The Peanut-Specific TCR Repertoire: A Statistical Approach to Define Homology Among CDR3β Sequences in Peanut-Allergic Individuals

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:42004158

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
The Peanut-specific TCR Repertoire:

A Statistical Approach to Define Homology Among CDR3β Sequences in Peanut-allergic Individuals

Neal Smith

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2019
Abstract

The induction and persistence of peanut allergies is fundamentally dependent on T cells’ ability to recognize allergenic antigens. Peanut-specificity of a T cell is conferred by its TCR and more specifically, the CDR3 of the β-chain. In this study, we elucidated 7345 putatively peanut-specific CDR3βs from 27 peanut-allergic individuals by applying a novel statistical approach to an existing TCRβseq data set that consisted of in vitro peanut-activated and memory resting T cell sequences. These putatively-specific sequences displayed enhanced homology and reduced diversity as compared to randomly sampled CDR3βs, which is consistent with independently published antigen-specific populations. Motif analysis revealed 149 unique motifs that were both highly enriched and public within the psCDR3s. Given peanut is a polyantigenic stimulant, we illustrated the local homology of the psCDR3s with a network analysis that clustered sequences based on the presence of either an enriched motif, a low Levenshtein distance (distance = 1), or evidence of convergent recombination as defined by multiple nucleotide rearrangements that corresponded to the same amino acid sequence. This network was broken into 990 distinct neighborhoods that were individually probed for their tendency to represent T_{eff} or T_{reg} sequences. Using bulk T_{eff} and T_{reg} data from 9 of the 27 individuals in this study, we discovered 253 T_{eff}- and 74 T_{reg}-exclusive neighborhoods, suggesting differences in antigen-recognition across the phenotypic compartments. The analysis approach described here can be a template for future studies aimed at understanding TCR repertoires where epitope and/or HLA genotype information is unknown and has the potential to define antigen-specific repertoire features that could be used as biomarkers of disease.
Dedication

This thesis is dedicated to my wife Julie McCarthy, who has been a constant source of support and motivation through the entire process.
Acknowledgments

I’d like to thank everyone in the Shreffler Laboratory at MGH who have provided resources and guidance that made the completion of this thesis possible. Bert Ruiter was instrumental in data generation and helped with interpretation of the analysis. In addition, Bert provided to me most of the current literature in the fields of allergy and immunology that would be useful for this analysis, helping to provide context to the work presenting here. Yamini Virkud assisted with the implementation of the statistics that were applied to the data. Sarita Patil provided thoughtful feedback regarding repertoire analysis and how to interpret the data. Lastly, I’d like to thank Wayne Shreffler for being an excellent mentor and helping guide the analysis in the most meaningful possible way. Wayne’s support and feedback has made this a very rewarding experience.
Table of Contents

Abstract .......................................................................................................................... iii
Dedication ......................................................................................................................... iv
Acknowledgments .......................................................................................................... v
Table of Contents ........................................................................................................... vi
List of Tables .................................................................................................................. viii
List of Figures ............................................................................................................... ix

I. Introduction ................................................................................................................. 1
   T cell receptor repertoire diversity ............................................................................. 1
   T cell phenotypes ......................................................................................................... 3
   Examination of the public and private T cell repertoire ............................................. 4
   TCR homology and motif discovery ............................................................................ 7
   Peanut allergies: Basic information and immune mechanism .................................. 12
   CD154 up-regulation Assay ....................................................................................... 15
   TCRβseq ....................................................................................................................... 17

II. Research methods ...................................................................................................... 19
   TCRβseq ....................................................................................................................... 19
   Subjects and samples .................................................................................................. 20
   Elucidation of psCDR3s .............................................................................................. 21
   Examination of the most expanded and public psCDR3s .......................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence distance metrics</td>
<td>23</td>
</tr>
<tr>
<td>Shannon Entropy</td>
<td>23</td>
</tr>
<tr>
<td>Generation of nmers and enrichment determination</td>
<td>24</td>
</tr>
<tr>
<td>Network analysis</td>
<td>25</td>
</tr>
<tr>
<td>Examination of $T_{eff}$ and $T_{reg}$ CDR3$\beta$ sequences</td>
<td>26</td>
</tr>
<tr>
<td>Code for data analysis</td>
<td>27</td>
</tr>
<tr>
<td>III. Results</td>
<td>28</td>
</tr>
<tr>
<td>psCDR3s are more homologous than the repertoires they are selected from</td>
<td>28</td>
</tr>
<tr>
<td>Motif analysis of psCDR3s</td>
<td>32</td>
</tr>
<tr>
<td>Network analysis of psCDR3s</td>
<td>35</td>
</tr>
<tr>
<td>Examination of $T_{eff}$ and $T_{reg}$ psCDR3$s$</td>
<td>36</td>
</tr>
<tr>
<td>IV. Discussion</td>
<td>41</td>
</tr>
<tr>
<td>Appendix 1</td>
<td>49</td>
</tr>
<tr>
<td>Appendix 2</td>
<td>52</td>
</tr>
<tr>
<td>References</td>
<td>55</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Nmer enrichment information ................................................................. 35
Supplemental table 1. Summary of Unique CDR3β sequences for each subject .......... 49
Supplemental table 2. Total Read counts of CDR3β sequences for each subject .......... 50
Supplemental table 3. Total CDR3β and psCDR3 Teff and Treg data ..................... 51
List of Figures

Figure 1: Simulated TCR power law distribution ................................................................. 5
Figure 2: generation of nmers. ............................................................................................... 25
Figure 3: Percent of total CDR3β sequences to meet psCDR3 parameters .................. 30
Figure 4: Distance metrics of psCDR3s............................................................................. 31
Figure 5: Entropy of psCDR3s. ......................................................................................... 33
Figure 6: Dominant nmers in the psCDR3s. ....................................................................... 34
Figure 7: Network analysis of psCDR3s. ............................................................................ 38
Figure 8: Evaluation of network structure. ....................................................................... 39
Figure 9: psCDR3s in Teff and Treg samples. ................................................................. 39
Figure 10: neighborhood assessment of Treg and Treg. .................................................. 40
Supplemental Figure 1: distribution of public psCDR3s vs. number of subjects .......... 52
Supplemental Figure 2: Overlap across top 100 sequences in each CDR3β pool ........ 52
Supplemental Figure 3: Dominant nmers in the psCDR3s. ............................................. 53
Supplemental Figure 4: Examination of subsets within network analysis. .................... 54
T cells are a group of adaptive immune cells necessary for effective host-response to a wide range of pathogens and they are defined by their T cell receptor (TCR), which confers specificity for a given target. In addition, T cells can also play a pathogenic role, eliciting immune responses to self-antigens causing autoimmunity or innocuous substances that drive allergic diseases. Peanut allergies, which affect more than 1% of the US population, are driven by a pathogenic T cell response, and thus understanding the peanut-specific TCR repertoire could lead to insights regarding disease pathology while also discovering biomarkers and therapeutic targets (Sicherer & Sampson, 2018). The goal of this study was to adapt current TCR repertoire analysis methods to find novel features of the peanut-specific repertoire in allergic individuals.

T cell receptor repertoire diversity

The adaptive immune system is necessary for an effective host response to pathogens in vertebrates and is defined by populations of antigen-specific cells including B and T cells. T cells are defined by their antigen-specific TCR which, when bound to its target, will cause activation of the cell (Dempsey, Vaidya, & Cheng, 2003). The TCR is a heterodimeric protein that is made up of an α- and β-chain that together confer a cell’s antigen-specificity. The array of all TCRs expressed by an individual is referred to as the
TCR repertoire, and its diversity is essential in host defense (Hou, Wang, Ding, Xie, & Diao, 2016) (Laydon, Bangham, & Asquith, 2015). There are two distinct mechanisms that contribute to repertoire diversity, the first of which is gene rearrangement during T cell development. Both chains of a functional TCR are comprised of the combination of multiple gene segments known as the variable (V), joining (J), and constant (C) genes. The β chain also includes an additional diversity (D) gene. In the genome, there are multiple unique copies of each one of these genes, and during T cell development, one of each (for both chains) is chosen to create a functional protein in a process known as gene rearrangement. Gene rearrangement contributes significantly to the diversity of the repertoire, as there are 2679 possible segment combinations of the α chain and 2808 possible combinations of the β chain (Hou et al., 2016) (Laydon et al., 2015). The diversity increases exponentially during the process of gene rearrangement when random nucleotides are added in the V region and at the V(D)J junction to create complementarity determining regions (CDRs). CDR1 and CDR2 reside in the V region, while the very important CDR3 sits in the gene junction (Laydon et al., 2015). Recent estimates indicate a healthy human TCR repertoire could comprise as many as $10^{11}$ unique sequences, but the accuracy of this number is still a topic of debate (Laydon et al., 2015). The difficulty in estimating repertoire diversity stems from the quasi-random manner that it is created. All T cells undergo selection in the thymus to eliminate sequences that have the potential to be auto-reactive, and this trimming is not uniform, as it will depend at least partially on the major histocompatibility complex (MHC) genotype of an individual. The MHC is a cell surface protein complex that is made up of human leukocyte antigen (HLA) proteins which are highly polymorphic. Differences in HLA
alleles create distinctions among the antigens presented to T cells as well as the T cell affinity for a given antigen across different genetic backgrounds, thus adding extra layers of complexity when trying to estimate repertoire diversity (Wieczorek et al., 2017).

