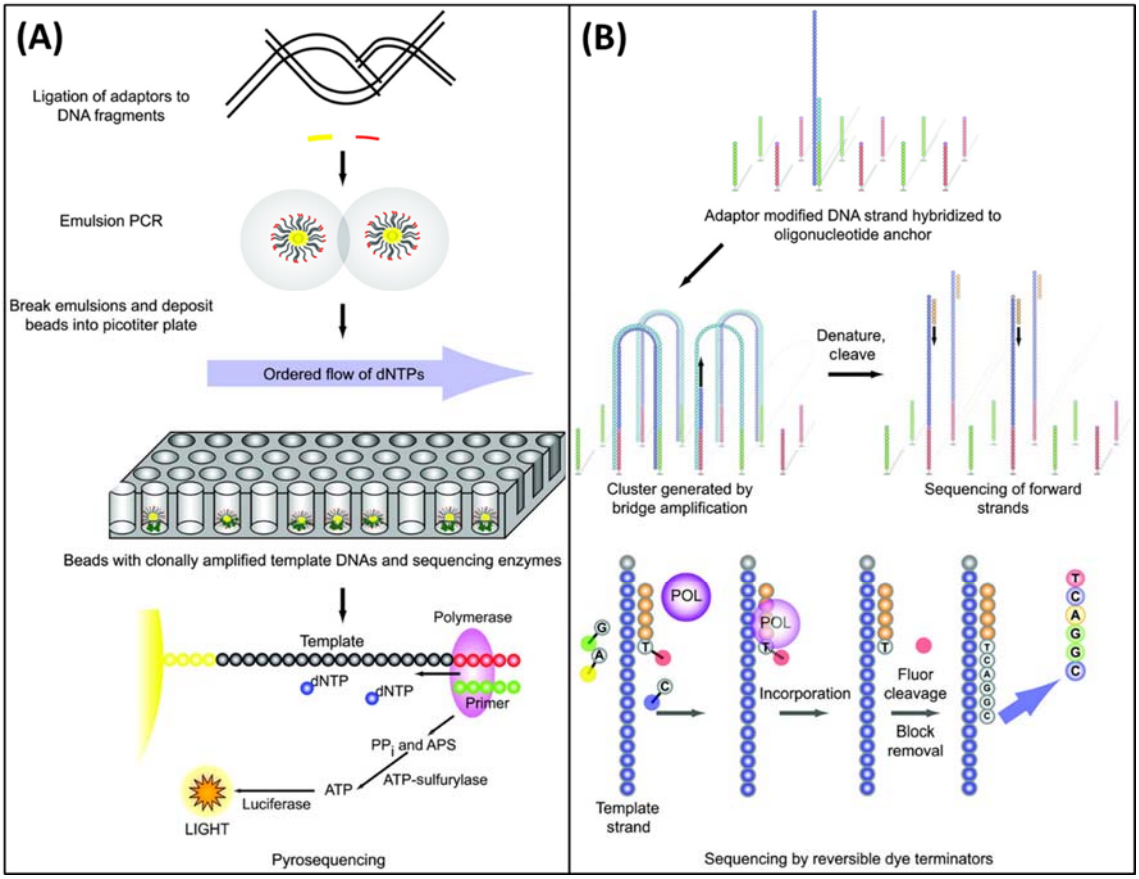


Figure 2.1: Diagrams of the HTS technologies

(**A**) 454 pyrosequencing occurs through several steps of DNA modification and enzyme chemistry. DNA fragments are ligated to adapters and fused to beads for clonal amplification. The beads are then distributed to a picotitre plate and sequenced using pyrophosphate chemistry and luciferin oxidation[60] and (**B**) Similar to 454 sequencing, Illumina technology uses a DNA adapters, but to hybridize template DNA to a flow channel surface instead of beads. Template DNA is clonally amplified and the sequence is read by fluorophore cleavage from the nucleotides[60].

In recent years, a second-generation HTS technology licensed by Illumina has replaced 454 pyrosequencing. Illumina sequencing uses different chemistry and sample processing to collect many millions of sequence reads with similar sequence quality to prior technology[61]. Sequence reads for Illumina are shorter than other technologies so

additional sample processing is required if the DNA sample is longer or if sample coverage is necessary for both forward and reverse directions[59].

To set up reactions, transposons add adaptors for flanking index motifs to template DNA using reduced cycle amplification. Next, the DNA samples are added to a flow channel to hybridize with complementary oligo adapters on the channel surface. Bridge amplification between nearby complementary adapters clonally amplifies the DNA as clusters of reactions. The 3' ends are then blocked and sequences are read using fluorescently labeled nucleotides which emit discrete wavelengths of photons based on the incorporated nucleotide base (**Figure 2.1B**). The light emitted by each clonal cluster during sequencing is collected using a fiber optic bundle monitoring the entire flow cell. This is called sequencing by synthesis since the sequence is read during second strand synthesis. Sequence clusters are read for the desired length before further processing to read the reverse direction of the template DNA. After sequencing in the forward direction, the template DNA is bridge amplified to a nearby complementary adapter and the sequence is read in the reverse direction. Processing and aligning the two read directions for each DNA amplification cluster allows a sequence validation as each read direction should be complementary of one another if the template DNA is the appropriate length.

These high-throughput technologies have dramatically increased the variety of analyses available to scientist including measuring gene expression with cDNA libraries from mRNA, the diversity of the microbiome using ribosomal DNA[40], and diverse cellular systems with polyclonal repertoires. Aside from the shorter read length from HTS, the error rate is also higher than for older Sanger technologies. These technological shortcomings are mitigated by the overwhelming amount of sequence coverage. The sheer number of total sequences that are read allow sequencing overlap to identify erroneous data[60].

2.2.2: Ig/TCR repertoire analysis

HTS studies of gene repertoires provide scientists with deeper insight into the mechanisms generating the diversity found in the adaptive immune. The sequences found in productive and non-productive TCR chains show the probability of certain gene rearrangements and junctions occurring with success during lymphocyte maturation selection[62]. Before HTS analysis became more prevalent, immune repertoire studies consisted of relatively small (<100) sample sizes of selected or sorted populations which comprises a minuscule fraction of the total expected diversity in an actual repertoire[63].

After genome sequencing of humans, mice, and other species with adaptive immune systems, scientists could study the genes and mechanisms which generate antibody and TCR diversity. Compiling this genetic information into a database which could model the gene rearrangements, Lefranc *et al*, constructed the IMunoGeneTics (IMGT) database[64]. This database can identify the genes and PN-junctions from both single entries and high-throughput data for DNA and protein sequences. The expanding number of repertoire analyses from other studies have begun to roughly define the TCR background repertoires[65]. This background repertoire in mice has helped elucidate mechanisms which produce inherent biases within the genetic rearrangements and junctional CDR3 sequences[19,66,67].

While sequencing and analysis of B-cell and T-cell repertoires is now a standard method, identifying TCR-antigen specificity requires additional considerations. The highly regulated T-cell maturation process selects for a degree of cross-reactivity within foreign antigen identification from the low affinity binding kinetics[68]. The cross-reactive nature of TCRs against a particular antigen suggests proper protection requires recognition by multiple antigen receptors. This requirement for polyclonal TCR recognition may also mitigate antigen escape, providing antigen mutation tolerance[69].

TCRs interact with antigens processed and presented by MHC molecules, which vary between individuals. Therefore describing the TCR repertoire response for any antigen requires an additional analysis of the pMHC generating that response. MHC molecules are highly polymorphic between individuals and the variations in both the α-helices and the β-sheet binding groove floor can cause differential presentation of disease antigens and TCR recognition[4]. For example certain human leukocyte antigen genes which recombine to form the MHC can provide enhanced protection against diseases like HIV due to antigen presentation compatibility[70].

Antigen specific TCR sequences shared across multiple individuals are known as public sequences. These types of sequences can occur with some frequency due to a stereotyped antigen interaction or from higher probabilities of generation during gene TCR gene rearrangment[71,72]. Both types of shared sequences occur by different immune mechanisms in normal and diseased populations. Individuals with public sequences have identical receptors thought to interrogate the MHC and particular peptide antigen nearly identically. These shared TCR sequences are found with diseases such as Epstein bar virus[73], CMV[72], and multiple cancers[74]. Identified public TCR sequences highlight the propensity to form particular receptors from the possible pool of gene combinations and junction rearrangments[74,75]. Public receptor sequences could also help develop new vaccines and therapeutics to target specific disease antigens with enough MHC tolerance shared across a broad population with limited nonspecific interactions.

Natural immune processes such as aging can also lead to biased T-cell repertoires. In these repertoire biases, an initially small portion of highly active disease-specific thymocytes proliferate until they comprise a much larger portion of the population termed memory inflation. Specifically the process of aging slows thymic output of naïve cells placing a higher burden and responsibility to protect on existing T-cells leading to

stagnating or decreasing receptor population diversity[62]. These more homogenous thymocyte populations lack the repertoire diversity necessary for proper protection and can cause complications similar to being immunocompromised.

In order to determine a peptide-derived basis for autoimmunity, cancers, and other T-cell mediated diseases, we need a better understanding of TCR VDJ-region gene utilization, hypervariable junction diversity, prevalent structural conformations in the TCR binding region, and receptor functionality. With recent protein developments utilizing TCRs as therapeutics against cancerous and viral targets, mapping the repertoires will identify those effective TCR gene segments frequency and CDR3 sequences to engineer increased antigen-specificity[76].

2.3: METHODS

2.3.1: Primer set design and validation

Using the IMGT database to identify the TCR variable region genes in the mouse genome, a primer set was designed to amplify the expected TCR sequences produced in the mouse repertoire[77]. The sense strand primers were designed to anneal within the signal sequence of each functional variable region. The antisense strand primer for TCR gene amplification was designed to anneal 12 nucleotides into the respective constant domain. To reduce the number of primers in this set a degree of nucleotide degeneracy was used to combine similar signal sequences. This way a single degenerate primer could contain up to 4 different sequences. The average melting temperature of the primer set was 55°C with an overall range of 51.5 – 59.2°C. The exception to these design parameters was TRBV13, which was instead used to target a more conserved sequence within the framework of the variable region. The mouse TCR primer set (**Table 2.1**) includes 29 and 19 primers for TRAV-regions and TRBV-regions respectively. Similar design principles have been used

to generate a primer set for the amplification of 83 mouse TCR genes from the more conserved region, but were not able to successfully target all the TRAV and TRBV genes[78].

Primer	Gene Targets	Tm (°C)	Sequence	HTS	TOPO	50°C	53°C
TRAV01	1	52.7	GGT TTG TTC TCT ATC TCT TCC TG		X	2	5
TRAV02	2	53.7	ATG AAG CAG GTG GCA AAA G			0	5
TRAV03	3-1,3,4; 3D-3	54.3±0.4	GAC CTT TGT TSC TGT GCT TC	X	X	0	2
TRAV04	4-3,4; 4D-3,4	55	GCT GTG CTG GGG ATT CT	X		2	5
TRAV04-2	4-2	59.2	CTG CTG TTG GTG CCG CT	X		4	2
TRAV05-1	5-1	54.8	CTT CTG GCT ACA GAT GGA CT	X		0	5
TRAV05x-4	5-4;5D-4	51.5	ATT ATT CAT GTT TCT ATG GCT GC			0	5
TRAV06-1	6-1:3; 6D-3	53.8	GCT TTA GTG ACT GTG ATG CTG	X		2	2
TRAV06-4	6-4; 6D-4	56.5±2.1	GTT TTA GTR ACT GYG ATG CTG CTG	X		3	4
TRAV06x-5	6-5; 6D-5	55.4	CCT GAA CTG GGK ATT CTA CTC TTC	-	-	-	-
TRAV06-6	6-6; 6D-6	54.4±0.8	CTY TTC TCC AGG CTT CGT G	X	X	2	5
TRAV06-7	6-7; 6D-7	52.5±1	ATG AYT GTG ATG CTC CTC AT	X		0	5
TRAV07	7-1:5; 7D-2:5	54.3±0.3	CTA GTG GTC CTG TGG CTY	X		2	5
TRAV07x-6	7-6; 7D-6	52.2	CTA GTG TTC CTC TGG CTT C			0	4
TRAV08x-1	8-1; 8D-1	53.9	GGT TGT TGA TGG TGT CAC TG	X	X	2	3
TRAV08x-2	8-2; 8D-2	52.8	GGA ATA TCT TTG GTG ACT CTA TGG		X	0	2
TRAV09x-1	9-1; 9D-1	54.3	TCG TTC CTC GGG ATA CAT TTC	X		0	5
TRAV09x-2	9-2:4; 9D-2:4	53.7±0.5	CAG TSC TGG GGA TAC ACT	X	X	0	5
TRAV10	10; 10D	54.5	ACA TCC CTT CAC ACT GTA TTC C	X		0	5
TRAV11	11; 11D	57.9±1.5	GCC TKA GTG CCT GCT GG	X		2	3
TRAV12	12-1:3; 12D-1:3	56.5±1.2	SCT CAG TTC TYG TGC TCC TC	X		0	5
TRAV13	13-1:5; 13D-1:4	56.9±2.6	TCT CTG YTG GGG CTY CTG	X		5	5
TRAV14	14-1:3; 14D-1:3	54.5±1.5	GAC AMG ATC CTG ACA GCA WC	X		3	5
TRAV15	15-1:2; 15D-1:2	57.1±0.3	ATG CCT CCT CAS AGC CTG	X		3	4
TRAV16	16; 16D	54.2±1.1	CTG ATT CTA AGC CTG YTG GG	X		2	4
TRAV17	17	55	AGT GAC CAT TCT GCT GCT C	X		0	4
TRAV18	18	54.5	ATG CTC CTG AAA CTC TCT GTG		X	2	4
TRAV19	19	54.4	GCC TTG CTG TTG GTT CTG			3	5
TRAV21	21	54.1	GAT GTG TGA GTG GAA TTG CC		X	0	5
TRBV01	1	56.6	TTC TGT GCC TCT GTG TAC TCA T	X		0	0
TRBV02	2	56.6	GGC TCC ATT TTC CTC AGT TGC	X	X	0	0
TRBV03	3	53.9	CTG GCT TCT AGG TTG GAT AAT TTT TAG	X		0	3
TRBV04	4	56.6	GCT GTA GGC TCC TAA GCT GT			0	4
TRBV05	5	55.4	GCA GGC TTC TCC TCT ATG TTT C	X	X	0	4
TRBV12	12-1:2	56.4±1.5	CTG CTA TCT TGG GTT RCT STC TTT C	X		0	3
TRBV13	13-1:3	56.2±2	RCA AGG TGR CAG TAA CAG GAG	X	X	2	0
TRBV14	14	56.7	CTT GGC TGG GCA GTG TTC	X		2	0
TRBV15	15	54.2	AGA CCC TCT GTT GTG TGA TC	X		0	0
TRBV16	16	56.4	CTT TTC TGT CTG GTT CTT TGC TTC		X	0	0
TRBV17	17	55.4	CTA GAC TTC TTT GCT GTG TGA TCT TC	X		0	3
TRBV19	19	56.1	ATG AAC AAG TGG GTT TTC TGC TG	X		2	4
TRBV20	20	55.2	ATG TTM CTG CTT CTA TTA CTT CTG G	X	X	0	0
TRBV23	23	53.7	CAC GGC TCA TTT GCT ATG TAG			1	0
TRBV24	24	55.4	CAA GAC TGC TCT GCT GTG TAG			1	0
TRBV26	26	56.7	ATG GCT ACA AGG CTC CTC TG	X		2	5
TRBV29	29	56.6	TTA GGC TCA TCT CTG CTG TGG	X		2	2
TRBV30	30	53.5	GAC ATT CCT GCT ACT TCT TTG G	X	X	2	4
TRBV31	31	56.5	TGC TGT ACT CTC TCC TTG CC	X	X	2	4

Table 2.1: Primer set for mouse TRAV and TRBV genes.

The TRAV6-5x primer was added Detailed list of the V-region primer set used to amplify the cDNA. The columns are primer name, V-region gene targets, melting temperature, nucleotide sequence, validated with TOPO cloning, and amplicon quality at 50°C and 53°C annealing temperatures respectively.

These primers were validated by amplification and sequencing of mouse lymph node (**Figure 2.2A**) and T-cell hybridoma (**Figure 2.2B**) cDNA from mRNA. Individual amplifications for each primer are shown below (**Figure 2.3 A** and **B**), all showed appropriate bands at the expected size except TRBV23. Samples of α and β bands were purified and subcloned with a topoisomerase vector for sequencing to confirm the targeted amplification of the TCR gene. These confirmations are reported in **Table 2.1**.

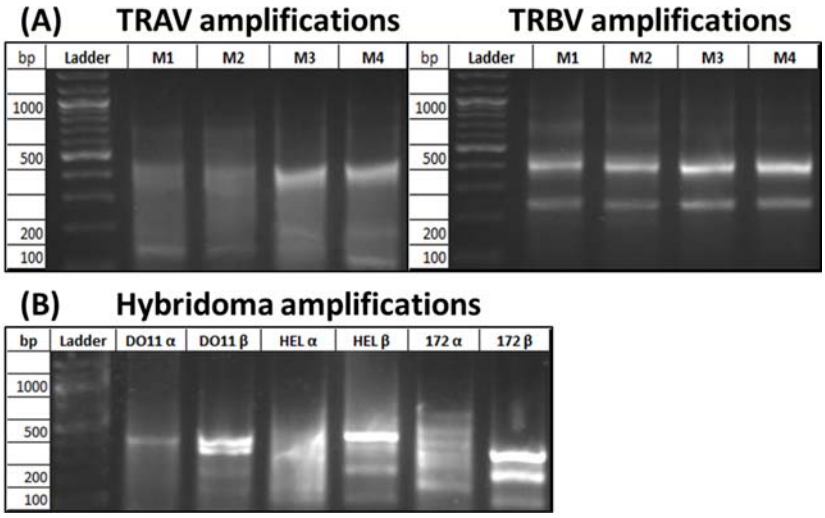


Figure 2.2: Amplified TCRs from mouse and hybridoma cDNA.

When the primer set are pooled and used to amplify cDNA from the treated groups of mice the products looked (**A**) similar to a combination of all the bands in (**Figure 2.3.1.2A** and **B**). Additionally these pooled primer sets worked on DO11.10, HEL, and 172.10 hybridoma lines (**B**) for the α- and β-chains respectively.

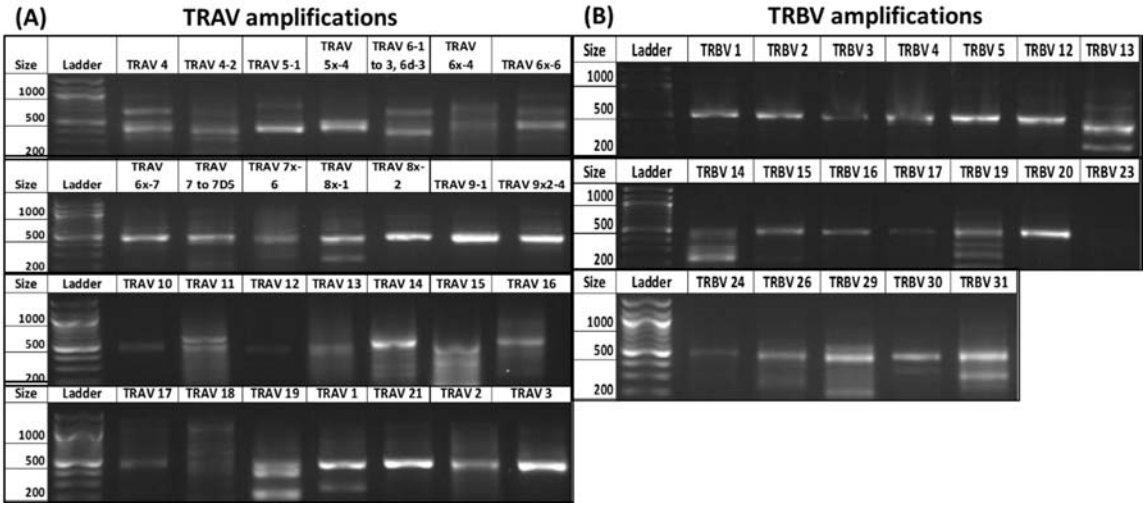


Figure 2.3: TCR V-regions amplified by the primer set.

PCR amplifications for the mouse TCR primer designed to optimize reaction conditions to amplify and prepare the samples for sequencing: (**A**) TRAV-regions, (**B**) TRBV-regions.

To optimize the cDNA PCR annealing temperatures were tested for each TRAV or TRBV primer from pooled mouse mRNA. The reaction quality is scored based on the resulting band(s) when run on an agarose gel (**Figure 2.3 A** and **B**). A strong single band received a **(0)**, a weak single band **(1)**, a doublet **(2)**, a triplet **(3)**, four or more bands **(4)**, and a smear received the worst score of **(5)**. This method found the most primers could successfully amplify the TCR sequences using an annealing temperature of 50°C.

2.3.2: Mouse inoculation

Four healthy 25-27 week old BALB/C (*C57/BL6)* mice immunized with 100 µg of OVA. Two weeks after inoculation, the mice were sacrificed and their spleen and lymph nodes were harvested collecting approximately 335 mg of organ material per mouse, which was then placed in a solution of *RNAlater* for one hour. The organ tissue was then homogenized the using a sonicator at the highest setting for 1 minute to suspend individual cells. The suspended cellular solution was then sorted using magnetic Dynabeads (Invitrogen #114.45D, 114.47D) for separating $CD4^+$ and $CD8^+$ t-cells.

2.3.3: TCR gene amplification

Total mRNA was purified from the sorted cells. TRIzol (Invitrogen) was added to the cell solutions and then lysed with a homogenizer (OMNI GLH) or mechanical sheering in a 23-gauged syringe. Following the manufacturer's protocols, RNA was then isolated from the aqueous phase after 1-bromo 3-chloropropane addition and sample centrifugation. The RNA samples were purified using QIAgen's RNEasy kit. Any contaminating DNA genomic was digested using a DNA-free DNA Removal Kit (Invitrogen). Purified samples immediately used in cDNA preparation or were stored at -80 °C until further use.

cDNA was generated from purified mRNA using the protocols prescribed SuperScript RT II (Invitrogen) with an extension of a PolyA Tail oligo (5' – $[T]_{23}$VN – 3') for 50 minutes at 42°C. The resulting cDNA was immediately amplified by the TRAV and TRBV primer sets separately mentioned in **Chapter 2.3.1** using Accuprime pfx polymerase (Invitrogen) with 25-30 PCR cycles with an annealing temperature of 50°C and elongation time of 30 seconds. In order to mitigate bias introduced by the primer set during cDNA amplification, a reduced number of PCR cycles was used to amplify TCR genes before sequencing. PCR samples were purified by gel extraction and either subcloned with a TOPO-TA kit (Invitrogen) or submitted to the Genome Sequencing and Analysis Facility at the University of Texas at Austin for further processing and sequencing with a Roche/454 Titanium Series Instrument.

2.4: RESULTS

2.4.1: Data processing with IMGT and SQL

The IMGT/HighV-Quest analysis tool was used to examine the sequence data for VDJ-region identity, CDR3 sequence, homology to germline sequences, and relative amount of error in the data. In a pilot study we performed, the IMGT processing returned a larger than expected amount of identifiable but unproductive receptor rearrangements

due to either single frameshifts in the coding region or stop codon mutations. Otherwise IMGT identified the majority of the gene regions in the sequence data and determined the CDR3 sequence errors were below 2% as determined by the total amount of mutations found and the anticipated errors per sequence length. We next set up a database using Microsoft's Server Query Language to further process the IMGT sequence data to calculate differences in sequence populations.

2.4.2: Mouse TCR populations

Two separate sets of sequencing submissions, the first was a pilot study with one immunized mouse and the second with two immunized mice and two control mice produced 4,882 and 460,796 sequences respectively. The TCR sequences in these studies were harvested and synthesized identically, but the smaller sample only included $CD4^+$ cells with pooled α- and β-chain amplifications whereas the larger study was sorted for both $CD4^+$ and $CD8^+$ T-cells with separate α and β amplifications. After IMGT processing and removal of sequences with low quality or truncated length, there were 3,922 (80 %) and 276,508 (60 %) identified TCRs in the pilot and second submission respectively.

In the pilot sequencing experiment with 3922 identified TCRs by IMGT with 55 of 82 different TCR V-region genes with TRAV-genes 3, 6, 7, 12, and 13 which individually appear greater than 5 % with diverse CDR3 sequences and TRAJ genes. TRAV-7 in particular comprised nearly 44 % of the data (**Figure 2.4A**). Further analysis of the TRAV-7 alleles (**Figure 2.4B**) identified 1288 unique and surprisingly diverse sequences. Additionally all clonal sequences each comprised much less than 1 % of the total sequences in the data. The diversity and distribution seen in this smaller sample with only 11 V-region genes identified more than 5 times (greater than 0.1 % frequency) also indicated the need for a deeper sequencing run to identify rarer genes.

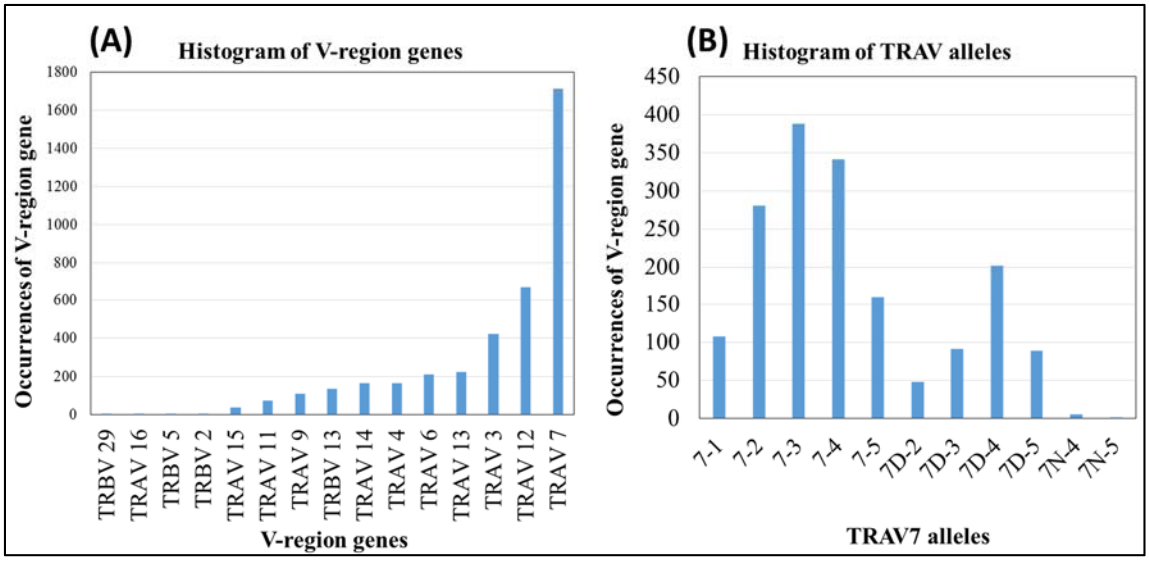


Figure 2.4: V-region and TRAV7 gene histograms.

These histograms show the distribution of the v-region gene frequencies for (A) the smaller sequencing experiment and (B) the identified TRAV7 genes by allele for the productive IMGT-identified sequences. We determined a lack of V-region gene because the distribution and primer melting temperature do not correlate.

For the larger sequencing experiment with 4 mice, 16 DNA samples of PCR amplified cDNA were submitted. Samples were distinct for the individual mouse, $CD4^+$ / $CD8^+$ sorting, and then separately amplified α- or β-chain samples. This study of four mice found 272,508 (60 %) sequences containing an identified CDR3 junction and the IMGT analysis returned 87 % of the sequences in the following breakdown: Productive (39.8 %), Unproductive (22.8 %), Unknown (30.0 %), and No Results (7.4 %). Our data showed the TCR sequence repertoire composition of control and OVA challenged mice are similar with approximately 9.7 %, 1.2 %, and 6.4 % being individually unique to each mouse, unique to the treatment group, and shared between treatment groups respectively. The remaining majority of sequences were shared between one or two mice with no correlation to treatment.

There were no specific V-regions, CDR3 sequences, or J-regions that were found in higher frequency in one mouse treatment group over the other. TRBV amplifications had higher rates of productive sequences with 52 % of sequences productive compared with 37 % for the α-chain counterparts. In fact, the strongest difference between the two sets was a TRBV29 deficiency in the treated mice with a 2.0 % difference between the groups. Otherwise, this slice of these mice TCR was too limited in size to determine any significant differences between the treatment groups.

Between the sorted populations of CD4 and CD8 cells, there was little measured variation in the utilization of specific TCR genes or CDR3 sequences. Aside from the variation occurring from the different sequencing depths, the relative amounts of unique and shared gene and CDR3 sequence combinations (**Table 2.2**) did not indicate the clear development of a specific immunization response.

Chain	Lineage	Mouse	Total	V,J,CDR3	V-J
α	CD4+	1	46532	28436	2058
		2	32777	20622	1921
		3	32426	21117	1866
		4	22050	15076	1628
	CD8+	1	5542	4385	1021
		2	44963	29100	2017
		3	18505	13098	1627
		4	19411	13906	1708
β	CD4+	1	5793	4946	135
		2	8332	6469	145
		3	7664	5900	151
		4	9199	7194	174
	CD8+	1	3301	2665	135
		2	7274	5860	171
		3	7122	5793	151
		4	5617	4532	119

Table 2.2: The total sequences for each sample pool.

These data show the number of identical sequences found when grouped by the various IMGT identified TCR characteristics. The number of TRAV sequence combinations is many times larger than the TRBV groups partially due to sample preparation and due to the fewer number of V- and J-region genes.

There were 1894 identical sequences in all 4 mice with matching V-region, CDR3 sequence, and J-region returned from the IMGT analysis. Additionally there were 14745 sequences shared across the pooled treatment groups. Using the three TCR traits this way was considered the most effective method to identify duplicate sequences and possible bias introduced in the handling of material or processing of the DNA for sequencing preparation. Setting a threshold quality score of 200 and combing the identified gene and junction information showed a varied amount of identical sequences in the data. Similar to the smaller pilot experiment there was a significant amount of TRAV7 gene sequences recovered in over 130,000 sequences across all the 16 samples. Additionally there was a strong TRAV11 and TRAJ18 combination with over 4,000 sequences found in the sequence set.

While certain sequences appear in dramatically different amounts in the data this does not necessarily mean they reflect a repertoire of the same composition. This could be the case if the sequences have increased representation due to PCR bias, but correlating the amplified sequence with the corresponding primer melting temperature did not find any trend between a lower melting temperature or primer degeneracy and the total number of identical sequences in the sample. Compiling the data in a histogram (**Figure 2.5**) shows the degree of sequence repeats fits well to the power law equation shown in the legend below. If the model holds for variable depths of sequencing this allows one to calculate the total number of sequences required to find a particular sequence of a low frequency.

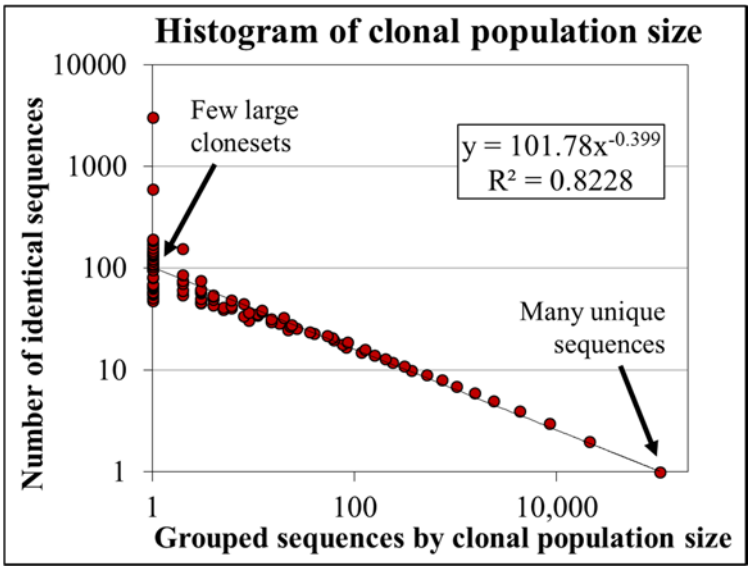


Figure 2.5: Plot of identical TCR sequence occurrence.

This plot shows the relative amount of identical sequences found within all the sequence data. The x-axis is comprised of the number of sequences based on the clones for that sequence. The y-axis is the number of clones within a group. From the plot, the x-intercept represents the 105,394 unique sequences that only occur once in the data. Similarly, the y-intercept represents the largest clonal population with 3029 identical TCRs and it occurs for one particular sequence in the data.

The sequence data showed no statistical difference between our control and immunized mice groups in terms of gene usage or CDR3 junction. However there is a slightly increased frequency of a few variable-regions in the OVA treated mice, but these data do not conclusively show a distinct shift in the overall TCR repertoire. PCR bias in the sequencing data could be partially seen in the abundance of the TRAV7 and TRBV13 gene sequences from all mice (**Figure 2.6**). Amplifications lacking these primers for these two regions still recovered these V-region sequences to a high frequency (Ben Roy, personal communication).

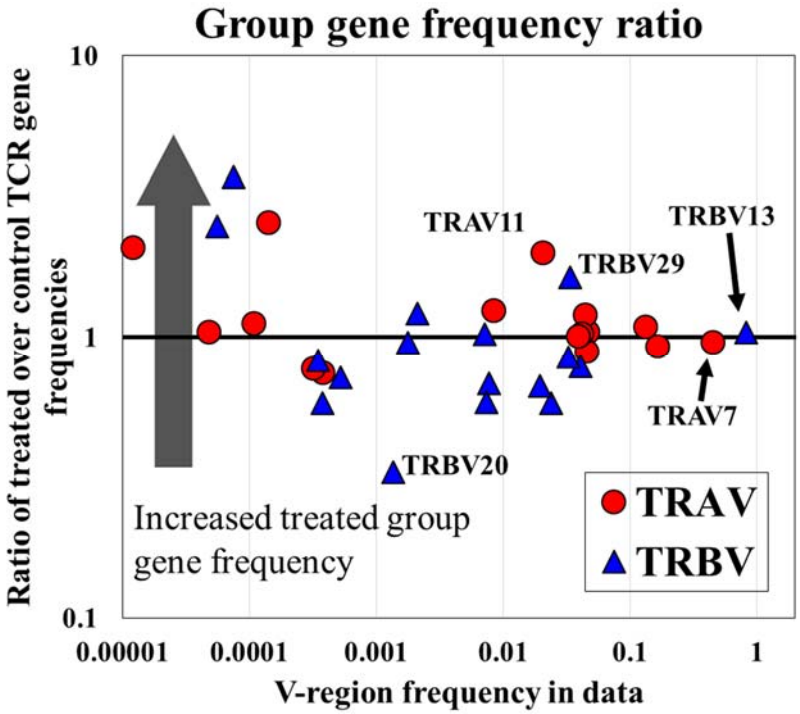


Figure 2.6: Treatment group V-region frequency

This graph shows the ratio of frequency for each V-region in the sequenced mice. V-regions appearing greater than 100 times in the data set (~0.03 % of the total identified sequences or ~0.2 % in each mouse) were measured for an increased frequency ratio between the treatment groups. A frequency ratio value above 1 means an increased frequency of a particular V-region in the OVA-immunized mouse repertoires.

The established OVA-specific and DO11.10-like hybridoma sequences were not found in any of the mouse sequences[45]. V-regions expected to have the most similarity by analysis using clustalΩ to the DO11.10 sequences include TRAV10 and TRBV29. A V-region gene analysis for Searches using Markov sorts measuring amino acid residue similarity of the V-region, J-region, and CDR3 sequence showed no strong similarity to the model hybridomas either.

Of the 61 identified TRAJ genes, 21 are non-functional, either as pseudogenes, vestigial genes, or open reading frames. Most have multiple errors in recombination sequences and elsewhere such that they do not seem to be able to recombine. Of 9 that appear to possibly be capable of recombination, IMGT predicted they would not form productive genes. For example TRAJ-36, -54, -55, -60, and -61 have stop codons leading to nonfunctional rearrangements, but appear in 320 sequences.

Additionally TRAJ-7, -44, and -47 which appear in over 5000 sequences, are is not recognized by IMGT as a functional TCR sequence from a lack of the conserved amino acid sequence motif of F (or W) followed by G-X-G, where X can be any amino acid. In the case of TRAJ47, our sequences conform to the consensus sequence. Instead of matching the reference sequence CGLG, our sequences are FGLG. This means that either their reference sequence is wrong or we have identified a distinct allele, TRAJ47*02, that is functional. The case of TRAJ-7 and -44 is a bit more interesting. Again, we have a sequence discrepancy. However, their reference sequences show LGAG for TRAJ44, while our results are LGTG. TRAJ7 matches in both our sequences and the IMGT reference. In other words, while there is a discrepancy, our genes still do not fit with the conserved FGXG motif.

