Recovery of T cell receptor variable sequences from 3' barcoded single-cell RNA sequencing libraries

by

Ang A. Tu

B.S. Biomedical Engineering
Johns Hopkins University, 2013

SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL ENGINEERING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2020

Signature of Author: ____

Ang A. Tu
Department of Biological Engineering
April 30th, 2020

Certified by: ____

J. Christopher Love, Ph.D.
Raymond A. (1921) and Helen E. St. Laurent Professor of Chemical Engineering, MIT
Thesis Supervisor

Accepted by: ____

Katharina Ribbeck, Ph.D.
Associate Professor of Biological Engineering, MIT
Chair, Graduate Program Committee of Biological Engineering

The following committee has evaluated this doctoral thesis:

Professor J. Christopher Love, Ph.D.
Thesis Supervisor
Professor of Chemical Engineering, MIT

Professor Paul Blainey, Ph.D.
Thesis Co-Supervisor
Associate Professor of Biological Engineering, MIT

Professor Michael Birnbaum, Ph.D.
Chair, Thesis Committee
Assistant Professor of Biological Engineering, MIT

Professor Kai W. Wucherpfennig, M.D. Ph.D.
Member, Thesis Committee
Chair, Department of Cancer Immunology and Virology, DFCI
Professor of Microbiology and Immunobiology, HMS
Professor of Neurology, BWH & HMS
Director, Center for Cancer Immunotherapy Research, DFCI

Recovery of T cell receptor variable sequences from 3’ barcoded single-cell RNA sequencing libraries

by

Ang A. Tu

Submitted to the Department of Biological Engineering on April 30th, 2020
In partial fulfillment of the requirement for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Heterogeneity of the immune system has increasingly necessitated the use of high-resolution techniques, including flow cytometry, RNA-seq, and mass spectrometry, to decipher the immune underpinnings of various diseases such as cancer and autoimmune disorders. In recent years, high-throughput single-cell RNA sequencing (scRNA-seq) has gained popularity among immunologists due to its ability to effectively characterize thousands of individual immune cells from tissues. Current techniques, however, are limited in their ability to elucidate essential immune cell features, including variable sequences of T cell antigen receptors (TCRs) that confer antigen specificity. Incorporation of TCR sequencing into scRNA-seq data could identify cells with shared antigen-recognition, further elucidating dynamics of antigen-specific immune responses in T cells.

In the first part of this thesis work, we develop a strategy that enables simultaneous analysis of TCR sequences and corresponding full transcriptomes from 3′ barcoded scRNA-seq samples. This approach is compatible with common 3′ scRNA-seq methods, and adaptable to processed samples *post hoc*. We applied the technique to identify transcriptional signatures associated with clonal T cells from murine and human samples. In both cases, we observed preferential phenotypes among subsets of expanded T cell clones, including cytotoxic T cell states associated with immunization against viral peptides.

In the second part of the thesis, we apply the strategy to a 12-patient study of peanut food allergy to characterize T helper cell responses to oral immunotherapy (OIT). We identified clonal T cells associated with distinct subsets of T helper cells, including Teff, Treg, and Tfh, as well as Th1, Th2, and Th17 signatures. We found that though the TCR repertoires of the patients were remarkably stable, regardless of their clinical outcomes, Th1 and Th2 clonotypes were phenotypically suppressed while Tfh clonotypes were not affected by therapy. Furthermore, we observed that highly activated clones were less likely to be suppressed by OIT than less activated clones. Our work represents one of the most detailed transcriptomic profiles of T helper cells in food allergy.

In the last part of the thesis, we leverage the simplicity and adaptability of the method to recover TCR sequences from previously processed scRNA-seq samples derived from HIV patients and a nonhuman primate model of TB. In the HIV study, we recovered expanded clonotypes associated with activated T cells from longitudinal samples from patients with acute HIV infections. In the TB study, we modified the primers used in the method to T cells from TB granulomas of cynomolgus macaques. We identified not only expanded clonotypes associated with cytotoxic functions, but also clonotypes shared by clusters of activated T cells. In total, these results demonstrate the utility of our method when studying diseases in which clonotype-driven responses are critical to understanding the underlying biology.