T cell phenotypes

Once activated by antigen-presentation through professional antigen-presenting cells, T cells are driven to differentiate, inducing significant phenotypic changes that allow for subsequent function. Historically, differentiated cells have been classified into groups based on their function. Effector T (T_{eff}) cells are those that are generally defined by causing a pro-inflammatory response, which may be of several distinct types, while Regulatory T (T_{reg}) cells, broadly speaking, modulate immune responses, suppressing T_{eff} and other pro-inflammatory cells and helping maintain homeostasis (Josefowicz et al., 2012; Wambre et al., 2017). While a diverse repertoire is important for immunity to pathogens, alterations and abnormalities in both the entire and T cell subset specific repertoires have been associated with human diseases. For example, the repertoire of the T_{reg} compartment is restricted in individuals with different types of severe autoimmunity (Rowe et al., 2017). In contrast, an increase in global repertoire diversity was observed in individuals with atopic dermatitis, a chronic inflammatory skin disease (Brunner et al., 2017). Studies such as these demonstrate that TCR repertoire makeup is likely to affect disease pathogenesis in multiple contexts.

One important group of T cells are helper T (T_{H}) cells, which can be categorized into multiple subsets that will be selectively activated (and thus be driven towards a T_{eff} phenotype) based on the type of infection present and are defined by the cytokines they
secrete. For example, TH1 cells most often respond to intracellular pathogens and produce IFN-γ, TH2 cells are responsible for defense against helminth parasites and defined by the production of IL-4, IL-5 and IL-13, while TH17 cells are important in bacterial and fungal infections, secreting IL-17 (Walker & McKenzie, 2018). Although TH cells are necessary for host defense, dysregulation of these cells can have pathogenic consequences. When left unchecked by Treg cells, an expansion of self-reactive TH1 and TH17 populations can lead to autoimmune disorders (Jager & Kuchroo, 2010). TH2 cells have been found to drive immune reactions to innocuous foreign substances, contributing to the inflammation associated with allergy and allergic asthma (Walker & McKenzie, 2018) (Sicherer & Sampson, 2018). Understanding the TCR repertoire of TH cells in all contexts could lead to insights about normal immune responses to infectious diseases as well as pathogenic responses to self-antigens and allergens.

Examination of the public and private T cell repertoire

One goal of current research is to associate certain TCR sequences with specific clinical phenotypes. The way in which repertoires are currently analyzed to elucidate associations however is not universal, with researchers taking a variety of approaches. A general goal of repertoire analysis studies has been to define the “public” and “private” compartment, with a public TCR defined by either an identical DNA or amino acid sequence across multiple individuals (Venturi et al., 2006) (Ruggiero et al., 2015) (Hou et al., 2016). Given the difficulty of obtaining matched TCRα and TCRβ sequences until recent developments in single cell analysis technology, most studies examine the public nature of a single chain. In addition, the majority of TCR repertoire analyses have
focused solely on the CDR3 because it is the part of the receptor that most closely interacts with the presented antigen (Hou et al., 2016). Public TCR sequences have been described in a wide variety of immune responses, which is immediately intriguing given the vast possibility of recombination events possible during T cell development would suggest their prevalence would be much lower than what has been observed (Li, Ye, Ji, & Han, 2012). While the immunologic explanation for public T cell responses is yet to be completely elucidated, it is thought that studying these sequences and their corresponding specificities can reveal insights about dominant antigens in an immune response which could help direct the development of vaccines, therapies and diagnostics, while also enhancing our understanding of biases in the process of recombination (Li et al., 2012) (Miles, Douek, & Price, 2011).

Statistically, the distribution of TCR sequences in a repertoire can be approximated by a power law, where the number of clones of a given frequency decays as a power of the number of clones in the repertoire (figure 1). The majority of clones

Figure 1: Simulated TCR power law distribution. TCR sequences in a repertoire can be estimated by a power law, where a small proportion of sequences are hyper-expanded.
are found at a very low frequency and a small minority of clones dominate the repertoire, which is driven by rapid proliferation of select T cells taking part in an active immune response (Koch, Starenki, Cooper, Myers, & Li, 2018). Interestingly, the structure of this distribution at homeostasis or early in life is also strongly right-skewed, indicating that during an immune response, the TCRs that make up the dominant space change, but the distribution of the overall repertoire remains the same. It also suggests underlying biases in the process of TCR selection in the absence of an immune response.

In an attempt to analyze the dominant CDR3s that are present in the repertoire, Madi et al. analyzed the top 1000 sequences by frequency of inbred mouse repertoires at homeostasis and discovered that early in life, dominant clones are often public, which is likely due to the similar thymic selection inbred mice with the same MHC genotype will experience.

However, through immune progression and aging, the dominant sequences of the repertoires were more likely to be made up of private clones that had expanded based on exposure to foreign antigens (Madi et al., 2017). To test the hypothesis that antigen-exposure leads to the expansion of mostly private clones, the researchers immunized mice with either a foreign (OVA) or self (HSP60) antigen and examined the top 1000 clones by frequency. The dominant CDR3s that populated the tail of the distribution were more likely to be private than what was seen pre-immunization, implicating the private repertoire to be a significant part of a normal immune responses (Madi et al., 2017). However, public clones were also present within the most expanded CDR3βs post-immunization, suggesting the recognition of dominant epitopes is at least partially responsible for public TCR sequences. Researchers in the tumor biology realm have also tried to leverage the observed power-law distribution to look for differences between
healthy individuals and patients with glioma, a type of tumor that develops in the central nervous system. General Perturbations in the TCR repertoire of glioma patients have been linked to immune progression and therefore, Sims. et al. hypothesized monitoring patients’ repertoires was a viable method to predict particular immunophenotypes (Sims et al., 2016). The researchers analyzed the TCR repertoire of peripheral blood mononuclear cells (PBMCs) of 14 glioma patients and examined the top 1000 CDR3s from both the α and β chains from each individual. This analysis revealed a subset of sequences that were highly public among the cohort. Interestingly, the presence of the public sequences seems to segregate the patients into two groups, with a subset of them having many dominant shared CDR3s. In addition, the groupings of the patients remained the same when examining just the CDR3α or CDR3β (Sims et al., 2016). As a comparison, the researchers performed the same analysis on 6 healthy individuals and in every case, they would cluster with the group of patients that had a dominant public repertoire, suggesting a divergence in the peripheral TCR repertoire in a subset of individuals with glioma. In their search for a non-invasive biomarker, the researchers suggest that looking for divergence in the dominant repertoire observed in PBMCs could be an early indicator of glioma, although given there were no CDR3s that distinguished healthy vs. all glioma patients, more work would need to be done to increase the sensitivity of the proposed diagnostic (Sims et al., 2016).

TCR homology and motif discovery

Working under the assumption that identical TCR amino acid sequences in distinct individuals bind the same epitope, examination of public TCRs can garner
insights into shared antigen-recognition. However, it has been demonstrated that distinct TCR sequences have the capability to bind the same epitope and therefore limiting analysis to just public sequences will likely underestimate the amount of shared antigen-recognition across repertoires (Glanville et al., 2017) (Ryan et al., 2016). To better understand shared antigen-recognition, recent work has been aimed at defining features of unique TCRs that confer the same specificity. The general approach has been to look for homology among different TCR amino acid sequences, however defining biologically relevant homology is not straightforward. This metric is likely different for distinct antigens and individuals as HLA genotypes will significantly alter the interaction between a TCR and the presented epitope.

In recent years, researchers have been trying to find TCR “motifs”: common amino acid sequences that likely confer antigen-specificity. Creating libraries of antigen-specific motifs has many useful applications. Clinically, motifs could be biomarkers for T cell related disorders (e.g., autoimmunity) or an indicator of an infection. In addition, tracking of motifs could indicate repertoire divergence that is associated with a clinical phenotype, as was the goal for Sims et al. when examining the TCR repertoires of Glioma patients (Sims et al., 2016). From a research perspective, knowledge of well-defined antigen-specific motifs could be used to implicate the specificity of a TCR without a priori knowledge of the epitope. More easily defining the specificity of TCRs would enhance basic immunological investigation, as certain phenotypes would be able to be connected with a given antigen without the need for functional binding assays. Given antigen-specific TCR motifs could be used to advance clinical readouts as well as
help basic immunology research, their identification for any given antigen is an important and worthwhile endeavor.

Researchers studying autoimmunity have made significant progress in defining self-specific motifs that drive pathology. Stadinski et al. defined a common biochemical motif when analyzing 53 self-reactive CDR3βs, in which there were hydrophobic amino acids at positions 6 and 7 (IMGT positions 109 and 110) of the sequences, no matter the length of the CDR3β (Stadinski et al., 2016). An independent report by Rowe et al. confirmed the expansion of this biochemical motif in the CD4+ T cell repertoire of 5 patients with combined immunodeficiency and 1 patient with atypical severe combined immunodeficiency syndrome (Rowe et al., 2017). Interestingly, the T_{reg} compartment of these patients showed no expansion of the motif, implicating T_{eff} cells in auto-reactivity.

In another study in the autoimmunity realm, Komech et al. recently tried to find similarities among the CDR3βs in the CD8+ compartment of patients with ankylosing spondylitis, a type of arthritis that affects the spine and causes inflammation of synovial membranes. A common feature of the repertoires of these patients was a group of public CDR3βs that were enriched in the synovial fluid (Komech et al., 2018). In addition, these sequences were all the same length and shared exact amino acid residues at 11 of the 15 positions in the CDR3β, with clear amino acid preferences at the remaining 4 positions (ex. position 8: 50% phenylalanine, 50% Tyrosine). The sequences were also biochemically similar when considering the hydrophilicity of the residues as measured by the Hopp-Woods index, suggesting similar antigen specificity. While these publications report observations of potential motifs and overall homology within the repertoire of
certain clinical phenotypes, they do not define a methodological approach that would be useful when searching for TCR motifs that confer antigen-specificity.