As a result of this mismatch with consensus sequences, we cannot definitively claim that TRAJ7, -44, -47 are functional genes, but we can provide evidence that they might

be. We have over 5000 sequences that include TRAJ7, -44, -47 regions, but we haven't found a single productive TRAJ-4, -36, -54, -55, -60, or -61. If all of these genes are non-functional, we might expect them to show up occasionally, just as we accurately predict that CDR3s with stop codons show up occasionally, as excluded alleles. The fact that there are none, and numerous TRAJ7 and TRAJ44 containing sequences indicates this gene selected as functional over and over again. The occurrence of these genes has also been found in other HTS studies with mouse repertoires[67].

Based strictly on the IMGT analysis we have 272,508 sequences, of which there are 170,184 unique V-Region-CDR3 pairings and 114,791 unique CDR3 sequences (when grouped solely by CDR3). This means there are 55,393 CDR3s shared by more than one V-region. The total number of identified productive CDR3s available to us could be theoretically increased with an additional analysis of the identified TCR sequences in our data to parse the CDR3 errors as they appear in large abundance as frameshift mutations or stop codons. We might be able to remedy these incorrect bases by identifying if the originates from the native gene sequence or formed in the PN junction during gene rearrangement.

To do this analysis we modified the CDR3 analysis to isolate the junction rearrangement using the most likely V-region and J-region sequences, which occur in the appropriate order. The program searches for either α or β CDR3 with the flanking conserved **Y–[YFLI]–(C104)** and **F118 (or W)–G–X–G** motifs. We could determine the accuracy of our filter since both chains have these highly conserved residues and because the 3' primer used to amplify the variable region sits at nearly the exact same location at the C-terminal end of the translated gene. Some determined junctions had obvious deficiencies in that they were too long or short. The average productive junction length for both CDRα3 and CDRβ3 was approximately 15 amino acids between IMGT

residues 104 to 118. The distributions of CDR3 lengths also matches previous spectratyping estimates of for the hypervariable loop. Monitoring these estimated junction lengths can limit the amount of error in the IMGT analysis, which from this limited data set appears to be a significant portion of the returned sequences.

2.4.3: DO11.10 and OT-II similar TCR sequences

In an effort to find OVA specific sequences in a high background of nonspecific T-cells in the sequenced samples, the data were passed through a filter for DO11.10 homology. The DO11.10 TCR consists of gene regions TRAV5D-4 and TRBV13-2. This assumes the single OVA immunization would be enough to generate both a stereotyped DO11.10-like sequence complete with gene regions and CDR3 sequence as well as produce this antigen specific response in a measurably significant number of sequences. To do this analysis, all the mouse V-regions were sorted through the residue homology analyzer clustalΩ[79]. Sequences with over 90% sequence homology were then analyzed for either CDR3 residue sequence conservation or similarity to published OVA-specific T-cells[45]. To measure the residue similarity we generated a Markov scoring algorithm to measure and compare the sequences for amino acid similarity or a consensus or to a particular sequence of interest[80]. Sequences with length differences greater than two residues from the average or published sequence were penalized. Results from the Markov algorithm are shown as the Markov score versus the ranked sequences based on that score (**Figure 2.7A**). There were few results or generalities gleaned from the scoring partially because of the variable CDR3 lengths shown as a histogram below (**Figure 2.7B**).

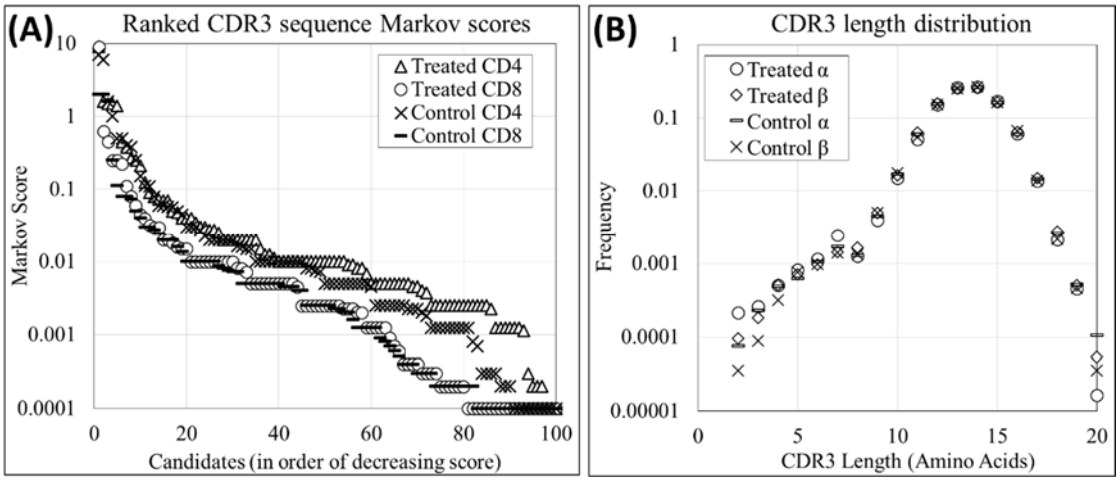


Figure 2.7: Markov scoring and CDR3 length histogram

(**A**) Plotting the ranked Markov scores shows the low frequency of DO11.10 like sequences in the data set and the overall abundance of diversity in CDR3. Additionally the (**B**) length distributions within the data showed no differences between samples or between different genes.

2.4.4: α-chain variant TCRs

While the sequenced repertoire was not expected to produce an identical TCR sequence to the well-studied DO11.10 or other OVA-specific T-cells it was surprising to find no sequences with TRAV5D-4. It was also unexpected to have none of the many TRBV13 sequences close to the DO11.10 β-chain sequence. As such, the data was then analyzed to find similar sequences based on homology found by clustalΩ and an increased frequency in the treated mice. Examining the sequence data for TRAV regions found more frequently in the treated mice than the untreated found the TRAV3-3 and TRAV7-3 genes with increased frequency usage across the other allele in the TCR gene in the treated mice when compared to the average in control mice.

Sequences with increased frequency were then cross-referenced with a clustalΩ analysis of TRAV-regions to measure the similarity to the DO11.10 CDRα3 sequence. Compilations of the CDR3 sequences and J-regions in V-region genes were further

examined to determine the most frequent residues in the CDR3 sequence and J-regions associated with the TRAV3-3 and TRAV7-3 gene sequences in each mouse. These TRAV-region, CDRα3 sequence, and TRAJ-region combinations from the treated mice are reported in **Table 2.3**. The CDRα3 sequences calculated this way had a high degree of similarity to wildtype DO11.10. Since the sequencing data provides information on α- and β-chain frequencies independent of each other this analysis focused on finding a relevant α-chain while holding the DO11.10 β-chain constant.

	V-region	CDR3	J-Region
WT	TRAV 5D4	CAASPNYNVLYF	TRAJ 21
1	TRAV3-3*01	CAVMPNYNVLYF	TRAJ 21*01
2	TRAV3-3*01	CAVSRNNNNRIFF	TRAJ 31*01
3	TRAV7-3*01	CAVSGPNYNVLYF	TRAJ 21*01
4	TRAV7-3*01	CAANNYAQGLTF	TRAJ 26*01

Table 2.3: scTCR variants cloned for expression

V-regions found with higher frequencies in the treated group were cloned into the well-expressing DO11.10 scTCR construct replacing the wildtype α-chain. The CDR3 sequences were additionally compiled from the most frequent CDR3 length and amino acids at particular positions.

With the data set analyzed, we decided to test if these sequence candidates obtained from the challenged mice could be both cloned into protein expression and bacteriophage constructs for further characterization. Since these candidates showed the highest frequency of utilization in the data, they were cloned using a combination of overlap extension PCR and specific V-region amplification from the pooled mRNA. The α-chain in the DO11.10 scTCR variant was exchanged with these 4 variants and cloned into MoPAC24 and pAK400 as a M13 pIII fusion construct for phage display and soluble expression using the high-yield DO11.10 β-chain variant[27,81].

Phage displayed and soluble scTCR variants were tested for binding a clonotypic anti-DO11.10 antibody using an enzyme-linked immunosorbent assay (ELISA), but these assays failed to produce stable scTCR fusions on the phage. Additionally the soluble scTCR variants did not express to a high degree and when purified by periplasmic expression did not yield measurable quantities of the expected protein size from size exclusion chromatography in an FPLC S75 purification (ELISA and FPLC S75 trace data not shown). Based on the low protein yields and the lack of clear antigen specific sequences in the data set we did not perform further characterization and scTCR combinations.

2.5: DISCUSSION AND CONCLUSIONS

From the 460,896 sequences collected across the four mice and two control groups in larger run of this experiment, IMGT identified the V-region, CDR3 sequence, and J-region for 276,508 sequences. Similar to our analysis with the smaller sequencing experiment there were no sequences with a strong DO11.10 similarity. While this experiment does not guarantee recovery of sequences identical to the previous DO11.10 and OVA-specific T-cell work, there was the expectation of some consensus sequences shared in only the treated mice as a public sequence[74]. Additionally sequences both shared and unique to each mouse between the α- and β-chain amplifications and $CD4^+/CD8^+$ sorted T-cells did not contain measurable differences between treatments or cell lineages. Without a larger distinction between the treatment groups in terms of gene region frequency or CDR3 sequence bias, the sequence data may have too many background nonspecific sequences masking the immune repertoire signal.

The results of these experiments show the necessity of additional immunizations in the challenged mice to promote the proliferation of antigen-specific cells. Without these additional immunizations, the total number of responding cells cannot be determined through HTS alone. With only a small fraction of the total T-cells expected to respond to

initial antigen challenges, sequencing depths less than 10^6 sequences per mouse are not able to effectively realize the repertoire, especially from cell samples lacking additional processing with pMHC or activation sorting[82]. A combination of additional immunizations, sorting, and deeper sequencing should overcome the incredibly low frequency of responding specific T-cells over the background repertoire variation within individual mice and treatment groups.

Processing the data to identify more frequently used genes or CDR3 sequences in the treated group found four possible candidate sequences. These four α-chain sequence candidates were fused with our well-behaved DO11.10 β-chain single chain construct for binding assays using phage display techniques. These variants were unstable, as indicated by low soluble expression yields when compared to other scTCRs. Additionally ELISA and western blot methods found the α-chain variants unable to bind the clonotypic DO11.10 antibody. These selected TRAV genes and CDR3 sequences may be incompatible with the particular DO11.10 β-chain as seen in other studies with combinatorial fusions of antibody and TCR chains[83].

Cloning novel TCRs from T-cell pools requires methods to screen libraries of TCR α and β chain combinations for pMHC binding. Since the TCR mRNA in this experiment is pooled from the dissociated T-cells, the repertoire analysis must consider the two α- and β-chain TCR sequence set independent of one another. These independent amplifications prevent the linking of clonotypic α and β chains. This limits our analysis since half the receptor repertoire analyzed is performed independently of the other chain and both are shown to play an important and shared role in pMHC interactions. One method to overcome the independence of each chain population links the two chains using gene assembly with a linker region to generate a random library of scTCR or bicistronic constructs[84]. These gene assemblies from amplified variable domains can produce libraries

of greater than 10^7 bypassing the problem of mismatched TCR chains when paired with a rapid screening method like phage display.

Determination of the four genes (TRAV, TRAJ, TRBV, and TRBJ) and junction sequences (CDRα3 and CDRβ3) for individual T-cells is important for finding the complete diversity and frequency of sequences in the T-cell repertoire. Previously, HTS and single-cell studies have been unable to obtain gene and junction data for both TCR chains from T-cell clones on a high-throughput basis. The single cell studies use RT-PCR techniques to obtain both α- and β-chain sequences using nested PCR reactions. Until recently, these single cell studies were time consuming, labor intensive, and could not scale to the depths necessary for repertoire sequencing.

In the last few years, new developments in microfluidics have dramatically increased the scale of single cell sequencing analysis. Two new HTS methods using different techniques generate isolated single cell reactors for cloning the two immune receptor chains. The first involves sorting and sealing individual cells into 125 pl wells on a printed microscope slide[85]. Thermocycling the slide then amplifies and links the receptor sequences for downstream HTS. This process works on scales of 1.7×10^5 well per slide and can generate sequences for more than 5×10^4 cells per experiment. The second method uses single cell emulsion droplets to isolate cells instead of the pl well slide for downstream gene amplifications and chain linkages[86]. This method reports sequencing rates between 3×10^5 – 1.2×10^6 quality, paired in reads per emulsion reaction. Implementation of these methods with other polymorphic or hypervariable sets of genes can measure natively unlinked diversity in a myriad of new applications. Coupling increased sequencing depths with the ability to sequence both chains of sorted cells would have greatly increased the utility of our experiment, allowing us to accurately pair TCR chains for antigen specific cells above the background repertoire.

In conclusion, supplemental experiments to pursue antigen-specific sequences in mouse TCR repertoires should take additional steps to increase the *in vivo* of OVA-specific T-cells in the mouse by giving the mice an additional immunization one to two weeks after the initial exposure. The use of Illumina technology will also require a redesign of the TCR primer set for the shorter sequencing read lengths used in this process. A more interesting question may be how the repertoire expands and contracts with time before during and after encountering various disease antigens. Characterization of the antigen-responding cells from HTS alone was hampered by non-specific background. A more extensive study of the repertoire either by single cell assays or with antigen stimulation and sorting would more easily identify the antigen specific T-cells.

Chapter 3: Evaluation of canonical loops in TCRs

3.1: CHAPTER SUMMARY

Canonical CDR loops in the majority of TCR and antibody variable regions have well-defined key residues that play an important and conserved role in loop structure. Previous work by Chothia *et al* with seven unique TCR structures predicted the presence of several canonical classes for each CDR loop, based on preexisting antibody canonical loop analysis[23]. Increased interest in analyzing TCR binding kinetics has motivated the ongoing investigation of canonical loop structures in TCRs[23,87,88]. The ability to accurately predict TCR CDR structures will aid in the effort to design functional TCRs against target peptide-MHC complexes and engineer specific immune responses[89,90].

In this study, we expanded Chothia's work and evaluated CDR loops from 249 solved TCR and TCR-like antibody structures deposited in the Protein Data Bank (PDB). Loops were grouped into canonical classes based on a pairwise root mean squared distance (RMSD) of ≤1.5 Å and analyzed the TCR-pMHC binding orientation and interfacial properties. This analysis found classifications for the majority of CDR loops, 70% and 81% in the α- and β-chains respectively and structures for 3 new canonical classes in CDR1 and CDR2. CDR structure prediction from sequence data provides insight into TCR-pMHC interactions. Protein modelers can use sequence data to design mutations to alter the CDR loop structure to change the TCR function. This TCR structural prediction analysis from sequence data has implications protein and immune-receptor engineering. Engineering receptors for affinity and specificity is important for immunotherapies and many biological processes. Antibody and TCR models can use key residues similar to this study to predict the general structural conformation and surface characteristics.

3.2: INTRODUCTION

The adaptive immune system plays a critical role in the recognition and neutralization of diseases with a complex network of chemicals, proteins, receptors, and cells. T-cells in particular are tasked with discerning between self and foreign antigen peptides presented by other cells in a rapid, robust fashion with a clonotypic T-cell receptor (TCR). Successful antigen peptide recognition by the TCR initiates a signaling cascade via a series of stereotyped events including T-cell differentiation, cytokine expression, and targeted cell lysis depending on T-cell type.

TCR/ pMHC function

To generate the level of binding diversity required to target the myriad of foreign peptides in the body, TCRs rely on genetic recombination and imprecise junction formation. The resulting diversity in the structure of CDR3 allows the fine distinction of the numerous self-peptides from the nearly infinite number of foreign antigen peptides indicative of illness. While TCR peptide-MHC (pMHC) interactions are more often considered homologous to antibody-antigen interactions generated in B-cells, several major differences warrant attention. Unlike high affinity soluble-antigen binding antibodies, TCRs engage pMHCs as membrane-bound molecules with a more constrained interaction. B cells further maturate stronger antibody binding affinities by somatic hypermutation for improved function while TCRs undergo no further mutations upon exiting the thymus[91,92]. As opposed to antibody-antigen interactions, the structural and kinetic mechanisms governing TCR/pMHC interactions are less well understood. T-cell maturation involves a coordinated selection of TCRs with a narrow binding affinity range balanced between no binding and cross-reactive aberrant MHC binding. Many TCR-pMHC kinetic studies describe the receptor's specificity as driven by a slow k_{on} and a rapid k_{off}[13,93]. A current hypotheses for high specificity and low affinity of this peculiar protein-

protein interaction results from the plasticity of all the interacting components wherein proper TCR deformation exposes internal co-receptor ITAMs for downstream signaling activation[94].

TCR/ pMHC structure

The surface TCR complex involves a host of accessory proteins (CD3-γ, -δ, -ε, & -ζ, CD4, CD8, etc.) to guide and stabilize pMHC interrogation[95]. As mentioned above the flexibility, low affinity and generally soluble-format instability of the TCR and pMHC has prevented a more thorough analysis and fewer solved unique TCR structures of these immunologically important proteins.[81,96,97] TCRs are structurally homologous to antibodies, with an α/β hetero-dimer, each having a constant and variable domain. The variable domain consists of an immunoglobulin (Ig) β-sheet barrel structure with three CDRs comprising the binding paratope, responsible for pMHC ligand interaction. The third CDR on both chains forms by the gene combinations and imprecise hypervariable junctions generating considerable sequence and structural diversity required to recognize the vast array of potential antigen peptide. MHC molecules are polygenic and highly polymorphic hetero-dimers with two structurally conserved α-helices atop a β-sheet "floor" forming a peptide binding groove to accommodate a myriad of short peptides (8-17 amino acids in length), lipids, and other small molecules depending on the MHC class. The polymorphic MHC molecules require a custom repertoire of TCR αβ-chain combinations and CDR3 sequences to engage this antigen presenting protein.

Biochemical studies suggest TCR-pMHCs interactions occur as a two-step process, with αβ CDR1 and CDR2 first engaging the MHC α-helices peripheral to the CDR3s and antigen peptide to stabilize the recruitment of accessory receptors and TCR deformation interrogation[95,98–100]. Many TCR-pMHC structures show a mostly conserved and concerted

diagonal engagement with the binding groove forming a pocket for the CDR3s to directly interrogate the peptide, giving the CDRs specific and stable binding[99,101].

Canonical Structures in Antibodies and TCRs

Previous work by Chothia *et al* identified, in both antibodies and TCRs, conserved Ig CDR residues predicted to make canonical structural conformations underscoring the balance in the variable domain's conserved framework structure and variable CDR loops[23,80,88,102,103]. While the variable genes in the entire Ig repertoire contain considerable sequence diversity, the majority contain key canonical loop forming residues in CDR1 and 2[88,104]. These key residues and subsequent canonical classes were initially identified by structural comparison and sequence alignment of antibody CDR loops and flanking frameworks. The hypervariable CDR3 loops, which vary significantly in length and sequence, have so far defied structural classification and prediction due to the dramatic structural deformations between bound and unbound states[105]. This early work identified multiple classes for each CDR1 and 2 loops, as well as key residues in the majority of variable gene regions. Ward *et al* later solved the structure for the predicted α1-4 canonical class, from an isolated α-chain homodimer with shorter CDR than observed in most other TCRs[24]. The utility of canonical classes, apart from describing structural diversity and predicting specificity-determining residues, allows engineering of CDR grafts on chimeric TCRs and the development of more accurate algorithms to predict CDR structure, residue interactions, and hypothesize specificity and function from sequence data[87,103,106,107].

In the years since the Chothia *et al* study, the TCR structural database has expanded significantly from the seven initial structures to over 200 today (**Figure 3.1**). TCR structural analyses has proceeded more slowly than antibodies due to the lack of TCR diversity in the solved structures with less than 100 novel TCRs defined on IMGT and PDB. This stems from difficulties in the production of soluble, recombinant TCR protein,

as TCRs have evolved for stable expression at the outer membrane surface. Recent advances in recombinant TCR expression technology resulted in an increase in the available TCR structures, providing a larger data set for a statistically significant analysis of TCR canonical classes. TCRs are structurally analogous to antibodies, but functionally distinct immunological receptors, and are more difficult to study biochemically.

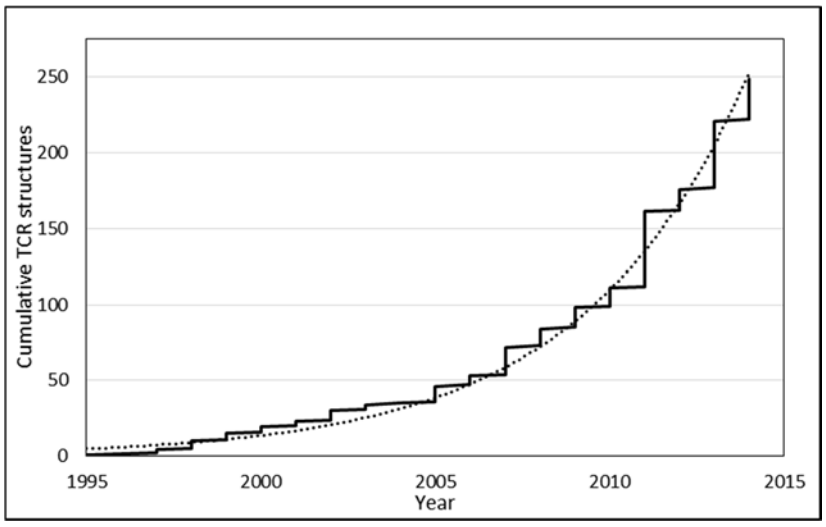


Figure 3.1: Published TCR structures in the PDB over time.

Starting from the initial discovery of thymocytes in 1985 the number of solved TCR structures remains only a fraction of the total antibody structures due in part to the instability of soluble receptors and the low affinity interactions with pMHC which can make complexation difficult.

For example, characterization of the 179 pMHC-bound TCRs have yet to elucidate a clear role for the conserved diagonal binding footprint and whether the orientation is forced by the co-accessory proteins in the immune complex or if the angle is a genetically encoded "interaction codon" or opposing charged residues key to regulating TCR-PMHC specificity[100,108]. The conserved interrogation of the MHC by CDR1 and CDR2 may cause the high amount of structural similarity between the various TCRs regardless of sequence.

3.3: STRUCTURAL ANALYSIS METHODS

TCR data set

In this study, we analyzed the prevalence of canonical classes in the available αβ, γδ, and TCR-like Ig structures, consisting of 249 PDB structures and 64 uniquely paired TRAV/TRBV TCR structures. Discussion of α- and β-chain properties applies to the respective chains of γδ TCRs and the Light/Heavy chains of antibodies. On March 12, 2015, all 249 available αβ and γδ TCR and TCR-like antibody structures were retrieved from the PDB in their various complex configurations (unbound, pMHC bound, IgG bound, single-chain variant, super-antigen bound, etc.). The CDR locations and sequences within each TCR chain were identified using the IMGT unique numbering system for Ig and T-cell receptors[77]. This work used slightly modified loop selections to give additional structural context to the loops within the Ig structure. CDR1 loops were expanded to include residues from the conserved cysteine at IMGT position 23 to the tryptophan at position 41. CDR3 loops also spanned from the conserved cysteine through to the aromatic -FG or -WG groups on the carboxyl-terminal end of the loop. CDR2 identification was slightly modified from prior rules established for antibody and TCR canonical loops. The variation in the CDR2 locations required additional modifications to the selected definition of CDR2. To overcome the possible variability of the loop location in the chains the CDR2 the loop was isolated using the residues of the C' and C'' β-sheet strand flanking the CDR approximately 20 residues after the C-terminal Trp41 residue of CDR1 and 36 residues before the conserved Cys104 in CDR3 on both chains.

CDR alignment comparison script

A program (**Align.py**) to align and measure the structural similarity between every identified CDR loop was written for pyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.) using a python-based scripting language. For each TCR

structure, the program isolates the CDR main-chain backbone atoms (N, C_α, C, & O) with the CDR identification parameters established above to remove the flanking framework residues. If PDB contained multiple identical TCR structures within the unit cell the program (**Aligned Unit Cells.py**) then compared each loop for its variability measuring the average distance between each backbone atom along the loop backbone. Often, the CDR3 loops varied significantly between identical chains within a unit cells. Additionally if the PDB structure contained atoms with alternate location indicators the program considered the atom listed first in the further analysis. After isolating one of the six sets of CDRs from all the TCRs in this fashion, the script performed a pairwise superposition alignment to calculate the root mean squared difference (RMSD) used to measure the aggregate structural difference between individual backbone atoms. If the alignment function excluded backbone atoms in this way, a further analysis was performed to determine if the exclusion loses integral structural information and the RMSD penalty that should be associated with the exclusion. Similarly this alignment program compared loops between the six regions for any structural similarity, but with lowered atomic exclusion parameters to allow for differences in loop sizes between the mobile and reference objects.

Identification of canonical classes and key residues

With all the RMSD values calculated between every individual CDR loop in a diagonally symmetric pairwise table, another program (**Data Processer.py**) minimized the distance between clustered loops by their mutual RMSD values excluding loops with RMSD values larger than 1 Å to any other clustered loop. Groups lacking a plurality of variable regions were rejected as definitive clusters since sequence and structural identity is not necessarily indicative of a canonical class. Any remaining loop structures not clustered or classified were considered unclassified. In most cases, CDR loops from identical TCRs fit into the same cluster except for a few exceptions for point mutations or

peculiar complexation configurations. To determine and verify the structural role of key residues as defined by Chothia *et al* (2000) the CDR sequences of clustered structures were aligned as seen in Table 2 showing the conserved physiochemical residues. Additionally alignments of the clustered CDRs showed definitively how the key residues had an incredibly low amount of structural variability in both the mainchain backbone and their orientation of the sidechain.

Comparisons between interaction codon

Structural co-variation with combinations of multiple CDRs used a similar alignment script as above but with expanded region selections. These combinations (α1α2, β1β2, α1α2+β1β2, or α1α2α3+β1β2β3) illustrate the conserved and coordinated nature of the CDR loop structures in the V-region sequences since certain key residues and canonical classes co-vary as features in their shared function and pMHC interactions.

TCR-pMHC structural interactions

To analyze any structural trends in the orientation and spatial arrangement of TCR-pMHC complexes CCP4 subroutines AREAIMOL, SC, and CONTACT calculated the buried surface area, shape complementarity, and the atomic contacts respectively. Another script (**Docking.py**) calculated the TCR-pMHC docking angle and spatial orientation between the long axial MHC binding groove and the line between the TCR α and β chains. This binding geometry was found between each TCR chain CDR mainchain atoms reduced to centers of mass and the linear and planar approximations for the peptide and MHC binding groove. The coordinate system centered on the flanking α-helices with the x- and y-axis parallel and perpendicular to the linear approximation of the peptide respectively the positive z-axis extends into the complex of the TCR. While the TCR chain centers of mass may be biased by the size and arrangement of the CDRs it more accurately

representatives the TCR footprint because it is closer to the actual pMHC contacts. This is varies from previous calculations that used the line through the center of the di-sulfide bonds for each chain by differing amounts based on the TCRs orientation in relation to MHC plane[96]. An alternative method for determining the crossing angle described as a residue-specific definitions of the TCR-pMHC binding-angle by identifying complementary opposing charged residues between the TCR CDRs and MHC alpha helices[108]. Surprisingly the separate linear approximations of both the peptide antigen and the α-helices showed little variability as linear vectors.

3.4: CANONICAL STRUCTURE RESULTS

The percentages of classifiable TCRs and their miscellaneous properties are summarized in Table 1. The overall average RMSD values for the 17 classes reported in **Table 3.2A-D** are based on comparisons to the average loop conformation depicted in **Figure 3.2A-D** compiled from all the TCR structures within the class. Within each class, the RMSD of the backbone alpha carbons position varied less than 1.5 Å, with each CDR having 3-6 observed classes[104,109]. Of the total CDR1 and CDR2 loop structures analyzed 76% fit into canonical classes with 82% and 73% for human and mouse TCRs respectively. The percentage of classified pMHC-bound TCR and unbound structures was 77% and 60% respectively with an additional 33 loops in Ig or super-antigen bound complexes classified. Many previously proposed canonical classes based on the seven original structures (PDBs: 1AO7, 1BD2, 1BEC, 1KB5, 1NFD, 1TCR, & 2CKB) and limited number of known human and mouse TCR v-regions had additional structures added in this analysis. Most classified structures maintained the proposed key residues and conformations predicted by Chothia *et al* and Ward *et al* (for class α1-4) with several new conformations found.

Class	Count	Length	TCRs	Sequence	V-region	RMSD	B(%)	U(%)	H(%)	M(%)	PDB
A1-01	63	13	1.D9.B2 (1.6); 172.1 (1.6); 1E6 (3.2); 1F1E8hu (1.6); 1G4 (14.3); 2W20 (1.6); 809.B5 (1.6); A6 (22.2); B7 (1.6); CD8 (1.6); D10 (3.2); DMF5 (3.2); ELS4 (1.6); H-Va12-Vb30 (6.3); H27-14 (4.8); Hy.1B11 (3.2); JKF6 (1.6); KB5-C20 (3.2); M-Va14 (1.6); RA14 (1.6); RL42 (3.2); T36-5 (4.8); clone42 TCR (1.6)	CDYTNSMFDYFLW (1.6), CSFPSSNFYALHW (1.6), CSFTDSAIYNLQW (19.0), CSYEDSTFDYFPW (4.8), CSYEDSTFNYFPW (3.2), CSYENSAFDAFPW (1.6), CSYENSAFDYFPW (11.1), CTYQTSGFNGLFW (3.2), CTYSDRGSQSFFW (31.7), CTYSDSASNYFPW (4.8), CTYSFLGSQSFFW (6.3), CTYSNSAFQYFMW (7.9), CTYSNSASQSFFW (3.2)	H-VA1-2 (3.2); H-VA12-1 (3.2); H-VA12-2 (36.5); H-VA12-3 (7.9); H-VA13-1 (4.8); H-VA21 (19.0); H-VA24 (1.6); H-VA29_DV5 (1.6); H-VA39 (1.6); M-VA14-1 (3.2); M-VA14-2 (1.6); M-VA14-3 (4.8); M-VA14D-2 (3.2); M-VA14D-3_DV8 (7.9)	0.31±0.32	76.2	23.8	79.4	20.6	1AO7; 1B88; 1BD2; 1BWM; 1D9K; 1KB5; 1KJ2; 1QRN; 1QSE; 1QSF; 1U3H; 2BNQ; 2BNR; 2BNU; 2F53; 2F54; 2GJ6; 2NX5; 2P1Y; 2P5E; 2P5W; 2PYE; 2PYF; 3C6L; 3D39; 3D3V; 3GSN; 3H9S; 3HG1; 3MFF; 3PL6; 3PWP; 3QDG; 3QDJ; 3QFJ; 3QH3; 3RDT; 3SJV; 3SKN; 3UTP; 3UTS; 3VXQ; 3VXR; 3VXS; 3VXT; 3VXU; 3W0W; 4EUP; 4FTV; 4G8F; 4GRL; 4GRM; 4JFD; 4JFE; 4JFF; 4JFH; 4MAY; 4P23; 4P46; 4P5T; 4X6B; 4X6C; 4X6D
A1-02	32	13	1934.4 (3.1); 2C (21.9); 3A6 (3.1); 42F3 (12.5); AHIII (12.5); B3K506 (3.1); C1-28 (3.1); H-Va8-Vb19 (3.1); HA1.7 (9.4); M-Va6-Vb13 (9.4); NB20 (3.1); YAE62 (6.2)	CKYSYSATPYLFW (40.6), CNYSASGYPALFW (15.6), CNYSSSVPPYLFW (9.4), CNYSYGATPYLFW (6.2), CTYQSTYSPFLFW (12.5), CTYSTTGYPTLFW (6.2), CTYSVSGNPYLFW (3.1), CTYTATGYPSLFW (6.2)	H-VA3 (3.1); H-VA8-3 (6.2); H-VA8-4 (9.4); H-VA9-2 (6.2); M-VA12D-2 (12.5); M-VA6-7_DV9 (3.1); M-VA6D-3 (6.2); M-VA6D-7 (12.5); M-VA9-4 (25.0); M-VA9D-3 (12.5); M-VA9D3 (3.1)	0.38±0.60	81.2	18.8	25	75	1AC6; 1FYT; 1G6R; 1I9E; 1J8H; 1LP9; 1MWA; 1TCR; 1ZGL; 2CKB; 2J8U; 2JCC; 2OI9; 2PXY; 2UWE; 2Z31; 2Z35; 3C5Z; 3C60; 3E2H; 3REV; 3RGV; 3TF7; 3TFK; 3TJH; 3TPU; 3VXM; 4GKZ; 4H1L; 4MS8; 4NHU; 4QRP
A1-03a	12	14	9C2 (16.7); BM3.3 (25.0); C12C (16.7); CA5 (8.3); CF34 (8.3); MICA (8.3); SB27 (16.7)	CLYETSWWSYYIFW (25.0), CTYDTSDPSYGLFW (8.3), CTYDTSDQSYGLFW (16.7), CVYETQDSSYFLFW (25.0), CVYETRDTTYYLFW (25.0)	H-VA14_DV4 (25.0); H-VA19 (25.0); H-VD1 (16.7); H-VG4 (8.3); M-VA16D_DV11 (25.0)	0.05±0.03	83.3	16.7	75	25	1FO0; 1NAM; 2AK4; 2OL3; 3FFC; 3KXF; 3OMZ; 4G8G; 4G9F; 4JRX; 4LFH; 4LHU
A1-03b	9	14	6218 (11.1); DM1 (22.2); KK50.4 (11.1); LC13 (33.3); MS2-3C8 (22.2)	CNHSTISGNEYVYW (33.3), CNHSTISGTDYIHW (55.6), CSHATISGNEYIYW (11.1)	H-VA26-1 (33.3); H-VA26-2 (55.6); M-VA21_DV12 (11.1)	0.82±0.14	88.9	11.1	88.9	11.1	1MI5; 2ESV; 3DX9; 3DXA; 3KPR; 3KPS; 3O6F; 3PQY; 3T0E
A1-04	25	12	21.3 (4.0); 226 (8.0); 2B4 (8.0); 5E (4.0); 5c.c7 (4.0); DMF4 (8.0); JM22 (28.0); M-Va10-Vb13 (4.0); M-Va4 (Av11S5) (4.0); SB47 (4.0)	CNFSDSVNNLQW (12.0), CNFSTTMKSVQW (4.0), CNFTTTMRAVQW (32.0), CNFTTTMRSVQW (4.0), CNSSSVFSSLQW (32.0), CNYSTTSDRLYW (4.0), CTSSSIFNTWLW (8.0), CTYSTTLNSMQW (4.0)	H-VA22 (12.0); H-VA27 (32.0); H-VA35 (8.0); H-VA39 (4.0); M-VA13D-3 (4.0); M-VA4-2 (4.0); M-VA4-4_DV10 (4.0); M-VA4D-4 (32.0)	0.05±0.04	76	24	56	44	1H5B; 1OGA; 2CDF; 2VLJ; 2VLK; 2VLM; 2VLR; 2XN9; 2XNA; 3MBE; 3QDM; 3QEQ; 3QIB; 3QIU; 3QIW; 3QJF; 3QJH; 3RUG; 4C56; 4E41; 4E42; 4JRY; 4P2O; 4P2Q; 4P2R
A1-new	61	13	2A3-D (1.6); C20:2 (3.3); ELS4 (1.6); H-Va1-Vb6 (3.3); H-Va10-Vb25 (4.9); H-Va11-M-Vb13 (1.6); M-Va11-Vb12 (1.6); M-Va11-Vb13 (13.1); M-Va11-Vb19 (3.3); M-Va11-Vb29 (1.6); M-Va14-Vb6 (4.9); MAIT (8.2); NKT15 (3.3); SMC124 (14.8); TK3 (4.9); clone18 TCR (1.6)	CNYSVTPDNHLRW (42.6), CNYTVSPFSNLRW (11.5), CSYTVSGLRGLFW (9.8), CTYQTSGFNGLFW (36.1)	H-VA1-2 (36.1); H-VA10 (11.5); H-VA11 (3.3); H-VA20 (9.8); M-VA11 (39.3)	0.59±0.67	93.4	6.6	60.7	39.3	2NW2; 2PO6; 2Q86; 3ARB; 3ARD; 3ARE; 3ARF; 3ARG; 3HE6; 3HE7; 3HUJ; 3MV7; 3MV8; 3MV9; 3O8X; 3O9W; 3QI9; 3QUX; 3QUY; 3QUZ; 3RTQ; 3RZC; 3SCM; 3SDA; 3SDC; 3SDD; 3SDX; 3TA3; 3TN0; 3TVM; 3TYF; 3TZV; 3VWJ; 3VWK; 4APQ; 4G8E; 4IIQ; 4IRJ; 4IRS; 4L4T; 4L4V; 4L8S; 4L9L; 4LCC; 4LCW; 4PJ5; 4PJ7; 4PJ8; 4PJ9; 4PJA; 4PJB; 4PJC; 4PJD; 4PJE; 4PJF; 4PJH; 4PJI; 4PJX; 4PRH; 4PRI; 4PRP