Thesis Supervisor: J. Christopher Love **Title**: Professor of Chemical Engineering

Acknowledgements

Much ink has already been spilled about the challenges of graduate school. Needless to say, my studies, and the completion thereof, would not have been possible without the support of numerous individuals whose kindness, intelligence, and thoughtfulness never cease to amaze me. I count myself fortunate to have known them during my time as a graduate student.

First and foremost, I am indebted to my advisor, Chris Love, for not just giving me the chance to join his group, but also for the tremendous patience he has shown for my development as a scientist over the years. Through all the setbacks, mistakes, and hardships, Chris never gave up on my projects, inspiring me to do the same. Even during periods of little progress, Chris never ceased to encourage me to continue my work, and always treated my results with enthusiasm. My co-advisor, Paul Blainey, and my thesis committee members, Michael Birnbaum and Kai Wucherpfennig, were also an instrumental source of guidance. For that, I am thankful.

I am also blessed to have had the pleasure to work with Todd Gierahn, whose combination of scientific acuity and perseverance gave me the confidence to tackle obstacles that often seemed too great to overcome. None of the work in this thesis would have been possible otherwise.

I would like to thank the members of the Love lab, for becoming my second family and supporting me through good times and bad. This included listening to me as I vented about my failed experiments and disappointing results, as well as celebrating with me my small victories and successes at the bench. I reminisce fondly about our lunchtime card games and outings at the Muddy Charles. Our camaraderie, in and out of the lab, made the trials and tribulations of science that much more bearable.

In particular, I would like to recognize Brinda Monian, for being a reliable teammate and a source of intellectual support; Duncan Morgan, for his fresh scientific perspectives; and Gary Shea and Patrick Petrossian, for their dedication to their work at the bench. I would also like to recognize Noelani Kamelamela and Stuart Levine, as well as all of the student technicians, from the BioMicro Center for their indispensable contribution of technical know-how and services.

Much of my research would not have been possible without our collaborators. This included the Shreffler lab (Wayne Shreffler, Bert Ruiter, and Sarita Patil) and the Shalek lab (Alex Shalek, Marc Wadsworth, Travis Hughes, and Sam Kazer). I am grateful for their insightful guidance and continuous support.

I wish to show my gratitude to the mentors that fostered my scientific curiosities and set me off on my journey as a scientist: Brad Spellberg, who entrusted me, at the age of 17, with the laboratory resources to carry out my first experiment at the bench; Jonathan Schneck, who treated my fledgling scientific inquiries with all the seriousness and respect that I did not deserve; and Karlo Perica, who inspired my interest in immunology through thoughtful discussions and painstaking instructions in the lab.

Finally, I would like to thank my family for supporting me all these years during my graduate studies: my siblings, John and Rose, who always made our time together joyful and memorable; my uncle Albert, who always took an interest in my studies growing up; and finally my parents, who made great sacrifices to immigrate to the States almost twenty years ago, allowing me to be where I am today, and instilled in me a sense of responsibility that has always guided me through the different stages of my life.