To enhance the study of antigen-specific repertoires, Ritvo et al. developed a homology-based method for motif discovery called the TCR neighbor enrichment test (TCRNET), where two CDR3s are defined as “neighbors” if they are within a single amino acid substitution of each other (Ritvo et al., 2018). This procedure compares the number of neighbors a given CDR3 has within a population of interest (i.e., thought to be enriched for specificity to a particular antigen) and determines the statistical significance of this count by comparing the number of neighbors generated from a given control (e.g., randomly selected) dataset. The sequences in neighborhoods that met their statistical criteria (q < 0.05 after FDR correction) were evaluated for noticeable amino acid sequence patterns that could be implicated in specificity. Using this method, researchers examined the repertoires of different T cell subsets in mice post-immunization with either a foreign (OVA) or self (insulin) antigen to better understand specificity in an immune response. The authors were able to find CDR3β “neighborhoods” that were restricted to particular T cell subsets (e.g. Teff cells) and immunizations, where clear amino acid patterns that suggested the presence of motifs emerged (Ritvo et al., 2018). It is important to note though that this approach simply creates neighborhoods of oligoclonal sequences likely to contain motifs, however they never define any particular amino acid motifs from the OVA- or insulin-immunized mice. There are inherent benefits to the TCRNET approach, as it creates clusters of homologous CDR3s likely to reveal amino acid patterns that confer specificity. However, it is somewhat restrictive, given a neighborhood will only contain CDR3s of the same length. Other reports have
demonstrated that CDR3s of different lengths are capable of having the same specificity and thus if a particular analysis uses sequence length as an initial prerequisite for the clustering of CDR3s, it is likely underestimating certain motifs that could confer specificity (Glanville et al., 2017) (Stadinski et al., 2016) (Ryan et al., 2016).

Glanville et al. recently published a method of finding antigen-specific CDR3 amino acid motifs from tetramer selected CD4 T cells (Glanville et al., 2017). In short, tetramer selection uses four fluorescently tagged MHC-peptide antigen complexes that will bind to antigen-specific T cells and used to physically isolate them (e.g., by FACS) despite their low frequency ($<10^{-3}$). Because the affinity of TCR-peptide/MHC binding is low, a multimeric construct such as this is necessary for sufficient overall avidity (and therefore the sensitivity/specificity characteristics of the assay). In this report, Glanville et al. isolated cells specific for common pathogen epitopes (Epstein-Barr virus, cytomegalovirus, and influenza) and sequenced the TCR CDR3 ($\alpha$ and $\beta$). They then used unselected naïve T cell CDR3s as a control sample set to look for enrichment of amino acid sequences of lengths 2, 3 and 4. They restricted their approach by looking only at those amino acid residues likely to make antigen contact by virtue of their position in the CDR3 as was determined by examination of 52 public TCR-pMHC crystal structures. This approach identified motifs in antigen-specific CDR3s shared across multiple individuals that were statistically enriched relative to naïve cells. Their use of an unselected naïve pool as the control population, versus the tetramer negative memory pool from the same donors may be a weakness in study design, given the data referenced above regarding naïve populations, which have not undergone antigen selection and expansion events, and therefore more closely conforming to the random power law
distribution. Against that naive population, even TCR clones pulled out by non-specific binding (i.e., false positives), could differ if they represent an expanded clone from a given patient. That is likely significantly less of a concern when selecting putatively selective cells by tetramer than by other means. The benefit of their approach is, in fact, based upon that high degree of antigen specificity, as using tetramer selection allows you to use a single well-defined antigen. The costs of such a method come in the form of technical difficulty and limited scope. Tetramers need to be developed in an MHC-specific way, and must be matched for use to the corresponding MHC alleles (which are present at several highly polymorphic loci) expressed by the T cell donor. Given the polymorphic nature of MHC alleles, the cost and time it takes to develop tetramers for many people across different genetic backgrounds becomes very burdensome. In addition to the MHC complexity, most immune responses are polyantigenic. Each tetramer molecule holds a single antigenic peptide, while exogenous proteins contain many immunogenic peptides and epitope mapping information of even the most dominant antigens is often limited. Limiting the scope of TCR repertoire analysis to only known epitopes will likely provide an incomplete understanding of T cell mediated immune responses.

Peanut allergies: Basic information and immune mechanism

Peanut allergy is a rising public health concern in the US, now affecting more than 1% of the population with the prevalence steadily increasing. There have been many risk factors associated with the development of peanut allergies, both environmental (e.g., microbial exposure, diet, etc.) and genetic, but we still do not have a
complete understanding as to why the incidence is increasing at such a rapid rate (Sicherer & Sampson, 2018). At the moment, there is no approved treatment for peanut allergies, and the current standard of care is strict dietary avoidance of the allergen. In order to make effective therapeutics for the treatment of peanut allergies, we must continue to research the molecular mechanisms involved in the pathology of the disease.

Once the allergic response is established, acute and potentially severe reactions are dependent upon allergen-specific IgE antibody responses that can lead to the degranulation of mast cells and basophils (Kulis, Patil, Wambre, & Vickery, 2018). Activation of effector Th2 cells are also associated with allergic reactions, although their association with the production and/or maintenance of allergen-specific IgE (if any) is unclear at this point (Kulis et al., 2018; Sicherer & Sampson, 2018; Wambre et al., 2017). Once activated, Th2 cells are driven to a T_{eff} cell phenotype, secreting the cytokines IL-4, IL-5, and IL-13 which will promote an IgE-mediated response. In contrast to the Th2 effector-function, T_{reg} cells are crucial for tolerance to allergens. Previous work in murine models has shown that T_{reg} deficiencies result in exacerbated Th2-driven inflammation, while in humans, it has been demonstrated that individuals who outgrow milk allergies have higher rates of milk-specific T_{reg} cells in circulation (Josefowicz et al., 2012; Shreffler, Wanich, Moloney, Nowak-Wegrzyn, & Sampson, 2009). Given the prominent role T cells play in allergic immune responses, understanding aspects of the peanut-specific repertoire could give us a better understanding of the pathology of peanut allergies that could help develop future therapeutics.

The majority of published TCR repertoire analyses have been in the context of infectious disease and autoimmunity, while literature that analyzes either the peanut- or
other allergen-specific repertoires of individuals afflicted with allergies is lacking. In one of the only reports on this topic, Ryan et al. examined the T cell response of individuals who successfully completed peanut oral immunotherapy, a treatment aimed to desensitize individuals to the allergen by giving them incrementally higher amounts of peanut over time (Ryan et al., 2016). The authors selected out T cells by TCR affinity to MHC loaded with a known peptide epitope from the peanut allergen, Ara h 2 and performed an enrichment analysis similar to that of Glanville et al. described above (Glanville et al., 2017). They identified four potential amino acid motifs of length 3 that were statistically enriched in comparison to those found in an unselected naïve TCR library. However, they only had 13 Ara h 2-specific CDR3s from a single subject to use for this analysis, making it very difficult to draw any conclusions about the prevalence and importance of these motifs in antigen-binding. In addition, like any approach relying on identification of specific cells by binding, it is limited by the authors’ a priori knowledge of specific antigenic epitopes (in this case, a single epitope), on given MHC backgrounds, despite our knowledge that the T cells response is polyantigenic. In fact, the polyantigenic nature of T cell responses to grass pollen was demonstrated by Schulten et al., as they described 56 distinct proteins that were able to elicit a Th2 response as measure by IL-5 secretion from PBMC cultures (Schulten et al., 2013). In addition, the authors found that only 34 proteins were reactive to allergic individuals’ IgE. The epitopes used in all current research that selects peanut-specific T cells by affinity have been from proteins known to elicit IgE responses, which likely represents a subset of the entire peanut-specific repertoire.
CD154 up-regulation Assay

CD154 (also known as CD40L) is a cell-surface protein induced to be expressed during T cell activation, which functions to provide costimulatory signals that can activate CD40+ cells, including B cells, DCs and monocyte/macrophages (Chattopadhyay, Yu, & Roederer, 2006) (Frentsch et al., 2005). With this phenotypic information in mind, researchers have developed a CD154 up-regulation assay to study antigen-specific Th cells. In this assay, T cells (as well as antigen-presenting cells) are cultured in the presence of a given antigen for anywhere from 6 to 20 hours. The cells then undergo fluorescence activated cell sorting (FACS) to isolate the cells activated in culture based on their CD154 expression (Chattopadhyay et al., 2006) (Frentsch et al., 2005). The CD154 up-regulation assay has been used in multiple contexts as a means to isolate antigen-activated T cells. Since it has been described, its validation has been based on phenotypic data, where the stimulated cells express expected cytokines in the context of the antigen used. For example, when it was first used with birch pollen as the stimulant, the CD154+ T cell pool had up-regulated IL-4, the hallmark Th2 cytokine (Frentsch et al., 2005). More recently, Archila et al. used this assay to examine cashew-specific T cells in cashew allergic individuals. The CD154+ population they isolated when using cashew allergens as stimulants expressed IL-4, IL-5 and IL-13 which are all Th2 promoting cytokines. Interestingly, they also saw a subpopulation of cells that had aspects of both Th2 and Th17 phenotypes, indicating the likely heterogenous nature of an allergic response (Archila et al., 2016). Importantly, similar results have been shown using peanut allergens, where a predominant Th2 population is found in the CD154+
activated pool, enabling researchers to better define the phenotype of peanut-activated T cells (Wambre et al., 2017) (Weissler et al., 2018) (Chiang et al., 2018).