Table 3.1A: Summary table for CDRα1

Class	Count	Length	TCRs	Sequence	V-region	RMSD	B(%)	U(%)	H(%)	M(%)	PDB
A2-01	27	14	1E6 (3.7); A6 (40.7); CD8 (3.7); DMF5 (11.1); H-Va12-Vb30 (14.8); JKF6 (3.7); T36-5 (11.1)	LIMFIYSNGDKEDG (14.8), LIMFTYREGDKEDG (14.8), LIMSIYSNGDKEDG (55.6), LLMYTYSSGNKEDG (14.8)	H-VA12-2 (81.5); H-VA12-3 (14.8); H-VA39 (3.7)	1.35±0.66	77.8	22.2	100	0	1QRN; 1QSE; 2GJ6; 3D39; 3D3V; 3H9S; 3HG1; 3PWP; 3QDG; 3QDJ; 3QEU; 3QFJ; 3QH3; 3UTT; 3VXT; 3VXU; 3W0W; 4EUP; 4FTV; 4GRM; 4JFD; 4JFE; 4JFF; 4JFH; 4X6B; 4X6C; 4X6D
A2-2a	26	14	2A3-D (3.8); H-Va11-M-Vb13 (3.8); M-Va11-Vb12 (3.8); M-Va11-Vb13 (30.8); M-Va11-Vb19 (7.7); M-Va11-Vb29 (3.8); M-Va14-Vb6 (11.5); SMC124 (34.6)	LTVLVDQKDKTSNG (100.0)	H-VA11 (7.7); M-VA11 (92.3)	1.48±0.06	96.2	3.8	7.7	92.3	2Q86; 3ARB; 3ARD; 3ARE; 3ARF; 3ARG; 3HE6; 3HE7; 3O8X; 3O9W; 3QI9; 3QUX; 3QUY; 3QUZ; 3RTQ; 3RZC; 3SCM; 3SDA; 3SDC; 3SDD; 3TA3; 3TN0; 3TVM; 4APQ; 4IRJ; 4IRS
A2-2b	8	14	1934.4 (12.5); B3K506 (12.5); M-Va6-Vb13 (37.5); YAE62 (25.0)	LFRASRDKEKGSSR (62.5), LLKATKADDKGSNK (12.5), LLQVTTANNKGSSR (25.0)	H-VA9-2 (12.5); M-VA6-7_DV9 (12.5); M-VA6D-3 (25.0); M-VA6D-7 (50.0)	0.55±0.01	75	25	12.5	87.5	1AC6; 2PXY; 2Z31; 2Z35; 3C5Z; 3C60; 3RGV; 4QRP
A2-2c	2	14	RA14 (50.0); SB47 (50.0)	LFVLLSNGAVKQEG (50.0), LFVMTLNGDEKKKG (50.0)	H-VA24 (50.0); H-VA39 (50.0)	1.33±0.04	100	0	100	0	3GSN; 4JRY
A2-3a	19	14	1G4 (36.8); 5F (5.3); C20:2 (10.5); H-Va10-Vb25 (15.8); H-Va17-Vb25 (5.3); H27-14 (15.8); NKT15 (10.5)	LILIRSNEREKHSG (10.5), LLLIPFWQREQTSG (5.3), LLLIQSSQREQTSG (47.4), LTIMTFSENTKSNG (36.8)	H-VA10 (36.8); H-VA17 (10.5); H-VA21 (52.6)	1.14±0.26	73.7	26.3	100	0	2BNQ; 2BNR; 2BNU; 2CDG; 2F53; 2F54; 2PO6; 2PYE; 2PYF; 3HUJ; 3SDX; 3TYF; 3TZV; 3VWJ; 3VWK; 3VXQ; 3VXR; 3VXS; 4EN3
A2-3b	10	14	1.D9.B2 (10.0); 172.1 (10.0); 1F1E8hu (10.0); 809.B5 (10.0); B7 (10.0); KB5-C20 (20.0); M-Va14 (10.0); OB (10.0)	LIAIRSVSDKKEDG (20.0), LILIRSNEREKHSG (20.0), LISILSVSDKKEDG (20.0), LISILSVSNKKEDG (10.0), LISIRSVSDKKEDG (20.0), LISISSIKDKNADG (10.0)	H-VA17 (20.0); H-VA29_DV5 (10.0); M-VA14-1 (20.0); M-VA14-3 (30.0); M-VA14D-3_DV8 (20.0)	0.53±0.24	70	30	30	70	1B88; 1BD2; 1KB5; 1KJ2; 1U3H; 2P1Y; 2WBJ; 3MFF; 3RDT; 4MJI
A2-4a	10	14	21.3 (10.0); 226 (20.0); 2B4 (20.0); 5c.c7 (10.0); E8 (30.0); M-Va4 (Av11S5) (10.0)	NLFYIPSGTKQNGR (30.0), NLFYLASGTKENGR (50.0), SLFYLASGTKENGR (10.0), TLFYLAQGTKENGR (10.0)	H-VA22 (30.0); M-VA4-2 (10.0); M-VA4-4_DV10 (10.0); M-VA4D-4 (50.0)	1.20±0.31	60	40	30	70	1H5B; 2IAL; 2IAM; 2IAN; 3MBE; 3QIB; 3QIU; 3QIW; 3QJF; 3QJH
A2-4b	7	14	DM1 (28.6); KK50.4 (14.3); LC13 (42.9); MS2-3C8 (14.3)	YIIHGLKNNETNEM (42.9), YVIHGLTSNVNNRM (57.1)	H-VA26-1 (42.9); H-VA26-2 (57.1)	0.16±0.10	85.7	14.3	100	0	1MI5; 2ESV; 3DX9; 3DXA; 3KPR; 3KPS; 3T0E
A2-new1	15	14	2C (60.0); 42F3 (20.0); C1-28 (6.7); H-Va8-Vb19 (6.7)	LLKYFSGDTLVQGI (13.3), LLKYYSGDPVVQGV (86.7)	H-VA8-3 (13.3); M-VA9-4 (66.7); M-VA9D-3 (20.0)	0.95±0.14	86.7	13.3	13.3	86.7	1G6R; 1I9E; 1MWA; 1TCR; 2CKB; 2E7L; 2ICW; 2OI9; 3E2H; 3TF7; 3TJH; 3TPU; 3VXM; 4H1L; 4NHU
A2-new2	3	14	9C2 (66.7); MICA (33.3)	FLIRQGSDEQNAKS (100.0)	H-VD1 (66.7); H-VG4 (33.3)	0.05±0.02	33.3	66.7	100	0	3OMZ; 4LFH; 4LHU

Table 3.1B: Summary table for CDRα2

Class	Count	Length	TCRs	Sequence	V-region	RMSD	B(%)	U(%)	H(%)	M(%)	PDB
B1-01	165	12	1.D9.B2 (0.6); 14.3.D (3.0); 172.1 (0.6); 1F1E8hu (0.6); 1G4 (5.5); 21.3 (0.6); 2A3-D (0.6); 2C (4.8); 2W20 (0.6); 42F3 (2.4); 5E (0.6); 5F (0.6); 809.B5 (0.6); A6 (7.9); AHIII (2.4); B3K506 (0.6); B7 (0.6); C1-28 (0.6); C12C (1.2); C20:2 (1.2); CA5 (0.6); D10 (0.6); DMF5 (1.8); E8 (1.8); ELS4 (1.2); H-Va1-Vb6 (1.2); H-Va10-Vb25 (1.8); H-Va11-M-Vb13 (0.6); H-Va17-Vb25 (0.6); H-Va8-Vb19 (0.6); HA1.7 (1.8); JKF6 (0.6); JM22 (4.2); M-Va10-Vb13 (0.6); M-Va11-Vb12 (0.6); M-Va11-Vb13 (4.8); M-Va11-Vb19 (1.2); M-Va11-Vb29 (0.6); M-Va14-Vb6 (1.8); M-Va6-Vb13 (1.8); M-Vb13 (0.6); MAIT (3.0); NB20 (0.6); NKT15 (1.2); RA14 (0.6); RL42 (1.2); S54N (0.6); SB27 (1.2); SF4 (0.6); SMC124 (5.5); T36-5 (1.8); YAE62 (1.2); clone18 TCR (0.6); clone42 TCR (0.6); iNKT (0.6)	CAQDMNHEYMSW (17.0), CAQDMNHEYMYW (4.8), CAQDMNHNSMYW (12.7), CDQTNNHNNMYW (0.6), CEQHMGHRAMYW (0.6), CEQNLNHDAMYW (6.1), CGQDMSHETMYW (1.2), CHQTENHRYMYW (1.2), CHQTNNHDYMYW (4.2), CNQTNNHNNMYW (24.2), CQQNFNHDTMYW (3.6), CQQTNNHNNMYW (4.2), CRQTNSHNYMYW (3.0), CSQNMNHEYMSW (1.8), CSQTMGHDKMYW (6.7), CTQDMNHNYMYW (2.4), CTQDMRHNAMYW (3.0), CVQDMDHENMFW (2.4)	H-VB10-3 (1.2); H-VB19 (6.7); H-VB25-1 (6.7); H-VB27 (1.8); H-VB28 (2.4); H-VB4-1 (0.6); H-VB5-1 (1.2); H-VB6-1 (12.7); H-VB6-2 (1.8); H-VB6-3 (3.0); H-VB6-4 (3.0); H-VB6-5 (15.8); H-VB6-6 (2.4); M-VB12-1 (1.2); M-VB13-1 (1.8); M-VB13-1* (0.6); M-VB13-2 (27.9); M-VB13-3 (4.2); M-VB19 (3.0); M-VB29 (1.2); R-VB5 (0.6)	0.07±0.13	80	20	60	40	1AO7; 1BD2; 1BEC; 1D9K; 1FYT; 1G6R; 1J8H; 1JCK; 1L0X; 1L0Y; 1LP9; 1MWA; 1OGA; 1QRN; 1QSE; 1QSF; 1SBB; 1TCR; 1U3H; 2AK4; 2APB; 2AXJ; 2BNQ; 2BNR; 2BNU; 2CDE; 2CDF; 2CDG; 2CKB; 2E7L; 2F53; 2F54; 2GJ6; 2IAL; 2IAM; 2IAN; 2ICW; 2J8U; 2JCC; 2NW2; 2NX5; 2OI9; 2P1Y; 2P5E; 2P5W; 2PO6; 2PXY; 2PYE; 2PYF; 2Q86; 2UWE; 2VLJ; 2VLK; 2VLM; 2VLR; 2XN9; 2XNA; 2Z31; 2Z35; 3ARB; 3ARD; 3ARE; 3ARF; 3ARG; 3C5Z; 3C60; 3C6L; 3D39; 3D3V; 3E2H; 3GSN; 3H9S; 3HE6; 3HE7; 3HUJ; 3KXF; 3MBE; 3MC0; 3MFF; 3O8X; 3O9W; 3PWP; 3QDG; 3QDJ; 3QEU; 3QFJ; 3QH3; 3QI9; 3QUX; 3QUY; 3QUZ; 3RDT; 3REV; 3RGV; 3RTQ; 3RUG; 3RZC; 3SCM; 3SDA; 3SDC; 3SDD; 3SDX; 3SJV; 3SKN; 3TA3; 3TF7; 3TFK; 3TJH; 3TN0; 3TPU; 3TVM; 3TYF; 3TZV; 3VWJ; 3VWK; 3VXM; 3VXT; 3VXU; 3W0W; 4APQ; 4C56; 4EN3; 4EUP; 4FTV; 4G8E; 4G8F; 4G8G; 4G9F; 4GKZ; 4GRM; 4H1L; 4IIQ; 4IRJ; 4IRS; 4JRX; 4L4T; 4L4V; 4L8S; 4L9L; 4LCC; 4LCW; 4MNQ; 4MS8; 4NHU; 4NQC; 4ONH; 4P23; 4P46; 4P5T; 4PJ5; 4PJ7; 4PJ9; 4PJA; 4PJB; 4PJC; 4PJD; 4PJE; 4PJF; 4PJG; 4PJH; 4PJI; 4PJX; 4X6B; 4X6C; 4X6D
B1-02	42	12	1E6 (7.1); 226 (4.8); 2B4 (4.8); 3A6 (2.4); 5c.c7 (2.4); A6 (2.4); CF34 (2.4); DM1 (4.8); H27-14 (7.1); Hy.1B11 (4.8); Hy19.3 (7.1); KK50.4 (2.4); LC13 (9.5); N15 (4.8); SB47 (2.4); TK3 (7.1)	CDPISEHNRLYW (11.9), CDPISGHDNLYW (2.4), CDPISGHTALYW (9.5), CDPISGHVSLFW (9.5), CDPVSNHLYFYW (7.1), CIPEKGHPVVFW (19.0), CIPISGHSNVVW (4.8), CKPISGHDYLFW (7.1), CNPISGHATLYW (4.8), CSPISGHRSVSW (2.4), CSPISGHTSVYW (4.8), CSPKSGHDTVSW (2.4), CSPRSGDLSVYW (14.3)	H-VB11-2 (4.8); H-VB12-4 (7.1); H-VB14 (2.4); H-VB5-1 (2.4); H-VB5-6 (2.4); H-VB5-8 (4.8); H-VB7-3 (9.5); H-VB7-8 (9.5); H-VB7-9 (11.9); H-VB9 (14.3); M-VB12-1 (4.8); M-VB26 (19.0); M-VB3 (7.1)	0.66±0.33	78.6	21.4	69	31	1KGC; 1MI5; 1NFD; 1ZGL; 2ESV; 3DX9; 3DXA; 3FFC; 3KPR; 3KPS; 3MV7; 3MV8; 3MV9; 3PL6; 3Q5Y; 3QIB; 3QIU; 3QIW; 3QJF; 3QJH; 3UTP; 3UTS; 3UTT; 3VXQ; 3VXR; 3VXS; 4E41; 4E42; 4EI5; 4ELK; 4ELM; 4GRL; 4JRY; 4MAY; 4MJI; 4P2O; 4P2Q; 4P2R; 4PRH; 4PRI; 4PRP; 4QRP
B1-new1	4	14	ES204 (25.0)	CLYETSWWSYYIFW (75.0), CTYDTVYSNPDLFW (25.0)	H-VD1 (75.0); H-VD3 (25.0)	0.01±0.001	25	75	100	0	1TVD; 4MNG; 4MNH; 4NDM
B1-new2	4	15	G115 (25.0); TCR-like IgG (75.0)	CAASGFTFDDYAMHW (25.0), CKASGYSFTNYGMHW (25.0), CKASGYTFTDYNMDW (25.0), CSMKGEAIGNYYINW (25.0)	H-VD2 (25.0); IgGH-IGHJ9-1 (25.0); IgGH-IGHV1-8 (25.0); IgGH-IGHV3-9 (25.0)	0.91±0.19	50	50	100	75	1HXM; 1W72; 3CVI; 3UBX

Table 3.1C: Summary table for CDRβ1

Class	Count	Length	TCRs	Sequence	V-region	RMSD	B(%)	U(%)	H(%)	M(%)	PDB
B2-01	159	14	1.D9.B2 (0.6); 14.3.D (3.1); 172.1 (0.6); 1F1E8hu (0.6); 1G4 (5.7); 21.3 (0.6); 2A3-D (0.6); 2C (5.0); 2W20 (0.6); 42F3 (2.5); 5E (0.6); 5F (0.6); 809.B5 (0.6); A6 (8.2); AHIII (2.5); B3K506 (0.6); B7 (0.6); C12C (1.3); C20:2 (1.3); CA5 (0.6); D10 (0.6); DMF4 (1.3); DMF5 (1.9); E8 (1.9); ELS4 (1.3); H-Va1-Vb6 (1.3); H-Va10-Vb25 (1.3); H-Va11-M-Vb13 (0.6); H-Va17-Vb25 (0.6); H-Va8-Vb19 (0.6); HA1.7 (1.9); JKF6 (0.6); JM22 (4.4); M-Va10-Vb13 (0.6); M-Va11-Vb12 (0.6); M-Va11-Vb13 (5.0); M-Va11-Vb19 (1.3); M-Va11-Vb29 (0.6); M-Va14-Vb6 (1.9); M-Va6-Vb13 (1.9); M-Vb13 (0.6); MAIT (3.1); NB20 (0.6); NKT15 (1.3); RA14 (0.6); S54N (0.6); SB27 (1.3); SF4 (1.3); SMC124 (5.7); YAE62 (1.3); clone42 TCR (0.6); iNKT (0.6)	LIHYSNTAGTTGKG (3.1), LIHYSVAIQTTDQG (0.6), LIHYSVGAGITDQG (13.8), LIHYSVGAGTTDQG (1.3), LIHYSVGEGTTAKG (3.1), LIHYSVSVGMTDQG (1.3), LIHYSYGAGNLQIG (2.5), LIHYSYGAGNTEKG (0.6), LIHYSYGAGSTEKG (25.2), LIHYSYGVKDTDKG (2.5), LIHYSYGVNSTEKG (6.3), LIHYSYVADSTEKG (4.4), LIYFSYDVKMKEKG (2.5), LIYISYDVDSNSEG (1.3), LIYYSASEGTTDKG (13.2), LIYYSQIVNDFQKG (6.9), LIYYSVGAGITDKG (2.5), LIYYSYGAGNLQIG (0.6), LIYYSYGAGSTEKG (8.2)	H-VB10-3 (2.5); H-VB19 (7.5); H-VB25-1 (6.3); H-VB28 (2.5); H-VB5-1 (1.3); H-VB6-1 (13.2); H-VB6-2 (1.9); H-VB6-3 (1.3); H-VB6-4 (3.1); H-VB6-5 (15.7); H-VB6-6 (2.5); M-VB12-1 (1.3); M-VB13-1 (1.9); M-VB13-1* (0.6); M-VB13-2 (28.9); M-VB13-3 (4.4); M-VB19 (3.1); M-VB29 (1.3); R-VB5 (0.6)	0.94±0.47	81.1	18.9	59.1	41.5	1AO7; 1BD2; 1BEC; 1D9K; 1FYT; 1G6R; 1J8H; 1JCK; 1L0X; 1L0Y; 1LP9; 1MWA; 1OGA; 1QRN; 1QSE; 1QSF; 1SBB; 1TCR; 1U3H; 2AK4; 2APB; 2AXH; 2AXJ; 2BNQ; 2BNR; 2BNU; 2CDE; 2CDF; 2CDG; 2CKB; 2E7L; 2F53; 2F54; 2GJ6; 2IAL; 2IAM; 2IAN; 2ICW; 2J8U; 2JCC; 2NW2; 2NX5; 2OI9; 2P1Y; 2P5E; 2P5W; 2PO6; 2PXY; 2PYE; 2PYF; 2Q86; 2UWE; 2VLJ; 2VLK; 2VLM; 2VLR; 2XN9; 2XNA; 2Z31; 2Z35; 3ARB; 3ARD; 3ARE; 3ARF; 3ARG; 3C5Z; 3C60; 3C6L; 3D39; 3D3V; 3E2H; 3GSN; 3H9S; 3HE6; 3HE7; 3HUJ; 3KXF; 3MBE; 3MC0; 3MFF; 3O8X; 3O9W; 3PWP; 3QDG; 3QDJ; 3QDM; 3QEQ; 3QEU; 3QFJ; 3QH3; 3QI9; 3QUX; 3QUY; 3QUZ; 3RDT; 3REV; 3RGV; 3RTQ; 3RUG; 3RZC; 3SCM; 3SDA; 3SDC; 3SDD; 3SDX; 3TA3; 3TF7; 3TFK; 3TJH; 3TN0; 3TPU; 3TVM; 3TZV; 3VWJ; 3VWK; 4APQ; 4C56; 4EN3; 4EUP; 4FTV; 4G8F; 4G8G; 4G9F; 4GKZ; 4GRM; 4H1L; 4IIQ; 4IRJ; 4IRS; 4JRX; 4L4T; 4L4V; 4L8S; 4L9L; 4LCC; 4LCW; 4MS8; 4NHU; 4NQC; 4ONH; 4P23; 4P46; 4P5T; 4PJ5; 4PJ7; 4PJ9; 4PJA; 4PJB; 4PJC; 4PJD; 4PJE; 4PJF; 4PJG; 4PJH; 4PJI; 4PJX; 4X6B; 4X6C; 4X6D
B2-2a	16	14	1E6 (18.8); 6218 (6.2); A6 (6.2); CF34 (6.2); Hy.1B11 (12.5); KK50.4 (6.2); LC13 (25.0); N30 (6.2)	LIQFQNNGVVDDSQ (12.5), LIYFNNNVPIDDSG (18.8), LIYFQGTGAADDSG (25.0), LLHFVKESKQDESG (6.2), LTYFQNEAQLDKSG (31.2), LVYFRDEAVIDNSQ (6.2)	H-VB11-2 (12.5); H-VB12-4 (18.8); H-VB14 (6.2); H-VB7-3 (25.0); H-VB7-8 (31.2); M-VB14 (6.2)	0.92±0.44	75	25	93.8	6.2	1KGC; 1MI5; 2ESV; 3FFC; 3KPR; 3KPS; 3OF6; 3PL6; 3Q5T; 3UTP; 3UTS; 3UTT; 4GRL; 4MAY; 4MJI; 4QRP
B2-2b	8	14	SB47 (12.5); TK3 (25.0)	IFQYYEEEERQRGN (12.5), LIHYYNGEERAKGN (12.5), LIQYYNGEERAKGN (50.0), LLWYDEGEERNRGN (25.0)	H-VB5-6 (12.5); H-VB5-8 (25.0); H-VB9 (62.5)	0.68±0.27	85.7	14.3	100	0	3MV7; 3MV8; 4E41; 4E42; 4JRY; 4PRH; 4PRI; 4PRP

Table 3.1D: Summary table for CDR

Summary tables for the canonical classes found in this study including the number of PDB structures in each class, the CDR length, the specific TCRs, CDR residue sequence, variable region, RMSD score to the central average structure, the percent of structures bound/unbound, from humans or mice, and the PDB numbers for each class.

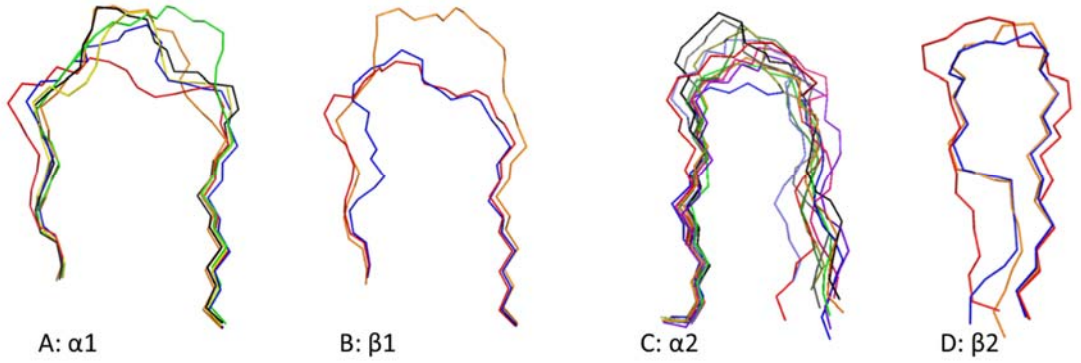


Figure 3. 1: Aligned CDR classes

Class and subclass alignments between the average structures compiled from the classified loop in each CDR show how similar the flanking positions of the loops are in CDR1 and more variable in CDR2. Average alignments for the loops include: (**A**) CDRα1, (**B**) CDRβ1, (**C**) CDRα2, and (**D**) CDRβ2.

The majority of loops in each CDR had similarity with at least one member of a different V-region leaving a minority of unclassified structures with 14%, 45%, 11%, and 23% unclassified for CDR α1, α2, β1, and β2 respectively. These unclassified loops are summarized in **Table 3.1A-D**. With additional solved structures, these distinct structures will be shared additional V-regions and possibly become a canonical class. While α loops had more canonical classes than β loops, the β structures had fewer unclassified structures, which is similar to the diversity of canonical classes found in antibody structures[110]. Due to the very different variable domain arrangements of the β-sheets no canonical classes were shared between antibodies and TCRs. There were 46 loop structures with identical chains in the unit cell, which did not converge on a single conformation. These structures were not included in subsequent analysis.

Previously established and predicted canonical classes with only one TCR type were further scrutinized to determine if the structures are a distinct canonical class. For example, the N15 TCR (PDB: 1NFD & 3Q5Y) α-chain was established as the

conformation for class α1-3 and had two solved structures forming distinctly different loop conformations. Although the N15 structures didn't agree on a similar conformation there were 21 other structures with those proposed key residues which now form the expected structural conformation for this canonical class.

Key residues, CDR length, and framework β-sheet arrangement all play an important role in the stabilization and conformation of the backbone loop in its interaction with other CDRs and the pMHC. The additional solved structures found 3 new canonical classes based on their dissimilarity to other clusters of structures and unique set of V-regions. In the structural comparisons, we noticed the occurrence of slight structural variations from canonical loops that also maintained the key residues. These groups while distinctly different from the canonical class were similar enough to warrant definition as a canonical subclass. The analysis determined a new set of 16 subclasses indicated in summary **Table 3.1A-D**, defined by identical and spatially overlapping key residues with backbone variations between residues. For example, the main conformation for α1-1 class had three mostly similar subclasses with an RMSD less than 1.5 Å.

Structures compared

By generating alignments by CDR sequence, key residues, V-region, or TCR identity, or CDR3 sequence showed the prevalence of certain loop conformation occurring and the slight variations occur between the bound and unbound receptors. As expected, the amount of backbone variation between these alignments increased as the grouped property became less specific. Alignments between identical V-regions and CDR sequences in both bound and unbound configurations showed low mainchain variation as seen in the differences between blue and red loops in figure 2. Coupled with the amount of structural similarity found in the β-sheet barrels comprising the V-region framework indicates a more conserved structure outside the hypervariable CDR3 region on both chains. This is an

interesting since other studies have found a higher degree of paratope plasticity expected from TCRs.[69,111,112]

The majority (72%, 179/249) of the solved TCR structures are bound to MHC molecules, with 83% (148/179) in complex with MHC class-I. TCR derived from either $CD4^+$ or $CD8^+$ T-cells have been shown to bind the pMHC differently between what MHC α-helix residues are contacted and the TCR affinity based on orientation over the pMHC and the lineage's function in the immune system. Several studies have noted the TCR affinity differences in $CD4^+$ and $CD8^+$ T-cells due in part to the need for the TCR to either specifically target an infected cell or serially scan for more peptide antigens for the two lines respectively, which may in fact be impacted by the binding of the CD4/CD8 co-receptor to the pMHC[113,114].

Predicted Key Residues

The previously predicted key residues were predominately accurate in nearly all classified structures between a variety of CDR sequences and V-regions. The key residues in many classified structures were functionally conserved, with physiologically similar amino acids, and pointed either into the CDR loop to contact or interlock with one another or towards a neighboring CDR loop. Loops without exactly matching key residues as defined by Chothia, had a similar functioned residue in their place, which did not seem to affect the overall conformation. The key residues predicted for classes TRBV1-5, TRBV2-4, and TRBV2-5 were not found to any structural relevance in this study did not cluster with any V-regions outside identical structures or sequences.

Note on mouse TRBV-13 with other V-region classes

The most prevalent V-region genes used in many TCR structures is mouse TRBV13 found in 62 structures, all of which were in the same CDR1 and 2 classes. The TRBV13 increased representation could be attributed to a more prevalent β-chain rearrangement

with this gene or due to the biased usage in cloning because its stability as a soluble protein in high quantities in *E. coli* or possibly the bias of this V-region found in mouse TCR repertoiress[27].

The α1 hypervariable regions

The CDRα1 domain has five canonical classes incorporating 31 V-regions disregarding the myriad of other polymorphisms. Of the 235 α1 structures analyzed 86% were classified.

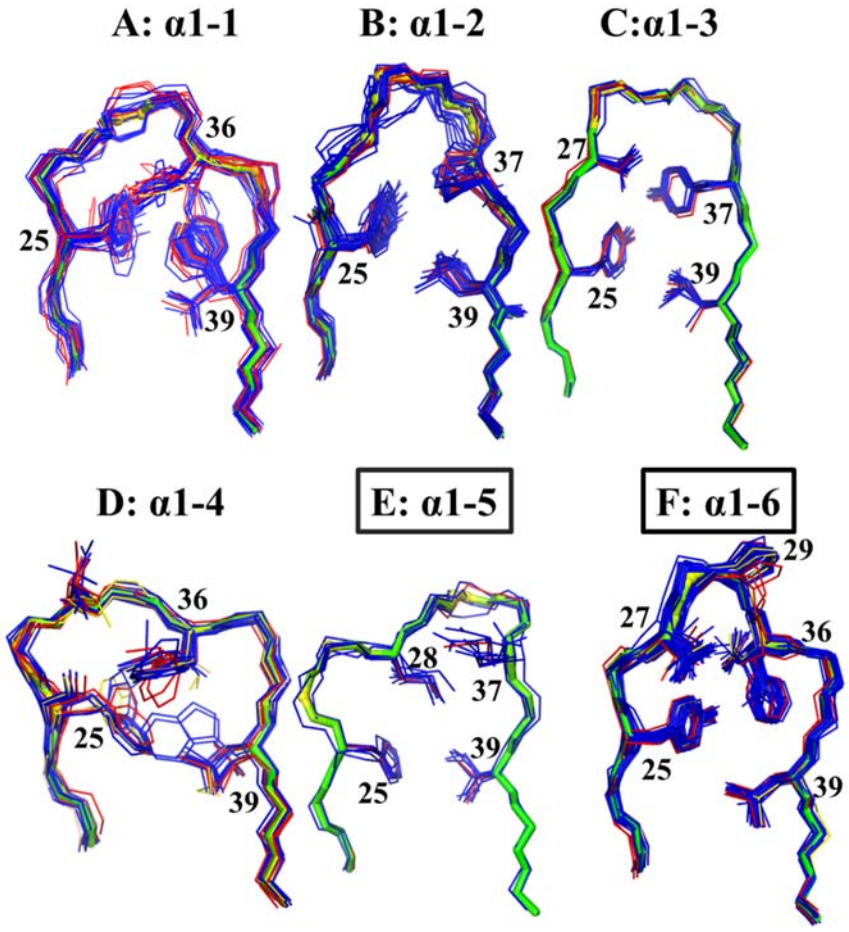


Figure 3. 2A-F: CDRα1 class alignments

Canonical loops with key residues shown for classes (**A**) α1-1, (**B**) α1-2, (**C**) α1-3, (**D**) α1-4, (**E**) α1-5, (**F**) α1-6.

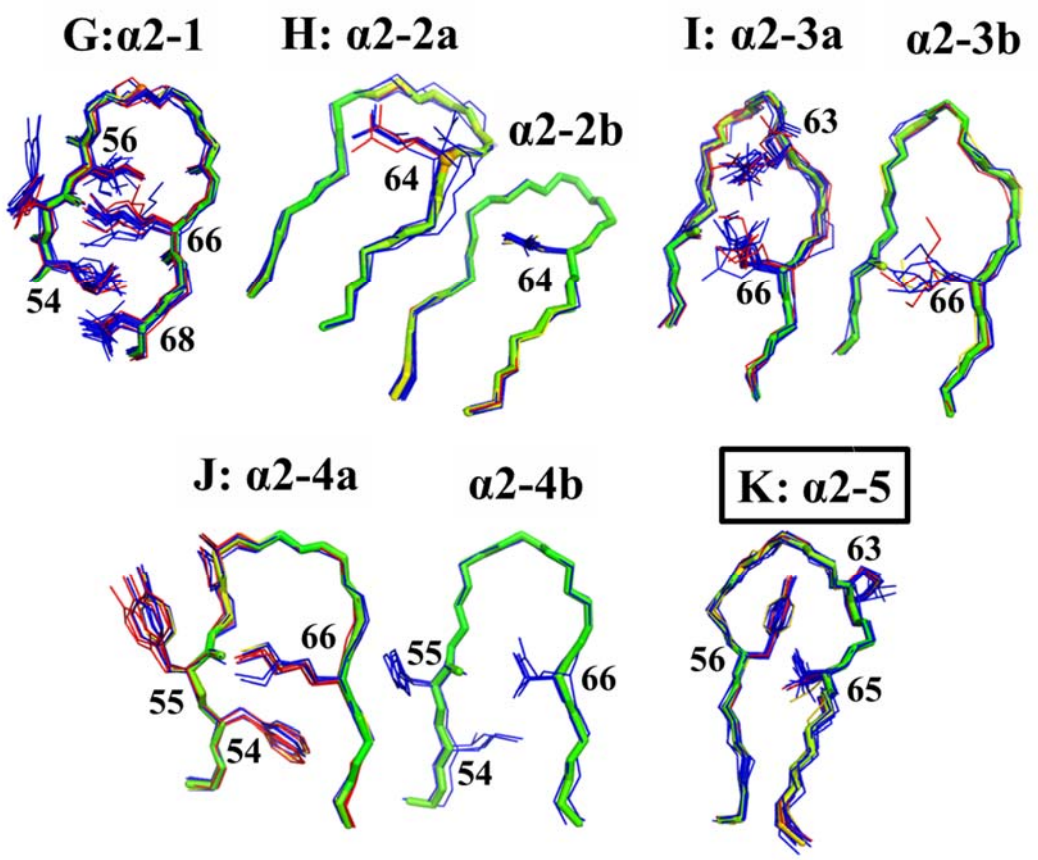


Figure 3.2G-K: CDRα2 class alignments

Canonical loops with key residues shown for classes (**G**) α2-1 (**H**) α2-2 with two subclasses, (**I**) α2-3 with two subclasses, (**J**) α2-4 with three subclasses, and (**K**) α2-5.

α1-1

From the 3 original structures in the class 60 more had the same CDRα1 conformation all within 0.31 Å of the average structure. The α1-1 class has the largest amount of genetic variability of the other α1 classes with 8 TRAV regions used (h-TRAV-1, -12, -13, and -29 and m-TRAV14) TRAJ region usage was also the most variable in this class using 19 gene regions. Residues important for the structure consist of residues 25 and 39, the majority of which are Tyr and Phe respectively. These two residues point inwards of the loop and if both are aromatic they stack. Residue 36 may also have structural

significance with either a Ser or Phe pointing into the loop as seen in **Figure 3.2A** causing an indent in the peptide backbone towards its center at that residue.

α1-2

The original CDRα1 TCRs 1AC6, 1TCR, and 2CKB barely fit into this class with an overall RMSD average of 0.793 Å for these three structures alone. The 29 new TCRs in class α1-2 all had a length of 13 residues with an overall average RMSD of 0.38±0.60 Å. This class contains 3 6TRAV-regions (human TRAV3, -8, & -9 and mouse TRAV6, -9, & -12) and uses 12 TRAJ-regions. The key residues in this class include inwardly pointing Tyr25, Pro37, and Lue39 which interact with one another with an approximate distance of 2.0 and 2.6 Å between the closest atoms of Try25:Pro37 and Try25:Leu39 respectively. The kinked and slightly variable structure in the loop apex highlighted in **Figure 3. 2B** is caused by a combination of Pro37 and it pushing against the side chain at position 27 out of the plane of the loop setting it apart from other canonical classes. Contained in this class are three subclasses which have matching positions at the key residues but very slightly at residue positions 27-36.