Table of Contents

List of figures and tables 9

1. Introduction 11

1.1 Motivation for characterizing the TCR repertoire 13

1.2 Conventional methods of bulk and single-cell TCR sequencing 15

1.3 Challenges and Opportunities 18

1.4 Thesis objectives 20

2. Size selection enrichment of TCR sequences from single-cell sequencing libraries 23

2.1 Motivation 24

2.1.1 Challenges in recovery TCR CDR3 from massively parallel single-cell libraries 24

2.2 Methods 26

2.3 Results 29

2.3.1 Enrichment of TCR transcripts through biotin-streptavidin pull-down 29

2.3.2 Enrichment of TCR variable region through size selection 31

2.3.3 Sequencing of size-selected libraries 33

2.3.4 Modification to the size-selection process 36

2.4 Discussion 38

3. Recovery of paired TCR sequences from single-cell Seq-Well libraries reveals clonotypic T cell signatures 40

3.1 Motivation 41

3.1.1 Learning from difficulties in previous TCR enrichment approaches 41

3.2 Methods 42

3.3 Results 46

3.3.1 V gene selection using multiplex primer sets 46

3.3.2 Sequencing using custom sequencing primers 49

3.3.3 Assessing accuracy by OT1 spike-in 51

3.3.4 TCR recovery reveals clonal expansion in immunized mice 53

3.3.5 Stimulated cells show clonotype-associated transcriptional profiles 57

3.3.6 Public clonotypes exhibit similar CDR3 sequences 59

3.3.7 Clonally expanded T cells detected in peanut-allergic patients 60

3.3.8 Expanded clonotypes exhibit varied expression of Th2 genes 63

3.4 Discussion 66

4. Application of Seq-Well and TCR recovery to the study of peanut oral immunotherapy 68
4.1 Motivation 69
4.1.1 Immunotherapy is effective for some, but not all, peanut-allergic patients 69
4.2 Methods 70
4.3 Results 75
4.3.1 Single-cell transcriptomic landscape of the patients undergoing peanut OIT 75
4.3.2 TCR recovery from peanut OIT samples 78
4.3.3 T helper cells comprise six clonally distinct subtypes 81
4.3.4 Th1 and Th2, but not Tfh, expression is suppressed by OIT 83
4.3.5 Clonotypes associated with increased expression are more resistant to OIT 86
4.4 Discussion 87
5. Application of Seq-Well and TCR recovery to the studies of other diseases 90
5.1 Motivation 91
5.1.1 Retroactive study of TCR from processed clinical samples 91
5.2 Methods 92
5.3 Results 93
5.3.1 Recovery of TCR sequences from HIV samples 93
5.3.2 Recovery of cynomolgus monkey TCR sequences 97
5.4 Discussion 101
6. Conclusion and outlook 102
7. Acknowledgement of contributions 106
8. References 107
A. Appendix 113

List of figures and tables

Figure 1-1. Protein structure of TCRα and TCRβ 14

Figure 1-2. Example schematic of TCR amplification using 5’ RACE 17

Figure 1-3. Schematic of conventional CDR3 amplification applied to 3′ barcoded libraries 20

Figure 2-1. Single-cell RNA-seq processing by Seq-Well...25

Figure 2-2.qPCR enrichment of TCRα and β chain by biotin-streptavidin pulldown..................31

Figure 2-3. Size selection of larger fragments allows paired-sequencing of cell barcode and CDR3...32

Figure 2-4. Size-selection using SPRI selects for fragments larger than 1kbp..........................33

Figure 2-5. Size-dependent mapping of TCR regions 34

Figure 2-6. Mapping of TCR sequences onto transcriptomic data in a murine sample 35

Figure 2-7. Examples of low mapping of TCR variable regions due to incomplete size selection 36

Figure 2-8. Improvement in V and J gene mappings after Pippin Prep size selection................37

Figure 2-9. An example of sequencing reads mapping to introns of J genes............................38

Figure 3-1. Strategy for TCR recovery from 3′ barcoded single-cell sequencing library............48

Figure 3-2. Incorporation of V gene primers improves mapping efficiency of TCR-enriched library...49

Figure 3-3. Sequencing quality of data produced with constant region sequencing primers......51

Figure 3-4. Recovery of OT-I *Tcra* and *Tcrb CDR3s* 52

Figure 3-5. scRNA-Seq and TCR analysis of HPV-E7 immunized mice 55

Figure 3-6. Stimulated and *ex vivo* cells are transcriptionally distinct 56

Figure 3-7. Group 1 and 2 clonotypes differ in expansion and gene expression upon stimulation 59

Figure 3-8. Analysis of shared clonotypes across four E7-HPV immunized mice 60

Figure 3-9. ScRNA-Seq and TCR analysis of peanut-dependent activated T cells from one of the peanut-allergic individuals (patient 77) combined with pseudotemporal analysis.................62

Figure 3-10. Distinct patterns of gene expression correlate with pseudotime 64

Figure 3-11. Psuedotime correlates with effector T cell signatures..65

Figure 4-1. Peanut-reactive T cells from individuals undergoing oral immunotherapy have diverse transcriptional signatures. 77

Figure 4-2. Top differentiating genes of the three sorted subsets of memory T cells 78

Figure 4-3. Functional gene modules are associated with clonal expansion and enriched expression in activated cells......80