There are limitations associated with the CD154 up-regulation assay, namely concerns with bystander activation. Bystander activation is a phenomenon where the antigen-specific cells activated in vitro will secrete IL-2, which non-specifically activates cells in a paracrine manner leading to contamination of what some researchers consider an antigen-specific pool (Di Genova, Savelyeva, Suchacki, Thirdborough, & Stevenson, 2010). When trying to analyze the TCRs of in vitro activated CD154+ cells, it is necessary to account for this contamination as to not include non-specific sequences. In addition, the inability to definitively elucidate the antigen/epitope specificity of the activated cells is another limitation associated with the CD154 up-regulation assay. When using a more stringent method such as tetramer selection, cells are selected based on their ability to bind a single specific MHC-bound epitope. In contrast, it is impossible to know the exact nature of antigen processing/presentation by the antigen presenting cells in vitro, nor are you able to associate an antigen-presentation event with a specific CD154+ cell. This issue is magnified when the CD154 up-regulation assay is used with a polyantigenic stimulant (e.g. whole peanut extract), as there will be diversity in the CD154+ TCR repertoire beyond that seen with approaches that select cells based on their affinity to a single MHC-bound epitope. Therefore, applying a method to group CD154+ TCRs in a biologically meaningful way is important when trying to examine the repertoires of antigen-activated cells.
TCRβseq

TCRβseq is the parallel sequencing of an entire library of TCRβ rearrangements to obtain the nucleotide and subsequent amino acid sequences of the CDR3 region. TCRβseq is a popular method for studying TCR repertoire diversity because the CDR3β is the most polymorphic region of either TCR chain (Hou et al., 2016). In addition, the CDR3β most closely interacts with the presented epitope, containing more amino acid residues that are found within 5 Å of the antigen during presentation than other CDRs, likely conferring specificity of a TCR (Glanville et al., 2017). To obtain TCRβseq data, genomic DNA or mRNA can be used as input and both methodologies have certain benefits (Rosati et al., 2017). Genomic DNA (gDNA) is altogether more stable than mRNA and because each T cell has a single rearranged template, there is theoretically perfect correspondence to cell number and therefore not only repertoire diversity (unique TCRs) but their relative expansion as well. However, a gDNA approach necessitates the PCR amplification of all rearranged VDJ genomic regions with complex sets of paired primers that may introduce amplification biases and skew this quantification. In contrast, the associated PCR amplification with mRNA is technically less challenging as all transcripts are rearranged and the unique sequences obtained can represent well the diversity of the repertoire. However, uneven numbers of TCR transcripts from each T cell makes elucidation of clonal expansion impossible, thus limiting the ability to determine the dominant sequences in a repertoire.

The objective of this thesis was to develop an approach to TCR CDR3β repertoire analysis, using previously obtained TCRβseq data from peanut allergic subjects, to
improve the specific identification and characterization of an antigen-specific polyclonal TCR repertoire to a complex antigen. Over the past 3 years, the CD154 up-regulation assay has been used with modern sequencing methods to collect peanut-activated and resting memory T cell CDR3βs from the same subjects. In the work presented here, this data was used to create a statistical approach to define likely antigen-specific sequences which were then interrogated for possible motifs that could confer specificity. Using a combination of sequence homology and motif discovery, CDR3βs were binned into groups that are likely to bind the same antigen. Lastly, the T_{eff} and T_{reg} compartments of a subset of the same peanut-allergic individuals were surveyed with the binned sequences to determine if there are any differences in antigen-recognition in each compartment.

The new methodology described in this study can be applied to research examining specific repertoires of T cell-mediated diseases beyond just allergies, and it has the ability to provide new insights to antigen-specificity that could lead to future therapeutics or biomarkers of disease.
Chapter II
Research methods

The following section describes the data generation protocols that preceded this study as well as detailing the data analysis used to examine the TCR repertoires of peanut allergic individuals. Briefly, TCRβseq data had been obtained from the peanut activated and resting memory T cell compartment as well as bulk T\textsubscript{eff} and T\textsubscript{reg} subsets with a kit from Adaptive Biotechnologies. Statistically enriched sequences in the peanut-activated compartment were elucidated and examined for features that could confer specificity. The enriched sequences were then clustered using a network analysis and these clusters were probed for the presence of T\textsubscript{eff}- and T\textsubscript{reg}-specific CDR3βs.

TCRβseq

To obtain rearranged CDR3β nucleotide and amino acid sequences, a proprietary method by Adaptive Technologies (Seattle, WA; TCRB Immunoseq kit) was used. gDNA was taken and underwent multiplex PCR with forward primers designed against all possible V-genes and reverse primers against all J-genes. The amplified products were then sequenced using the Illumina sequencing platform (Robins et al., 2009). Sequencing reads were corrected based on a baseline developed from a suite of synthetic templates, primer concentrations and primer bias’ known to be common among multiplex PCR reactions. Sequences were then filtered on the basis of TCRβ V-, D-, and J-gene
sequences as defined in the IMGT database (www.imgt.org) and binned with a modified nearest-neighbor algorithm to merge closely related sequences as a way to remove PCR and sequencing errors.

Subjects and samples

The TCRβseq data used for this study were derived from research participants that were a part of a peanut oral immunotherapy trial (NCT01750879) at the Food Allergy Center at Massachusetts General Hospital. Criteria to be screened for the study included having a previous diagnosis of peanut allergy, a history of peanut-induced reactions consistent with immediate hypersensitivity and peanut-IgE serum levels greater than 0.35 kU/L (ImmunoCAP; Thermo Fisher). The subjects then underwent a double-blind placebo-controlled food challenge to confirm peanut allergy. 27 individuals reacted during peanut challenge and their data was used in this study.

The TCRβseq data used for this study was derived from CD154 up-regulation assays aimed to isolate and examine peanut-specific T cells of allergic subjects. In short, peripheral blood mononuclear cells (PBMC) were isolated and cultured overnight in 1mL AIM-V medium (Gibco) with 100 µg/mL peanut extract. The next morning, cells were harvested and prepared for FACS by staining with AF700-conjugated anti-CD3 (clone UCHT1), APC-Cy7-conjugated anti-CD4 (RPA-T4), FITC-conjugated anti-CD45RA (HI100), PE-conjugated anti-CD154 (all from BD Biosciences) and Live/Dead Fixable Violet stain (L34955; Thermo Fisher). Live CD3+CD4+CD45RA-CD154+ cells were sorted to obtain the peanut-activated memory T cell population while live CD3+CD4+CD45RA-CD154- cells were sorted to isolate the resting memory T cell
population. Sorted T cells were lysed in buffer RLT+ with 1% β-mercaptoethanol (Qiagen), DNA was isolated with the AllPrep DNA/RNA Micro Kit (Qiagen) and TCRβseq data was generated for both the peanut-activated and resting populations as described above using the Adaptive Biotechnologies TCRB Immunoseq kit. Use of this pre-existing data was approved by the Massachusetts General Hospital IRB under protocol 2018P001824/PHS and the Harvard IRB under protocol IRB18-1416.

Elucidation of psCDR3s

TCRβseq data from activated (CD154+) and resting (CD154-) memory T cell samples from all 27 peanut allergic individuals was parsed using the tcR package (version 2.2.3) in R (Nazarov et al., 2015). The CD154 up-regulation assay has been used to study the phenotype of peanut-activated cells in vitro, however there are concerns with this method regarding bystander activation that could contaminate the activated CD154+ CDR3βs. Therefore, we applied a statistical method to define putatively-specific CDR3βs (psCDR3s), sequences that are likely to be specific and not be present due to bystander activation. This methodology begins by performing a G-test of independence on every CDR3 found in an individual’s activated CD154+ sample, determining if the proportion of a given sequence is higher in the activated CDR3s versus the resting CD154- CDR3s. A G-test (also known as a likelihood ratio test) can be described by the formula:

$$G = 2 \sum_i O_i \ln \left( \frac{O_i}{E_i} \right)$$
where $O_i$ is the observed read count of a given CDR3 in a particular population (activated or resting) and $E_i$ is the expected count based on the proportion of this CDR3 in the entire memory T cell population (pooled activated and resting CDR3s). With the $G$ value for each CDR3, the probability that the proportion a CDR3 in the activated compartment is derived by chance from the proportion in the resting compartment can be calculated. Given the sizable number of CDR3s that were analyzed from each individual’s data, a false discovery rate (FDR) correction was used to generate adjusted $P$-values (q-value) for each sequence, and those that met a cutoff of q < 0.05 were considered significant for this study. To enhance the stringency and reduce type 1 error, we also filtered out those with a read count of 1 in the activated compartment ($n = 24520$) and those who’s proportion in the activated compartment was less than that in the resting compartment ($n = 4$). All sequences that met these criteria were deemed putatively specific (ps)CDR3s.

Examination of the most expanded and public psCDR3s

Public psCDR3s were defined as any amino acid sequence in the psCDR3 pool that met our psCDR3 statistical criteria in more than one individual. Public sequences were defined on an amino acid level (as opposed to nucleotide level) because our goal was to examine them in the scope of shared antigen-recognition. Therefore, the translated CDR3β sequence is the most informative for our purposes.

To better understand if the most expanded sequences in peanut allergic individual’s repertoires were likely peanut specific, we first combined the activated and resting CDR3βs and defined this as the memory T cell repertoire. The top 100 sequences by read count in the memory, total activated and psCDR3 populations were determined
and the overlap between these sequences was displayed with a Venn diagram created with the VennDiagram package in R (version 1.6.2).

**Sequence distance metrics**

To determine global levels of similarity, the minimum Hamming distance (number of amino acid differences among CDR3βs of same length) and minimum Levenshtein distance (minimum number of insertions/deletions/substitutions between CDR3βs) of each psCDR3 against all other psCDR3s was determined using the programming language R (version 3.5.1) with the package stringdist (version 0.9.5.1). The percentage of psCDR3s at each minimum Hamming/Levenshtein distance was calculated. As a control comparison, the minimum Hamming and Levenshtein distances of 100 equal-sized random resamplings of CDR3βs from the total activated and resting pools were determined. The median percent of sequences at each minimum Hamming and Levenshtein distance for the 100 resamplings was calculated. A 2-sided Fisher’s exact test was used to compare the frequency of psCDR3s at each Hamming/Levenshtein distance to the median frequency across 100 random resamplings of total activated or resting CDR3β sequences.