α1-3 (With novel subclass recognized as α1-5)

The initially proposed α1-3 conformation using the N15 (PDB: 1NFD) CDRα1 loop did not have any additional structures added in this expanded set of TCRs even though 21 other structures had the prescribed key residues. These other loops also formed two subclass structures with a loop length of 14 residues and an average RMSD value of 0.05±0.03 Å and 0.82±0.14 Å respectively. The N15 in structure 1NFD might be a structural anomaly because it is IgG-bound possibly altering the CDR structures since antibody affinities are much higher than pMHC interactions. This class contains 7 V-regions (human TRAV14, -19, -26, human TRDV1, human TRGV4, mouse TRAV16, and -21) and 11 TRAJ-regions. The subclasses split and for a different configuration at residues

28 and 29 with the human TRAV26 and mouse TRAV21 regions using Ile at residue 29 to interact with the other key residues instead of Thr28 in all the other classified structures. Many of the residues in this group are either charged or polar allowing various intra-loop interactions shown in **Figure 3.2C and E**. The second subclass shown on the right has residue 30 interacting with the nitrogen of residue 27, bringing the top portion of the backbone into a hairpin turn. Variations between the two subclasses occur at residue positions 26 and 28-30 while all the other loop positions and key residues are in nearly identical positions.

α1-4

This class was originally predicted based on the following the 12 residues sequence pattern found in 5 V-region sequences:

C-X-S-S-X-X-[FL]-X-X-[VLIW]-X-W[23]**.**

Ward *et al* then solved the first structure with a similar sequence motif with mouse TRAV4-4/DV10 (PDB: 1H5B). The α1-4 class contains 8 different CDR sequences and 6 TRAV-regions (human TRAV22, -27, -35, -39, mouse TRAV4, and -13) and 11 TRAJ-regions. The overall average RMSD value for this class is 0.05±0.04Å, the lowest of all the classes. The three key residues in this class all strongly associate and cause the distinctly square loop shape seen in the **Figure 3.2D**.

α1-6 (New canonical class)

This new conformation with an average RMSD value of 0.59±0.67Å includes many structures with class α1-1 key residues, but has an incredibly different backbone shape well beyond what could be considered a subclass. There are 4 unique V-regions in this class (human TRAV1, -10, -11, 20, and mouse TRAV11) and 8 TRAJ-regions. Looking at the α1-1 and α1-5 average structure in **Figure 3.2F** for these loops shows a large deformation between residues 27 – 29 and 36 with a maximum distance of approximately 3.4 Å at

residue 28. The key difference of this class to the α1-1 structure is the small residue at position 29 allowing a sharper turn and increasing the gap at that location.

The α2 hypervariable regions

Just over half (54%) of CDRα2 loops fit into a canonical class, with the largest amount of unclassified loops of all the CDRs. This CDR showed the largest amount of structural diversity with higher average RMSD values and the most classes and subclasses likely due to the variability caused by the location of the CDRα2 loop at the far end of the 2 sets of parallel β-sheets and the size of the residues oriented into the loop. In this study the CDRα2 loop consisted of the residues between flanking β-turns on either side of β-sheet strands C' and C''.

α2-1

The α2-1 class includes 27 human TCR loops 12 residues in length and an overall average RMSD of 1.35±0.66 Å. The majority (96 %) of the TCRs utilized human TRAV12 except ELS4 (PDB: 4GRM) which used human TRAV39 and 6 TRAJ-regions. The Lys66 residue points towards the C' β-strand where it pushes between residues 55 or 57. The loop apex flattens at the top with the residue side chains pointing nearly perpendicular (see **Figure 3.2G**) to the plane of the loop, which is interesting since this occurs in both bound and unbound structures in this class.

α2-2

The α2-2 class includes 36 structures in three subclass conformations 13 residues long and higher average RMSD values of 1.48±0.06 Å, 0.55±0.01 Å, and 1.33±0.04 Å respectively. The α2-2 class had 6 TRAV-regions (human TRAV9, -11, -24, -39, mouse TRAV6, and -11) and 8 TRAJ-regions. Alternating, inwardly pointing, medium-sized residues (54, 56, 58, 64, and 66) in this class help give this loop its almost zipper-like shape

in subclass α2-2a. The Lys66 in α2-2a subclass further brings the two sides of the loop together with a hydrogen bond with the carboxyl oxygen on residue 55. The variations between the three classes occur in the different sizes of the residues at these positions, which slightly alter the distance between the flanking ends of the loops shown in **Figure 3.2H**.

α2-3

The α2-3 class includes 29 structures and has 2 nearly identical subclasses which only vary at the C'' strand by a distance of 1 Å. This class has 5 different TRAV-regions (human TRAV10, -17, -21, -29, and mouse TRAV14) and 12 TRAJ-regions with an overall average RMSD of 1.14±0.26 and 0.53±0.24 Å for the two subclass structures respectively. The bulky residues positioned at the apex cause greater deformation of the backbone which allows the charged or polar residue at position 66 to point into the loop to form hydrogen bonds with the backbone carboxyl group of residue 55. This structure is visualized in **Figure 3.2I**.

α2-4

The α2-4 canonical class was proposed by Ward *et al* (S Ward) containing Vα11 (1H5B). Class α2-4 has 17 TCR structures, a length of 11 residues, and comprises two slightly varying subclasses **Figure 3.2J** with overall RMSD averages of 1.20±0.31 and 0.16±0.10 Å respectively. There are 3 TRAV-regions (human TRAV22, -26, and mouse TRAV4) and 14 TRAJ regions. This class is structurally similar to α2-3 class with residue 66 pointing toward the backbone of residue 55 forming a bond if the residue is long enough as in the case of a Lys66 in mouse TRAV4. Residues at positions 54, 56, 66, and 68 alternately almost stack beside one another to help give the loop a planar shape.

α2-5 (New canonical class)

This new class conformation found in 15 structures was the only new unique conformation found in CDRα2 using 2 TRAV-regions (human TRAV8 and mouse TRAV9) and 4 TRAJ-regions and an overall RMSD average of 0.95±0.14. The Tyr56 in this class pushes itself within the apex of the loop forcing residues 57-59, 62, and 63 to arrange themselves away from the loop center with help from the kink provided by residue 63 shown in **Figure 3.4.2K**. The hydrophobic residues at 54 and 65 additionally narrows the distance between the flanking ends of the loop. Based on these structural features described the key residues for this class include hydrophobic residues at 54 and 65, a large aromatic residue at 56, and a small to medium-sized amino acid at position 63.

The β1 hypervariable regions

CDRβ1 had the highest number of classified structures of all CDRs with 86% of structures analyzed fitting into the two established classes. Additionally there was another new conformation with 4 structures found in human TRDV-regions. Similar to antibodies there was less structural variability in CDRβ1 and β2 possibly because evolutionary pressure has focused on CDRβ3 since it has a much larger diversity generated during VDJ gene rearrangement. CDRβ1 classes 3-5 that were initially proposed in 2000 did not have any structures cluster beyond identical TCRs even though there are now over 14 TCR structures with TRBV-regions.

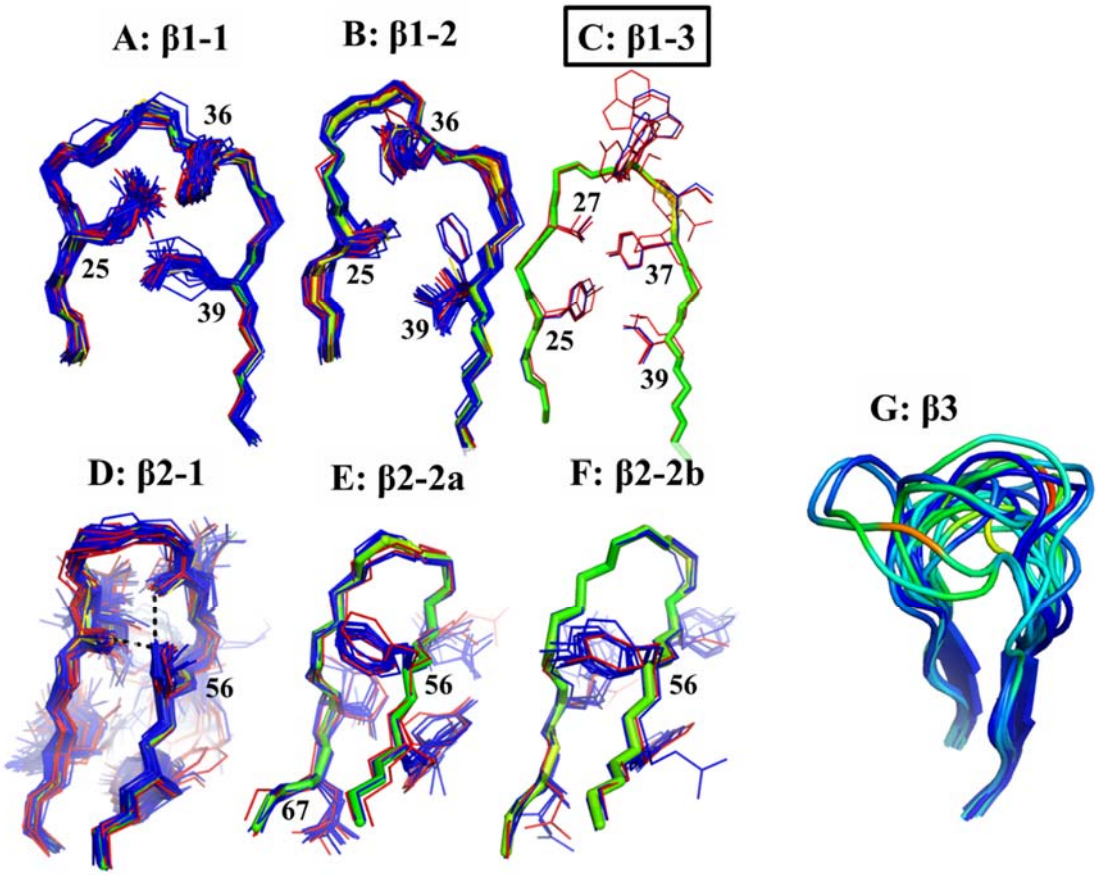


Figure 3.3A-G: CDRβ class alignments

Alignments for the CDRβ1 and CDRβ2 classes show the majority of structures conform to a single class. Canonical loops with key residues shown for classes (**A**) β1-1, (**B**) β1-2, (**C**) β1-3, (**D**) β2-1, (**E,F**) β2-2 with two subclasses, and (**G**) CDRβ3 alignments showing the high structural diversity these loops.

β1-1

As predicted by the key residues in CDRβ1, the majority (68%) of the β1 structures examined were classified as β1-1 loops. This class had 13 V-regions from human, mouse and rat TCRs (human TRBV4, -5, -6, -10, -19, -25, -27, -28, mouse TRBV12, -13, -19, -29, and rat TRBV5) and 20 TRBJ-regions, a 12 residue length, and an overall average RMSD of 0.07±0.13 Å. The β1-1 class is one of the most structurally consistent classes,

seen in **Figure 3.3A**, even with so many different TCRs, structures, and sequences. This consistency hinges upon the close interactions between residues 25, 27, 29, and 32.

β1-2

This CDRβ1 class also contains a large number of structures with 42 (17 %) loops, 9 TRBV-regions (human TRBV5, -7, -9, 11, -12, -14, mouse TRBV3, -12, and -26), 11 TRBJ-regions, and an average RMSD value of 0.66±0.33 Å. The key difference of this loop from β1-1 is the Pro25 residue which varies the backbone at residues 25-28 shown in aligned classes of **Figure 3.3C**. This backbone alteration narrows the loop at residues 26-29 with close interactions between 26 and 29 shown in **Figure 3.3B**. Outside of this variation, the remainder of the loop residues are nearly identical to β1-1.

β1-3 (New canonical class)

The previously proposed β1-3 class is structurally most likely unique to its variable region since the 10 structures with matching β1-3 key residues (PDB: 1FO0, 1KB5, 1KJ2, 1NAM, 2OL3, 3HG1, 4JFD, 4JFE, 4JFF, and 4JFH) failed to form any structural consensus seen by **Figure 3.3C**. Instead of a third β1 class around mouse TRBV1 another smaller set of 4 TRDV-region structures (human TRDV1 and -3), 14 residues long, and an average RMSD value of 0.01±0.01 Å showed a distinctly unique and interesting conformation. This class has residues 25, 27, 37 and 39 all pointing into the center of the loop in direct opposition to one another. The loop apex also has residues 29 and 30 oriented either perpendicular to the plane of the loop or directly away from the base of the loop. With only two CDR sequences in this class the key residues are not especially distinct.

The β2 hypervariable regions

Similar to CDRβ1, this analysis found larger but fewer β2 classes than originally proposed in the 2000 canonical study with only two classes with well-defined

conformations. CDRβ2 found classifications for 75% of the loops. The remaining smaller groups of unclassified CDRβ2 loops lacked diversity within the groups to be definitive classes. Similar to CDRα2 loop, this loop is the antigen engaging portion between β-sheet strands C' and C'' which can vary depending on the β-chain sheet composition or α-chain V-domain interaction.

β2-1

As expected based on the co-variation of classes, most of the TCRs classified as β1-1 had β2-1 CDRβ2 loops. This class had 159 structures in it with 11 TRBV-regions (human TRBV5, -6, -10, -19, -25, mouse TRBV12, -13, -19, -29, and rat TRBV5), 22 TRBJ-regions, a length of 13 residues, and an average RMSD of 0.94±0.47 Å. All the loops have a Ser56 in the same orientation forming hydrogen bonds with the backbone carboxyl oxygen in residues 58 and 65 labeled in **Figure 3.3D**. These hydrogen bonds help fold the loop apex into the center of the β-chain and orient the side chains of residues 55, 57, 65, and 67 into solvent exposed positions near CDRα3.

β2-2

Similar to the previously mentioned classes formed around TCR N15 there were no additional structures added to its particular loop conformation. The key residues proposed for the β2-2 class otherwise had 24 structures in 2 subclasses 7 TRBV regions (human TRBV 5, -7, -9, -11, -12, -14, and mouse TRBV14) and 7 TRBJ-regions, all 12 residues long, and an average RMSD values of 0.92±0.44 Å and 0.68±0.27 Å for each subclass respectively. This loop class is tightly bound with 5 sets of hydrogen bonds connecting the backbone nitrogen and carboxyl oxygen atoms in residues 55, 57, 64, 66, and 67 indicated in **Figure 3.3E & F**. Similar to β2-1 the side chains at residues 55, 57, 66, and 67 into solvent exposed positions near CDRα3.

Additional results

CDR3

We performed the same structural analysis with CDR3 and found the loops mostly unclassifiable due in part to the substantial variability in length, sequence, and structure. Depending on the complex conditions, the longer and often structurally ambiguous CDR3 structures are either solvent exposed or pressed against the pMHC causing a high degree of structural variability between the solved structures of nearly identical TCR complexes. The only consistent feature of CDR3 structures was the conserved framework and take-off portions highlighted in **Figure 3.3G**, which contrasts with the central pMHC-interrogating portion that contains a large amount of variability. Studies attempting to obtain parameters for CDR3 structural predictions have found limited success in both TCR and antibody hypervariable regions[102,105].

γδ CDR Loops

γδ TCRs comprise a small minority of lesser studied structures with 7 γδ dimers (G115, G8, and MICA-specific human delta scFv) (PDBs: 1HXM, 3OMZ, 4LFH, 4LHU, 4MNG, 4MNH, and 4NDM) and one δ chain fragment (PDB: 1TVD) analyzed similarly to αβ TCRs for possible canonical structures. Due to the different overall framework structure of these γδ TCRs only 3 CDR structures fit into established canonical classes. It could be the different functional role of γδ TCRs binding to a variety of antigens, MHC-bound or otherwise, could give them CDR conformations similar to both αβ TCRs and antibodies.

Antibody Comparison

After establishment of these detailed canonical classes, they were then compared to known antibody classes[110]. The two sets of canonical classes had no overlapping similarities even though both are structurally similar. It could be the higher binding

affinities in antibodies change the framework and CDR backbone structure when compared to TCRs. TCR-like antibodies have only been reported for a few pMHC molecules, 25-D1.16 in particular binds to pOV8-K^b with nearly the same orientation as its TCR counterpart KB5-C20 (PDB: 1KB5)[115,116]. Even with the nearly identical epitope found between the two proteins none of the CDR loops of the 25-D1.16 overlapped with the KD5-C20 TCR.

3.5: DISCUSSION AND CONCLUSIONS

With the advancement of technology in high throughput sequencing and increased reliability of protein modeling and structural prediction algorithms, protein engineers can more accurately hypothesize the TCR structures capable of involvement in specific antigen interactions. The hypervariable CDR3 regions defy conventional means of determining their structure and predicting their antigen engagement, limiting the accuracy of these structural models. While this study confirmed these challenges for CDR3, the strong prominence of key residues in the conformations of the other CDRs in TCRs allowed us to describe the canonical loops of TCRs as has previously been done for antibodies.

In this study with 249 TCR and TCR-like structures it was shown the majority of TCR CDR conformations correlate with a set of proximal key residues. These residues tend to provide either structural stability when they point inwards or a packing stability involving hydrogen bonds with other residues or with the loop backbone. Additionally three new canonical class conformations not previously proposed were found in the larger dataset. The original *de novo* predictions of canonical classes from the V-region sequences available in 2000 provided a limited insight into the various key residues found in total TCR gene repertoire. Many of the class predictions have not been tested from a lack of solved crystal structures for respective V-regions. Of the 18 classes proposed based on the

sequence and structure data, 12 had accurately unique conformations with the remaining needing either additional solved structures or reevaluation with an updated V-region gene repertoire.

The total number of TCR canonical classes still lags behind the number of diverse structural classes in antibodies (a total of 42 classes across all the heavy and light chain CDRs[110]). A recent examination of antibody CDR clustering in over 12,000 loops found a clustering rate of over 90% for all CDRs except H3, similar rates may be expected with an increasing diversity and number of solved TCRs[117]. In antibodies approximately a fifth to a third of the CDR residues participate in antigen binding, where van der Waals radii atomic contact is designated by a distance of 0.5 Å[107]. Antibody structure modeling using solved structures and established canonical class templates have increased the theoretical accuracy of structural and functional experiments[118].

Homogeneous β-chain structures

The predictions for the TRAV-regions were not as accurate as for the TRBV-regions. Different attributes of the two chains may cause these phenomena such as the β-chain's larger reliance on the VDJ junction to generate the necessary the TCR structural diversity. Since most bound TCR structures indicate a shared pMHC engagement with the TCR αβ chains the structural differences between the chains may be more nuanced. The increased structural homology found in the large canonical classes in the β-chain may indicate some sort of evolutionary bias either in the TCRβ structure. Rearrangement and stable membrane display of the β-chain occurs with a generic pre-TCR α-chain (pTα). The pTα when associated with the β-chain does not interact with the TRBV-domain selecting for more stable soluble TRBV-regions[119].

Unclassified and other overlapping loops

While the large majority of the loops were classifiable, 14%, 45%, 11%, and 23% of CDR α1, α2, β1, and β2 loops remained unclassified. Many of these unclassified loops were slightly above the RMSD threshold required for clustering or involved smaller, single V-region groups and a few truly distinct structures when compared to identical TCRs such as 1NFD. Certain V-regions had a majority represented in a canonical class with a few remaining structures left unclassified or in another class with an alternative conformation such as human TRAV39 with 2 loops in α1-1 and 22 loops in the new α1-5 class. These exceptions were in the minority such as CDRα1 in TCR 2C where 3 of 7 loops diverged at residues 28, 29, and 36 by 4.3 Å generating the subclass structure in α1-2.

Affinity matured TCRs

Affinity maturated TCRs have engineered pMHC interactions for altered affinity and specificity and while these mutations usually occur at the interfacial positions they do not always significantly alter the structure of CDR loops. Most TCR engineering aimed at increasing TCR affinity without cross reactivity focuses on CDR3 mutations to increase antigen interactions. For example, the well-studied 1G4 has undergone several studies to increase the TCR affinity from ~15 μM to less than 10 nM[120]. The α-chains vary in CDRα2 with the point mutations of S58P and S59W in structures 2P5E and 2P5W causing the loop apex to fold further upon itself. The solved structures for these mutations show slight (~1.15 Å between 2F54 and 2P5E) CDRα2 alterations from the wildtype structure and more significant loop changes in CDR3 structure[121,122].

Mouse versus human?

The percent of human and mouse CDR structures classified was roughly the same at 73% and 82% respectively based on the current analysis. Humans have a larger number

of variable region genes than mice which could lead to increased structural variability. Mice also have a limited pool of T-cells at any given time which could restrict the diversity found in their TCRs. Conclusions about the differences between species remains fairly limited because the data set encompassing unique TCRs was generated from high-affinity variants, autoimmune models, or soluble-engineered and solvable proteins. Additionally the TCRs of many predominant disease models like multiple sclerosis are extensively studied and comprise many of the solved TCR crystal structures in protein databases[123,124]. TCRs that have increased soluble stability or express in higher concentrations may be overrepresented in the available PDB structures due to relative ease of crystallization, and any bias of this type would carry over into our analysis.

Structural Mechanisms

There are various theories about the evolution of T-cell maturation and how germline VDJ-regions have evolved their specific function and sequential order on the chromosome[100] . Determining correlations within TCR rearrangement events and V- and J-region combinations by analyzing their sequences may shed insight into the co-evolution of diseases and cellular immune systems. The pMHC interaction may play a distinct role in determine the whether a TCR's CDR fits into a class because over twice as many classified structures come from pMHC-bound TCRs than unbound. While some loops change conformation significantly when pMHC-bound (B7 with structures 1B2D and 1BWM having RMSD values 1.01, 1.70, 0.98, & 1.39 Å for CDRs α1, α2, β1, & β2 respectively), most others show little to no structural change between the pMHC-bound/unbound structures seen in many TCRs like 2C for example. Using these TCR CDR conformations, the pMHC interacting residues can be determined similarly to studies of antibody specificity-determining residues. These are identified by combining consensus

sequence data with structurally important positions in CDRs and neighboring key residues[107]. The TCR-pMHC contacts analyzed in this study found a set of focused contacts on the binding groove MHC α-helices as reported in previous studies, but no strong correlation of contacts with canonical classes[11,111]. Additional studies with HTS and restricted pMHC sorting molecules could determine if these canonical classes are stereotyped to particular pMHC epitopes. This analysis could show if there is a limited sequence space generated from the same structural epitope.

Analysis of TCR canonical loop classes enable prediction of TCR structure from sequence data, which will be particularly useful in conjunction with HTS of *in vivo* TCR repertoires. This study with an expanded set of TCR structures validated 12 of the conformations proposed in the initial Chothia *et al* study and 3 new classes not previously proposed. While this work helps to further characterize the structures of TCRs there are still considerable gaps in our structural modeling capability. There are a couple factors limiting the expansion and accuracy of these structural models from V-region sequences including conformational difference between pMHC-bound/unbound structures, TCR-pMHC contacts, and predominantly CDR3 sequence and structure variability

Chapter 4: Structural repertoires for *de Novo* antibody engineering

4.1: CHAPTER SUMMARY

In a similar application of immune-receptor and CDR structural analyses described in Chapter 3, we used existing loop conformations and key residues in antibody and TCR structures for structural predictions in a *de novo* structure. This study used OptCDR a structure predicting algorithm for proposing a novel antibody sequence and structure for interacting with a flag-peptide tag[103]. *In vitro* characterization of these antibodies proved difficult due to low expression yields and binding affinity with the flag peptide. Here we detail the use of an antibody structural repertoire to predict and analyze the inherent feasibility of generating the model antibody structures.

The designed set of single-chain antibody structures used a common framework mouse IGHV1-82*01 and mouse IGKV1-117*01 with CDR sequence variants to bind a flag-peptide (**Figure 4.1**). By aligning the either the framework structure or the framework-CDR loop interface between the many available structures the variability between loops and structures was measured and compared to the *de novo* models originally proposed in the antibody study. Searching the much larger repertoire of antibody structures found the conformational loop limits and sequence variability permitted in these structures on this specific framework.

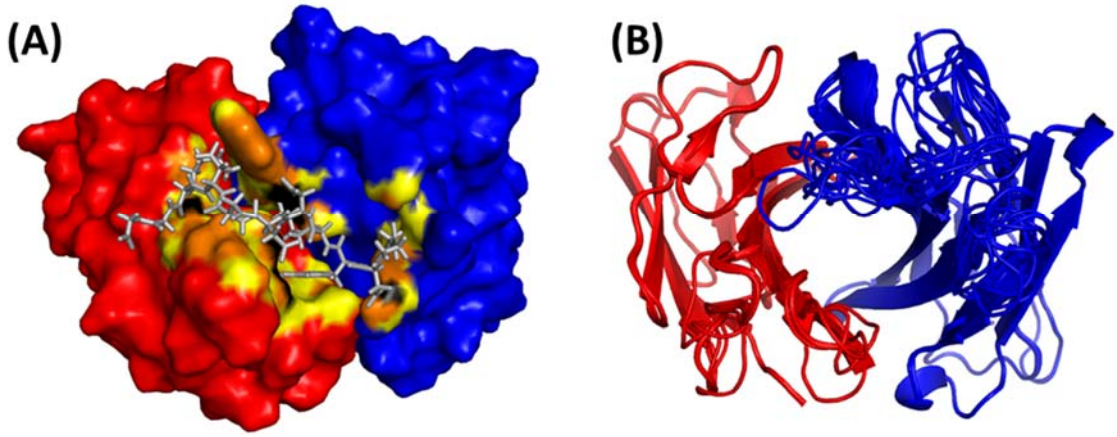


Figure 4.1: Overview of the OptCDR model.

(**A**) Depiction of the model generated for the OptCDR structure's binding groove with the heavy, light, and flag-peptide colored blue, red, and gray respectively. Binding distances between peptide and scFv of 2-, 3-, and 4-Å contacts highlighted in yellow, orange and black respectively. (**B**) Alignment of the 14 antibody structures predicted to bind to flag peptide showing the consensus framework with the more variable CDR loops.

These structural comparisons helped determine why the single chain variants from these models failed to express or bind as predicted. The structural analysis on the structures in the PDB measured the sequence diversity in compared variable domain frameworks which maintain the overall shape and framework/CDR interface position, orientation, and arrangement predicted in the model. Variable domain frameworks matching the model were then found, and the binding region CDR loops for conformation and sequence diversity were analyzed. Finally, we compared the flag-peptide structure with other peptides and peptide binding antibodies to characterize the interactions between peptide and antibody. This analysis was performed over all the OptCDR models to understand the experimental results showing poor expression and peptide binding. The results proposed specific CDR changes to leverage the known structural repertoire for these *de novo* models.

4.2 METHODS

The OptCDR model system uses a framework input to generate an interface, between the variable region framework and CDRs, to fix the orientation and position of CDR loop termini. CDRs from existing PDB structures are grafted *in situ* to the framework and residues are assigned, rotated, and tested for possible clashes to determine the best surface complementarity and local chemistry to encourage peptide binding. Within the CDR fitting the rigid flag peptide input is probed along the surface of the paratope searching for the highest binding energy.

Based on analysis from TCR canonical structures it was hypothesized this modelling system makes a few assumptions which may overlook some key features common to antibody CDRs and canonical loop structures. First, main-chain loop structures are highly variable even with fixed N- & C- termini. Second, slight CDR sequence variations can have a significant effects on loop conformation as shown in studies of affinity maturated structures and key residues.[23,87,105,125] CDR key residues usually point internally or towards other CDR loops for stabilizing overall structure. Finally, peptide binding occurs cooperatively when both the antibody and peptide interact which alters the structure of both. Antibody CDR deformation to surround the much smaller peptide within a binding groove come at too high an energetic penalty between bound and unbound structures[126].

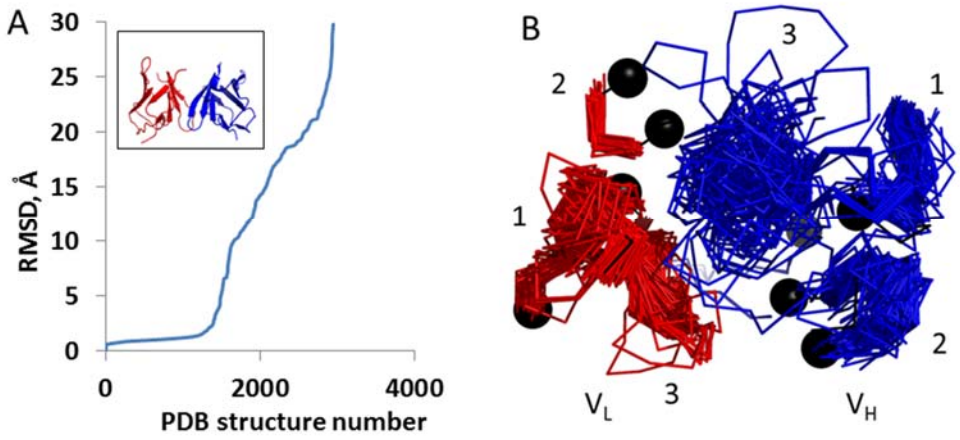
4.3: RESULTS AND DISCUSSION

A search of PDB associated with the search terms ‘antibody’ and ‘TCR’ found 3056 structures of antibodies (2844) and TCRs (212) to compare to the modeled complex. We aligned these structures with the proposed framework input for the model. Aligned PDB structures with RMSD values less than 1Å were further compared to a set of 36 residues comprising the framework/CDR interface (residues: H 24-26, 37-39, 53-55, 65-67, 102-

104, & 118-120; L 24-26, 39-41, 53-55, 66-68, 102-104, & 118-120). Three residues were used at the N- and C-termini of each CDR to help guide the alignment program to orient the compared structures. Alignment calculations were allowed to reduce the overall number of atoms up to half the total residues in the reference framework. Any reduction in the compared number of atoms was also factored into the alignment metric score (RMSD / (% of framework atoms aligned)). The alignments were labeled partial and full if the framework or both the framework and interface had RMSD values less than 1Å respectively. The program then isolated the CDRs from the matching structures to analyze the compatible CDR loop structures and sequences for comparison to the predicted model structure and sequences. Additionally the depth of the flag binding peptide into the V_H/V_L interface was measured as the z-axis distance from the plane made from the N and C terminal residues of the heavy and light CDR3s. This analysis was also run with the total distance from the centroid of the four residues, but was considered less informative than z-axis depth due to the variable orientation of the rigid flag peptide over the paratope.

4.3.1: Structural repertoire analysis

Of the 3056 structures analyzed, 326 had a similarity to the framework and 42 had a full framework/CDR interface compatibility with the predicted flag peptide-binding model. RMSD scores showed a distinct gap between aligning and dissimilar frameworks shown in **Figure 4.2A and Figure 4.2B** below. The input framework is distinct in shape and highly represented in the PDB structures with over 10% of the total structures having RMSD values <1Å. The 326 candidate structures were then analyzed for their CDR similarity, length, sequence, and key residue variation. Additionally the 42 interface compatible structures were analyzed for their CDR similarity to the models and general alignment overall (**Figure 4.3A**)



Figures 4.2: Alignment to model framework

From the screen of PDB frameworks (**A**) the RMSD alignment scores to the entire candidate framework used in the OptCDR model and (**B**) the subsequent alignment score to only the interfacial residues, shown as spheres, describing the arrangement of the CDRS.

4.3.2: Light chain

There were 45 light chain CDRs with at least one of the model's proposed sequence including two antibodies with all 3 matching light CDR sequences (PDB: 3NN8 and 4NKD), both being from previous flag peptide structure work within the Maynard Lab. These matches on CDRs L1, L2, and L3 have limited mainchain loop structural variability. Light chain CDR lengths were more limited in conformation (**Figure 4.3C**) with many compared structures having the same CDR length. For example, the majority of the L2 loops had 3 residues, which is probably an artifact of L2 in general. Further analysis found the majority of the models predicted a greater interaction of the light chain with the peptide than the heavy chain.

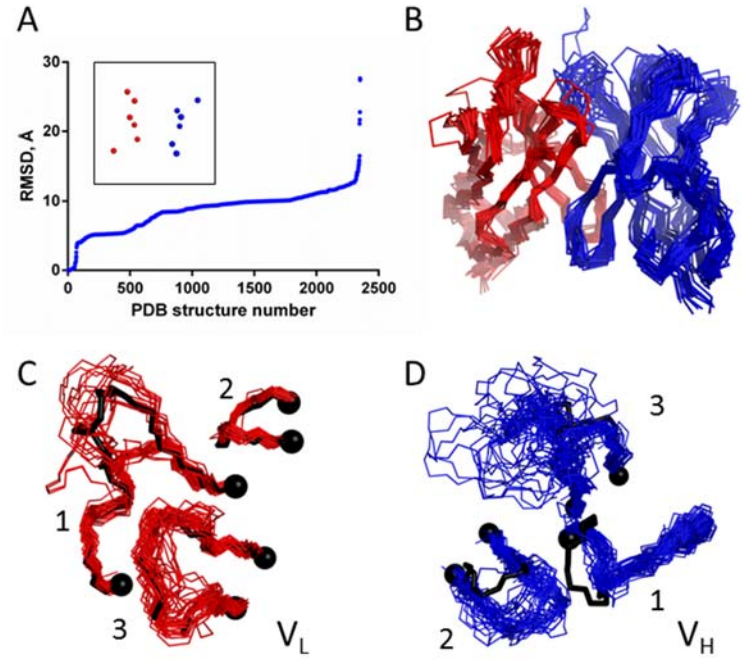


Figure 4.3: Alignments to framework interface.

(**A**) For the EEh14.3 parent model, three residues at *N*- and *C*-termini of the framework/CDR interface for all six CDRs (*inset*) were compared to all antibody-related PDB structures. 58 structures possessed interface alignment RMSD values <1 Å. (**B**) This subset shows low overall variability in the framework light chain (red) and heavy chain (blue), excluding the CDRs shown at the top. (**C**) Isolated light chain CDRs of the interface-aligned structures, shown from above, show low structural variation, and EEh14.3 parent model CDRs (black) exhibit similar conformations. Interface residues used for alignment are shown as black spheres. (**D**) Isolated heavy chain CDRs, rendered similarly, show large structural and length variability, and EEh14.3 parent model CDR1 and CDR2 show distinctly different conformations from the interface-aligned population.

4.3.3: Heavy chain

While the heavy chain framework and interface is maintained in many other solved antibodies, the model's CDRs themselves are unique in sequence, length, and structure (**Figure 4.3D**). The heavy chain had no identically matching CDR sequences for any of the regions, indicating OptCDR may have attempted to create and fit *de novo* sequences and structures to the framework. The design of the heavy chain may have had a much larger

influence on the overall performance of the antibody compared to the light chain. As modeled, the heavy chain contributes closer, but fewer contacts to the flag binding than the light chain. Additionally the gap in the heavy chain required for flag binding may be leading to the overall instability of the antibody.

CDR length also plays an important role in the ability of the antibody to interact effectively with the peptide. Compared to other peptide binding antibodies the 14a complex may have CDR3 lengths well outside of the stable range for the framework (**Figure 4.4A**).

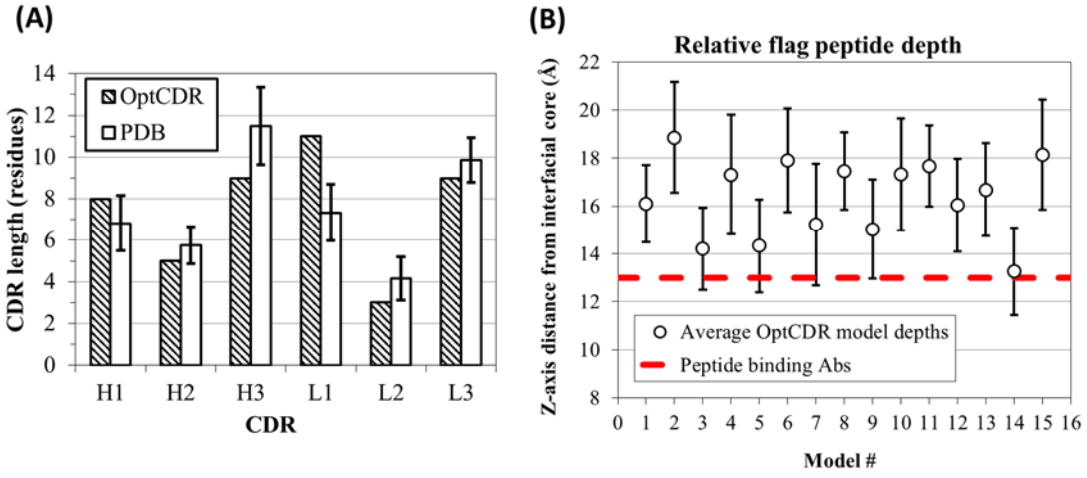


Figure 4.4: CDR length variations and flag peptide depth

Calculations of the average CDR length for the structurally similar candidates and the lengths for the proposed 14a structure showing the variations of CDR lengths outside of the average may contribute to the instability and (**B**) a plot of the variation of z-axis distance from the interfacial core of the complex along with the minimum and maximum distances.