Figure 4-4. Subsets of T helper are clonotypically distinct......82

Figure 4-5. Th1 and Th2, but not Tfh, subsets are suppressed by OIT......84

Figure 4-6. Highly Th2 clonotypes are more likely to remain highly activated at both BL and MN timepoints......87

Figure 4-7. Clonotypes that are Th2+ at BL and MN are not more clonally expanded or more likely to be of a particular Th subset......87

Figure 5-1. Recovery rate of TCRα and TCRβ sequences from human HIV samples......95

Figure 5-2. Clonal expansion of T cells from each of the HIV patient samples. 96

Figure 5-3. Clonal expansion by T cell cluster......97

Figure 5-4. T cells of granulomas form distinct clusters......99

Figure 5-5. Clonal expansion detected in T cells from granulomas......100

Figure 5-6. Number of TCRβ sequences shared between clusters of T cells shown in Figure 5-4 100

Figure A3-1. 113

Table A3-1. 115

Figure A3-2. 115

Figure A3-3. 116

Table A3-2. 117

Figure A3-4. 117

Figure A3-5. 118

Figure A3-6. 119

Figure A3-7. 120

Figure A4-1. 121

Figure A4-2. 122

Figure A4-3. 122

Figure A4-4. 123

Figure A4-5. 124

Figure A5-1. 125

Figure A5-2. 126

1. Introduction

Intuitions regarding the adaptive immune response, namely immunological memory, likely existed before the recognition of immunology as a field of study. Even before the introduction of the first successful vaccine (or at least, one of the firsts) by Edward Jenner, many had likely observed that survivors of infectious diseases were granted virtual immunity from those diseases[1]. Almost a hundred years later, experiments pinpointed humoral immunity by transfer of serums from immunized to unimmunized animals[2]. These experiments, which postulated existence of "antitoxin" that could confer protection, constituted an early proof-of-concept of antibody therapy, an industry now worth over $250 billion globally[3].

Study of the adaptive immune system accelerated in the 20th century. Organs associated with B and T cells were identified first in birds (e.g. bursa fabricus), then in other animals[4]. Despite early advances made by Burnet and others regarding clonal selection and maturation of antibody producing B cells, the exact mechanism by which these cells originate without available precedent was only partially known[5]. It was not until the works of Hozumi and Tonegawa, using the newly available techniques of restriction enzyme digest and gel electrophoresis, did evidence of distant gene fragments joining together become evident[6]. This process was later termed V(D)J Recombination. In this process, segments of variable genes (V, or variable; D, or diversity; and J, or junction) randomly join together in an imprecise manner to generate random and diverse antigen receptors in both B and T cells[7].

Analogous processes were identified in a wide variety of animals, including most vertebrates. It is clear that the ability for the body to maintain memory of past exposures to pathogens confers tremendous evolutionary advantage. Despite the energy costs of producing a large number of nonfunctional clones and cells that ultimately have to be deleted via apoptosis, the adaptive immune system, in one form or another, remains preserved across species[5]. The generated diversity is staggering. It is estimated that V(D)J recombination could generate up to

10^{18} different antigen receptors[8]. The realistic repertoire is likely much smaller in practice, as the number of possible combinations outnumbers the number of available cells in the body. Were the repertoire to be “hardcoded” in the human genome, it would take roughly the same number of nucleotides as the entire genome.

Because the antigen receptor repertoire is not hardcoded in the genome, it also means that the repertoire of any individual could change and evolve in response to exposures to pathogens. Early attempts at characterizing the T and B cell repertoires depended on basic molecular biology techniques, such as spectrographing, wherein the sizes (that is, the lengths of the nucleotide sequences) of the re-arranged T or B cell receptors (TCRs and BCRs, respectively) are visualized via gel electrophoresis, resulting in a gaussian distribution of bands in a normal repertoire, and a skew distribution in those that have gone through clonal selection and expansion[9]. Advances in qPCR further allowed researchers to quickly identified the presence of specific V or J gene segments in their samples, though still without exact nucleotide sequences[10]. Later development in Sanger sequencing allowed for investigation of the exact nucleotide sequence of the junctional region in a selected T or B cell, one at a time.