**Shannon Entropy**

To estimate the diversity of our psCDR3s, the Shannon entropy index was calculated for all sequences. Shannon entropy is often used as a statistic of species diversity in ecology but may be applied to evaluate the diversity of immune repertoires as
it quantifies the uncertainty of correctly predicting a sequence that is randomly selected from a population (such as a pool of CDR3β sequences). The Shannon entropy index is calculated with the following formula:

\[ SI = - \sum \left( \frac{n_i}{N} \right) \log_2 \left( \frac{n_i}{N} \right) \]

Where \( i \) represents a given CDR3β sequence, \( n_i \) is the number of reads of that sequence, and \( N \) is the total number of reads in the population (or sample) that sequence came from (Ruggiero et al., 2015). Utilizing the DescTools package (version 0.99.27) of R (version 3.5.1), the Shannon entropy index of psCDR3s from each subject was calculated and then compared to the median entropy index of 100 equal-sized random re-samplings of autologous activated and resting CDR3βs. In order to normalize for baseline differences and evaluate the enrichment from selection, the distribution of \( \Delta \) resting:activated and \( \Delta \) resting:psCDR3 was determined. Differences in those distributions were statistically evaluated with a two-sided Wilcoxon signed rank test.

**Generation of nmers and enrichment determination**

A schematic of the process to generate nmers can be found in figure 2. First, CDR3β amino acid sequences were trimmed to the likely antigen contact region (IMGT positions 106-117) as the stem positions of CDR3βs are not predicted to be involved with antigen binding (Glanville et al., 2017). The trimmed sequences were then broken into all possible continuous nmers of size 3, 4, and 5. In addition, discontinuous motifs were generated of size 4 and 5, allowing for gaps in the sequences so long as there were still 3 conserved residues. The proportion of each nmer in the psCDR3s was the number of
reads containing that nmer divided by the total number of psCDR3 reads. To determine fold-enrichment, this proportion of psCDR3 reads with a given nmer was divided by the proportion of reads in the resting CDR3βs with the nmer. To mimic the approach Glanville et al. used when comparing tetramer-selected sequences to a naïve TCR pool, we isolated all nmers that had a fold-enrichment $\geq 10$. These nmers were further filtered to those that were present in the psCDR3s of at least three subjects and at least three unique psCDR3s. All nmers that met these criteria were deemed enriched motifs.

Network analysis

To better understand antigen-specific diversity, we performed network analysis to

![Network analysis diagram](image)

Figure 2: generation of nmers. A CDR3β is trimmed to the likely antigen-contact region (IMGT positions 106-117). The trimmed sequence is then broken into continuous nmers of lengths 3-5 and discontinuous nmers of length 4 and 5 where there are at least 3 amino acid residues still contributing to the defined sequence.
cluster homologous psCDR3s. Each unique psCDR3 sequence was represented as a node and edges between nodes were made if there was a Levenshtein distance of 1 between them or if they shared an enriched motif. Levenshtein distances were determined using the R package stringdist (version 0.9.5.1). A self-edge was created on a node for every additional nucleotide sequence that corresponded to the subsequent amino acid sequence. Network object (gml) files were created in R using the igraph package (version 1.0.1) and network visualization was performed with Cytoscape 3.7.0, using a force-directed open-CL layout. To evaluate the amount of structure there was among psCDR3s, the median number of edges was determined in 50 equal-sized random resamplings of resting and activated CDR3β sequences, creating edges between the sequences of each resampling if they were within a Levenshtein distance of 1 or had an enriched motif. In addition, the number of clusters ≥ 5 sequences was determined at each resampling. The median number of edges and clusters ≥ 5 sequences across the 50 resamplings was calculated.

**Examination of T_{eff} and T_{reg} CDR3β sequences**

For 9 of the 27 subjects in this study, available bulk T_{eff} and T_{reg} TCRβseq data was used to determine if there are differences in TCR repertoire use in different phenotypic compartments. This data was generated by using the Adaptive Biotechnologies TCRB Immunoseq kit on sorted live CD3+CD4+CD45RA-CD25+CD127+ (T_{eff}) and live CD3+CD4+CD45RA-CD25+CD127- (T_{reg}) cells. The resulting sequences were then interrogated for the presence of psCDR3s to understand differences in antigen-recognition between the two compartments.
To analyze differences in specificity of the T\textsubscript{eff} and T\textsubscript{reg} compartments, an approach similar to the TCR neighborhood enrichment test (TCRNET) described by Ritvo et al. was employed (Ritvo et al., 2018). The neighbors of a CDR3β were defined as any unique amino acid sequences that were connected to it by an edge (within Levenshtein distance of 1 or shared an enriched motif). First, the number of neighbors for each psCDR3 was determined, followed by determining the number of neighbors that existed in all resting CDR3β sequences. The proportion of neighbors in the psCDR3s was compared to that proportion in the resting CDR3β sequences by using a G-test of independence as described in equation 1. The resulting \( P \)-values were then adjusted with an FDR correction and the psCDR3s with a \( q < 0.05 \) were considered core psCDR3s. Core psCDR3s along with their corresponding neighbors were considered core neighborhoods. The proportion of T\textsubscript{eff} and T\textsubscript{reg} CDR3β reads from each of the 9 samples that correspond to each neighborhood was determined. The heatmap used to display this data was created with the \textit{gplots} package (version 3.0.1.1) in R.

**Code for data analysis**

All R code written for this data analysis is publicly available within a github repository at [https://github.com/nealpsmith/thesis](https://github.com/nealpsmith/thesis).
Chapter III

Results

The goal of this study was to apply a statistical approach to previously curated peanut-activated TCRβseq data as well as bulk T_{eff} and T_{reg}-TCRβseq data to better understand the peanut-specific repertoire of allergic individuals. First, psCDR3s were selected based on their relative enrichment in the peanut-activated pool when compared to an autologous resting memory pool. The overall similarity and homology of the psCDR3s were examined using Hamming/Levenshtein distances as well as Shannon entropy. Utilizing motif and network analyses, psCDR3s were then clustered to group homologous sequences and these clusters were then examined for their propensity to represent T_{eff} or T_{reg}-specific sequences. This analysis helped gain insights regarding differences in antigen-recognition across T cell subsets.

psCDR3s are more homologous than the repertoires they are selected from.

A summary of the total unique amino acid- and nucleotide-sequences as well as the total reads recovered from every sample can be found in Appendix 1. When utilizing an in vitro T cell activation system, antigen-specific cells can produce IL-2, which will non-specifically activate cells in a paracrine manner that is referred to as bystander activation (Di Genova et al., 2010). To eliminate sequences that had the highest probability of being present in the activated population due to bystander activation, we
used a statistical enrichment strategy, which defined 7345 out of the 53205 unique activated sequences across our 27 peanut-allergic individuals as psCDR3s. While the vast majority (n = 7309) of psCDR3s were private, 36 unique amino acid sequences (0.49% of psCDR3s) were found in more than one individual (i.e., public). 33 of the public sequences were found in just two individuals, while one psCDR3 (CASSFRFLSGRALNEQFF) was statistically enriched in 8 of the 27 subjects, suggesting a strong selective pressure for entrance of this sequence into the memory T cell repertoire (appendix 2, supplemental figure 1). The CDR3β amino acid- and nucleotide-sequences that met our psCDR3 criteria across the 27 subjects was only a minority among the entire activated pool (amino acid: median 14.05%, nucleotide: median 14.46%), indicating the high level of stringency with this statistical approach (figure 3A). When comparing to the total sequences found in each individual’s entire memory T cell compartment (total unique activated and resting CDR3βs), psCDR3s made up less than 1% (amino acid: median 0.81%, nucleotide: median 0.86%) (figure 3B). Of the 100 sequences with the highest read count in the memory pool, only 6 were found in the psCDR3s, while 84 of the top 100 sequences in the activated pool were found to be among the psCDR3s (Appendix 2, supplemental figure 2). At the time of sample collection, subjects were avoiding the antigen, which is likely the reason the dominant sequences in the memory repertoire did not include many psCDR3s. In addition, this avoidance can also explain the lack of overall abundance of psCDR3s from the memory T cell population.
If our enrichment strategy is narrowing the activated CDR3βs to those most likely to be specific, psCDR3s should exhibit more similarity than unselected activated or resting sequences. Hamming distance compares sequences of the same length, counting the number of positions that have different amino acids, while the more permissive Levenshtein distance determines the minimum number of edits (insertions, deletions or substitutions) required to make two sequences the same. Both minimum Hamming and Levenshtein distances were measured by comparing each psCDR3 to all other psCDR3s and taking the minimum value among from these comparisons. As a control, the minimum Hamming and Levenshtein distances of equal-sized random repeated sampling of resting and activated CDR3βs was also measured (figure 4). There are significantly more sequences that had a minimum Hamming distance of 0-2 and Levenshtein distance

Figure 3: Percent of total CDR3β sequences to meet psCDR3 parameters. (A) Distribution of the percent of the total activated CDR3β sequences to meet psCDR3 parameters across all 27 samples. (B) Distribution of the percent of unique memory sequences (pooled activated and resting CDR3β sequences) that meet psCDR3 criteria.
of 0-1 in the psCDR3s vs. resampling of either resting or activated CDR3βs, indicating the statistically enriched sequences show enhanced similarity. Importantly, these distance metrics show no differences between random activated or resting CDR3βs, supporting our hypothesis that there are non-specific sequences in the activated pool and that our statistical enrichment strategy is effective in removing them.