4.3.4: Peptide Binding

The depth of the peptide when penetrating the antibody structure may be the largest difference of the model to other PDBs. The majority of the antibodies with framework matches have CDRs stably filling the gap wedged between the H/L domains. The H/L interfacial domain opens very wide for flag peptide interaction with the majority of the

deformation occurring in the Heavy chain. It was indeterminable if the residues comprising the binding interface were stable enough to tolerate solvent exposure during peptide binding. It is possible the groove can stably open wide enough to accept the peptide, but this would probably reduce the overall efficiency of the binding kinetics even if it binds tightly. Surprisingly when compared to the other models the average flag peptide depth was the deepest for complex 14 (**Figure 4.4B**). Looking at highly variable rigid flag peptide arrangements over the paratope 15 predicted complex structures it is clear the peptide structure should have been allowed to flex or been modelled along with the CDRs to further minimize the binding energy and increase the model accuracy.

4.4: DISCUSSION AND CONCLUSIONS

In the future, library screening efforts may be more successful if the sequences are capable of conforming to a canonical CDR structure typical of peptide binding and the statistical makeup of beta-turns potentials in terms of the probability of having a certain amino acid in the turn position is taken into account. Altering the structures of solved antibodies may provide the best insight when grafting new sequences and arrangements. Antigens that bury deeper into the chain may be tighter binders if the antibody can stably expose the interface between the chains to form the binding groove. Finally, rigid peptides should not be considered accurate unless they have additional structural context excluding the antibody.

Chapter 5: Activation studies of engineered TCRs

5.1: Chapter Summary

Here we report our work to generate a surface display system for testing the activation of engineered TCRs on mouse α^-/β^- T-cells. These affinity maturated engineered TCRs were generated in two separate bacteriophage display studies by other members in the Maynard lab. Complete TCR surface expression in eukaryotic and thymocyte cells will also verify the *in vitro* binding assays by mitigating any biases occurring in soluble bacterial expression and phage display[122,127]. Affinity maturation of TCRs using phage display or soluble expression systems has shown to have variable effects on activation and antigen specificity when reconstituted on accepting thymocytes[13]. While *in vitro* assays can measure altered antigen specificity, determination of differences in T-cell activation requires the reconstitution of the engineered receptor into a T-cell immune complex. Progress for collecting display and activation data for these TCRs confronted multiple problems including receptor stability, transfection efficiency, and the improper trafficking of the protein to the outer membrane. Additional work is ongoing to collect T-cell activation data for the TCR variants and to test variant signal sequences for intracellular trafficking and display.

The discovery of the improper TCR trafficking due to the antibody signal sequence on the expression vector led to a brief investigation of how these N-terminal peptides influence TCR expression and trafficking. This study varied the signal sequences on full-length TCRs as individual chain plasmids, a bicistronic IRES, a single transcription 2A peptide, or a CAR. The sequence variations examined how residue mutations in the hydrophobic (h) and the peptidase cleavage (c) regions alter the display stability and secretion of proteins trafficking through the secretion pathway.

5.2: INTRODUCTION

Mutations in TCR CDRs can modify the receptor's overall activity through slight secondary structural alterations carried through to the bound CD3 proteins in the immune complex[128]. To test variant TCRs generated *in vitro* we developed an expression system to display full length TCRs including the variable, constant, and transmembrane domains[129]. TCRs displayed this way can bestow activated $\alpha^-\beta^-$ T-cells with additional antigen specificity and activity dependent on the formation of the immune complex with the CD3 co-receptors[130]. TCR display and activity was measured by transfecting these constructs into the TCR deficient $58^{-/-}$ thymocyte cell line. This immortalized cell line was derived from the BW5147 murine T-cell line and has non-functional α- and β-chains[131]. The TCR chains tested in this display and activation study included DO11.10, 172.10, and TRBV5 (**Table 5.1**). The mouse TRBV5 chain was unexpectedly recovered from the DO11.10 and BW5147 hybridoma fusion[132,133].

TCR	Leader/Signal Sequence	V region gene	CDR3 sequence	D	J	C
DO11.10	**MKTYAPTLFMFLWLQLDGMSQ**	**mTRAV 5D-4**	AASPNYNVLY	-	21	1
172	**MDKILTASFLLLGLHLAG**	**mTRAV 14-3**	AASANSGTYQR	-	13	1
DO11.10	**MGSRLFFVLSSLLCSKHM**	**mTRBV 13-2**	ASGSGTTNTEYF	1	1-1	2
172	**MGSRLFFVLSSLLCSKHM**	**mTRBV 13-2**	ASGDAGGGYEQY	2	2-7	1
BW5147	**MGSRLLLYVSLCLVETALM**	**mTRBV 5**	ASSQITSNQDTQYF	2	2-5	2

Table 5.1: TCR constructs tested in this display study.

The TCR chain compositions for this study included DO11.10, 172.10, and mouse TRBV5. The signal sequence oligo was also able to subclone a functional mouse TRBV5 gene from the BW5147 hybridomas with the genes and CDR3 sequence. Since the DO11.10 and 172.10 β-chains are so similar, the mouse TRBV5 construct was used as a negative control when transfected with different α-chain.

5.2.1: Phage display directed evolution and affinity maturation

Previous work with the directed evolution and affinity maturation of a DO11.10 single-chain TCR using phage display found two CDRα3 variants with higher binding affinities to a clonotypic antibody specific for the anti-mouse DO11.10 TCR (clone: KJ-26, BD Biosciences). The two higher affinity TCRs had CDRα3 sequences varying slightly from the wildtype DO11.10 sequence by only a few residues (**Table 5.2**). Each variant was selected as a fusion with different M13 phage coat proteins.

					Randomized residues								
	Fusion	104	105	106	107	108	109	113	114	115	116	117	118
WT	-	C	A	A	S	P	N	Y	N	V	L	Y	F
176	pIII	C	A	A	S	T	N	W	H	N	L	Y	F
817	pVIIIo2	C	A	A	S	P	R	F	D	V	L	Y	F

Table 5.2: Sequence data for DO11α-chain variants 176 and 817.

The two DO11.10 variants were selected from a CDRα3 library using different C-terminal fusions on different M13 bacteriophage coat proteins. The expected scTCR local concentration on these different coat proteins alters the binding through relative avidity affects.

The M13 phage used in this experiment has 5 coat proteins with different arrangements along the shaft and termini of the phage (**Figure 5.1A** and **B**)[134]. Fusing scTCR constructs to these coat proteins allows the rapid screening of libraries in excess of 10^{10} variants[135].The two higher affinity variants called 176 and 817 were selected on coat proteins pIII and pVIII respectively (**Figure 5.1C** and **D**). These coat proteins have strikingly different roles and arrangements on the phage and as such, screening of scTCR fusions to these coat proteins can lead to different binding kinetics based on avidity affects. Characterization of the two selected variants also found differences in the expression yields, stability, and binding kinetics of each protein.

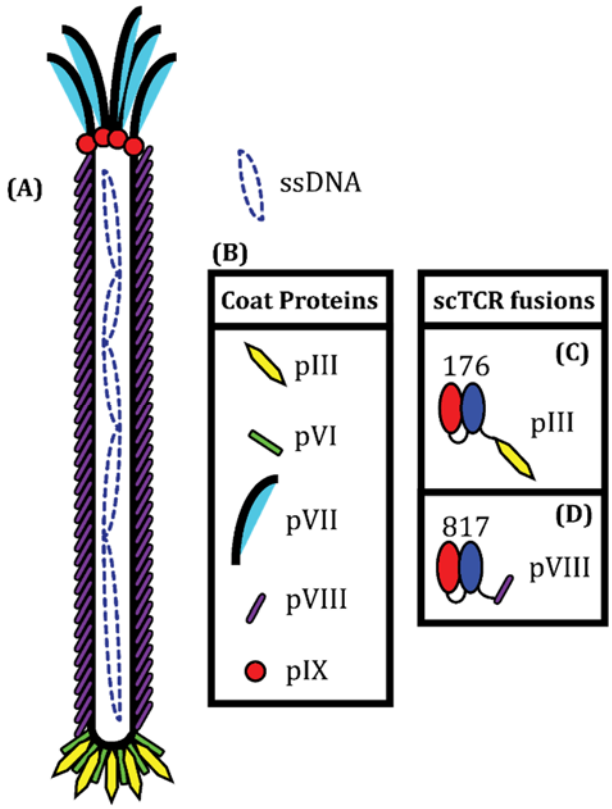


Figure 5.1: M13 phage with coat proteins labels.

(**A**) M13 is a filamentous bacteriophage with (**B**) 5 coat proteins encasing its single stranded genomic DNA. Affinity maturation with scTCR C-terminal coat protein fusions found two DO11.10 CDRα3 variants with higher binding compared to the wild type TCR. Variant 176 (**C**) is a pIII fusion found in 1-5 copies at the end of the bacteriophage and 817 (**D**) is a pVIIIo2 fusion found in several thousand copies comprising the central shaft of the phage. The difference in copy number and arrangement on the phage selected for different secondary binding properties in each variant.

5.2.2: Eukaryotic protein expression

Expression of engineered proteins in eukaryotic cells requires many complex techniques and genetic components. First, the genetic material must become transfected into both the cell and the nuclear envelope. Transfection techniques for eukaryotic cells are more complex than for bacterial transformations due to less permeable cell membranes and resistance to foreign genetic material[136]. Depending on the cell type, a transfection

method's efficiency can vary dramatically. Several transfections techniques include viral transduction, lipofection (Invitrogen and others), nucleofection (Lonza), electroporation, or calcium phosphate precipitation. Additionally the cell only transiently maintains and expresses the introduced DNA unless additional selection pressures, such as antibiotic resistance selection for example, force the cell to integrate the material into its chromosome.

Depending on the cell assay or protein expression system, transient transfection or stable cell line generation may be necessary. For example, small-scale antibody expression in CHO cells often uses transient transfection to produce large quantities of antibodies without the more involved, lengthy stable selection process. Transient transfections only last for a few cell generations because the plasmid DNA is not maintained in the nucleus and becomes diluted with every cell replication. Alternatively, some assays require a stable homogenous population selected from a single cell to guarantee uniformity in the results. Stable cell lines take weeks to months to generate, as the desired gene must stably incorporate into the genomic DNA followed by single cell or polyclonal selection to remove cells lacking the desired phenotype.

After transfection into the nucleus, transcription factors transcribe the genetic material into mRNA. Next, the mRNA processing removes any introns and adds stabilizing components including a 5' mRNA cap and 3' polyA tail before transport out of the nucleus into the cytoplasm. A ribosome next associates with the processed mRNA, recognizing the Kozak sequence immediately 5' of the first start codon in the open reading frame before beginning translation of the genetic information into a polypeptide sequence[137]. As the nascent peptide chain emerges from the ribosome, it acts as a signaling mechanism to direct the trafficking of the synthesizing protein-ribosome complex to the proper receptors and organelles in the cell for further post-translational modifications.

5.2.3: Signal sequence in protein trafficking

This signal sequence N-terminal to the gene of interest serves a critical role in secreted or displayed proteins requiring transport through the ER, Golgi, and eventually the outer cell membrane. Signal sequences while diverse share 3 distinct peptide regions designated the n-, h-, and c-regions[138]. These regions are responsible for the recruitment of the signal recognition particle (SRP) and after translocation into the ER, encourage signal peptide peptidase activity.

For secreted and transmembrane proteins destined for the ER, the signal sequence recruits an SRP by a coordinated interaction with a conserved central h- region of the signal sequence and a cluster of methionine residues on the SRP[139]. Recent work has shown the interaction between the SRP and signal sequence h-region is aided by a series of suboptimal amino acid codons approximately 40 – 60 residues after the nascent chain first emerges from the ribosome. These codons slow translation to allow time for proper interaction between the SRP and signal sequence and proper trafficking to the ER[140]. The SRP changes shape after binding to the signal sequence and guides the nascent chain-ribosome-mRNA complex to the SRP-receptor on the outer surface of the ER for translocation of the synthesizing protein.

During protein translocation into the ER, the signal sequence can remain anchored within the ER membrane where it awaits cleavage by signal peptidase[141]. Signal sequences with weak or cryptic cleavage regions become unstable and targeted for degradation as incompletely processed proteins[142]. Signal sequence cleavage in general is thought to occur as the protein translocates into the ER, but this varies between proteins. Depending on the protein's final destination, this ER anchoring and cleavage event may cause downstream effects in expression.

Antibodies and TCRs alike both use strong signal sequences imperative for proper folding and incorporation into the ER membrane for eventual transport to the Golgi and outer membrane. Each variable region gene has a corresponding signal sequence which follows the n-, h-, and c-region motif with some clear distinctions between antibody and TCRs signals. Comparisons of the signal sequences for these secreted and membrane-bound proteins in humans and mice using weblogos of sequences 19 residues long show the slight variations the signal sequences of these classes of proteins (**Figure 5.2**), which may give some insight into their functionality as soluble proteins and membrane-bound receptors[143]. The noticeable occurrence of cysteine residues and more variable hydrophobic region of the TCR signal sequences are curious as aberrant di-sulphide bond formation could significantly influence overall expression of the receptors. The ClustalΩ web server compiled consensus signal sequences generated from all functional antibody and TCR V-regions are reported in **Table 5.3**. Surprisingly there is no overlap in signal sequences between antibodies and TCR and each consensus sequence is 21 residues long.

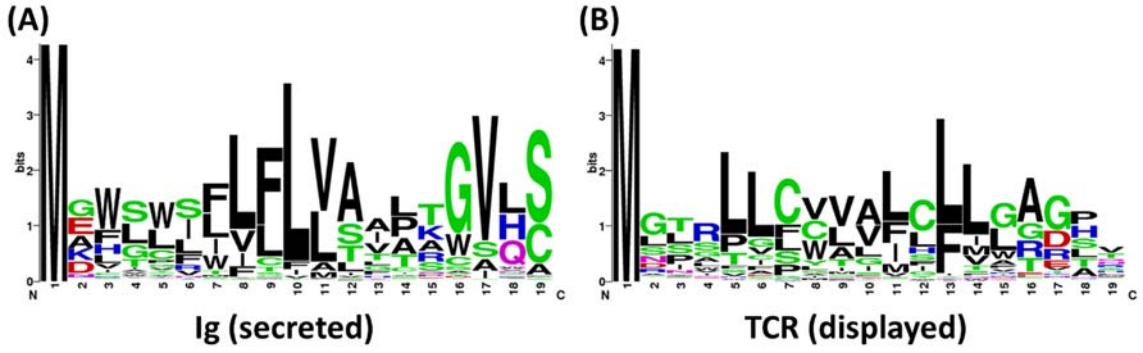


Figure 5.2: Weblogos for antibody and TCR signal sequences.

Signal sequences from human and mouse (**A**) antibodies and (**B**) TCRs 19 residues long were aligned to show the n-, h-, and c-regions along with the variations between the two protein classes. These variations involve a very hydrophobic h-region in the antibody signal sequence and an interesting number of cysteine residues in the h-region of TCRs.

	Consensus Signal Sequence	Genes	Range	Mean	StDev
Ig	MGWSW**IFLFLVAAPTGV**HSLS	390	16-24	19.46	1.02
TR	MLMGTR**LLCWVALCLLG**AGHV	256	14-39	20.06	2.97

Table 5.3: Ig and TCR gene consensus signal sequences.

Compiling all the signal sequences for antibody and TCR genes for humans and mice using the ClustalΩ web server found two distinctly consensus sequences[79]. The h-region in the consensus signal sequences is in bold, red text and the cysteine residues in the TCR peptide are underlined. The Genes column reports the total number of functional gene alleles with unique signal sequences.

The critical signal sequence components required for proper display have been vaguely described for several protein models without a more knowledgeable understanding of the mechanisms involved[144]. Analysis of many signal sequences in immune-proteins (**Table B.4**) found an incredible diversity of sequences to convey the level of variability in length and sequence for leaders among different secreted vs membrane bound proteins.

A review of different TCR display constructs finds most use the cognate signal sequence for each variable TCR gene without modification[122,127,136,145–149]. Rarely are these sequences modified in expression vectors unless it is, for example, to use engineered signal sequences such as the mouse H2-K^{b} signal sequence to control stoichiometric compilation and targeting to the ER microsomes[150]. Another examination of CARs found constructs using human CD8α signal sequence for protein expression[151,152] or the *Gaussia* luciferase signal sequence from *Vargula* luciferase to maximize the soluble protein secretion[153].

5.2.4: T-cell activation and the 58$^{-/-}$ cell line

As mentioned before in **Chapter 1.2.4**, T-cells use clonotypic receptors to recognize and respond to disease antigens presented on the surface of other cells. When a TCR interacts with its cognate antigen the receptor initiates an activation cascade through peripheral receptors in the immune complex. During T-cell activation, early activation marker CD69, a C-type lectin protein becomes highly expressed on the cell surface. This

increased receptor expression down-regulates Sphingosine 1-Phosophate Receptor-1 ($S1P_1$) to prepare the cell for proliferation and further activation in the context of APCs, pathogens or apoptotic cells[154]. As an early indicator of T-cell activation, $CD69^{hi}$ cells can be stained for cytometry to show the relative reactivity of the TCR-pMHC interaction. The induced $CD69^{hi}$ expression can occur in as little as 2 – 4 hours depending on the antigen peptide[12,155].

An alternative method to detect T-cell activation involves measuring the concentration of the activation cytokine, IL-2. The production of IL-2 in activated T-cells can be measured directly using intracellular staining of permeabilized cells or indirectly by an ELISA of the activated cell culture supernatant. The predominant method to measure T-cell activation levels is the IL-2 ELISA. For activation assays involving polyclonal or transient TCR transfections, the cytokine concentration in the supernatant may be below the sensitivity of ELISA, in which case the intracellular IL-2 staining or early activation marker CD69 staining should be an appropriate substitute.

5.3: METHODS

5.3.1: Expression vector constructs for TCR display

The vector used in this study is derived from a κ-chain antibody expressing system with a CMV promoter, cloning sites EcoRI and AgeI before and after the signal sequence respectively, a HindIII site 3' of the stop codon, and a SV40 PolyA tail signal 3' to the coding region[156]. These vectors also have a CMV promotor for immediate early expression of proteins compatible with many eukaryotic cell types. A Kozak sequence with an optimal ribosome binding site was used in these vectors as well. To make cassette vectors for testing different combinations of TCR α and β chains, we cloned restriction sites SspI and

EcoO109I for each chain respectively into the linker between the variable and constant domains.

The genetic material for the constant and transmembrane regions was amplified using the RT-PCR method described in **Chapter 2.3.3** from the DO11.10 and 172.10 immortalized hybridomas in the lab. Additional CD3 and CD28 receptors were recovered from these mRNA samples in the same way. Point mutations for the different TCR variants were generated using the Kunkel mutagenesis protocol from single stranded DNA described by (Sidhu and Weiss)[157].

5.3.2 Lipofectamine transfection protocol for CHO cells

This is almost identical to the standard protocol with the Lipofectamine 2000 reagent.
Note: this transfection reagent does not work with the immortalized T-cells $58^{-/-}$

Day 1- Plate cells for transfection.

1. Trypsinize and count cells from a previously growing CHO cell culture.
2. Seed the plate with the appropriate number of cells (**Table 5.4**) with the normal media WITHOUT antibiotic to make the correct final volume.

Culture vessel	Cells	Total media	Low IgG media	Opti-MEM (2 tubes)	DNA	Lipo-fectamine-2000	Complex added	
24-Well	1.0E+05	0.5 mL	0.4 mL	50 ul	0.6 ug	1.5 ul	100 ul	*per well*
12-Well	2.0E+05	1 mL	0.8 mL	100 ul	1.2 ug	3 ul	200 ul	*per well*
6-Well	5.0E+05	2 mL	1.5 mL	250 ul	3 ug	7.5 ul	500 ul	*per well*

Table 5.4: Lipofectamine transfections materials.

This table lists the amounts of cells, DNA, Lipofectamine and culture volumes required for the different CHO transfections used in this study.

3. Return the plates to the incubator to sit overnight.

Day 2- Transfection.

4. Remove and discard media from cells. Replace with the appropriate volume of media with low IgG FBS, no antibiotic (**Table 5.4**).
5. Prepare 2 tubes with the appropriate amount of Opti-MEM reduced serum media.
6. To one tube add the appropriate amount of DNA (0.5-5 ug/ul solution) and mix.

7. Add to the second tube the appropriate amount of Lipofectamine-2000 reagent and mix.
8. Add the diluted DNA to the tube with Lipofectamine and vortex to mix.
9. Let the DNA-reagent complexes form for 15 minutes at room temperature.
10. Add the appropriate amount of complex to each well or flask and tilt gently to mix.
11. Return plates to incubator.

5.3.3: Electroporation of Hybridoma Cells

*Protocol adapted from Nath Shubhankar from the Martin Ponie lab

This protocol is for transient or stable selection of eukaryotic cells using a ratio of DNA and selection vector. For each cell type and antibiotic used a kill curve must be generated to determine the amount of time required to kill cells lacking the selection vector.

Materials:

- 4mm electroporation cuvette (BD Biosciences)
- 2 x 10^7 healthy cells for transfection
- 15µg DNA with the antibiotic resistance gene
- Pre-warmed Opti-mem (Life Technologies)
- 10 ml warmed culture without antibiotics in a T75 flask
- Selection antibiotic

Protocol:

1. Rinse cells in Optimem with at least ½ volume of medium with a final volume of 400 µl (do not exceed the metal plates of the cuvette.
2. Add DNA to the cell sample and incubate at room temperature for 10 – 15 minutes.
3. Electroporate cells using the parameters 250V and 950 µF which should have a ~27.5 ms time constant.
4. Let the cells rest for 2 – 3 minutes, dead cells will float to the top of the sample and cling to the sides.
5. Carefully remove the cells in solution to avoid the aggregate floating dead cells.
6. Gently pipette cells into the T75 flasks and place the cells in the incubator for 24 – 48 hours before adding the antibiotic to begin selection.
7. Check cell counts over the next few weeks depending on the strength and stringency of the antibiotic based on kill curves.
8. Maintain the concentration of dead cells at a minimum to promote cell growth.

5.3.4: Peptide antigen presentation and T-cell activation

Activation assay for DO11.10 hybridomas and $OVA_{323-339}$ presented by A20 cells.
T-cell IL-2 activation assays secretion and immunofluorescence staining combined (Combined from Ben, Jamie, Edith)

Materials

- OVA whole protein solution in RPMI (**200 µg/ml**)
- OVA peptide in RPMI (**20 µM or 3.55 µg/ml**) for a *MW of 1774 g/mol*
- 96-well plate

Solution Prep

Dissolve the protein and antigen peptide in RPMI and filter sterilize.
Make only enough for the study.

Cells (use >90% confluent cells)

- A20 cells (15,000 cells/well)
- DO11.10 Hybridoma cells (15,000 cells/well)
- Transfected $58^{-/-}$ cells (15,000 cells/well)

Controls

- A20 and DO11.10 s cells without peptide
- DO11.10 cells with peptide
- A20 cells incubated with peptide for ≥12 hours

Protocol

1. Put 75 µl containing 2 x 10^4 A20 cells into each well of a 96 well plate.
2. Add 25µl of RPMI or protein solution to the respective wells.
3. Incubate cells and protein in the Euk incubator for 12 hours.
4. Add 100 µl containing 3 x 10^4 Tcells to the respective wells.
5. Incubate cells in the Euk incubator for the necessary amount of time for the assay.

IL-2 release assay with ELISA

1. Coat a 96-well plate with the IL-2 capture antibody from the ELISA kit with the configuration required based on the samples and standards.
    a. Use a 1:250 dilution of the Capture antibody in the Coating buffer (100 µl/well) and incubate ON at 4°C
    b. Rinse with wash buffer before use and allow the plate to come to RT.
2. After 24-48 hour incubation of Tcells with APCs harvest 100 µl of sample well supernatant by spinning plates. Immediately freeze the supernatant or begin ELISA.
3. Add the samples to the coated plate and incubate for 2-4 hours.
4. Wash plate and add 100 µ of detection mix (detection antibody at 1:1000 and SA-HRP 1:250).
5. Incubate for 1 hour then thoroughly wash
6. Add 100 µl substrate solution and incubate until color change occurs.
7. Add 100 µl stop solution and read plate at 450 nm.

Surface and intracellular staining (~8 hours)

1. 4-6 hours before cell staining add an appropriate amount of Brefeldin A such that the final concentration in the well is 5 µg/ml (for the 1000X stock of 5 mg/ml)
2. Pellet cells and wash in PBS
3. Stain samples for surface markers for <60 min with:
   CD69 –APC-Cy7 ($_{Ex}$640nm, $_{Em}$660nm), (clone FN50, BD BioLegends)
   KJ-26-PE ($_{Ex}$561nm, $_{Em}$575nm) (clone KJ1-26 Biosciences), and
   mTCR-C$_\beta$-FITC ($_{Ex}$488nm, $_{Em}$530nm) (clone H57-597, BD Biosciences)
4. Stain for 1 hr then pellet cells, remove SN, add 100 µl **Fixation buffer**, incubate for 20-60 min
5. Wash cells and add 100 µl **Permeabilization buffer**, incubate for 20 minutes
6. Stain with intracellular antibody for 30 min with:
   IL-2 – APC ($_{Ex}$405nm, $_{Em}$450nm) (clone JES6-5H4, BioLegends)
7. Add 450 µl PBS and run sample on the cytometer with the correct detection parameters.

5.4: RESULTS

5.4.1 TCR reconstitution

To reconstitute TCR on a T-cell for measuring receptor-mediated activation, we generated a TCR display system by cloning full-length αβ TCR genes into a eukaryotic expression vector previously used for antibody expression in CHO cells. Initial transfection tests with these expression vectors in CHO cells failed to display the receptor even with positive controls which had transfection efficiencies of greater than 50 % of live cells. After a more detailed examination of the vector construct components and comparing these to literature, we discovered most TCR display studies use the native signal sequence associated with TCR variable region gene. The signal sequences for these constructs were then swapped from the mouse IgHV antibody signal sequence to the native TCR peptide for additional transfection experiments.

With this single change, CHO cell transfections with these new constructs had nearly 20 % positive staining for antibodies specific for both the DO11.10 epitope and the

TCR β-constant domain (clone H57-597, BD Bioscience) indicating surface display of the DO11.10 TCR. Contrary to previous reports, the charged residues in the TM domains of the α and β chains did not prevent TCR display on the surface of CHO cells which lack the charge-balancing CD3 co-receptors imperative for proper receptor display on thymocytes[158]. Additionally, the positively transfected CHO cells had higher surface concentrations of receptor than DO11.10 hybridoma cells based on MFI values from flow cytometry with KJ-26 staining in the positive transfected populations (**Figure 5.3A**). This display difference between DO11.10 hybridomas and CHO cells may be a result of the more regulated TCR display in thymocytes due to immune complex formation and receptor recycling mechanisms to control surface receptor concentration[149].

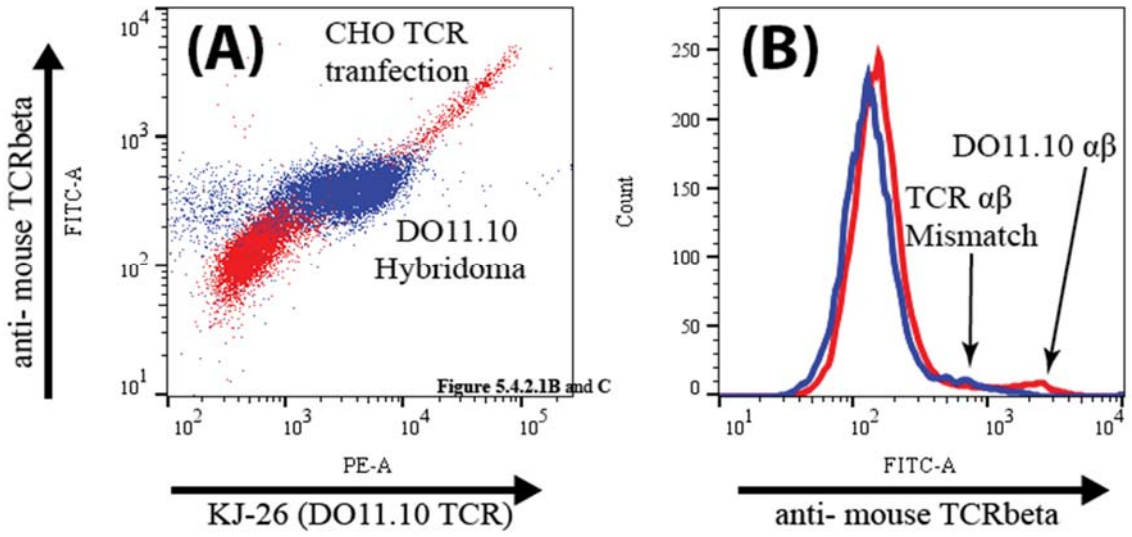


Figure 5.3: Surface concentrations of displayed TCR

(**A**) Transient transfection of the DO11.10 TCR on CHO cells shows a range of receptor concentrations. (**B**) Dramatically reduced overall display for a transfection of DO11.10α + mouse TRBV5 indicated by low staining with the anti- mouse TCRβ-constant domain.

CHO transfections with mismatched α and β chains (i.e. DO11.10α + 172.10β, or DO11.10α + TRBV5) showed decreased display of the receptor as measured by staining

of the TCR β-constant domain (**Figure 5.3B**). This result was expected as several studies have reported the interfacial instability occurring between non-native TCR chains similar to what was described in **Chapter 2.4.4** which attempted express novel pairs of TCR chains[83].

After the success of CHO cells to transiently display the various TCRs, we next tried to replicate this display on the $58^{-/-}$ T-cells. This proved harder than expected due to thymocytes being highly resistant to lipofection, the method successfully used for CHO cells. After many attempts to transfect cells with the TCR constructs, we determined generation of stable cell lines using electroporation and antibiotic selection would likely be necessary (Nath Shubhankar, personal communication).

For stable selection, the TCR chains were cloned into the antibiotic selection vectors using G418 and puromycin resistance genes. These vectors, pEGFP-N1 (GenBank Accession #U55762) and pIRESPURO (Clonetech Cat. #631619), were a generous gift from the Martin Poenie lab (**Appendices B.1.2.2**). Additionally these vectors were modified to replace EGFP with CFP, YFP, mCherry, or to remove the fluorescent protein fusion. Similarly, the pIRESPURO vector was modified for our purposes to display TCRs. After cloning full-length TCRs into the new plasmids for stable selection, electroporations were performed with the DO11.10α-EGFP or DO11.10β-EGFP. These cells grew well under selecting conditions and tracked to a two-phase selection kill curve (**Figure 5.4 black solid line**).

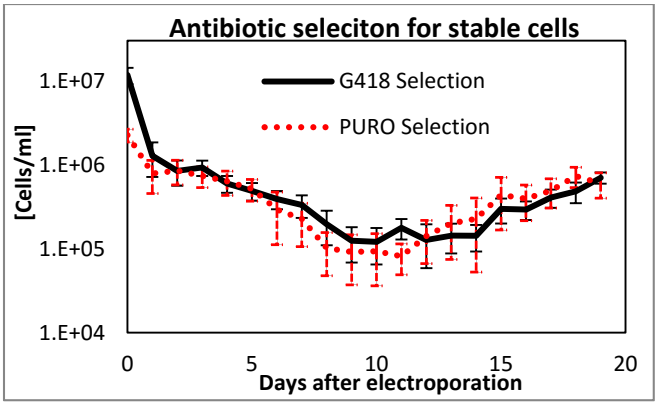


Figure 5.4: Selection curves for stable cells.

Compiled data from multiple cell selections details the expected kill curve and resistant population growth in antibiotic medium. Sorting for positive cells through either the fluorochrome fusions or antibody staining occurs once cells reach confluence and the expected positive population is over 10^6 total cells, which is approximately 3-4 weeks after initial electroporation.

After 28 days under G418 antibiotic selection, $EGFP^+$ cells were sorted with a BD FACSAria cell sorter. Stable populations continued to grow in the presence of antibiotic and maintained EGFP expression (**Figure 5.5A**). Additionally genomic DNA amplifications showed the TCR construct remained intact during selection as seen by the strong bands corresponding with TCR DNA (**Figure 5.5B**). With these encouraging results, the second chain on a plasmid with a puromycin resistance gene for each respective TCR fusion was additionally electroporated into the EGFP fusion cell lines to generate stable cell lines with both chains. Antibiotic selection tracked nearly identically to the first round of selection (**Figure 5.4 red dashed line**) with slightly lower cell counts during the selection process possibly due to puromycin toxicity. After an additional 28 days under selection with both G418 and puromycin, the cells were stained with KJ-26 to test for TCR display (**Figure 5.5C**). These samples also failed to stain for CD3ε indicating a lack of TCR α and β chains trafficked to the surface to form the immune complex.

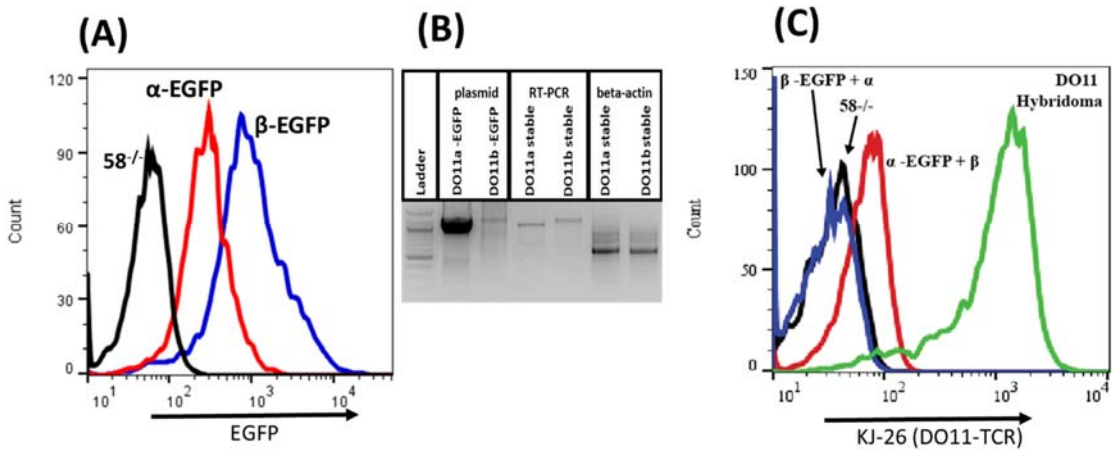


Figure 5.5: Stable cells generation and tests

(**A**) Flow cytometry plot of the transiently electroporated cells displaying the DO11.10 TCR as stained by the anti-DO11.10 antibody. The overall TCR display is limited to the expression of the chain fused to the EGFP (α-chain-EGFP fusions have a lower stability than the β-chain). (**B**) Tests for the presence of the DO11.10 TCR genes in purified genomic DNA in the antibiotic resistant samples showed positive amplification of each TCR chain. The gel image shows plasmid controls, each TCR-EGFP fusion, and a β-actin control. (**C**) After antibiotic selection with both G418 and Puromycin for selection of both TCR chains, the cell cultures failed to display TCR in concentrations similar to the DO11.10 hybridoma cells.

The lack of proper TCR display on stably selected cells indicated either a problem with the antibiotic selection or the some other problem in the proper transport of the receptor and immune complex to the receptor to the surface. Western blots with cell lysate of these populations also indicated proper formation of the TCR dimer at the 100 kDa band for the non-reduced lanes for stably transfected cells with EGFP-3X-Flag tag fusions (**Figure 5.6A**). The $58^{-/-}$ cells failed to display the receptor even with both genes thought to be stably transfected based on antibiotic selection, protein expression, and TCR heterodimer formation.