Breakthroughs in massively parallel DNA sequencing, or Next Generation Sequencing (NGS), in the 2000s presented a significant technological leap for immunologists studying the immune repertoire. No longer were researchers restricted to investigating a single, isolated clone at a time, but rather it was now possible to assess the repertoire as a whole in a single experiment with high accuracy. For the remainder of the chapter, we review the various methods that have been used to query the immune repertoire at various resolutions, including at the bulk (i.e. population) level and at the single-cell level. We also discuss the motivation for the subsequent chapters, which describe our attempts at combining TCR sequencing with existing scRNA-seq platforms and the application of our technique to several immunological studies, including the those of food allergy and infectious diseases.

This thesis represents our endeavors to improve upon available technologies in a rational way that can be easily adapted by others without significant investment of labor and financial resources. This work focuses entirely on the TCR, though we anticipate many of the methods detailed here can also be applied to BCR. Along with the development of our technique, we were also faced with the challenge of integrating and analyzing TCR repertoires with their corresponding single-cell transcriptomic data. In the latter half of the thesis, we describe our attempt to make sense of such data to better understand clonal T cell responses in our samples.

1.1 Motivation for characterizing the TCR repertoire

Due to the nature of the adaptive immune system, the ability for a single T cell to recognize a specific antigen is of great interest to researchers. Antigen-specific T cells play key roles in a number of diseases including autoimmune disorders and cancer[10–12]. T cells recognize antigens presented in major histocompatibility complexes (MHCs) through their TCRs. In response to cognate recognition of antigen, T cells can respond through activation and clonal expansion, resulting in a progeny of daughter cells of identical TCR sequences. Therefore, characterizing TCR sequences and their relative proportions, or the TCR repertoire, in populations of relevant T cells can illuminate the dynamics and diversity of antigenic responses against pathogens, and how these characteristics associate with different disease states. TCR repertoire measurements, such as clonal expansion and selection, have been shown to be predictive of vaccine and therapy efficiencies, in a wide range of conditions and diseases[12–16]. The dynamics of clonal expansion and the diversity of the TCR repertoire have been correlated to autoimmune and other hypersensitivity conditions such as food allergy[17]. Accurate characterization of the TCR repertoire has also proven to be pivotal in designing new immunotherapeutic modalities, where expansion of antigen-specific T cells is crucial to treatment efficacy[18].

The TCR comprises two subunits: an alpha chain (TCRα) and a beta chain (TCRβ). Each chain contains three complementarity-determining regions (CDRs) that are directly involved with recognition of antigens presented on MHCs. Of the CDRs, the CDR3 contains the highest amount of sequence variability, and is directly in contact with the presented antigen (whereas the CDR1 and the CDR2 mainly contact the MHC protein itself). Together, the alpha and beta chains determine the antigen-specificity of the T cell (**Figure** 1-1). As such, TCR sequencing at the single-cell level is of particular interest to researchers. At the single-cell level, the exact pairing of alpha and beta chain can be quantified, making it possible to directly investigate the receptor's antigen target through expression of the TCR in cell lines[19]. The potential of identifying and recapitulating antigen-specific TCRs is particular important for development of new therapeutics, such as CAR (chimeric antigen receptor) T cell therapy, in which an antigen-specific receptor could be repurpose as potent treatment for diseases[20].

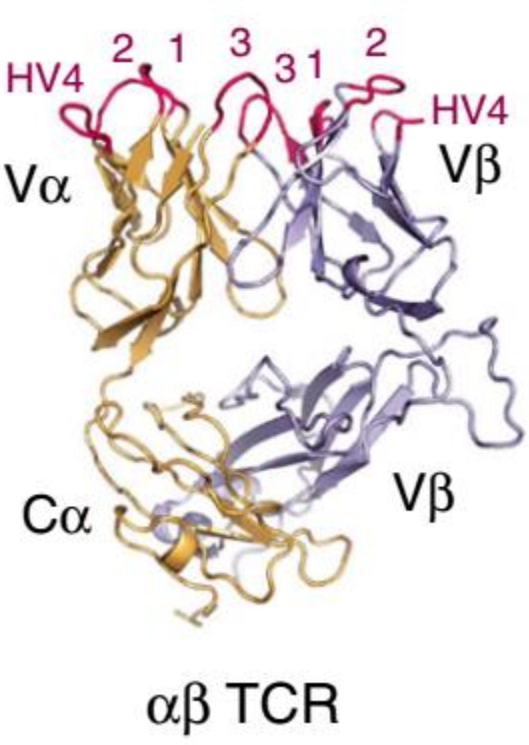


Figure 1-1. Protein structure of TCRα and TCRβ. The CDR regions are highlighted in pink. Figure adapted from Castro, C. et al[21].