In addition to enhanced similarity among sequences, enrichment should yield a selection of psCDR3s that are less diverse than the unselected activated or resting compartments. We tested this hypothesis by measuring the Shannon entropy index of different CDR3β populations, which gives a value to the richness and abundance of the
sequences present. Shannon entropy was measured in the psCDR3s from each of the 27 subjects and compared to the median entropy of equal-sized random repeated samplings of autologous resting and activated CDR3βs (figure 5A). In every subject, the entropy of the psCDR3s was lower than what was observed in the random resampling of either the activated or resting sequences. In conjunction with the similarity measurements, the difference in Shannon entropy between the randomly selected activated and resting CDR3β sequences was significantly less than the difference between resting CDR3βs and the psCDR3s, further reinforcing our prediction that some sequences found in the activated compartment are likely to be non-specific (figure 5B). In addition, this finding demonstrates the added utility of using a statistical enrichment to enhance the CD154 up-regulation assay.

**Motif analysis of psCDR3s**

Given the psCDR3s show enhanced similarity over randomly selected CDR3βs, we wanted to elucidate patterns among the sequences that could implicate the same or similar specificity. Using methodology analogous to that described in Glanville et al., motif analysis was performed on the psCDR3s, where sequences were trimmed to IMGT positions 106-117 (the likely antigen-contact region) and broken into continuous and discontinuous nmers (figure 2) (Glanville et al., 2017). To distinguish nmers that are common among the repertoire versus those that are more likely to be antigen-specific, the resting CDR3βs were searched for all nmers that were found in the psCDR3s. The fold-enrichment of each nmer was determined by dividing the proportion of psCDR3 reads to resting CDR3β reads with a given nmer. In addition to this metric, the number of unique
psCDR3s with an nmer as well as the number of subjects who contribute to those psCDR3s were determined. Based on these measures, many motif candidates were clearly enriched in the psCDR3 pool and were present across many unique sequences (Figure 6; appendix 2, supplemental figure 3). To focus the analysis on nmers most likely to be antigen-specific motifs, the nmer pool was filtered to those that were 10-fold enriched in the psCDR3s, found in at least 3 unique psCDR3s and found in the psCDR3s of at least 3 individuals. 148 nmers met these criteria, which constituted 0.14% of all discovered nmers, and these sequences were defined as enriched motifs (table 1). The contribution of each nmer size pool was fairly equal, with the exception of 3mers, where only 3 of the 4376 (0.06%) sequences met our stringent criteria. Overall, 399 psCDR3s (5.4%) were found to contain an enriched nmer which represents a much larger pool of sequences than what was found to be public by entire amino acid sequence

Figure 5: Entropy of psCDR3s. (A) The Shannon entropy of the psCDR3s along with equal-sized random resamplings of activated and resting sequences for each subject was determined (n = 27). (B) The Δ resting:activated and Δ resting:psCDR3 Shannon indexes was determined for all 27 subjects and compared by a Wilcoxon signed rank test (** P < 0.01).
Figure 6: Dominant nmers in the psCDR3s. Log-adjusted enrichment (proportion in psCDR3s / proportion in resting CDR3β) of continuous (A) and discontinuous (B) 4mers vs. number of subjects whose psCDR3s that contribute to the sequence. All labelled points meet enrichment criteria: 10-fold enriched in psCDR3s vs. resting CDR3β, present in psCDR3s of at least 3 individuals and present in at least 3 unique psCDR3s.
(0.49%). Given the criteria to define an enriched nmer includes being public within the psCDR3s, the motif analysis supports our hypothesis that there is more sharing across peanut-specific compartments than what would be revealed by binning CDR3βs into public and private sequences as has been the practice of many recent TCR repertoire studies.

Network analysis of psCDR3s

The CD154-assay used to derive the data presented here was with a polyantigenic stimulation via whole peanut extract and therefore, our psCDR3s likely contain sequences specific for a variety of epitopes and display “local” homology. With this in mind, network analysis was performed on our psCDR3s in an attempt to cluster very similar sequences that are likely specific for the same epitope. psCDR3s were connected by an edge when the Levenshtein distance between them was 1 or if they shared an enriched motif. Self-edges were also created when multiple nucleotide rearrangements contributed to the same amino acid sequence (figure 7). Overall, there Table 1.

<table>
<thead>
<tr>
<th>Nmer</th>
<th>3mer</th>
<th>4mer</th>
<th>5mer</th>
<th>Disc. 4mer</th>
<th>Disc. 5mer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique</td>
<td>4376</td>
<td>15559</td>
<td>23473</td>
<td>8279</td>
<td>52833</td>
<td>104520</td>
</tr>
<tr>
<td>Enriched</td>
<td>3</td>
<td>26</td>
<td>38</td>
<td>9</td>
<td>72</td>
<td>148</td>
</tr>
<tr>
<td>% enriched</td>
<td>0.06</td>
<td>0.16</td>
<td>0.16</td>
<td>0.11</td>
<td>0.14</td>
<td>0.14</td>
</tr>
</tbody>
</table>
were 1804 edges created from our psCDR3 pool, which is significantly more than what is observed when performing the same analysis on equal-sized random-resamplings of resting or activated CDR3βs (figure 8A). In addition, the abundance of clusters that contain 5 or more CDR3βs was significantly higher in the psCDR3 network than what was observed in the random-resamplings of activated and resting sequences, further demonstrating the psCDR3s have more structure that what is expected by chance (figure 8B).

Examination of T_{eff} and T_{reg} psCDR3s

There is evidence to suggest that antigen-specific responses from different T cell compartments are distinct, either through the binding of different epitopes or differences in affinity (Ritvo et al., 2018) (Bacher et al., 2016). We hypothesized there are differences in antigen-recognition across the peanut-specific T_{eff} and T_{reg} repertoires. For 9 of the 27 subjects in this study, TCRβseq data was obtained on bulk T_{eff} and T_{reg} cells and this data was interrogated for the presence of psCDR3s (Appendix 1). Overall, there were more unique CDR3βs and total reads recovered from the T_{reg} data than the T_{eff} data, however the opposite pattern was observed when interrogating the libraries for psCDR3s (figure 9). The median percent of unique T_{eff} CDR3βs that are psCDR3s is 1.95%, while the median of unique T_{reg} CDR3βs that are psCDR3s was only 0.84%. This difference was also reflected when examining the number of psCDR3 reads in both compartments (T_{eff} median: 5.23%, T_{reg} median: 1.00%), suggesting either a bias in the CD154 up-regulation assay to preferentially select effector cells or a physiologically relevant imbalance in peanut-specific T_{eff} vs. T_{reg} cells.
To assess specificity of the compartments, we used an approach similar to the TCR neighborhood enrichment test described by Ritvo et al., where clusters from our network analysis were segregated into neighborhoods of psCDR3s that had an unexpectedly high degree of similarity (Ritvo et al., 2018). In total, there were 995 neighborhoods with high degree of similarity and each neighborhood was interrogated for the presence of psCDR3s that were also found in the T_{eff} and T_{reg} data. While many neighborhoods had psCDR3s found in both compartments, there were 253 T_{eff}- and 74 T_{reg}-exclusive neighborhoods, suggesting there are biases in the peanut-specific repertoire of both compartments (figure 10). However, on a broader scale, many clusters in the network analysis contain psCDR3s found in both compartments, indicating differences in specificity is likely subtle, not involving complete differences in epitope recognition (figure 11; appendix 2, supplemental figure 4).
Figure 7: Network analysis of psCDR3s. Nodes (psCDR3s) were connected by an edge when they were either within a Levenshtein distance of 1 or shared an enriched motif. Self-edges were created to represent the number of additional nucleotide sequences that correspond to the psCDR3 amino acid sequence. Clusters were defined by the existence of a minimum of 2 nodes and one edge. All clusters are displayed in a force-directed open-CL layout.
Figure 8: Evaluation of network structure. (A) The number of total edges created among our psCDR3s as well as the median number of edges created from equal-sized random-resamplings of resting and activated CDR3βs was determined. (B) The number clusters that comprised of > 4 CDR3βs was determined in the psCDR3 network as well as equal-sized random resamplings of resting and activated sequences.

Figure 9: psCDR3s in Teff and Treg samples. 9 Bulk Teff and Treg samples from peanut-allergic individuals were interrogated for psCDR3s. Plots represent the percent of total reads and unique CDR3β sequences that are comprised of psCDR3s.
Figure 10: neighborhood assessment of Treg and Treg. Heatmap of the proportion of reads for each T_{eff} and T_{reg} sample that correspond to neighborhoods in the psCDR3 network analysis. All rows are scaled to have 0 mean and standard deviation of 1. Blue boxes highlight neighborhoods that are exclusive to a particular compartment.
Chapter IV
Discussion

Understanding antigen-specific T cell repertoires can help elucidate disease mechanisms and can lead to more effective diagnostics and therapeutics. In this study, our goal was to gain a better understanding of the peanut-specific repertoire of allergic individuals by applying a statistical approach to data derived from a biologically relevant in-vitro protocol. We successfully created a novel approach to define putatively peanut-specific CDR3βs by selecting sequences whose proportion in an individual’s peanut-activated pool was significantly higher than that of an autologous resting CDR3β pool. Motif analysis was then performed on these psCDR3s, defining dominant continuous and discontinuous patterns that were seemingly over-represented in our putatively-specific population. These potential motifs were used in conjunction with overall homology and evidence of convergence to cluster the psCDR3s and the resulting network analysis was applied to bulk T\textsubscript{eff} and T\textsubscript{reg} sequences of allergic individuals. The results of the T cell subset analysis support the hypothesis that there are differences in antigen recognition between the compartments, which could play a role in allergy pathology.