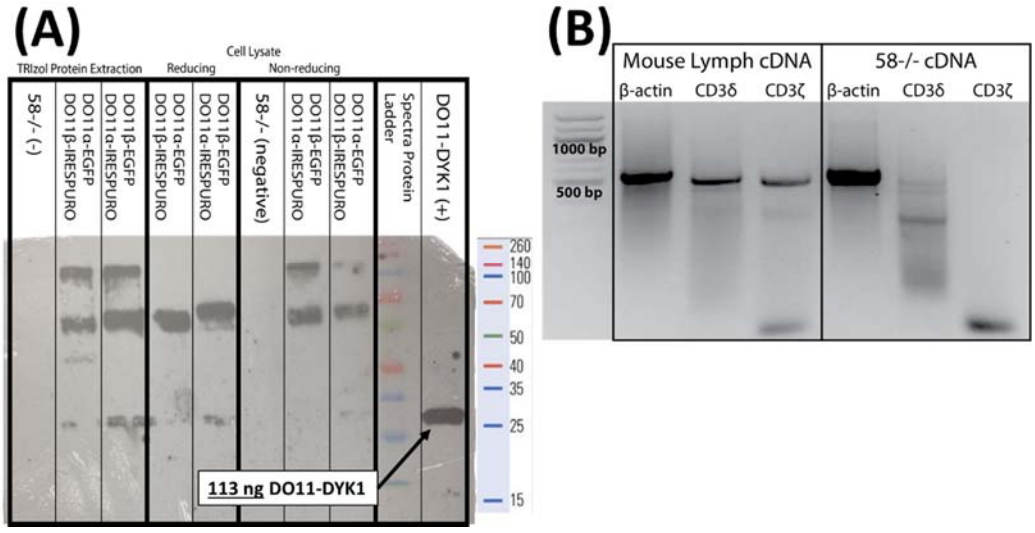


Figure 5.6: Cell lysate western blot and CD3δζ gene detection

(**A**) Cell lysate from stable cell cultures was run in a 12 % SDS-PAGE gel and transferred to a nitrocellulose membrane for a western blot detection with anti-flag peptide. The presence and size of the bands indicate the TCR αβ chains (100 kDa band in non-reducing lanes) are forming the TCR. (**B**) Comparing the presence of the CD3 mRNA between mouse T-cells and the $58^{-/-}$ hybridoma cells shows a distinctly lower amount mRNA for expression of CD3δ and CD3ζ.

The TCR αβ dimer should increase the stability of the chains to aid CD3 association and formation of the immune complex for transport to the outer cell membrane. Previous studies have mentioned a lack of CD3 co-receptor expression can significantly hinder overall display of the TCR on T-cells[159]. Interestingly Kranz *et al.* found CD4 and CD8 receptors did not significantly increase the binding and activation of cloned surface TCR to pMHC and in fact in some tests their addition decreased binding as measured by IL-2 release[160]. The cells were then tested for the presence of the CD3δ and CD3ζ co-receptors by RT-PCR (**Figure 5.6B**) (amplifications of CD3γ and CD3ε not shown). The $58^{-/-}$ cells lacked these CD3 mRNA sequences which suggest these genes may have been crippled during the generation of the cell line. Additional work is required to determine if

transfecting these two chains will reconstitute surface display in the antibiotic selected cell lines.

5.4.2 Signal sequence for receptor display and secretion

Based on BCR and TCR cellular display experimental data in the Maynard lab, it appeared that signal sequences play a significant role in both receptor expression and trafficking to the cell surface. Initial display experiments with a mouse IgHV signal sequence, failed to display BCRs and TCRs on CHO and thymocyte cells[161]. This could have been due to several reasons including: (1) lack of proper translocation to the ER lumen, (2) instability due to improper signal peptide cleavage, or (3) instability in the transmembrane domain.

Normally with antibody engineering and expression, the native signal sequence should be utilized to prevent complications to re-optimize signal sequences[162]. Even slight changes to this native signal sequences can often cripple expression and secretion of proteins (Michael Kaleko, personal correspondence). After the lack of display seen in transfected cells, we restored the native signal sequence corresponding to the TCR variable-region on the receptor chains, surface display was detected to densities on positive cells greater than in the positive control DO11.10 hybridoma cells (**Figure 5.3A** and **Figure 5.7A** and **D**). Additionally a heterogeneous signal sequence systems with at least one native signal sequence on either TCR α or β chains could properly display the receptor (**Figure 5.7B** and **C**). From these challenges and observations, we wanted to understand what signal sequence characteristics factor into this almost binary expression of membrane displayed proteins.

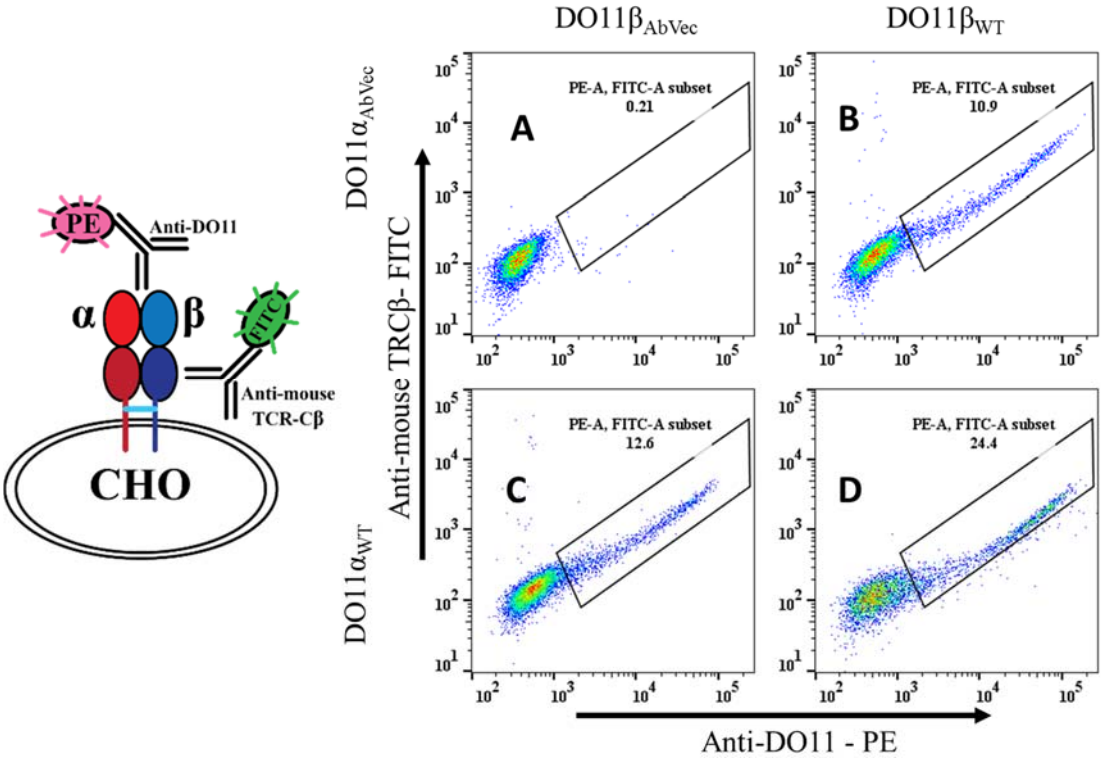


Figure 5.7: CHO cell TCR display with Ig/TCR signal sequences

Flow cytometry plots for staining with mouse TRBC-FITC and anti-DO11.10-PE on CHO transient transfections. Each transfection used two plasmids with signal sequences according to the labels on the left and top of the plots for the α- and β-chains respectively. The sample using the antibody secretion signal sequence (**A**) for both chains fails to display any TCR. All other samples with at least one wildtype signal sequence (**B**) display TCR as expected.

We designed a set of signal sequences modified from the native TRBV13-2 leader sequence with increment point mutations to lower the h-region hydrophobicity or change the polarity of c-region residues (**Table 5.5**). Additionally we fused the consensus Ig and TCR signal sequences and a novel sequence called SpMax to the N-terminus of the DO11.10 β-chain. The SpMax sequence was designed to maximize the predicted signal trafficking and peptide cleavage scores calculated by the SignalP server and reported in **Table 5.5**[138,163]. Tests with these variant leaders are currently underway. It is hypothesized to show decreases in the h-region hydrophobicity will correlate with diminished TCR

transport to the cell surface and similarly the loss of polar residues from the c-region will prevent signal peptide cleavage and negatively impact protein stability[153]. If these results show increased display in the consensus signal sequences, SpMax or any of the other variants, we will test signal sequence fusions with other TCRs and antibodies. This will determine if the sequences have utility as more general protein expression and trafficking peptides across many Ig and TCR variable regions or other ER trafficked proteins.

Variant	Sequence	Cleavage Position	Smean	D
TRBV13	No Leader	None	0.099	0.102
WT	MGSRLFFVLSSLLCSKHM	21	0.809	0.693
WT_1L	MGSRAFFVLSSLLCSKHM	21	0.724	0.610
WT_2L	MGSRAFFVASSLLCSKHM	21	0.548	0.465
WT_3L	MGSRAFFVASSALCSKHM	21	0.354	0.311
WT_4L	MGSRAFFVASSAACSKHM	21	0.233	0.221
WT_C	MGSRLFFVLSSLLSSKHM	21	0.740	0.610
AbVec SS.	MGWSCIILFLVATATGVHS	17	0.945	0.814
Ig_consensus	MGWSWIFLFLVAAPTGVHSLS	20	0.952	0.898
TCR_consensus	MLMGTRLLCWVALCLLGAGHV	24	0.954	0.941
SpMax	MLMGTLLLLLLLLLLLLAKHV	24	0.968	0.957

Table 5.5: Mouse TRBV13-2 leader variants and novel signal sequences.

This describes the signal sequences generated to test the variability of display when placed on the DO11.10 β-chain construct. Variants include a reduced h-region by removal of leucine residues, modified polarity with serine substitutions for cysteine residues (mutations shown by underline). The Ig and TCR consensus signal sequences were also cloned into the construct as well as a variant called SpMax designed to have the best SRP recognition and peptidase activity for all Ig and TCR v-region sequences as calculated by SignalP[138].

Analysis of signal sequences will help determine the characteristics necessary for proper trafficking of secreted and surface displayed proteins. Sequences with less affinity for the SRP would not be expected to translocate into the ER membrane as this process is highly regulated through both receptor proteins and transmembrane channels[141].

Additionally proteins tested with weak or cryptic cleavage-regions may cause the retention of the protein in the ER lumen if it is not targeted for degradation otherwise[164].

These experiments could also answer how certain signal sequences are selected for their capacity to direct the post-translational activities of proteins within the cells. These sequences already perform a multitude of roles with stabilizing the nascent protein, recruiting trafficking molecules for translocation, and possibly downstream stabilization during and after protein translocation through the ER membrane. If tuning the three regions of the signal peptide controls additional aspects of expressed proteins then optimization may increase the stability and yields.

5.5: CHIMERIC ANTIGEN RECEPTORS IN T-CELLS

A new class of engineered proteins called chimeric antigen receptors (CARs) have fused domains from various immuno-protein origins altering the protein's overall functionality[152]. CARs are currently under heavy development as they link the antigen binding regions of antibodies and TCRs with other stimulatory or effector domains to engineer a particular immune response[165]. As protein formulation and cellular therapeutics have advanced, highly functional immunological receptors comprising fusions of antibodies coupled with T-cell activation using CD3ζ ITAMs have proven their efficacy in a few trials[166]. The CD28 domain provides co-stimulatory interactions with CD80 and CD86 on the APC and causes the surface display of protein as a disulfide-linked homodimer. Additionally most CARs with published signal sequences use CD8 for stable display[152,167]. This preliminary work designed (**Figure 5.8A**) and expressed a DO11.10 second-generation CAR on $58^{-/-}$ (**Figure 5.8B**) and CHO (**Figure 5.8C**) cells, but similar to previously described results the transiently transfected cells could not be activated.

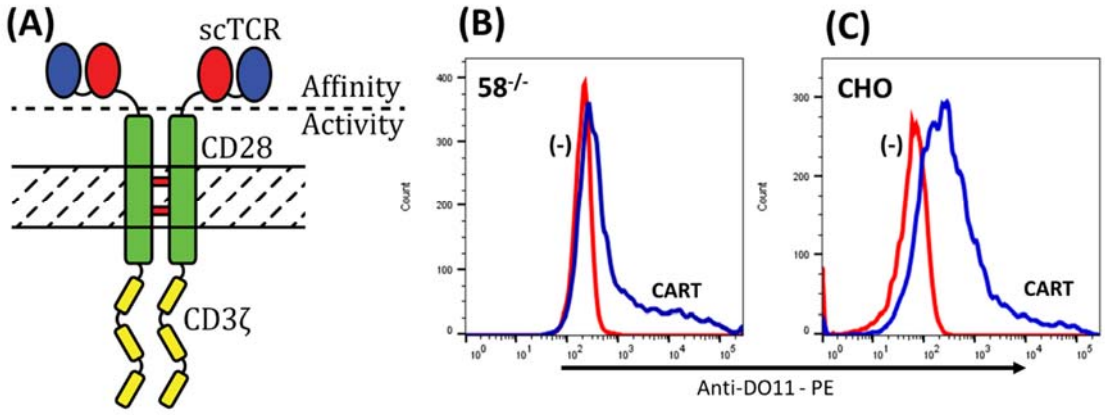


Figure 5.8: DO11.10 CART schematic and surface display.

(A) Diagram of the DO11.10 CART construct generated with scTCR, CD28, and CD3ζ fusion, (B) Histogram of CART display on the surface of CHO and $58^{-/-}$ cells showing proper surface display of the receptor.

5.6: DISCUSSION AND CONCLUSIONS

TCR engineering *in vitro* requires the reconstitution of the receptor onto the T-cell surface to test for any alterations in the activation of the receptor. Transient transfections with the TCR display constructs showed stable display on CHO cells, but failed to transiently transfect $58^{-/-}$ cells for an activation assay. Antibiotic selection for stable cell lines using G418 and puromycin also failed to display TCR at levels comparable to DO11.10 hybridomas. This was even after extensive antibiotic selection and subsequent sorting.

A deeper analysis of literature found some problems with the $58^{-/-}$ cell line which may cause the malformation of the immune complex and lack of TCR display seen in these tests. There are several conflicting reports about the cell line's ability to display particular TCR constructs as a complete immune complex. One study found $CD3\zeta^{-/-}$ mice have abolished TCR display shown in a figure that nearly identically matched the flow cytometry histogram seen in the DO11.10α-EGFP + β-PURO sample of **Figure 5.4.1.2C**

and other studies have noted the need to co-transfect CD3ζ or CD4 to reestablish the immune complex[159,168]. Meanwhile other researchers have shown in a variety of experiments the ability to stably transfect TCRs in $58^{-/-}$ cells for activation studies similar to the one attempted in this report without the additional CD receptor transfections[169–171].

Further experimentation will be required to increase the transfection efficiency for more robust TCR display on the $58^{-/-}$ T-cells. These experiments could include transfections with cloned full-length CD3 chains to reestablish the receptor expression required for the immune complex[172]. Additional transfection methods such as nucleofection and viral transduction could bypass the low efficiencies seen with electroporation. Alternatively, a separate activated T-cell line already expressing an immune complex could be transfected with our TCR constructs for duel TCR display, but with some complications if there the native TCR interferes with the transfected receptor. These efforts to display engineered TCRs on the surface of T-cells for activity assays has led to many interesting investigations on the myriad of complicated and important aspects for TCR expression in thymocytes and protein production in eukaryotic systems.

The current understanding of protein expression with SRP-mediated translocation is somewhat limited to certain a few classes of proteins and model system. The complications with surface display due to the signal sequences in this study show there is opportunity to optimize commonly used signal sequences on commercial vectors for broader applications with more general proteins. We hypothesize the SpMax, Ig, and TCR will increase display but with variability due to the different h-regions, including the two central cysteine residues in the TCR consensus variant. Finally, we expect the TRBV13-2L variants will have diminished display as their signal sequences depart from the established n-, h-, c-region motif.

Chapter 6: Future directions and concluding remarks

In this report, we present evidence linking TCR sequence, structure, and function to help elucidate the requirements for the robust recognition and activation that is characteristic of these molecules in the adaptive immune system. The high throughput sequencing experiments provided an opportunity to understand the types of analytical methods needed to probe TCR repertoires for background composition and changes due to antigen immunizations (**Chapter 2**). In addition, an examination of structural repertoires in antibodies and TCRs for conserved framework interfaces, CDR loops, aligned residue sequences, and key residues found residue sequence-based correlations which can guide structure modeling for novel receptor proteins (**Chapters 3 and 4**). Finally, we applied the insights of the sequence and structure analyses from the previous chapters to test the *in vivo* characteristics of reconstituted engineered TCRs (**Chapter 5**).

6.1: HIGH-THROUGHPUT SEQUENCING FOR DISEASE RESPONSIVENESS

While the data from the HTS study described in **Chapter 2** did not show a definitive antigen-specific shift in the TCR repertoire, this type of analysis to sort and screen sequences is important for determining the health of the immune system. In recent years, numerous studies have used HTS to study the many facets of the immune system and increased its utility as a platform technology[41,53,72]. Monitoring the development of T-cell repertoires and analyzing memory inflation with HTS in chronically infected individuals can both identify rare antigen specific T-cells and characterize antigen recognition within inflated T-cell populations. Identifying TCR sequences selected for memory cell maturation also gives insight into various components of disease progression leading to a narrowed TCR repertoire and memory inflation, which decreases immune responsiveness.

An investigation by Holtappels *et al.* showed the accumulation of public $CD8^+$ CMV-specific TCRs occurring in multiple chronically infected individuals[173]. Leveraging these public sequences on $CD8^+$ cells and transfecting them into $CD4^+$ cells can enhance the overall immune response by recruiting more diverse CMV-specific T-cells to sites of infection[52]. Production of these diverse T-cell responses is imperative for prolonged protection against the diverse antigen landscape. Correlating these public TCR repertoires with the corresponding MHCs also reveals which HLA genes present antigens to generate the best protection against infections[72,174].

While we ultimately chose not to pursue further HTS experiments at this time, we have gained the following insights into requirements for any future repertoire sequencing efforts. Developing finer cell isolation processes, gene amplification techniques, and sequence analysis algorithms will increase the utility of HTS as a platform technology. A comprehensive study of repertoire changes due to antigen challenges and environmental exposures allows for improved vaccines and immunotherapeutics for initiating proper immune responses. These advancements will allow scientists to analyze and characterize individual immune systems nearly down to the cellular level to identify and correct for gaps in the repertoire.

6.2: MODELING WITH STRUCTURAL REPERTOIRES

Characterization of TCR-pMHC interactions provides a quantitative insight into the structural mechanisms involved with peptide recognition and downstream activation. The conserved structural features described here with key residues and canonical classes, will allow us to assign structural meaning to the affinity-maturated variants selected in directed evolution studies or from HTS experiments. Structural modeling of antibodies and TCRs can help generate more effective vaccines to prevent viral escape as well as

immunotherapeutics to direct receptor specificity away or towards self-peptides in cases of autoimmune disorders and cancers respectively[123].

While the work discussed in **Chapter 3** found structural classifications in TCRs for the majority of CDR1 and CDR2 loops, much work remains to delineate the structural prediction capabilities for the hypervariable CDR3. This longer and more flexible CDR loop provides the majority of the important peptide contacts during TCR-pMHC interactions. In addition more structural analysis should be performed to elucidate how during pMHC engagement, the CDR3 translates the recognition signal through β-strand G to the constant domain and CD3 co-receptors[12].

The supplemental structural analysis in **Chapter 4** found potential areas for improvement in the modeling capacity of the OptCDR program[103]. By using a hybrid-modeling program to build novel CDR loops on a structurally well-defined framework, the program treated the CDRs as flexible transposable elements. This study found the modeling software should instead leverage key residues and canonical class conformations found in the majority of antibodies and TCRs to design the antigen-binding region *in silico*[23,25]. In this way, with enough engineering to control the chain interfaces and β-sheet composition a program could possibly design a universal framework to present, orient, and generate most antigen-binding regions.

6.3: FUTURE DIRECTIONS WITH TCR THERAPEUTICS

The success of adoptive transfer therapies with the antigen-binding regions of antibodies and TCRs has sparked the development of a suite of new cellular and receptor-mediated therapies[14]. These types of treatments use modified T-cells to stimulate the activity of the patient's own immune system. This also has the added benefit of producing modified memory cells for prolonged treatment efficacy.

Engineered BCRs, TCRs, and CARs combines affinity-maturated receptors with the biological activity associated with immunoproteins in B-cells and T-cells. For example, transfecting engineered TCRs with cancer specificity into T-cells can bypass negative selection since these self-peptide specific receptors are normally removed during T-cell maturation[167]. Another soluble CAR technology uses a single chain protein with both a CMV-specific scTCR fused to an IgG1 heavy chain Fc region to label infected cells for destruction[175]. This platform technology could be applied to a variety of soluble CTL receptors to target a range pMHCs for an improved ability to treat infections without possible harmful side effects. Further engineering of these receptors to fuse various domains together for their activities short-circuits signaling pathways to circumvent cell regulators that might dampen a desired immune response[176]. The preliminary work in **Chapter 5** for displaying engineered TCR constructs to study the effect of *in vitro* affinity maturation on T-cell activity provides deeper understanding into the complex nature cellular engineering. In addition, the design of signal sequence variants for tuned TCR and CAR display can further modify the activity these receptor-displaying cells.

6.4: CONCLUDING REMARKS

The critical role of TCRs in the immune system makes them an excellent target for protein engineering to aid disease response through affinity maturation, chimeric fusions, and adoptive cellular therapies. While antibody-based therapies have been the subject of research for decades, TCR-based therapies are relatively new and the research landscape remains fairly open. Linking the various aspects of TCRs discussed here can allow a more comprehensive macro- and micro-analysis of individual receptors and repertoires. With this greater understanding of TCR-pMHC complex, advancements in repertoire

sequencing, crystallography, and cellular manipulations will continue to allow scientists to make new discoveries and develop therapies to tackle challenging immune disorders.

Appendices

A: Selected Methods

A.1: Cell Suspension Preparation

1. Isolate spleens from mouse. Store in each spleen in 3 ml DMEM-5 on ice
2. Cut spleens with scissors in several places with EtOH washed scissors
3. Transport spleens back to the lab on ice
4. Disrupt tissue structure with the back of a sterile syringe grinding against the petri dish until mostly fibrous tissue remains
5. Draw up cells in media with needle/syringe and expel through a cell strainer into a 50 ml centrifuge tube
6. Wash petri dish with ~4 ml of DMEM-5, pass through strainer into centrifuge tube
7. Centrifuge 5 min at 250 x g, discard supernatant and store on ice until ready for red blood cell lysis.

A.2: M13 Phage Protocols

Protocols adapted from from Sidhu and Weiss (2004)

Make sure to use the correct phagemids **(MoPAC24)** and **XL1-Blue** or (another **F pilus**$^{+}$ line) cells cell lines that can produce phage and become infected by the M13 phage used in the lab. To further increase yields add **%1 glucose** to the media to give cells more energy as they consume sugar for the lac operon. Create M13 phage displaying the protein of interest. Used in studies with TCR:pMHC interactions. MoPAC24 has GP3 linker region attached to the scFv, the M13 grabs onto the GP3 and displays up to three scFv proteins on the phage surface. *Good yields with 100 ml of culture*. This protocol is also found in the Weiss paper #6.

A.2.1: Phage Production

1. Start an overnight 2 ml culture of either E15 or XL1Blue with the "MoPAC" plasmid containing the scFv of interest
2. Increase the culture to 100 ml and place in the 37°C shaker until an OD of ~0.5
3. Induce with 1 mM IPTG, incubate at room temperature for 1.5 hours
4. Add helper phage to final concentration of 10^{10} phage/ml, incubate at room temperature with no shaking for 15' then incubate for 1 hour with shaking.
5. Add Kanamycin and place in the 37°C shaker overnight.

A.2.2: Phage Purification

Harvest phage particles and determine the titer concentration.

1. Centrifuge culture for 10' at 10 krpm and 4°C
2. Transfer SN to a fresh tube and add 1/5 vol PEG/NaCl solution to precipitate the phage and incubate for 5' at room temperature
3. Centrifuge for 10' at 10 krpm and 4°C and aspirate all SN.
4. Resuspend the pellet in 1/20 volume of PBS. Pellet insoluble matter by centrifuging for 5' at 15 krpm and 4°C. Transfer SN to clean tube.
5. Estimate [phage] spectrophotometrically (OD_{268}= 1.0 for 5 x 10^{12} phage/ml)
6. Use immediately or store at -80°C

A.2.3: dU-ssDNA Isolation

Constructing Phage display libraries by oligonucleotide-directed mutagenesis. Phage Display. Eds. Clackson and Lowman

1. Transfer 1 colony of CJ236 with phagemid to 1mL TB + Amp, incubate at 37°C 250rpm 6-8h.
2. Add helper phage to ~10^{10} virons/mL (10µL at 10^{13-14} virons/mL). Incubate at RT 15min.
3. Transfer to 30 mL 2xYT + Amp + uridine (0.25µg/mL). Incubate at 37°C 250 rpm ON.

4. Centrifuge 15,000 rpm 10m in 4°C. Transfer SN to new tube with 1/5th vol. 20% (w/v) PEG-8000/2.5M NaCl. Incubate on rotor 4°C 20 min.
5. Harvest phage at 10k rpm 10min 4°C. Pour off SN and invert on paper towel to dry ~5min
6. Resuspend in 500µL PBS (pH 7.4), transfer to tube. Spin max 5min and transfer SN to new tube. Repeat to remove all cell debris (1X repeat typically OK).
7. Add 7µL Buffer MP (QIAgen Spin M13 Kit), mix and incubate at rt 2min
8. Add to spin column, spin 8,000rpm 15s, discard FT
9. Add 700µL Buffer MLB, repeat spin and discard FT
    a. Repeat with additional 1min incubation at rt
10. Add 700µL Buffer PE, repeat spin and discard FT
    a. Repeat and perform final spin 8,000rpm 30s to remove residual EtOH
11. Transfer to new tube, add 100µL Buffer EB. Incubate at rt 10min. Elute at 8,000rpm 30s.
12. Spec to determine concentration and run on gel.
    a. Should provide >20ug of dU-ssDNA.

B: OLIGO-NUCLEOTIDE AND PLASMID SEQUENCES

B.1: High-throughput sequencing

Mouse	Pre Sorting [cells/ml]	Post Sorting [cells/ml]
180	$1.79 * 10^7$	$1.6 * 10^6$
181	$2.00 * 10^7$	$2.1 * 10^6$
178	$1.45 * 10^7$	$1.4 * 10^6$
200	$2.22 * 10^7$	$2.2 * 10^6$

Table B.1: Cell concentrations for the various samples

T-cell counts taken before and after CD4 and CD8 sorting

Mouse	[RNA] (ng/µl)	260/280	260/230
178 Elution 1	475	2.15	1.84
178 Elution 2	45.2	2.01	0.38
200 Elution 1	401.3	2.13	1.49
200 Elution 2	53.7	1.85	0.19

Table B.2: Amount of sorted cDNA collected from the mice

After harvesting RNA using Triazol and Qiagen kit the yields are as follows

	$CD4^+$ Sorted Cells			**$CD8^+$ Sorted Cells**		
Mouse	**tRNA [ng/µl]**	**260/280**	**260/230**	**tRNA [ng/µl]**	**260/280**	**260/230**
1	57.7	1.93	0.97	60.9	1.93	1.12
2	72.3	1.86	0.31	73.1	1.79	0.40
3	62.8	1.94	0.70	70.1	1.89	0.74
4	105.8	1.86	0.28	87.7	1.92	0.84

Table B.3: Amount of sorted cDNA collected from the mice

This table shows the quality and yield of total RNA from the four mice in the larger sequencing experiment.

B.1.2: Generated HTS Variants

HTS Variant #1

```
   1: G  L  L  L  L  A  A  Q  P  A  M  A  E  A  A  V  T  Q  S  P  R  N  K  V  A  V  T  G  E  K
      GGATTGTTATTACTCGCGGCCCAGCCGGCCATGGCGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCAGTAACAGGAGAGAAG  90
   1: V  T  L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L  R  L  I  Y  Y
      GTGACATTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTGAGGCTGATCTACTAT 180
   1: S  Y  G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N  F  S  L  T  L
      TCATATGGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCTAGACCAAGCCAAGAGAACTTCTCCCTCACTCTG 270
   1: E  S  A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T  N  T  E  V  F  F  G  K  G  T
      GAGTCGGCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACAAACACAGAAGTCTTCTTTGGTAAAGGAACC 360
   1: R  L  T  V  V  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  E  Q  V  E  Q
      AGACTCACAGTTGTAGGTGGAGGCGGTTCAGGCGGAGGTGGCTCCGGAGGCGGTGGTTCTGGAGGAGGAGGATCCGAGCAGGTGGAGCAG 450
   1: R  P  P  H  L  S  V  R  E  G  D  S  A  V  I  T  C  T  Y  T  D  P  N  S  Y  Y  F  F  W  Y
      CGCCCTCCTCACCTGAGTGTCCGGGAGGGAGACAGTGCCGTTATCACCTGCACCTACACAGACCCTAACAGTTATTACTTCTTCTGGTAC 540
   1: K  Q  E  P  G  A  S  L  Q  L  L  M  K  V  F  S  S  T  E  I  N  E  G  Q  G  F  T  V  L  L
      AAGCAAGAGCCGGGGGCAAGTCTTCAGTTGCTTATGAAGGTTTTCTCAAGTACGGAAATAAACGAAGGACAAGGATTCACTGTCCTACTG 630
   1: N  K  K  D  K  R  L  S  L  N  L  T  A  A  H  P  G  D  S  A  A  Y  F  C  A  V  M  P  N  Y
      AACAAGAAAGACAAACGACTCTCTCTGAACCTCACAGCTGCCCATCCTGGGGACTCAGCCGCGTACTTCTGCGCGGTGATGCCGAACTAT 720
   1: N  V  L  Y  F  G  S  G  S  K  L  T  V  E  P  A  A  S  G  A  D  H  H  H  H  H  H  *  *  A
      AACGTGCTGTATTTTGGCAGCGGCAGCAAACTGACCGTGGAACCGGCGGCCTCGGGGGCCGATCACCATCATCACCATCATTAGTAAGCT 810
```

GLLLLAAQPAMAEAAVTQSPRNKVAVTGEKVTLSCNQTNNHNNMYWYRQDTGHGLRLIYYSYGAGSTEKGD
IPDGYKASRPSQENFSLTLESATPSQTSVYFCASGSGTTNTEVFFGKGTRLTVVGGGGSGGGGSGGGGSGG
GGSEQVEQRPPHLSVREGDSAVITCTYTDPNSYYFFWYKQEPGASLQLLMKVFSSTEINEGQGFTVLLNKK
DKRLSLNLTAAHPGDSAAYFCAVMPNYNVLYFGSGSKLTVEPAASGADHHHHHH*

HTS Variant #2

```
   1: L  L  L  A  A  Q  P  A  M  A  E  A  A  V  T  Q  S  P  R  N  K  V  A  V  T  G  E  K  V  T
      TTATTACTCGCGGCCCAGCCGGCCATGGCGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCAGTAACAGGAGAGAAGGTGACA 90
   1: L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L  R  L  I  Y  Y  S  Y
      TTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTGAGGCTGATCTACTATTCATAT 180
   1: G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N  F  S  L  T  L  E  S
      GGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCTAGACCAAGCCAAGAGAACTTCTCCCTCACTCTGGAGTCG 270
   1: A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T  N  T  E  V  F  F  G  K  G  T  R  L
      GCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACAAACACAGAAGTCTTCTTTGGTAAAGGAACCAGACTC 360
   1: T  V  V  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  E  Q  V  E  Q  R  P
      ACAGTTGTAGGTGGAGGCGGTTCAGGCGGAGGTGGCTCCGGAGGCGGTGGTTCTGGAGGAGGAGGATCCGAGCAGGTGGAGCAGCGCCCT 450
   1: P  H  L  S  V  R  E  G  D  S  A  V  I  T  C  T  Y  T  D  P  N  S  Y  Y  F  F  W  Y  K  Q
      CCTCACCTGAGTGTCCGGGAGGGAGACAGTGCCGTTATCACCTGCACCTACACAGACCCTAACAGTTATTACTTCTTCTGGTACAAGCAA 540
   1: E  P  G  A  S  L  Q  L  L  M  K  V  F  S  S  T  E  I  N  E  G  Q  G  F  T  V  L  L  N  K
      GAGCCGGGGGCAAGTCTTCAGTTGCTTATGAAGGTTTTCTCAAGTACGGAAATAAACGAAGGACAAGGATTCACTGTCCTACTGAACAAG 630
   1: K  D  K  R  L  S  L  N  L  T  A  A  H  P  G  D  S  A  A  Y  F  C  A  V  M  P  N  Y  N  V
      AAAGACAAACGACTCTCTCTGAACCTCACAGCTGCCCATCCTGGGGACTCAGCCGCGTACTTCTGCGCGGTGATGCCGAACTATAACGTG 720
   1: L  Y  F  G  S  G  S  K  L  T  V  E  P  A  A  S  G  A  D  H  H  H  H  H  H  *  *  A  *  P
      CTGTATTTTGGCAGCGGCAGCAAACTGACCGTGGAACCGGCGGCCTCGGGGGCCGATCACCATCATCACCATCATTAGTAAGCTTGACCT 810
```

LLLAAQPAMAEAAVTQSPRNKVAVTGEKVTLSCNQTNNHNNMYWYRQDTGHGLRLIYYSYGAGSTEKGDIP
DGYKASRPSQENFSLTLESATPSQTSVYFCASGSGTTNTEVFFGKGTRLTVVGGGGSGGGGSGGGGSGGGG
SEQVEQRPPHLSVREGDSAVITCTYTDPNSYYFFWYKQEPGASLQLLMKVFSSTEINEGQGFTVLLNKKDK
RLSLNLTAAHPGDSAAYFCAVMPNYNVLYFGSGSKLTVEPAASGADHHHHHH*

HTS Variant #3

```
   1: L  L  L  A  A  Q  P  A  M  A  E  A  A  V  T  Q  S  P  R  N  K  V  A  V  T  G  E  K  V  T
      TTATTACTCGCGGCCCAGCCGGCCATGGCGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCAGTAACAGGAGAGAAGGTGACA 90
   1: L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L  R  L  I  Y  Y  S  Y
      TTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTGAGGCTGATCTACTATTCATAT 180
   1: G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N  F  S  L  T  L  E  S
      GGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCTAGACCAAGCCAAGAGAACTTCTCCCTCACTCTGGAGTCG 270
   1: A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T  N  T  E  V  F  F  G  K  G  T  R  L
      GCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACAAACACAGAAGTCTTCTTTGGTAAAGGAACCAGACTC 360
   1: T  V  V  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  E  Q  V  E  Q  R  P
      ACAGTTGTAGGTGGAGGCGGTTCAGGCGGAGGTGGCTCCGGAGGCGGTGGTTCTGGAGGAGGAGGATCCGAGCAGGTGGAGCAGCGCCCT 450
   1: P  H  L  S  V  R  E  G  D  S  A  V  I  T  C  T  Y  T  D  P  N  S  Y  Y  F  F  W  Y  K  Q
      CCTCACCTGAGTGTCCGGGAGGGAGACAGTGCCGTTATCACCTGCACCTACACAGACCCTAACAGTTATTACTTCTTCTGGTACAAGCAA 540
   1: E  P  G  A  S  L  Q  L  L  M  K  V  F  S  S  T  E  I  N  E  G  Q  G  F  T  V  L  L  N  K
      GAGCCGGGGGCAAGTCTTCAGTTGCTTATGAAGGTTTTCTCAAGTACGGAAATAAACGAAGGACAAGGATTCACTGTCCTACTGAACAAG 630
   1: K  D  K  R  L  S  L  N  L  T  A  A  H  P  G  D  S  A  A  Y  F  C  A  V  S  R  N  N  N  N
      AAAGACAAACGACTCTCTCTGAACCTCACAGCTGCCCATCCTGGGGACTCAGCCGCGTACTTCTGCGCGGTGAGCCGCAACAACAACAAC 720
   1: R  I  F  F  G  D  G  T  Q  L  V  V  K  P  A  A  S  G  A  D  H  H  H  H  H  H  *  *  A  *
      CGCATTTTTTTTGGCGATGGCACCCAGCTGGTGGTGAAACCGGCGGCCTCGGGGGCCGATCACCATCATCACCATCATTAGTAAGCTTGA 810

LLLAAQPAMAEAAVTQSPRNKVAVTGEKVTLSCNQTNNHNNMYWYRQDTGHGLRLIYYSYGAGSTEKGDIPDGYKASRPSQENFSLTLE
SATPSQTSVYFCASGSGTTNTEVFFGKGTRLTVVGGGGSGGGGSGGGGSGGGGSEQVEQRPPHLSVREGDSAVITCTYTDPNSYYFFWY
KQEPGASLQLLMKVFSSTEINEGQGFTVLLNKKDKRLSLNLTAAHPGDSAAYFCAVSRNNNNRIFFGDGTQLVVKPAASGADHHHHHH*
```

HTS Variant #4

```
   1: L  L  L  A  A  Q  P  A  M  A  E  A  A  V  T  Q  S  P  R  N  K  V  A  V  T  G  E  K  V  T
      TTATTACTCGCGGCCCAGCCGGCCATGGCGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCAGTAACAGGAGAGAAGGTGACA 90
   1: L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L  R  L  I  Y  Y  S  Y
      TTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTGAGGCTGATCTACTATTCATAT 180
   1: G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N  F  S  L  T  L  E  S
      GGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCTAGACCAAGCCAAGAGAACTTCTCCCTCACTCTGGAGTCG 270
   1: A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T  N  T  E  V  F  F  G  K  G  T  R  L
      GCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACAAACACAGAAGTCTTCTTTGGTAAAGGAACCAGACTC 360
   1: T  V  V  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  G  G  G  G  S  Q  N  V  Q  Q  S  P
      ACAGTTGTAGGTGGAGGCGGTTCAGGCGGAGGTGGCTCCGGAGGCGGTGGTTCTGGAGGAGGAGGATCCCAGAATGTGCAGCAGAGCCCG 450
   1: E  S  L  I  V  P  E  G  A  R  T  S  L  N  C  T  F  S  D  S  A  S  Q  Y  F  W  W  Y  R  Q
      GAATCCCTCATTGTCCCAGAGGGAGCCAGGACCTCTCTCAACTGCACTTTCAGTGACAGTGCTTCTCAGTATTTCTGGTGGTACAGACAG 540
   1: H  S  G  K  A  P  K  A  L  M  S  I  F  S  N  G  E  K  E  E  G  R  F  T  I  H  L  N  K  A
      CATTCTGGGAAAGCCCCCAAGGCACTGATGTCCATCTTCTCCAATGGTGAAAAAGAAGAAGGGAGATTCACAATTCACCTCAATAAAGCC 630
   1: S  L  H  F  S  L  H  I  R  D  S  Q  P  S  D  S  A  L  Y  L  C  A  A  N  N  Y  A  X  G  L
      AGTCTGCATTTCTCCCTGCACATCAGAGACTCCCAGCCCAGTGACTCTGCTCTCTATCTCTGCGCGGCGAACAACTATGCGCANGGCCTG 720
   1: T  F  G  L  G  T  R  V  S  V  F  P  A  A  S  G  A  D  H  H  H  H  H  H  *
      ACCTTTGGCCTGGGCACCCGCGTGAGCGTGTTTCCGGCGGCCTCGGGGGCCGATCACCATCATCACCATCATTAG
 721 ---------!---------!---------!---------!---------!---------!---------!----- 795

LLLAAQPAMAEAAVTQSPRNKVAVTGEKVTLSCNQTNNHNNMYWYRQDTGHGLRLIYYSYGAGSTEKGDIPDGYKASRPSQENFSLTLESATPSQ
TSVYFCASGSGTTNTEVFFGKGTRLTVVGGGGSGGGGSGGGGSGGGGSQNVQQSPESLIVPEGARTSLNCTFSDSASQYFWWYRQHSGKAPKALM
SIFSNGEKEEGRFTIHLNKASLHFSLHIRDSQPSDSALYLCAANNYAXGLTFGLGTRVSVFPAASGADHHHHHH*
```

B.2: Designs for TCR display constructs

Primers to add the restriction sites (EcoRI/AgeI/HindIII) at either end of the inserts:
Initial primers to amplify the coding regions of the mRNA include:

Name	Tm	Sequence	Source
Va14-3 Leader	54.6	ATG GAC AAG ATC CTG ACA GC	http://www.imgt.org/ligmdb/annotation_view?id=L77154
Vb13-2 Leader	57.2	ATG GGC TCC AGG CTC TTC	http://www.imgt.org/ligmdb/annotation_view?id=M15617
Cα Exon3 3'	57.8	CTC AAC TGG ACC ACA GCC TC	http://www.imgt.org/ligmdb/annotation_view?id=M64239
Cβ*02 Exon	50.9	GTC TCA GGA ATT TTT TTT CTT GAC	http://www.imgt.org/ligmdb/annotation_view?id=M26060
Vα5D-4			http://www.imgt.org/cgi-bin/IMGTlect.jv AC005855

EcoRI site with ***Kozak sequence*** as well the start of the murine IgG1 signal peptide; derived from DEC205/IgG fusion protein in GenBank Accession Number DQ407610. This maintains the TCR Signal sequence before coding the V-region.