Nonetheless, both bulk and single-cell TCR sequencing can be useful, depending on the application. In the next section, we briefly review several conventional methods for both.

1.2 Conventional methods of bulk and single-cell TCR sequencing

Traditional techniques of TCR characterization have depended on gene specific amplification of the TCR mRNA or gDNA. In both cases, a multiplex pool of primers targeting the variable (V) genes on the 5' side of the transcript and the constant (C) or joining (J) regions on the 3' side are used to amplify the TCR sequences. Due to the complexity of the primers, and the nonspecific artifacts inherent to the approach, a nested PCR approach is usually used, wherein two sets of gene-specific primers are used to successively amplify the target of interest[22]. Flanking sequences that are necessary for NGS platforms (e.g. Illumina) are then attached to the amplicons with additional PCR amplification or ligation[23].

While TCR sequences can be amplified from mRNA or gDNA, each presents different advantages and challenges. mRNA contains the recombined TCR transcripts without introns between the V, D, J, and C genes. This allows for efficient amplification of the TCR transcripts that can be readily sequenced on the Illumina sequencers. The resulting insert length varies between 400-600bp, which is just under the sequencing length limit of major Illumina platforms (i.e. 600 cycle kit on the MiSeq). However, each T cell could contain a different amount of mRNA TCR transcripts. Therefore, at the population level in which the frequency of a TCR sequence is equated to the frequency of the T cell clone, the variable number of transcripts could confound the results, making comparing the frequencies of different T cell clones more difficult. On the other hand, while each cell only has a fixed number of gDNA molecules, the introns between the C genes and the recombined VDJ segments are not excised at the gDNA level, making amplification of the recombined receptor sequences more difficult. Furthermore, while mRNA sequences directly represent protein sequences that are being produced by the cells, gDNA sequences may not necessarily represent translated protein, since multiple loci are available within each cell[24].

PCR amplification can also introduce other artifacts in the resulting data. Firstly, a DNA polymerase can often introduce errors into the amplified products, though with the current generation of error correcting enzymes this issue is at least somewhat ameliorated[25]. Secondly, during a PCR reaction, different templates are often preferentially amplified at different rates, often due to the variable GC content of the template molecules. Templates can also be preferentially amplified randomly, a process termed PCR "jackpotting," wherein a particular template molecule is preferentially amplified in the early stages of the PCR process by chance[26]. Then due to the exponential nature of PCR, the jackpotted template exponentially amplifies over other template molecules.

Both issues are in large part ameliorated by the introduction of unique molecular identifiers (UMIs) into the reverse transcription of mRNA transcripts. In this process, a random nucleotide barcode is introduced to each transcribed cDNA product, resulting in an identifier for each unique mRNA molecule that is maintained in the subsequent steps of PCR amplifications. The barcode can be read out during sequencing, and therefore amplified products originating from the same cDNA transcript could be computationally grouped together. This process not only ameliorates the differential amplification of different cDNA templates, but PCR errors introduced into individual amplified product can also be corrected by comparing the sequences sharing the same UMI to derive a single consensus sequence[27].

While UMIs may be introduced using a variety of methods, they are most often incorporated using template switching oligos (TSOs) as part of a 5' rapid amplification of cDNA ends (RACE) reverse transcription reaction (**Figure** 1-2). In this process, a reverse transcriptase derived from the Moloney Murine Leukemia Virus is used to transcribe mRNA into cDNA. As the enzyme transcribes the mRNA molecule, it adds additional cytosines to the end of the transcribed molecule (on the 3' side of the cDNA, or 5' on the mRNA template). The cytosines act as a capture site for the TSO, which contains guanines on its 3' side. In the ideal situation, the TSO attaches to the newly transcribed cDNA, and the enzyme continues to transcribe the oligo (i.e. "switching" the template), incorporating the TSO sequence, including the UMI, into the cDNA product.