The goal of the statistical enrichment strategy we developed was to help focus our analysis on CDR3β sequences most likely to be peanut-specific. The stringency of our approach led us to focus on only a fraction (14%) of the activated and total memory
(0.8%) pool. While 0.8% is a seemingly small fraction of the total memory pool, this is likely an overestimation of the actual peanut-specific memory compartment in allergic individuals because of the sequencing protocol used for data generation. Sequence coverage of the CD154+ samples was complete because of the relatively small number of cells present. To the contrary, the sequence coverage of the CD154- samples was incomplete as data was only collected on a fraction of the total resting populations, likely missing a large portion of low-frequency CDR3βs. This difference in coverage across the activated and resting samples causes us to overestimate the activated sequences in the entire memory pool. Therefore, deeper sequencing of the resting compartment is necessary to better estimate the true proportion of peanut-specific TCRs in the memory T cell population of allergic individuals.

The enhanced overall similarity of the psCDR3s as shown by minimum Hamming and Levenshtein distances is an indication that our statistical selection of sequences has biological relevance. At a Hamming distance of 0-2 and a Levenshtein distance of 0-1, the proportion of psCDR3s is significantly higher than that of randomly selected activated or resting CDR3βs. When you also account for the fact that there is no difference in minimum Hamming or Levenshtein distances between randomly selected activated and resting CDR3βs, it becomes apparent that our statistical approach greatly enhances the CD154 up-regulation assay when trying to examine antigen-specific repertoires. Importantly, our minimum Hamming distance result is markedly similar to what was reported by Glanville et al. when they compared tetramer-selected CDR3βs to a randomly selected naïve pool (Glanville et al., 2017). This suggests using a statistically-
enhanced CD154 up-regulation assay is a feasible alternative to tetramer selection when either the specific antigens or HLA genotype is unknown.

In addition to overall similarity metrics, the entropy of our psCDR3s is significantly lower than that of randomly sampled activated or resting CDR3βs, further demonstrating the value of our statistical approach to enhance the CD154 up-regulation assay. While our data cannot directly implicate bystander activation as a major contaminant of the activated T cell population, it does suggest there are non-specific sequences in the CD154+ pool and that examination of all sequences might dilute information about the most prominent CDR3βs. Glanville et al. performed TCRβseq and motif analysis on the entire CD154+ population that was sorted from in vitro cultures with *Mycobacterium tuberculosis*, likely including non-specific sequences (Glanville et al., 2017). A re-evaluation of the activated CDR3β sequences reported that focuses on only those that are significantly enriched in the activated pool could enhance their analysis and yield motifs not initially recognized by the authors while removing any that are simply common among the entire repertoire.

When using the CD154 up-regulation assay with a polyantigenic stimulus such as whole peanut extract, it is difficult to elucidate the exact specificity of any activated CDR3βs. To mitigate this limitation, future studies could use the CD154 up-regulation assay with better defined peanut antigens, including the proteins Ara h 1, Ara h 2, Ara h 3, Ara h 6 and Ara h 8 (Wambre et al., 2017). While this approach would enhance our ability to pinpoint specificity, it would likely reduce our sensitivity for capturing peanut-specific T cells given we are yet to elucidate all peanut T cell antigens. To assess the specificity of the psCDR3s from whole-peanut stimulation, our data could be used in
conjunction with a single-cell approach, where you have the ability to reveal TCRα chains that are paired with a putatively specific TCRβ. With paired sequences, complete TCRs can be recombinantly expressed into immortalized cell lines that can be used to assess binding ability to different antigens (Guo et al., 2016). Following the present study with research that defines the specificity of the psCDR3s could determine the most dominant T cell antigens in the repertoire of peanut allergic individuals.

One goal of performing motif analysis on the psCDR3s was to find evidence of more overlap in antigen recognition across repertoires than what would be revealed by binning sequences into “public” and “private” pools as defined by the entire CDR3β amino acid sequence. One of the criteria to define an enriched motif was based on publicity, filtering out nmers that were in the psCDR3s of less than 3 individuals. Even with this filtering, we found 399 psCDR3s that contained an enriched motif as compared to just 36 psCDR3s that were public based on the entire amino acid sequence. However, we lack the ability to implicate the motifs as sequences that confer specificity, limiting our ability to draw conclusions about shared antigen recognition. To validate the hypothesis that shared antigen-recognition can be elucidated via motif analysis, future work should focus on defining the specificity of the enriched motifs. Cloning of multiple unique motif-containing CDR3β sequences could allow for parallel antigen-binding assays that would have the ability to implicate shared specificity (Guo et al., 2016).

An application of the enriched motifs that could be independent of known specificity would be as biomarkers for clinical phenotype, as was the goal by Sims et al. when examining the repertoire of Glioma patients (Sims et al., 2016). One example could be to examine if the abundance of the sequences in the memory CD4+ T cell
compartment successfully classify allergic from healthy individuals. A more focused application would be to segregate allergic individuals by their likelihood to have a favorable response to therapies. Currently, a therapeutic approach that is being evaluated for peanut allergy is immunotherapy, where individuals receive incrementally higher doses of peanut over the course of many months with the goal of becoming desensitized to the antigen (Burks, Sampson, Plaut, Lack, & Akdis, 2018). Some patients receive full tolerance to peanut from this therapy, maintaining their unresponsiveness even after long periods of avoidance. Conversely, other patients either fail the treatment by never becoming desensitized while others have transient desensitization, where tolerance is lost after a period of avoiding the antigen post-therapy. Determining if the abundance of the enriched motifs can classify patients pre-immunotherapy by their likelihood to have a favorable outcome could help clinicians make more informed decisions about how to treat allergic individuals.

Although the psCDR3s show enhanced overall similarity, we hypothesized that homology would be apparent in a “local” way, where sequences would cluster based on their specificity for one of the many epitopes in our polyantigenic stimulant. The network analysis tested this hypothesis by clustering psCDR3s based on homology, enriched motifs and evidence of convergence. While we expected to see some large clusters that likely correspond to dominant epitopes, we did not expect to find the high levels of convergent recombination that is apparent in our data. Convergent recombination is the process by which the recombination events from independent T cells produce different nucleotide sequences that correspond to the same amino acid sequence (Ruggiero et al., 2015) (Venturi et al., 2006). This phenomenon can significantly impact
the prevalence of certain CDR3β amino acid sequences and its presence among our psCDR3s suggests positive antigen-driven selective pressures that are driving these sequences to be in the memory T cell pool. If in fact these selective pressures exist, they could play a major role in the development of peanut allergy, driving specific T cells towards a pathogenic phenotype.

Sequence homology defined in this study was based on either matching amino acid sub-sequences (enriched motifs) across distinct CDR3βs or looking for minimal substitutions/insertions/deletions when aligning unique sequences. An alternative approach would be to define homology based on the likelihood substitutions across different amino acid sequences still confer the same specificity. Dash et al. published a method to group epitope-specific sequences that uses a modified BLOSUM matrix score to define homology among unique TCRs (Dash et al., 2017). A BLOSUM matrix is a substitution matrix that scores amino acid substitutions across two aligned amino acid sequences based on how conserved that specific change is evolutionarily (more conserved equates to higher score). Applying this approach to the psCDR3s could reveal homology that is based on biological reason that could be lost in the presented methods and thus should be a focus of future studies.

When examining the bulk T cell subset CDR3β data from 9 of the 27 subjects used in this study, there is a very apparent discrepancy between the psCDR3 sequences recovered from the Teff (median 5.23% of total reads) and Treg (median 1.00% of total reads) sequences. There have been recent reports that demonstrate allergen-activated CD154+ cells are predominantly a Teff phenotype, while activated Treg cells can be distinguished by expression of CD137 (Bacher et al., 2016) (Weissler et al., 2018).
However, Chiang et al. successfully found T_{reg} cells in the CD154+ population from peanut-stimulated cultures, suggesting that while our CD154-based approach likely favors the selection of T_{eff}, T_{reg} are not completely absent from the psCDR3s (Chiang et al., 2018).

Incorporating the T_{eff} and T_{reg} CDR3β data into our network analysis demonstrates the utility in defining the “local” homology of the psCDR3s as it allows us to test hypotheses that incorporate both phenotype and specificity. Examination of the overall clusters’ tendencies to skew to either T_{reg} or T_{eff} did not reveal any obvious patterns, suggesting differences in antigen-recognition across the compartments is subtle. Indeed, our TCRNET-like approach also supported the hypothesis that compartment-specific repertoires have subtle differences in recognitions, as there are 253 T_{eff}- and 74 T_{reg}-exclusive neighborhoods within the network of psCDR3s. However, given the likely propensity of the psCDR3s to be derived from a T_{eff} cell, it is possible more evident differences between the two compartments could be revealed if T_{reg} cells were better selected from the stimulated cultures using a marker such as CD137. In addition, we only had available T_{eff} and T_{reg} CDR3β data from 9 of the 27 individuals that contribute to the total psCDR3 pool. Completing this data set to include T_{eff} and T_{reg} CDR3β data from all subjects in the study could help solve any sampling biases that could exist in our current dataset.

Using the CD154 up-regulation assay in conjunction with TCRβseq is a feasible alternative to MHC-bound epitope selection when the exact antigens and/or HLA information is unknown. However, given the polyantigenic nature of the stimulation, efforts should be made to study the specificity of psCDR3s as well as determine the
ability of motifs to confer specificity. In addition, a more nuanced approach to isolate activated $T_{reg}$ cells from the stimulated cultures would allow for more confident conclusions regarding differences in antigen-specificity across phenotypic T cell compartments. However, the general approach described in this study can be used to broaden the discovery of antigen-specific TCR features as well as help define true levels of homology and oligoclonality within antigen-specific compartments.
Appedix 1

Supplemental table 1.