AbVec	CGA TTG AAT TCC ACC ATG GGA TGG TCA TGT ATC ATC CTT TTT CTA R L N S T M G W S C I I L F L
Va14 EcoRI	GA TTG AAT TCC ACC ATG GAC AAG ATC CTG
Va14 seq	TTC AGT CTA GGA GGA ATG GAC AAG ATC CTG
Va5D-4 EcoRI	ATT TAA TTG AAT TCC ACC ATG AAA ACA TAC GCT CC
Va5D-4 seq	ATT TAA TTG GGA AGA GCA ATG AAA ACA TAC GCT CCT ACA
Vb13 EcoRI	GA TTG AAT TCC ACC ATG GGC TCC AGG CTC
Vb13 seq	TGC CTT GGT CCC AAG ATG GGC TCC AGG CTC
pEGFP-N1	*TAT CGA TGA ATT CTC ACC ATG GGA TGG TCA TGT ATC ATC CTT TTT CTA* R * N L T M G W S C I I L F L

AgeI site at the end of the murine IgG1 signal peptide; derived from DEC205/IgG fusion protein in GenBank Accession Number DQ407610

AbVec	ACT GCA ACC GGT GTA CAC TCG GAT ATC CAG ATG A T A T G V H S
Va 14 AgeI	T GCA ACC GGT GTA CAC TCG GTG AGT GGC CAG CAG G
Va sequence	CTT CTA GGC CTT CAC CTA GCT GTG AGT GGC CAG CAG GAG
Vb13 AgeI	T GCA ACC GGT GTA CAC TCG CAC ATG GAG GCT GCA G
Vb 13-2 AgeI	CTC TCC ACC GGT GTA CAC TCG AAA CAC ATG GAG GCT GCA G
Vb mRNA seq	TG CTC TCC AGT CTC CTG TGT TCA AAA CAC ATG GAG GCT GCA GTC ACC C
Vb IMGT seq	CTC TCC AGT CTC CTG TGT TCA CAC ATG GAG GCT GCA GTC
Va 5D-4 AgeI	TTT CTA ACC GGT GTA CAC TCG **GAG CAG GTG GAG CAG CTT CCT TCC A**
Va 5D-4 seq	TTT CTA TGG CTG CAG CTG GAT **GAG CAG GTG GAG CAG CTT CCT TCC A**
Vb 5 AgeI	ACT GCA ACC GGT GTA CAC TCG AAC ACT AAA ATT ACT CAG TCA CC

HindIII site C-terminal of the Constant region; derived from DEC205/IgG fusion protein in GenBank Accession Number DQ407610

3'-Vector	TGA AGC TTG GCC GCC
Ca sequence	ACG CTG AGG CTG TGG TCC AGT TGA GGT CTG CAA GAC TGA CA
Ca HindIII	CG CTG AGG CTG TGG TCC AGT TGA AGC TTG CAA GAC TGA C
Ca XmaI	ACG CTG AGG CTG TGG TCC AGT TCC CGG G*TG GCC GCC (AbVec)*
VbC*01 seq	G GCT ATG GTC AAA AGA AAG AAT TCA TGA AGT CAG ATG TGA A
Cb01 HindIII	G GCT ATG GTC AAA AGA AAG AAC TCA TGA AGC TTG ATG TGA A
Cb02 HindIII	AA AAT TCC TGA AGC TTA CTT TTA TGC
VbC*02 seq	CAG GTC AAG AAA AAA AAT TCC TGA GAC AAA CTT TTA TGC ATC CTG AGC
Cb02 HindIII	G GTC AAG AAA AAA AAT TCC TGA AGC TTA CTT TTA TGC ATC CTG AGC
Cb XmaI	G GCC ATG GTC AAR ARA AAA AAT TCC CGG GTC GAG AAG

377 (same as 378 with one shift ATGAAGCTTG)
CGGTTCTATCGATTGAATTCCACCATGGGATGGTCATGTATCATCC……TTCTAGTAGCAACTGCAACCGGT
GTACACTCGGATATCC…GAGAGTGTTAGAAGCTTGGCCGCCAT

Table B.4: Descriptions of the display vector cloning sites.

B.2.1: Model TCRs 172 and DO11.10

172 α sequence (Leader-V-J-Constant):

```
 M  D  K  I  L  T  A  S  F  L  L  L  G  L  H  L  A  V  S  G  Q  Q  E  K  R  D  Q  Q  Q  V
ATGGACAAGATCCTGACAGCATCGTTTTTACTTCTAGGCCTTCACCTAGCTGTGAGTGGCCAGCAGGAGAAACGTGACCAGCAGCAGGTG

 R  Q  S  P  Q  S  L  T  V  W  E  G  E  T  A  I  L  N  C  S  Y  E  N  S  A  F  D  Y  F  P
AGACAAAGTCCCCAATCTCTGACAGTCTGGGAAGGAGAGACCGCAATTCTGAACTGCAGTTATGAGAACAGTGCTTTTGACTACTTCCCA

 W  Y  Q  Q  F  P  G  E  G  P  A  L  L  I  S  I  L  S  V  S  D  K  K  E  D  G  R  F  T  I
TGGTACCAGCAGTTCCCTGGGGAAGGTCCCGCTCTCCTGATATCCATACTTTCAGTGTCCGATAAAAAGGAAGATGGACGATTCACAATC

 F  F  N  K  R  E  K  K  L  S  L  H  I  A  D  S  Q  P  G  D  S  A  T  Y  F  C  A  A  S  A
TTCTTCAATAAAAGGGAGAAAAAGCTCTCCTTGCACATTGCAGACTCTCAGCCTGGAGACTCAGCCACCTACTTCTGTGCAGCAAGTGCA

 N  S  G  T  Y  Q  R  F  G  T  G  T  K  L  Q  V  V  P  N  I  Q  N  P  E  P  A  V  Y  Q  L
AATTCTGGGACTTACCAGAGGTTTGGAACTGGGACAAAACTCCAAGTCGTTCCAAACATCCAGAACCCAGAACCTGCTGTGTACCAGTTA

 K  D  P  R  S  Q  D  S  T  L  C  L  F  T  D  F  D  S  Q  I  N  V  P  K  T  M  E  S  G  T
AAAGATCCTCGGTCTCAGGACAGCACCCTCTGCCTGTTCACCGACTTTGACTCCCAAATCAATGTGCCGAAAACCATGGAATCTGGAACG

 F  I  T  D  K  T  V  L  D  M  K  A  M  D  S  K  S  N  G  A  I  A  W  S  N  Q  T  S  F  T
TTCATCACTGACAAAACTGTGCTGGACATGAAAGCTATGGATTCCAAGAGCAATGGGGCCATTGCCTGGAGCAACCAGACAAGCTTCACC

 C  Q  D  I  F  K  E  T  N  A  T  Y  P  S  S  D  V  P  C  D  A  T  L  T  E  K  S  F  E  T
TGCCAAGATATCTTCAAAGAGACCAACGCCACCTACCCCAGTTCAGACGTTCCCTGTGATGCCACGTTGACTGAGAAAAGCTTTGAAACA

 D  M  N  L  N  F  Q  N  L  S  V  M  G  L  R  I  L  L  L  K  V  A  G  F  N  L  L  M  T  L
GATATGAACCTAAACTTTCAAAACCTGTCAGTTATGGGACTCCGAATCCTCCTGCTGAAAGTAGCCGGATTTAACCTGCTCATGACGCTG

 R  L  W  S  S  *
AGGCTGTGGTCCAGTTGA
```

172 β sequence (Leader-V-J-Constant):

```
 M  G  S  R  L  F  F  V  L  S  S  L  L  C  S  H  M  E  A  A  V  T  Q  S  P  R  N  K  V  A
ATGGGCTCCAGGCTCTTCTTCGTGCTCTCCAGTCTCCTGTGTTCACACATGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCA

 V  T  G  G  K  V  T  L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L
GTAACAGGAGGAAAGGTGACATTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTG

 R  L  I  H  Y  S  Y  G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N
AGGCTGATCCATTATTCATATGGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCCAGACCAAGCCAAGAGAAC

 F  S  L  I  L  E  L  A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  D  A  G  G  G  Y  E  Q  Y
TTCTCCCTCATTCTGGAGTTGGCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGTGATGCGGGGGGGGGGTATGAACAGTAC

 F  G  P  G  T  R  L  T  V  L  E  D  L  R  N  V  T  P  P  K  V  S  L  F  E  P  S  K  A  E
TTCGGTCCCGGCACCAGGCTCACGGTTTTAGAGGATCTGAGAAATGTGACTCCACCCAAGGTCTCCTTGTTTGAGCCATCAAAAGCAGAG

 I  A  N  K  Q  K  A  T  L  V  C  L  A  R  G  F  F  P  D  H  V  E  L  S  W  W  V  N  G  K
ATTGCAAACAAACAAAAGGCTACCCTCGTGTGCTTGGCCAGGGGCTTCTTCCCTGACCACGTGGAGCTGAGCTGGTGGGTGAATGGCAAG

 E  V  H  S  G  V  S  T  D  P  Q  A  Y  K  E  S  N  Y  S  Y  C  L  S  S  R  L  R  V  S  A
GAGGTCCACAGTGGGGTCAGCACGGACCCTCAGGCCTACAAGGAGAGCAATTATAGCTACTGCCTGAGCAGCCGCCTGAGGGTCTCTGCT

 T  F  W  H  N  P  R  N  H  F  R  C  Q  V  Q  F  H  G  L  S  E  E  D  K  W  P  E  G  S  P
ACCTTCTGGCACAATCCTCGCAACCACTTCCGCTGCCAAGTGCAGTTCCATGGGCTTTCAGAGGAGGACAAGTGGCCAGAGGGCTCACCC

 K  P  V  T  Q  N  I  S  A  E  A  W  G  R  A  D  C  G  I  T  S  A  S  Y  H  Q  G  V  L  S
AAACCTGTCACACAGAACATCAGTGCAGAGGCCTGGGGCCGAGCAGACTGTGGAATCACTTCAGCATCCTATCATCAGGGGGTTCTGTCT

 A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  G  L  V  L  M  A  M  V  K  K
GCAACCATCCTCTATGAGATCCTACTGGGGAAGGCCACCCTATATGCTGTGCTGGTCAGTGGCCTAGTGCTGATGGCCATGGTCAAGAAA

 K  N  S
AAAAATTCC
```

DO11.10 *α* sequence (Leader-V-J-Constant):

```
 M  K  T  Y  A  P  T  L  F  M  F  L  W  L  Q  L  D  M  S  Q  G  E  Q  V  E  Q  L  P  S  I
ATGAAAACATACGCTCCTACATTATTCATGTTTCTATGGCTGCAGCTGGATATGAGCCAAGGCGAGCAGGTGGAGCAGCTTCCTTCCATC

 L  R  V  Q  E  G  S  S  A  S  I  N  C  T  Y  E  N  S  A  S  N  Y  F  P  W  Y  K  Q  E  P
CTGAGAGTCCAGGAGGGATCCAGTGCCAGCATCAACTGCACTTATGAGAACAGTGCCTCCAACTACTTCCCTTGGTATAAGCAAGAACCT

 G  E  N  P  K  L  I  I  D  I  R  S  N  M  E  R  K  Q  T  Q  G  L  I  V  L  L  D  K  K  A
GGAGAGAATCCTAAGCTCATCATTGACATTCGTTCAAATATGGAAAGAAAGCAGACCCAAGGACTCATCGTTTTACTGGATAAGAAAGCC

 K  R  F  S  L  H  I  T  D  T  Q  P  G  D  S  A  M  Y  F  C  A  A  S  P  N  Y  N  V  L  Y
AAACGCTTCTCCCTGCACATCACAGACACCCAGCCTGGAGACTCAGCCATGTACTTCTGTGCTGCAAGTCCTAATTACAACGTGCTTTAC

 F  G  S  G  T  K  L  T  V  E  P  N  I  Q  N  P  E  P  A  V  Y  Q  L  K  D  P  R  S  Q  D
TTCGGATCTGGCACCAAACTCACTGTAGAGCCAAACATCCAGAACCCAGAACCTGCTGTGTACCAGTTAAAAGATCCTCGGTCTCAGGAC

 S  T  L  C  L  F  T  D  F  D  S  Q  I  N  V  P  K  T  M  E  S  G  T  F  I  T  D  K  T  V
AGCACCCTCTGCCTGTTCACCGACTTTGACTCCCAAATCAATGTGCCGAAAACCATGGAATCTGGAACGTTCATCACTGACAAAACTGTG

 L  D  M  K  A  M  D  S  K  S  N  G  A  I  A  W  S  N  Q  T  S  F  T  C  Q  D  I  F  K  E
CTGGACATGAAAGCTATGGATTCCAAGAGCAATGGGGCCATTGCCTGGAGCAACCAGACAAGCTTCACCTGCCAAGATATCTTCAAAGAG

 T  N  A  T  Y  P  S  S  D  V  P  C  D  A  T  L  T  E  K  S  F  E  T  D  M  N  L  N  F  Q
ACCAACGCCACCTACCCCAGTTCAGACGTTCCCTGTGATGCCACGTTGACTGAGAAAAGCTTTGAAACAGATATGAACCTAAACTTTCAA

 N  L  S  V  M  G  L  R  I  L  L  L  K  V  A  G  F  N  L  L  M  T  L  R  L  W  S  S  *
AACCTGTCAGTTATGGGACTCCGAATCCTCCTGCTGAAAGTAGCCGGATTTAACCTGCTCATGACGCTGAGGCTGTGGTCCAGTTGA
```

DO11.10 β sequence (Leader-V-J-Constant):

```
 M  G  S  R  L  F  F  V  L  S  S  L  L  C  S  H  M  E  A  A  V  T  Q  S  P  R  N  K  V  A
ATGGGCTCCAGGCTCTTCTTCGTGCTCTCCAGTCTCCTGTGTTCACACATGGAGGCTGCAGTCACCCAAAGCCCAAGAAACAAGGTGGCA

 V  T  G  G  K  V  T  L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D  T  G  H  G  L
GTAACAGGAGGAAAGGTGACATTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGACACGGGGCATGGGCTG

 R  L  I  H  Y  S  Y  G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R  P  S  Q  E  N
AGGCTGATCCATTATTCATATGGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCCAGACCAAGCCAAGAGAAC

 F  S  L  I  L  E  L  A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T  N  T  E  V  F
TTCTCCCTCATTCTGGAGTTGGCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACAAACACAGAAGTCTTC

 F  G  K  G  T  R  L  T  V  V  E  D  L  R  N  V  T  P  P  K  V  S  L  F  E  P  S  K  A  E
TTTGGTAAAGGAACCAGACTCACAGTTGTAGAGGATCTGAGAAATGTGACTCCACCCAAGGTCTCCTTGTTTGAGCCATCAAAAGCAGAG

 I  A  N  K  Q  K  A  T  L  V  C  L  A  R  G  F  F  P  D  H  V  E  L  S  W  W  V  N  G  K
ATTGCAAACAAACAAAAGGCTACCCTCGTGTGCTTGGCCAGGGGCTTCTTCCCTGACCACGTGGAGCTGAGCTGGTGGGTGAATGGCAAG

 E  V  H  S  G  V  S  T  D  P  Q  A  Y  K  E  S  N  Y  S  Y  C  L  S  S  R  L  R  V  S  A
GAGGTCCACAGTGGGGTCAGCACGGACCCTCAGGCCTACAAGGAGAGCAATTATAGCTACTGCCTGAGCAGCCGCCTGAGGGTCTCTGCT

 T  F  W  H  N  P  R  N  H  F  R  C  Q  V  Q  F  H  G  L  S  E  E  D  K  W  P  E  G  S  P
ACCTTCTGGCACAATCCTCGCAACCACTTCCGCTGCCAAGTGCAGTTCCATGGGCTTTCAGAGGAGGACAAGTGGCCAGAGGGCTCACCC

 K  P  V  T  Q  N  I  S  A  E  A  W  G  R  A  D  C  G  I  T  S  A  S  Y  Q  Q  G  V  L  S
AAACCTGTCACACAGAACATCAGTGCAGAGGCCTGGGGCCGAGCAGACTGTGGGATTACCTCAGCATCCTATCAACAAGGGGTCTTGTCT

 A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  T  L  V  V  M  A  M  V  K  R
GCCACCATCCTCTATGAGATCCTGCTAGGGAAAGCCACCCTGTATGCTGTGCTTGTCAGTACACTGGTGGTGATGGCTATGGTCAAAAGA

 K  N  S  *
AAGAACTCATGA
```

TRBV5*03 chain sequence 303 AA 909 bp (Leader-V-D-J-Constant)
(http://www.nature.com/nature/journal/v316/n6028/pdf/316517a0.pdf)
BW5147 TCR selected for its non-immunosuppressive characteristics.

```
 M  S  C  R  L  L  L  Y  V  S  L  C  L  V  E  T  A  L  M  N  T  K  I  T  Q  S  P  R  Y  L
ATGAGCTGCAGGCTTCTCCTCTATGTTTCCCTATGTCTTGTGGAAACAGCACTCATGAACACTAAAATTACTCAGTCACCAAGATATCTA

 I  L  G  R  T  N  K  S  L  E  C  E  Q  H  L  G  H  N  A  M  Y  W  Y  K  Q  S  A  E  K  P
ATCCTGGGAAGAACAAATAAGTCTTTGGAATGTGAGCAACATCTGGGACATAATGCTATGTACTGGTATAAACAGAGCGCTGAGAAGCCG

 P  E  L  M  F  L  Y  N  L  K  Q  L  I  R  N  E  T  V  P  S  R  F  I  P  E  C  P  D  S  S
CCAGAGCTCATGTTTCTCTACAATCTTAAACAGTTGATTCGAAATGAGACGGTGCCCAGTCGTTTTATACCTGAATGCCCAGACAGCTCC

 K  L  L  L  H  I  S  A  V  D  P  E  D  S  A  V  Y  F  C  A  S  S  Q  I  T  S  N  Q  D  T
AAGCTACTTTTACATATATCTGCCGTGGATCCAGAAGACTCAGCTGTCTATTTTTGTGCCAGCAGCCAGATAACTAGTAACCAAGACACC

 Q  Y  F  G  P  G  T  R  L  L  V  L  E  D  L  R  N  V  T  P  P  K  V  S  L  F  E  P  S  K
CAGTACTTTGGGCCAGGCACTCGGCTCCTCGTGTTAGAGGATCTGAGAAATGTGACTCCACCCAAGGTCTCCTTGTTTGAGCCATCAAAA

 A  E  I  A  N  K  Q  K  A  T  L  V  C  L  A  R  G  F  F  P  D  H  V  E  L  S  W  W  V  N
GCAGAGATTGCAAACAAACAAAAGGCTACCCTCGTGTGCTTGGCCAGGGGCTTCTTCCCTGACCACGTGGAGCTGAGCTGGTGGGTGAAT

 G  K  E  V  H  S  G  V  S  T  D  P  Q  A  Y  K  E  S  N  Y  S  Y  C  L  S  S  R  L  R  V
GGCAAGGAGGTCCACAGTGGGGTCAGCACGGACCCTCAGGCCTACAAGGAGAGCAATTATAGCTACTGCCTGAGCAGCCGCCTGAGGGTC

 S  A  T  F  W  H  N  P  R  N  H  F  R  C  Q  V  Q  F  H  G  L  S  E  E  D  K  W  P  E  G
TCTGCTACCTTCTGGCACAATCCTCGAAACCACTTCCGCTGCCAAGTGCAGTTCCATGGGCTTTCAGAGGAGGACAAGTGGCCAGAGGGC

 S  P  K  P  V  T  Q  N  I  S  A  E  A  W  G  R  A  D  C  G  I  T  S  A  S  Y  H  Q  G  V

TCACCCAAACCTGTCACACAGAACATCAGTGCAGAGGCCTGGGGCCGAGCAGACTGTGGAATCACTTCAGCATCCTATCATCAGGGGGTT

 L  S  A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  G  L  V  L  M  A  M  V
CTGTCTGCAACCATCCTCTATGAGATCCTACTGGGGAAGGCCACCCTATATGCTGTGCTGGTCAGTGGCCTGGTGCTGATGGCCATGGTC

 K  K  K  N  S  *
AAGAAAAAAAATTCCTGA
```

B.2.2: Eukaryotic selection vectors

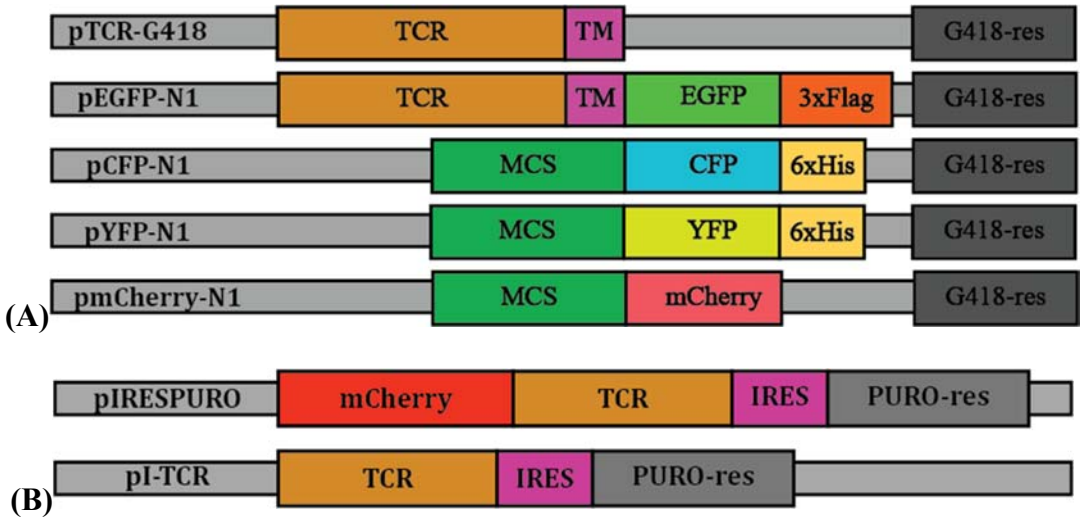


Figure B.1: Diagram of the expression vectors generated for this study

The constructs generated for this surface display study included fluorescent protein fusions with EGFP, CFP, YFP, and mCherry. Vectors with resistance genes were formed from vectors (**A**) pEGFP-N1 (**B**) and pIRESPURO3 with G418 and puromycin antibiotic resistance genes respectively. The backbone vectors were a gift from the Poenie lab.

B.2.3: IRES Format creation

Enzyme	Site	Location
EcoRI	GAATTC	5′ kozak sequence
BsiWI	CGTACG	First protein junction
BglII	AGATCT	5′ of Second protein, end of IRES
SalI	GTCGAC	Second protein junction
NheI	GCTAGC	3′ of First protein, start of IRES
XhoI	CTCGAG	3′ of Second protein
BssHII	GCGCGC	3′ EcoRI site

Table B.5: IRES cloning sites

B.2.4: 2A Peptide Construct

Gene 1– RAKRGSGVKQTLNFDLLKLAGDVESNPGP[177,178] – Gene 2 Signal sequence

```
L  N  S  T  M  G  S  R  L  F  F  V  L  S  S  L  L  C  S  K  H  M  E  A  A  V  T  Q  S  P
TTGAATTCCACCATGGGCTCCAGGCTCTTCTTCGTGCTCTCCAGTCTCCTGTGTTCAAAACACATGGAGGCTGCAGTCACCCAAAGCCCA 90
R  N  K  V  A  V  T  G  E  K  V  T  L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q  D
AGAAACAAGGTGGCAGTAACAGGAGAGAAGGTGACATTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAGGAC 180
T  G  H  G  L  R  L  I  Y  Y  S  Y  G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S  R
ACGGGGCATGGGCTGAGGCTGATCTACTATTCATATGGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCTAGA 270
P  S  R  E  N  F  S  L  T  L  E  S  A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T  T
CCAAGCCGAGAGAACTTCTCCCTCACTCTGGAGTCGGCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACAACA 360
N  T  E  V  F  F  G  K  G  T  R  L  T  V  V  E  D  L  R  N  V  T  P  P  K  V  S  L  F  E
AACACAGAAGTCTTCTTTGGTAAAGGAACCAGACTCACAGTTGTAGAGGACTTGAGAAATGTGACTCCACCCAAGGTCTCCTTGTTTGAG 450
P  S  K  A  E  I  A  N  K  Q  K  A  T  L  V  C  L  A  R  G  F  F  P  D  H  V  E  L  S  W
CCATCAAAAGCAGAGATTGCAAACAAACAAAAGGCTACCCTCGTGTGCTTGGCCAGGGGCTTCTTCCCTGACCACGTGGAGCTGAGCTGG 540
W  V  N  G  K  E  V  H  S  G  V  S  T  D  P  Q  A  Y  K  E  S  N  Y  S  Y  C  L  S  S  R
TGGGTGAATGGCAAGGAGGTCCACAGTGGGGTCAGCACGGACCCTCAGGCCTACAAGGAGAGCAATTATAGCTACTGCCTGAGCAGCCGC 630
L  R  V  S  A  T  F  W  H  N  P  R  N  H  F  R  C  Q  V  Q  F  H  G  L  S  E  E  D  K  W
CTGAGGGTCTCTGCTACCTTCTGGCACAATCCTCGAAACCACTTCCGCTGCCAAGTGCAGTTCCATGGGCTTTCAGAGGAGGACAAGTGG 720
P  E  G  S  P  K  P  V  T  Q  N  I  S  A  E  A  W  G  R  A  D  C  G  I  T  S  A  S  Y  H
CCAGAGGGCTCACCCAAACCTGTCACACAGAACATCAGTGCAGAGGCCTGGGGCCGAGCAGACTGTGGAATCACTTCAGCATCCTATCAT 810
Q  G  V  L  S  A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  G  L  V  L
CAGGGGGTTCTGTCTGCAACCATCCTCTATGAGATCCTACTGGGGAAGGCCACCCTATATGCTGTGCTGGTCAGTGGCCTGGTGCTGATG 900
A  M  V  K  K  K  N  S  R  A  K  R  G  S  G  V  K  Q  T  L  N  F  D  L  L  K  L  A  G  D
GCCATGGTCAAAAAAAAAAATTCCAGAGCCAAAAGGGGATCCGGCGTGAAGCAGACACTCAACTTCGACCTGCTGAAGCTTGCGGGCGAC 990
V  E  S  N  P  G  P  G  E  Q  V  E  Q  L  P  S  I  L  R  V  Q  E  G  S  S  A  S  I  N  C
GTGGAATCTAACCCTGGCCCTGGCGAGCAGGTGGAGCAGCTTCCTTCCATCCTGAGAGTCCAGGAGGGATCCAGTGCCAGCATCAACTGC 1080
T  Y  E  N  S  A  S  N  Y  F  P  W  Y  K  Q  E  P  G  E  N  P  K  L  I  I  D  I  R  S  N
ACTTATGAGAACAGTGCCTCCAACTACTTCCCTTGGTATAAGCAAGAACCTGGAGAGAATCCTAAGCTCATCATTGACATTCGTTCAAAT 1170
M  E  R  K  Q  T  Q  G  L  I  V  L  L  D  K  K  A  K  R  F  S  L  H  I  T  D  T  Q  P  G
ATGGAAAGAAAGCAGACCCAAGGACTCATCGTTTTACTGGATAAGAAAGCCAAACGCTTCTCCCTGCACATCACAGACACCCAGCCTGGA 1260
D  S  A  M  Y  F  C  A  A  S  P  N  Y  N  V  L  Y  F  G  S  G  T  K  L  T  V  E  P  N  I
GACTCAGCCATGTACTTCTGTGCTGCAAGTCCTAATTACAACGTGCTTTACTTCGGATCTGGCACCAAACTCACTGTAGAGCCAAACATC 1350
Q  N  P  E  P  A  V  Y  Q  L  K  D  P  R  S  Q  D  S  T  L  C  L  F  T  D  F  D  S  Q  I
CAGAACCCAGAACCTGCTGTGTACCAGTTAAAAGATCCTCGGTCTCAGGACAGCACCCTCTGCCTGTTCACCGACTTTGACTCCCAAATC 1440
N  V  P  K  T  M  E  S  G  T  F  I  T  D  K  T  V  L  D  M  K  A  M  D  S  K  S  N  G  A
AATGTGCCGAAAACCATGGAATCTGGAACGTTCATCACTGACAAAACTGTGCTGGACATGAAAGCTATGGATTCCAAGAGCAATGGGGCC 1530
I  A  W  S  N  Q  T  S  F  T  C  Q  D  I  F  K  E  T  N  A  T  Y  P  S  S  D  V  P  C  D
ATTGCCTGGAGCAACCAGACATCCTTCACCTGCCAAGATATCTTCAAAGAGACCAACGCCACCTACCCCAGTTCAGACGTTCCCTGTGAT 1620
A  T  L  T  E  K  S  F  E  T  D  M  N  L  N  F  Q  N  L  S  V  M  G  L  R  I  L  L  L  K
GCCACGTTGACTGAGAAATCCTTTGAAACAGATATGAACCTAAACTTTCAAAACCTGTCAGTTATGGGACTCCGAATCCTCCTGCTGAAA 1710
V  A  G  F  N  L  L  M  T  L  R  L  W  S  S  *  S
GTAGCCGGATTTAACCTGCTCATGACGCTGAGGCTGTGGTCCAGTTGAAGCTT 1763
```

B.3: scDO11-CAR

```
 L  N  S  T  M  G  W  S  C  I  I  L  F  L  V  A  T  A  T  G  V  H  S  E  A  A  V  T  Q  S
TTGAATTCCACCATGGGATGGTCATGTATCATCCTTTTTCTAGTAGCAACTGCAACCGGTGTACACTCGGAGGCTGCAGTCACCCAAAGC 90
 P  R  N  K  V  A  V  T  G  E  K  V  T  L  S  C  N  Q  T  N  N  H  N  N  M  Y  W  Y  R  Q
CCAAGAAACAAGGTGGCAGTAACAGGAGAGAAGGTGACATTGAGCTGTAATCAGACTAATAACCACAACAACATGTACTGGTATCGGCAG 180
 D  T  G  H  G  L  R  L  I  Y  Y  S  Y  G  A  G  S  T  E  K  G  D  I  P  D  G  Y  K  A  S
GACACGGGGCATGGGCTGAGGCTGATCTACTATTCATATGGTGCTGGCAGCACTGAGAAAGGAGATATCCCTGATGGATACAAGGCCTCT 270
 R  P  S  Q  E  N  F  S  L  T  L  E  S  A  T  P  S  Q  T  S  V  Y  F  C  A  S  G  S  G  T
AGACCAAGCCAAGAGAACTTCTCCCTCACTCTGGAGTCGGCTACCCCCTCTCAGACATCAGTGTACTTCTGTGCCAGCGGCTCCGGGACA 360
 T  N  T  E  V  F  F  G  K  G  T  R  L  T  V  V  G  G  G  G  S  G  G  G  G  S  G  G  G  G
ACAAACACAGAAGTCTTCTTTGGTAAAGGAACCAGACTCACAGTTGTAGGTGGAGGCGGTTCAGGCGGAGGTGGCTCCGGAGGCGGTGGT 450
 S  G  G  G  G  S  E  Q  V  E  Q  L  P  S  I  L  R  V  Q  E  G  S  S  A  S  I  N  C  T  Y
TCTGGAGGAGGAGGATCCGAGCAGGTGGAGCAGCTTCCTTCCATCCTGAGAGTCCAGGAGGGATCCAGTGCCAGCATCAACTGCACTTAT 540
 E  N  S  A  S  N  Y  F  P  W  Y  K  Q  E  P  G  E  N  P  K  L  I  I  D  I  R  S  N  M  E
GAGAACAGTGCCTCCAACTACTTCCCTTGGTATAAGCAAGAACCTGGAGAGAATCCTAAGCTCATCATTGACATTCGTTCAAATATGGAA 630
 R  K  Q  T  Q  G  L  I  V  L  L  D  K  K  A  K  R  F  S  L  H  I  T  D  T  Q  P  G  D  S
AGAAAGCAGACCCAAGGACTCATCGTTTTACTGGATAAGAAAGCCAAACGCTTCTCCCTGCACATCACAGACACCCAGCCTGGAGACTCA 720
 A  M  Y  F  C  A  A  S  P  N  Y  N  V  L  Y  F  G  S  G  T  K  L  T  V  E  P  I  E  F  M
GCCATGTACTTCTGTGCTGCAAGTCCTAATTACAACGTGCTTTACTTCGGATCTGGCACCAAACTCACTGTAGAGCCAATTGAGTTCATG 810
 Y  P  P  P  Y  L  D  N  E  R  S  N  G  T  I  I  H  I  K  E  K  H  L  C  H  T  Q  S  S  P
TACCCTCCGCCTTACCTAGACAACGAGAGGAGCAATGGAACTATTATTCACATAAAAGAGAAACATCTTTGTCATACTCAGTCATCTCCT 900
 K  L  F  W  A  L  V  V  V  A  G  V  L  F  C  Y  G  L  L  V  T  V  A  L  C  V  I  W  T  N
AAGCTGTTTTGGGCACTGGTCGTGGTTGCTGGAGTCCTGTTTTGTTATGGCTTGCTAGTGACAGTGGCTCTTTGTGTTATCTGGACAAAT 990
 S  R  R  N  R  L  L  Q  S  D  Y  M  N  M  T  P  R  R  P  G  L  T  R  K  P  Y  Q  P  Y  A
AGTAGAAGGAACAGACTCCTTCAAAGTGACTACATGAACATGACTCCCCGGAGGCCTGGGCTCACTCGAAAGCCTTACCAGCCCTACGCC 1080
 P  A  R  D  F  A  A  Y  R  P  S  R  A  K  F  S  R  S  A  E  T  A  A  N  L  Q  D  P  N  Q
CCTGCCAGAGACTTTGCAGCGTACCGCCCCAGCAGAGCAAAATTCAGCAGGAGTGCAGAGACTGCTGCCAACCTGCAGGACCCCAACCAG 1170
 L  Y  N  E  L  N  L  G  R  R  E  E  Y  D  V  L  E  K  K  R  A  R  D  P  E  M  G  G  K  Q
CTCTACAATGAGCTCAATCTAGGGCGAAGAGAGGAATATGACGTCTTGGAGAAGAAGCGGGCTCGGGATCCAGAGATGGGAGGCAAACAG 1260
 Q  R  R  R  N  P  Q  E  G  V  Y  N  A  L  Q  K  D  K  M  A  E  A  Y  S  E  I  G  T  K  G
CAGAGGAGGAGGAACCCCCAGGAAGGCGTATACAATGCACTGCAGAAAGACAAGATGGCAGAAGCCTACAGTGAGATCGGCACAAAAGGC 1350
 E  R  R  R  G  K  G  H  D  G  L  Y  Q  G  L  S  T  A  T  K  D  T  Y  D  A  L  H  M  Q  T
GAGAGGCGGAGAGGCAAGGGGCACGATGGCCTTTACCAGGGTCTCAGCACTGCCACCAAGGACACCTATGATGCCCTGCATATGCAGACC 1440
 L  A  P  R  *  A  W  P  P  R  P
CTGGCCCCTCGCTAAGCTTGGCCGCCGCGGCCGC 1474
```

Legend

EcoRI (GAATTC),

AbVec Signal sequence, (MGWSCIILFLVATATGVHS)
SH2 binding site, Tyr can become phosphorylated (PYAP)
ITAMS YXX(I/L)$X_{(6-8)}$YXX(I/L)
HindIII (AAGCTT)
NotI (GCGGCCGC)
Approximate size: 1460 bp, 485 AA, 53.5 kDa

B.4: Signal sequence modifications

TCR Chain	Leader/Signal Sequence	TM region (where applicable with charged residues underlined)
AbVec SS.	MGWSCIILFLVATATGVHS	
TCR consensus	MLMGTRLLCWVALCLLGAGHV	
Ig consensus	MGWSWIFLFLVAAPTGVHSLS	
mVα5D-4 (DO11)	MKTYAPTLFMFLWLQLDGMSQ	VMGLRILLLKVAGFNLLMTLRL
mVα14-3 (172)	MDKILTASFLLLGLHLAG	
TRAV24	METLLGVSLVILWLQLA	
mVβ13-2 (DO11)	MGSRLFFVLSSLLCSKHM	TILYEILLGKATLYAVLVSTLVVMAMV
mVβ13-2 (172)	MGSRLFFVLSSLLCSKHM	
mVβ5 (BW5147)	MGSRLLLYVSLCLVETALM	
TRBV16	MAPRLLFCLVLCFLRAEPT	
mCD3γ	MEQRKGLAGLFLVISLLQ	IGTISGFIFAEVISIFFLALGVYLIAG
mCD3δ	MEHSGILASLILIAVLPQ	GVIFIDLIATLLLALGVYCFA
mCD3ε	MRWNTFWGILCLSLLAV	DLTAVAIIIIVDICITLGLLMVIYYW
mCD3ζ	MKWKVSVLACILHVRFPGAE	LCYLLDGILFIYGVIITALY
mCD4	MCRAISLRRLLLLLLQLSQL	VFLACVLGGSFGFLGFLGLCILC
mCD8 α	MASPLTRFLSLNLLLLGESIILGSGEA	IWAPLAGICVAPLLSLII
mCD8 β	MQPWLWLVFSMKLAALWS	LSLLVVCILLLLAFLGVAVYFY
mCD28	MTLRLLFLALNFFSVQVT	FWALVVVAGVLFCYGLLVTVALCVIW
hCD8 α	MALPVTALLLPLALLLHAARP	IYIWAPLAGTCGVLLLSLVIT
hCD8 β	MRPRLWLLLAAQLTVLHGNSV	ITLGLLVAGVLVLLVSLGVAI
Gaussia Luciferase	MGVKVLFALICIAVAEA	

Table B.6: Sample signal sequences in immune-proteins

Consensus signal sequences for antibodies and TCRs generated using all functional v-region signal sequences and ClustalΩ.