This method has been utilized to good effect by various groups[28]. The template switching reaction has the added benefit of obviating the need for V region-specific primers. This allows for easy adaptation of the technique to different species without redesigning the multiplex primer pools, which can often contain an upwards of 30-60 different primers.

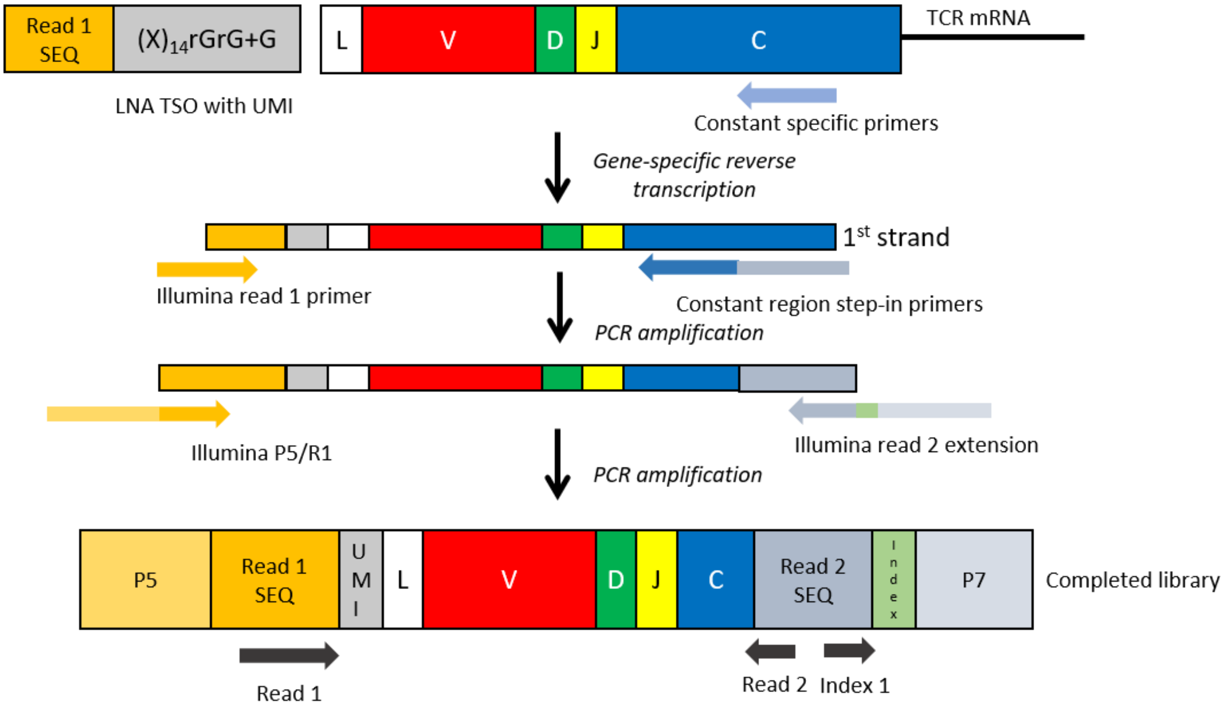


Figure 1-2. Example schematic of TCR amplification using 5' RACE. UMIs are introduced via the TSOs (grey) during reverse transcription. Subsequent PCR amplification is nested with primers specific to the constant regions. Final amplification completes the sequencing handles. Final library is commonly sequenced on the Illumina Miseq.

Template switching is an inherently inefficient process, however[29]. This presents a significant challenge especially when the starting RNA material is sparse, such as the case in single-cell TCR sequencing. Conventionally, single-cell TCR sequencing is done by isolating single T cells into separate compartments, most often by flow cytometry sorting (FACS), followed by amplification of the TCR genes in individual reactions using multiplex primer pools[30]. The process is often laborious and expensive, limiting throughput.

Nevertheless, efforts in TCR sequencing of sorted single T cells have resulted in better understanding of antigen-specific responses in diseases such as cancer[31,32]. Furthermore, because the pairings of alpha and beta chains are maintained, it is possible to clone the sequenced receptors into cell lines to validate their specificities. These approaches have led to better definition of factors necessary to elicit TCR-mediated activation, including putative binding affinity as well as the cellular context of the TCR-MHC interaction[33–35].