*Summary of Unique CDR3β sequences for each subject.*

<table>
<thead>
<tr>
<th>ID</th>
<th>Resting CDR3 sequences</th>
<th>Activated CDR3 sequences</th>
<th>psCDR3 sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nucleotide</td>
<td>Amino acid</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>100</td>
<td>36370</td>
<td>34788</td>
<td>1601</td>
</tr>
<tr>
<td>101</td>
<td>28813</td>
<td>27389</td>
<td>442</td>
</tr>
<tr>
<td>104</td>
<td>35539</td>
<td>34347</td>
<td>5015</td>
</tr>
<tr>
<td>106</td>
<td>27559</td>
<td>26759</td>
<td>2898</td>
</tr>
<tr>
<td>107</td>
<td>39469</td>
<td>37867</td>
<td>3425</td>
</tr>
<tr>
<td>27</td>
<td>32368</td>
<td>30993</td>
<td>2701</td>
</tr>
<tr>
<td>69</td>
<td>30850</td>
<td>29440</td>
<td>1845</td>
</tr>
<tr>
<td>80</td>
<td>28181</td>
<td>27060</td>
<td>2421</td>
</tr>
<tr>
<td>81</td>
<td>15455</td>
<td>14990</td>
<td>894</td>
</tr>
<tr>
<td>84</td>
<td>29375</td>
<td>28419</td>
<td>6548</td>
</tr>
<tr>
<td>85</td>
<td>28765</td>
<td>27729</td>
<td>1712</td>
</tr>
<tr>
<td>89</td>
<td>23926</td>
<td>23180</td>
<td>1009</td>
</tr>
<tr>
<td>93</td>
<td>25345</td>
<td>24641</td>
<td>5867</td>
</tr>
<tr>
<td>94</td>
<td>35993</td>
<td>34424</td>
<td>409</td>
</tr>
<tr>
<td>1</td>
<td>33969</td>
<td>31944</td>
<td>1253</td>
</tr>
<tr>
<td>105</td>
<td>46910</td>
<td>44712</td>
<td>344</td>
</tr>
<tr>
<td>111</td>
<td>46156</td>
<td>43877</td>
<td>172</td>
</tr>
<tr>
<td>22</td>
<td>26237</td>
<td>25123</td>
<td>2806</td>
</tr>
<tr>
<td>24</td>
<td>44896</td>
<td>41881</td>
<td>311</td>
</tr>
<tr>
<td>29</td>
<td>35985</td>
<td>35252</td>
<td>811</td>
</tr>
<tr>
<td>33</td>
<td>32546</td>
<td>31144</td>
<td>3671</td>
</tr>
<tr>
<td>95</td>
<td>32158</td>
<td>30374</td>
<td>118</td>
</tr>
<tr>
<td>97</td>
<td>28573</td>
<td>27762</td>
<td>1952</td>
</tr>
<tr>
<td>108</td>
<td>36144</td>
<td>34326</td>
<td>3175</td>
</tr>
<tr>
<td>19</td>
<td>30347</td>
<td>28852</td>
<td>45</td>
</tr>
<tr>
<td>90</td>
<td>60097</td>
<td>53606</td>
<td>2833</td>
</tr>
<tr>
<td>96</td>
<td>63352</td>
<td>61566</td>
<td>1431</td>
</tr>
</tbody>
</table>
Supplemental table 2.

*Total Read counts of CDR3β sequences for each subject.*

<table>
<thead>
<tr>
<th>ID</th>
<th>Resting CDR3β reads</th>
<th>Activated CDR3β reads</th>
<th>psCDR3 reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>53486</td>
<td>2058</td>
<td>718</td>
</tr>
<tr>
<td>101</td>
<td>35749</td>
<td>496</td>
<td>102</td>
</tr>
<tr>
<td>104</td>
<td>46916</td>
<td>8112</td>
<td>3461</td>
</tr>
<tr>
<td>106</td>
<td>37943</td>
<td>3618</td>
<td>1047</td>
</tr>
<tr>
<td>107</td>
<td>55754</td>
<td>4332</td>
<td>1491</td>
</tr>
<tr>
<td>27</td>
<td>52738</td>
<td>3570</td>
<td>1175</td>
</tr>
<tr>
<td>69</td>
<td>45457</td>
<td>3155</td>
<td>1547</td>
</tr>
<tr>
<td>80</td>
<td>38632</td>
<td>3092</td>
<td>1031</td>
</tr>
<tr>
<td>81</td>
<td>20116</td>
<td>1101</td>
<td>340</td>
</tr>
<tr>
<td>84</td>
<td>44118</td>
<td>8986</td>
<td>2867</td>
</tr>
<tr>
<td>85</td>
<td>41305</td>
<td>2308</td>
<td>863</td>
</tr>
<tr>
<td>89</td>
<td>30456</td>
<td>1228</td>
<td>368</td>
</tr>
<tr>
<td>93</td>
<td>35469</td>
<td>7896</td>
<td>1109</td>
</tr>
<tr>
<td>94</td>
<td>47815</td>
<td>471</td>
<td>116</td>
</tr>
<tr>
<td>1</td>
<td>52441</td>
<td>1610</td>
<td>495</td>
</tr>
<tr>
<td>105</td>
<td>71704</td>
<td>375</td>
<td>59</td>
</tr>
<tr>
<td>111</td>
<td>64387</td>
<td>198</td>
<td>46</td>
</tr>
<tr>
<td>22</td>
<td>33235</td>
<td>3259</td>
<td>678</td>
</tr>
<tr>
<td>24</td>
<td>74226</td>
<td>387</td>
<td>120</td>
</tr>
<tr>
<td>29</td>
<td>58504</td>
<td>1025</td>
<td>317</td>
</tr>
<tr>
<td>33</td>
<td>50324</td>
<td>4467</td>
<td>986</td>
</tr>
<tr>
<td>95</td>
<td>42454</td>
<td>133</td>
<td>24</td>
</tr>
<tr>
<td>97</td>
<td>53865</td>
<td>2289</td>
<td>506</td>
</tr>
<tr>
<td>108</td>
<td>49635</td>
<td>3917</td>
<td>860</td>
</tr>
<tr>
<td>19</td>
<td>45943</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>90</td>
<td>85114</td>
<td>3461</td>
<td>878</td>
</tr>
<tr>
<td>96</td>
<td>113320</td>
<td>1769</td>
<td>451</td>
</tr>
</tbody>
</table>
Supplemental table 3.

*Total CDR3β and psCDR3 T_{eff} and T_{reg} data.*

<table>
<thead>
<tr>
<th>ID</th>
<th>T_{eff}</th>
<th></th>
<th>T_{reg}</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>psCDR3</td>
<td>Total</td>
<td>psCDR3</td>
</tr>
<tr>
<td></td>
<td>Unique CDR3βs Reads</td>
<td>Unique CDR3βs Reads</td>
<td>Unique CDR3βs Reads</td>
<td>Unique CDR3βs Reads</td>
</tr>
<tr>
<td>100</td>
<td>10275</td>
<td>15575</td>
<td>239</td>
<td>827</td>
</tr>
<tr>
<td>106</td>
<td>14195</td>
<td>25452</td>
<td>313</td>
<td>1742</td>
</tr>
<tr>
<td>107</td>
<td>17626</td>
<td>27295</td>
<td>494</td>
<td>1428</td>
</tr>
<tr>
<td>69</td>
<td>7908</td>
<td>12415</td>
<td>143</td>
<td>319</td>
</tr>
<tr>
<td>84</td>
<td>8535</td>
<td>15754</td>
<td>442</td>
<td>1353</td>
</tr>
<tr>
<td>85</td>
<td>8017</td>
<td>13343</td>
<td>200</td>
<td>773</td>
</tr>
<tr>
<td>101</td>
<td>6311</td>
<td>9858</td>
<td>48</td>
<td>121</td>
</tr>
<tr>
<td>27</td>
<td>14293</td>
<td>30987</td>
<td>253</td>
<td>974</td>
</tr>
<tr>
<td>81</td>
<td>7119</td>
<td>11316</td>
<td>108</td>
<td>301</td>
</tr>
</tbody>
</table>
Appendix 2

Supplemental Figure 1: distribution of public psCDR3s vs. number of subjects. Public psCDR3s were defined as psCDR3 amino acid sequences that met all psCDR3 statistical criteria in multiple subjects. Displayed is the number of public psCDR3s vs. the number of subjects they are present in.

Supplemental Figure 2: Overlap across top 100 sequences in each CDR3β pool. The top 100 CDR3βs by read count were taken from the total memory, activated and psCDR3 sequences and the overlap between the sequences was determined.
Supplemental Figure 3: Dominant nmers in the psCDR3s. Log-adjusted enrichment (proportion in psCDR3s / proportion in resting CDR3β) of continuous 3mers and 5mers vs. number of subjects whose psCDR3s that contribute to the sequence. All labelled points meet enrichment criteria: 10-fold enriched in psCDR3s vs. resting CDR3β, present in psCDR3s of at least 3 individuals and present in at least 3 unique psCDR3s
Supplemental Figure 4: Examination of subsets within network analysis. Mean T_{eff} and T_{reg} proportion of total reads across 9 subjects was determined. The ratio of mean T_{eff}: mean T_{reg} was calculated and scaled to be used for node color within the network. Clusters were defined by the existence of a minimum of 2 nodes and one edge. All clusters are displayed in a force-directed open-CL layout.


56


58