C: CANONICAL AND HTS SQL PROGRAM SCRIPTS

Program scripts for the canonical loop structural analysis are located in the separate supplemental materials file due to their length.

Glossary

2A:	self-cleaving peptidase peptide signal
APC:	antigen presenting cell
CAR:	chimeric antigen receptor
CD:	cluster of differentiation
CDR:	complementarity determining region
FACS:	fluorescence-activated cell sorting
FR:	framework residue
HTS:	high-throughput sequencing
IL-2:	interleukin 2
IMGT:	Immuno-genetics Database
IRES:	internal ribosome entry site
OVA:	chicken-ovalbumin
PDB:	Protein Data Bank
pMHC:	peptide-major histocompatibility complex
RT:	Room temperature
RT-PCR:	Reverse transcription polymerase chain reaction
SPR:	surface plasmon resonance
scTCR:	single-chain T-cell receptor
TCR:	T-cell receptor
TM:	transmembrane
TRXV:	T-cell receptor Variable region for chain X

References

1. Nunes-Alves, C., Nobrega, C., Behar, S. M. & Correia-Neves, M. Tolerance has its limits: How the thymus copes with infection. *Trends in Immunology* **34,** 502–510 (2013).
2. He, X. *et al.* Structural snapshot of aberrant antigen presentation linked to autoimmunity: the immunodominant epitope of MBP complexed with I-Au. *Immunity* **17,** 83–94 (2002).
3. Cole, D. K. D. *et al.* Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. *J. Immunol.* **185,** 2600–2610 (2010).
4. Landais, E. *et al.* New design of MHC class II tetramers to accommodate fundamental principles of antigen presentation. *J. Immunol.* **183,** 7949–7957 (2009).
5. Gascoigne, N. R. J. Do T cells need endogenous peptides for activation? *Nat. Rev. Immunol.* **8,** 895–900 (2008).
6. Birnbaum, M. E. *et al.* Molecular architecture of the αβ T cell receptor–CD3 complex. *Proc. Natl. Acad. Sci.* 1–6 (2014). doi:10.1073/pnas.1420936111
7. Thoenes, G. *et al.* Structural analysis of low TCR-CD3 complex expression in T cells of an immunodeficient patient. *J. Biol. Chem.* **267,** 487–493 (1992).
8. Cohen, C. J., Zhao, Y., Zheng, Z., Rosenberg, S. a. & Morgan, R. a. Enhanced antitumor activity of murine-human hybrid T-cell receptor (TCR) in human lymphocytes is associated with improved pairing and TCR/CD3 stability. *Cancer Res.* **66,** 8878–86 (2006).
9. Kuhns, M. S., Davis, M. M. & Garcia, K. C. Deconstructing the form and function of the TCR/CD3 complex. *Immunity* **24,** 133–139 (2006).
10. Valitutti, S., Müller, S., Cella, M., Padovan, E. & Lanzavecchia, a. Serial triggering of many T-cell receptors by a few peptide-MHC complexes. *Nature* **375,** 148–151 (1995).
11. Tynan, F. E. *et al.* T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide. *Nat. Immunol.* **6,** 1114–22 (2005).
12. Beddoe, T. *et al.* Antigen ligation triggers a conformational change within the constant domain of the αβ T cell receptor. *Immunity* **30,** 777–88 (2009).

13. Stone, J. D., Chervin, A. S. & Kranz, D. M. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology* **126,** 165–176 (2009).

14. Humphries, C. Adoptive cell therapy: Honing that killer instinct. *Nature* **504,** S13–S15 (2013).

15. Ribeiro, R. M., Mohri, H., Ho, D. D. & Perelson, A. S. In vivo dynamics of T cell activation, proliferation, and death in HIV-1 infection: why are CD4+ but not CD8+ T cells depleted? *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15572–15577 (2002).

16. Venturi, V., Price, D. a, Douek, D. C. & Davenport, M. P. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8,** 231–8 (2008).

17. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334,** 395–402 (1988).

18. Parham, P. *The Immune System*. **4th,** (Garland Science, 2014).

19. Jackson, K. J. L., Kidd, M. J., Wang, Y. & Collins, A. M. The shape of the lymphocyte receptor repertoire: Lessons from the B cell receptor. *Front. Immunol.* **4,** 1–12 (2013).

20. Bartok, I. *et al.* T cell receptor CDR3 loops influence αβ pairing. *Mol. Immunol.* **47,** 1613–1618 (2010).

21. Rudolph, M., Stanfield, R. & Wilson, I. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* (2006). doi:10.1146/annurev.immunol.23.021704.115658

22. Garcia, K. C., Adams, J. J., Feng, D. & Ely, L. K. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat. Immunol.* **10,** 143–147 (2009).

23. Al-lazikani, B., Lesk, A. M. & Chothia, C. Canonical Structures for the Hypervariable Regions of T Cell a b Receptors. *J. Mol. Biol.* (2000).

24. Machius, M., Cianga, P., Deisenhofer, J. & Ward, E. S. Crystal Structure of a T Cell Receptor V alpha 11 (AV11S5) Domain: New Canonical Forms for the First and Second Complementarity Determining Regions. *J. Mol. Biol.* **11,** 689–698 (2001).

25. Chothia, C. *et al.* Conformations of immunoglobulin hypervariable regions. *Nature* **342,** 877–883 (1989).

26. Laugel, B., Boulter, J., Lissin, N. & Vuidepot, A. Design of soluble recombinant T cell receptors for antigen targeting and T cell inhibition. *J. Biol. Chem.* **280,** 1882–1892 (2005).

27. Maynard, J. & Adams, E. High-level bacterial secretion of single-chain alphabeta T-cell receptors. *J. Immunol. Methods* **306,** 51–67 (2005).

28. Bowers, P. M. *et al.* Coupling mammalian cell surface display with somatic hypermutation for the discovery and maturation of human antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 20455–60 (2011).

29. Weiss, G. a, Watanabe, C. K., Zhong, a, Goddard, a & Sidhu, S. S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 8950–4 (2000).

30. Dalby, P. a. Optimising enzyme function by directed evolution. *Curr. Opin. Struct. Biol.* **13,** 500–505 (2003).

31. Weidanz, J. a, Card, K. F., Edwards, A., Perlstein, E. & Wong, H. C. Display of functional alphabeta single-chain T-cell receptor molecules on the surface of bacteriophage. *J. Immunol. Methods* **221,** 59–76 (1998).

32. Li, Y., Moysey, R., Molloy, P. & Vuidepot, A. Directed evolution of human T-cell receptors with picomolar affinities by phage display. *Nat. Biotechnol.* **23,** 349–354 (2005).

33. Mazor, Y., Van Blarcom, T., Iverson, B. L. & Georgiou, G. E-clonal antibodies: selection of full-length IgG antibodies using bacterial periplasmic display. *Nat. Protoc.* **3,** 1766–77 (2008).

34. Shusta, E. V, Holler, P. D., Kieke, M. C., Kranz, D. M. & Wittrup, K. D. Directed evolution of a stable scaffold for T-cell receptor engineering. *Nat. Biotechnol.* **18,** 754–9 (2000).

35. Codamo, J., Munro, T. P., Hughes, B. S., Song, M. & Gray, P. P. Enhanced CHO cell-based transient gene expression with the Epi-CHO expression system. *Mol. Biotechnol.* **48,** 109–115 (2011).

36. Kappler, B. J. W., Skidmore, B., White, J. & Marrack, a N. P. Antigen-inducible, H-2-restricted, interleukin-2-producing T cell hybridomas. *J. Exp. Med.* **153,** 1198–1214 (1981).

37. Shimonkevitz, R., Colon, S. & Kappler, J. Antigen recognition by H-2-restricted T cells. **158,** 303–316 (1984).

38. Roy, B. M., Zhukov, D. V & Maynard, J. a. Flanking residues are central to DO11.10 T cell hybridoma stimulation by ovalbumin 323-339. *PLoS One* **7,** e47585 (2012).

39. McFarland, B. J., Sant, a J., Lybrand, T. P. & Beeson, C. Ovalbumin(323-339) peptide binds to the major histocompatibility complex class II I-A(d) protein using two functionally distinct registers. *Biochemistry* **38,** 16663–70 (1999).

40. Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P. & Reid, G. High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods* **95,** 401–414 (2013).

41. Wu, D. *et al.* High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* **4,** 134ra63 (2012).

42. Miles, J. J. J. *et al.* Antigen-driven patterns of TCR bias are shared across diverse outcomes of human hepatitis C virus infection. *J. Immunol.* **186,** 901–12 (2011).

43. Britanova, O. V *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192,** 2689–98 (2014).

44. Yagüe, J. *et al.* The T cell receptor: the α and β chains define idiotype, and antigen and MHC specificity. *Cell* **42,** 81–87 (1985).

45. Robertson, J. M., Jensen, P. E. & Evavold, B. D. DO11.10 and OT-II T cells recognize a C-terminal ovalbumin 323-339 epitope. *J. Immunol.* **164,** 4706–4712 (2000).

46. Jenkins, M. K. & Moon, J. J. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. *J. Immunol.* **188,** 4135–40 (2012).

47. Relland, L. M. *et al.* Affinity-based selection of regulatory T cells occurs independent of agonist-mediated induction of Foxp3 expression. *J. Immunol.* **182,** 1341–1350 (2009).

48. Zehn, D., Lee, S. Y. & Bevan, M. J. Complete but curtailed T-cell response to very low-affinity antigen. *Nature* **458,** 211–4 (2009).

49. Furmanski, A. L. *et al.* Public T Cell Receptor b-Chains Are Not Advantaged during Positive Selection. *J. Immunol.* **180,** 1029–1039 (2008).

50. Shugay, M. *et al.* Huge Overlap of Individual TCR Beta Repertoires. *Front. Immunol.* **4,** 466 (2013).

51. Acha-Orbea, H. *et al.* Limited heterogeneity of T cell receptors from lymphocytes mediating autoimmune encephalomyelitis allows specific immune intervention. *Cell* **54,** 263–73 (1988).

52. Smithey, M. J. M., Li, G., Venturi, V., Davenport, M. P. & Nikolich-Žugich, J. Lifelong Persistent Viral Infection Alters the Naive T Cell Pool, Impairing CD8 T Cell Immunity in Late Life. *J. Immunol.* **189,** 5356–66 (2012).

53. Karrer, U. *et al.* Memory inflation: continuous accumulation of antiviral CD8+ T cells over time. *J. Immunol.* **170,** 2022–2029 (2003).

54. Bernardin, F. *et al.* Estimate of the total number of CD8+ clonal expansions in healthy adults using a new DNA heteroduplex-tracking assay for CDR3 repertoire analysis. *J. Immunol. Methods* **274,** 159–75 (2003).

55. Campregher, P. V, Srivastava, S. K., Deeg, H. J., Robins, H. S. & Warren, E. H. abnormalities of the alphabeta t cell receptor repertoire in advanced myelodysplastic syndrome. *Exp. Hematol.* **38,** 202–212 (2011).

56. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci.* **111,** 13139–13144 (2014).

57. Bashford-Rogers, R. J. *et al.* Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* **15,** 29 (2014).

58. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281,** 363, 365 (1998).

59. Krell, P., Reuther, S., Fischer, U. & Keller, T. Next-generation-sequencing-spectratyping reveals public T-cell receptor repertoires in pediatric very severe aplastic anemia and identifies a beta chain CDR3 sequence. *Haematologica* **98,** (2013).

60. Voelkerding, K. V., Dames, S. a. & Durtschi, J. D. Next-generation sequencing:from basic research to diagnostics. *Clin. Chem.* **55,** 641–658 (2009).

61. Dames, S., Durtschi, J., Geiersbach, K., Stephens, J. & Voelkerding, K. V. Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes. *J. Biomol. Tech.* **21,** 73–80 (2010).

62. Braber, I. Den *et al.* Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* **36,** 288–97 (2012).

63. Barth, R., Kim, B., Lan, N. & Hunkapiller, T. The murine T-cell receptor uses a limited repertoire of expressed Vβ gene segments. *Nature* (1985).

64. Ruiz, M. *et al.* IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **28,** 219–21 (2000).

65. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135,** 183–191 (2012).

66. Ndifon, W. *et al.* Chromatin conformation governs T-cell receptor J gene segment usage. *Proc. Natl. Acad. Sci.* **109,** 15865–15870 (2012).

67. Genolet, R., Stevenson, B. J. B., Farinelli, L., Østerås, M. & Luescher, I. F. Highly diverse TCRα chain repertoire of pre-immune CD8+ T cells reveals new insights in gene recombination. *EMBO J.* **31,** 1666–1678 (2012).

68. Rubtsova, K. *et al.* Many different Vbeta CDR3s can reveal the inherent MHC reactivity of germline-encoded TCR V regions. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7951–6 (2009).

69. Colf, L. a *et al.* How a single T cell receptor recognizes both self and foreign MHC. *Cell* **129,** 135–46 (2007).

70. Martin, M. P. & Carrington, M. Immunogenetics of HIV disease. *Immunol. Rev.* **254,** 245–264 (2013).

71. Turner, S. J., Doherty, P. C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6,** 883–94 (2006).

72. Wang, G. C., Dash, P., McCullers, J. a, Doherty, P. C. & Thomas, P. G. T cell receptor αβ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Science translational medicine* **4,** 128ra42 (2012).

73. Silins, S. L. *et al. Development of Epstein-Barr virus-specific memory T cell receptor clonotypes in acute infectious mononucleosis. The Journal of experimental medicine* **184,** 1815–1824 (1996).

74. Li, H., Ye, C., Ji, G. & Han, J. Determinants of public T cell responses. *Cell Res.* **22,** 33–42 (2012).

75. Gras, S., Kjer-Nielsen, L., Burrows, S. R., McCluskey, J. & Rossjohn, J. T-cell receptor bias and immunity. *Curr. Opin. Immunol.* **20,** 119–25 (2008).

76. Alonso-Camino, V. *et al.* CARbodies: Human Antibodies Against Cell Surface Tumor Antigens Selected From Repertoires Displayed on T Cell Chimeric Antigen Receptors. *Mol. Ther. Nucleic Acids* **2,** e93 (2013).

77. Lefranc, M., Pommié, C. & Ruiz, M. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev.* … **27,** 55–77 (2003).

78. Broeren, C. P. *et al.* Conserved nucleotide sequences at the 5' end of T cell receptor variable genes facilitate polymerase chain reaction amplification. *Eur. J. Immunol.* **21,** 569–575 (1991).

79. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** (2011).

80. North, B., Lehmann, A. & Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406,** 228–56 (2011).

81. Holler, P. D. *et al.* In vitro evolution of a T cell receptor with high affinity for peptide/MHC. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 5387–92 (2000).

82. Estorninho, M. *et al.* A Novel Approach to Tracking Antigen-Experienced CD4 T Cells into Functional Compartments via Tandem Deep and Shallow TCR Clonotyping. *J. Immunol.* (2013). doi:10.4049/jimmunol.1300622

83. Manuscript, A., Disorders, I. & Technology, A. R. A New Twist in TCR Diversity Revealed by a forbidden alpha beta TCR. **27,** 417–428 (2009).

84. Marks, J. D. *et al.* By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J. Mol. Biol.* **222,** 581–597 (1991).

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

86. Turchaninova, M. a *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43,** 2507–15 (2013).

87. Decanniere, K., Muyldermans, S. & Wyns, L. Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes? *J. Mol. Biol.* (2000). doi:10.1006/jmbi.2000.3839

88. Chothia, C. & Lesk, a M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196,** 901–17 (1987).

89. Lynch, J. J. N., Donermeyer, D. D. L., Weber, K. S., Kranz, D. M. & Allen, P. M. Subtle changes in TCRα CDR1 profoundly increase the sensitivity of CD4 T cells. *Mol. Immunol.* **53,** 283–94 (2013).

90. Manning, B. T. C., Parke, E. A., Teyton, L. & Kranz, D. M. Effects of Complementarity Determining Region Mutations on the Affinity of an alpha / beta T Cell Receptor: Measuring the Energy Associated with CD4/CD8 Repertoire Skewing. *J. Exp. Med.* **189,** (1999).

91. Maynard, J. a *et al.* Protection against anthrax toxin by recombinant antibody fragments correlates with antigen affinity. *Nat. Biotechnol.* **20,** 597–601 (2002).

92. Foote, J. & Eisen, H. N. Breaking the affinity ceiling for antibodies and T cell receptors. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 10679–10681 (2000).

93. Lyons, D. S. *et al.* A TCR binds to antagonist ligands with lower affinities and faster dissociation rates than to agonists. *Immunity* **5,** 53–61 (1996).

94. Hawse, W. F. *et al.* TCR Scanning of Peptide/MHC through Complementary Matching of Receptor and Ligand Molecular Flexibility. *J. Immunol.* (2014). doi:10.4049/jimmunol.1302953

95. Irvine, D. J., Purbhoo, M. a, Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419,** 845–9 (2002).

96. Rudolph, M. M. G., Stanfield, R. L. R. & Wilson, I. a. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24,** 419–466 (2006).

97. Marrack, P., Scott-Browne, J. P., Dai, S., Gapin, L. & Kappler, J. J. W. Evolutionarily conserved amino acids that control TCR-MHC interaction. *Annu. Rev. Immunol.* **26,** 171–203 (2008).

98. Manning, T., Schlueter, C. & Brodnicki, T. Alanine scanning mutagenesis of an αβ T cell receptor: mapping the energy of antigen recognition. *Immunity* **8,** 413–425 (1998).

99. Kranz, D. M. T cell receptor CDRs: starring versus supporting roles. *Nat. Immunol.* **6,** 130–2 (2005).

100. Feng, D., Bond, C., Ely, L., Maynard, J. & Garcia, K. Structural evidence for a germline-encoded T cell receptor–major histocompatibility complex interaction'codon'. *Nat. Immunol.* **8,** 975–983 (2007).

101. J L Jorgensen, P A Reay, E W Ehrich, and M. M. D. Molecular components of T-cell recognition. *Annu. Rev. Immunol.* **10,** 835–875 (1992).

102. Choi, Y. & Deane, C. M. Predicting antibody complementarity determining region structures without classification. *Mol. Biosyst.* **7,** 3327–34 (2011).

103. Pantazes, R. J. & Maranas, C. D. OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng. Des. Sel.* **23,** 849–58 (2010).

104. Chothia, C., Lesk, A. & Tramontano, A. Conformations of immunoglobulin hypervariable regions. *Nature* (1989).

105. Oliva, B., Bates, P. a, Querol, E., Avilés, F. X. & Sternberg, M. J. Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.* **279,** 1193–210 (1998).

106. Sircar, A. & Gray, J. J. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.* **6,** e1000644 (2010).

107. Padlan, E., Abergel, C. & Tipper, J. Identification of specificity-determining residues in antibodies. *FASEB J.* 133–139 (1995).

108. Khan, J. M. & Ranganathan, S. Understanding TR binding to pMHC complexes: How does a TR scan many pMHC complexes yet preferentially bind to one. *PLoS One* **6,** (2011).

109. MacCallum, R. M., Martin, A. C. & Thornton, J. M. Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262,** 732–45 (1996).

110. Marcatili, P., Olimpieri, P. P., Chailyan, A. & Tramontano, A. Antibody structural modeling with prediction of immunoglobulin structure (PIGS). *Nat. Protoc.* **9,** 2771–2783 (2014).

111. Burrows, S. R. *et al.* Hard wiring of T cell receptor specificity for the major histocompatibility complex is underpinned by TCR adaptability. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10608–10613 (2010).

112. Gagnon, S. J. *et al.* T cell receptor recognition via cooperative conformational plasticity. *J. Mol. Biol.* **363,** 228–43 (2006).

113. Stone, J. D., Harris, D. T. & Kranz, D. M. TCR affinity for p/MHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. *Curr. Opin. Immunol.* **33,** 16–22 (2015).

114. David K. Cole, N. J. P. *et al.* Human TCR-binding affinity is governed by MHC class restriction. *J. Immunol.* **178,** 5727–5734 (2007).

115. Mareeva, T., Martinez-Hackert, E. & Sykulev, Y. How a T cell receptor-like antibody recognizes major histocompatibility complex-bound peptide. *J. Biol. Chem.* **283,** 29053–29059 (2008).

116. Dunbar, J., Knapp, B., Fuchs, A., Shi, J. & Deane, C. M. Examining Variable Domain Orientations in Antigen Receptors Gives Insight into TCR-Like Antibody Design. *PLoS Comput. Biol.* **10,** e1003852 (2014).

117. Nikoloudis, D., Pitts, J. E. & Saldanha, J. W. A complete, multi-level conformational clustering of antibody complementarity-determining regions. *PeerJ* **2,** e456 (2014).

118. Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* **37,** W474–9 (2009).

119. Von Boehmer, H. Unique features of the pre-T-cell receptor alpha-chain: not just a surrogate. *Nat. Rev. Immunol.* **5,** 571–577 (2005).

120. Dunn, S., Rizkallah, P. & Baston, E. Directed evolution of human T cell receptor CDR2 residues by phage display dramatically enhances affinity for cognate peptide-MHC without increasing apparent cross-reactivity. *protein Sci.* 710–721 (2006). doi:10.1110/ps.051936406.rearrangements

121. Sami, M. *et al.* Crystal structures of high affinity human T -cell receptors bound to peptide major histocompatibility complex reveal native diagonal binding geometry. *Protein Eng. Des. Sel.* **20,** 397–403 (2007).

122. Robbins, P. F. *et al.* Single and dual amino acid substitutions in TCR CDRs can enhance antigen-specific T cell functions. *J. Immunol.* (2008).

123. Nicholson, M. J., Hahn, M. & Wucherpfennig, K. W. Unusual features of self-peptide/MHC binding by autoimmune T cell receptors. *Immunity* **23,** 351–60 (2005).

124. Wucherpfennig, K. W., Call, M. J., Deng, L. & Mariuzza, R. Structural alterations in peptide-MHC recognition by self-reactive T cell receptors. *Curr. Opin. Immunol.* **21,** 590–595 (2009).

125. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196,** 901–17 (1987).

126. Chen, J., Sawyer, N. & Regan, L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **22,** 510–515 (2013).

127. Chervin, A. A. S., Aggen, D. H. D., Raseman, J. M. & Kranz, D. D. M. Engineering higher affinity T cell receptors using a T cell display system. *J. Immunol. Methods* **339,** 175–184 (2008).

128. C, M. *et al.* A New Twist In TCR Diversity Revealed By A Forbidden αβ TCR. *J. Mol. Biol.* **375,** 1306–1319 (2008).

129. Smith, K. *et al.* Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat. Protoc.* **4,** 372–84 (2009).

130. Call, M. E., Pyrdol, J., Wiedmann, M. & Wucherpfennig, K. W. The organizing principle in the formation of the T cell receptor-CD3 complex. *Cell* **111,** 967–79 (2002).

131. Letourneur, F. & Malissen, B. Derivation of a T cell hybridoma variant deprived of functional T cell receptor alpha and beta chain transcripts reveals a nonfunctional alpha-mRNA of BW5147 origin. *Eur. J. Immunol.* **19,** 2269–2274 (1989).

132. Maynard, J. *et al.* Structure of an autoimmune T cell receptor complexed with class II peptide-MHC: insights into MHC bias and antigen specificity. *Immunity* **22,** 81–92 (2005).

133. White, J., Haskins, K. M., Marrack, P. & Kappler, J. Use of I region-restricted, antigen-specific T cell hybridomas to produce idiotypically specific anti-receptor antibodies. *J. Immunol.* **130,** 1033–7 (1983).

134. Russel, M., Lowman, H. B. H. & Clackson, T. Introduction to phage biology and phage display. *Approach to Phage Disp.* 1–26 (2004).

135. Sidhu, S. S., Weiss, G. a & Wells, J. a. High copy display of large proteins on phage for functional selections. *J. Mol. Biol.* **296,** 487–95 (2000).

136. Zhao, Y. *et al.* High-efficiency transfection of primary human and mouse T lymphocytes using RNA electroporation. *Mol. Ther.* **13,** 151–159 (2006).

137. Kozak, M. & Limited, I. R. L. P. Nucleic acids research. *Nucleic Acids Res.* **38,** ii (2010).

138. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6 (1997).

139. Akopian, D., Shen, K., Zhang, X. & Shan, S. *Signal recognition particle: an essential protein-targeting machine. Annual review of biochemistry* **82,** (2013).

140. Pechmann, S., Chartron, J. W. & Frydman, J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat. Struct. & Mol. Biol.* **21,** 1–9 (2014).

141. Martoglio, B. & Dobberstein, B. Signal sequences: More than just greasy peptides. *Trends Cell Biol.* **8,** 410–415 (1998).

142. Hoover, T. R. *et al.* Signals and salvage sequences. *Nature* **329,** 855–857 (1987).

143. Crooks, G. & Hon, G. WebLogo: a sequence logo generator. *Genome Res.* 1188–1190 (2004). doi:10.1101/gr.849004.1

144. Vatakis, D. & McMillan, M. The signal peptide sequence impacts the immune response elicited by a DNA epitope vaccine. *Clin. Vaccine Immunol.* **18,** 1776–1780 (2011).

145. Furmanski, A. L. *et al.* public t cell receptor beta-chains are not advantaged during positive selection. *J. Immunol.* (2011). doi:10.4049/jimmunol.180.2.1029

146. Kessels, H. W., van Den Boom, M. D., Spits, H., Hooijberg, E. & Schumacher, T. N. Changing T cell specificity by retroviral T cell receptor display. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 14578–83 (2000).

147. Berntzen, G. *et al.* Prolonged and increased expression of soluble Fc receptors, IgG and a TCR-Ig fusion protein by transiently transfected adherent 293E cells. *J. Immunol. Methods* **298,** 93–104 (2005).

148. Kessels, H. & Wolkers, M. Immunotherapy through TCR gene transfer. *Nat.* … **2,** 957–961 (2001).

149. Favier, B., Burroughs, N. J., Wedderburn, L. & Valitutti, S. TCR dynamics on the surface of living T cells. *Int. Immunol.* **13,** 1525–1532 (2001).

150. Berry, R. *et al.* Structure of the chicken CD3ed/g heterodimer and its assembly with the abT cell receptor. *J. Biol. Chem.* **289,** 8240–8251 (2014).

151. Shirasu, N., Yamada, H., Shibaguchi, H., Kuroki, M. & Kuroki, M. Molecular characterization of a fully human chimeric T-cell antigen receptor for tumor-associated antigen EpCAM. *J. Biomed. Biotechnol.* **2012,** (2012).

152. Barrett, D. M., Singh, N., Porter, D. L., Grupp, S. a & June, C. H. Chimeric antigen receptor therapy for cancer. *Annu. Rev. Med.* **65,** 333–47 (2014).

153. Stern, B., Olsen, L. C., Tröße, C., Ravneberg, H. & Pryme, I. F. Improving mammalian cell factories : The selection of signal peptide has a major impact on recombinant protein synthesis and secretion in mammalian cells. *Trends Cell Mol. Biol.* **2,** 1–17 (2007).

154. Corse, E., Gottschalk, R. & Allison, J. Strength of TCR–peptide/MHC interactions and in vivo T cell responses. *J. Immunol.* (2011). doi:10.4049/jimmunol.1003650

155. Caruso, a *et al.* Flow cytometric analysis of activation markers on stimulated T cells and their correlation with cell proliferation. *Cytometry* **27,** 71–6 (1997).

156. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453,** 667–71 (2008).

157. Sidhu, S. S. & Weiss, G. A. Chapter 2 Constructing phage display libraries by oligonucleotide-directed mutagenesis. *Small* 27–42 (1991).

158. Call, M. E. & Wucherpfennig, K. W. Molecular mechanisms for the assembly of the T cell receptor-CD3 complex. *Mol. Immunol.* **40,** 1295–305 (2004).

159. Malissen, M. *et al.* T cell development in mice lacking the CD3-zeta/eta gene. *europepmc.org* **1,** 4347–4355 (1993).

160. Holler, P. D. & Kranz, D. M. Quantitative analysis of the contribution of TCR/pepMHC affinity and CD8 to T cell activation. *Immunity* **18,** 255–64 (2003).

161. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321,** 522–525 (1986).

162. George, a. J. T. *Antibody engineering*. *Endeavour* **18,** (1994).

163. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–786 (2011).

164. Kimata, Y. *et al.* KDEL and KKXX Retrieval Signals Appended to the Same Reporter Protein Determine Different Trafficking between ER, Intermediate Compartment, and Golgi Complex. *Mol. Biol. Cell* **14,** 2559–2569 (2003).

165. Porter, D. L., Levine, B. L., Kalos, M., Bagg, A. & June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *N. Engl. J. Med.* **365,** 725–33 (2011).

166. Maher, J., Brentjens, R. J., Gunset, G., Rivière, I. & Sadelain, M. Human T-lymphocyte cytotoxicity and proliferation directed by a single chimeric TCRzeta /CD28 receptor. *Nat. Biotechnol.* **20,** 70–75 (2002).

167. Park, T. S., Rosenberg, S. a & Morgan, R. a. Treating cancer with genetically engineered T cells. *Trends Biotechnol.* **29,** 550–7 (2011).

168. Schleicher, U., Röllinghoff, M., Gessner, a., Rollinghoff, M. & Gessner, a. A stable marker for specific T-cells: a TCR alpha/green fluorescent protein (GFP) fusionprotein reconstitutes a functionally active TCR complex. *J. Immunol. Methods* **246,** 165–174 (2000).

169. Thatte, J., Qadri, a, Radu, C. & Ward, E. S. Molecular requirements for T cell recognition by a major histocompatibility complex class II-restricted T cell receptor: the involvement of the fourth hypervariable loop of the Valpha domain. *J. Exp. Med.* **189,** 509–520 (1999).

170. Holler, P. D., Lim, a R., Cho, B. K., Rund, L. a & Kranz, D. M. CD8(-) T cell transfectants that express a high affinity T cell receptor exhibit enhanced peptide-dependent activation. *J. Exp. Med.* **194,** 1043–52 (2001).

171. Stone, J. D., Chervin, A. S., Aggen, D. H. & Kranz, D. M. *T cell receptor engineering. Methods in enzymology* **503,** (Elsevier Inc., 2012).

172. Szymczak, A. L. *et al.* Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat. Biotechnol.* **22,** 589–594 (2004).

173. Holtappels, R. *et al.* Subdominant CD8 T-cell epitopes account for protection against cytomegalovirus independent of immunodomination. *J. Virol.* **82,** 5781–96 (2008).

174. Messaoudi, I. & Patino, J. Direct link between mhc polymorphism, T cell avidity, and diversity in immune defense. *Science (80-.).* **298,** 1797–1800 (2002).

175. Mosquera, L. a *et al.* In vitro and in vivo characterization of a novel antibody-like single-chain TCR human IgG1 fusion protein. *J. Immunol.* **174,** 4381–4388 (2005).

176. Jensen, M. C. & Riddell, S. R. Design and implementation of adoptive therapy with chimeric antigen receptor-modified T cells. *Immunol. Rev.* **257,** 127–144 (2014).

177. Szymczak-Workman, A. A. L., Vignali, K. M. & Vignali, D. A. A. a. Design and construction of 2A peptide-linked multicistronic vectors. *Cold Spring Harb. …* **7,** 199–204 (2012).

178. Fang, J. *et al.* Stable antibody expression at therapeutic levels using the 2A peptide. *Nat. Biotechnol.* **23,** 584–90 (2005).