In recent years, computational pipelines have also been developed to reconstruct TCR and BCR sequences from full-length RNA-seq results[36,37]. The pipelines usually first identify short reads that have been mapped onto the TCR transcripts, then perform *de novo* assembly to generate the likely full-length TCR sequences. This approach relies on good collective coverage of the TCR sequences by the short reads, and as such is best applied to full-length RNA-seq results, such as SmartSeq2. To avoid erroneous construction of unrelated sequences, this approach is best applied to single-cell sorted datasets, though it is still possible to construct multiple sequences at the bulk level.

1.3 Challenges and Opportunities

While characterization of TCR repertoire can elucidate the dynamics of antigen recognition by the immune system, RNA sequencing (RNA-seq) in contrast, can reveal novel states and functions of disease-relevant T cells through unique patterns of gene expression, albeit without determination of whether those cells are recognizing common antigens[38–40]. Coupling these two types of data is of great interest for modeling T cell responses and isolating those cells most relevant to disease states[30,36,41]. Currently, the preferred method for linking these measures relies on sorting single T cells into multi-well plates by flow cytometry, performing full-length single-cell

RNA-seq (scRNA-seq), and then reconstructing the sequences of rearranged TCRα and TCRβ genes. This strategy is limited in throughput (~10–1,000 cells) by cost, labor and time[13,42,43].

Recently developed high-throughput scRNA-seq methods can profile the transcriptomes of 10^3–10^5 single cells at once, but accomplish this task by first barcoding mRNAs on their 3′ ends during reverse transcription followed by quantification of gene expression by sequencing only those 3′ ends[44–46]. While sufficient to enumerate mRNA abundances, this process hinders precise, direct sequencing of recombined TCR genes because the variable regions of those transcripts—particularly the CDR3 regions closer to the 5′ end of those mRNAs—are not captured efficiently by 3′ library preparation and sequencing protocols[39]. Primer-based approaches that target constant regions of the TCR transcripts to directly enrich CDR3 sequences eliminate reverse-transcription-appended cellular barcodes and UMIs positioned on the 3′ ends of transcripts during amplification, and thus obscure the single-cell resolution of the data (**Figure** 1-3).

New approaches have emerged to determine clonotypes from high-throughput 3′ or 5′ scRNA-seq libraries. These typically rely on specialized RNA-capture reagents (e.g., the customized TCR transcript capture beads of DART-seq or specific kits for InDrop, Dolomite and 10X), limiting their adoption and application to previously archived samples. Some also require combinations of different sequencing technologies (e.g., Illumina and Nanopore in RAGE-seq), complicating their implementation[41,47–51]. Methods that allow for cost-efficient and simple recovery of TCR sequences from 3′ scRNA-seq libraries would enable the study of clonotypic T cell responses with confidence.

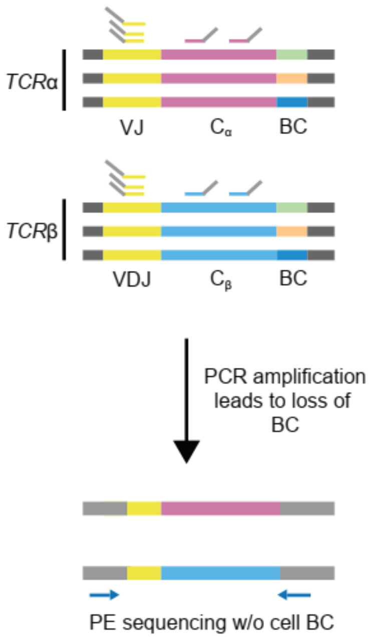

TCRα
VJ
C_α
BC
TCRβ
VDJ
C_β
BC
PCR amplification leads to loss of BC
PE sequencing w/o cell BC

Figure 1-3. Schematic of conventional CDR3 amplification applied to 3′ barcoded libraries. The use of primers specific to the constant regions results in efficient amplification of the TCR CDR3 region but leads to loss of single-cell barcodes.

1.4 Thesis objectives

The aims of this thesis work are twofold. The first is to develop a method that can reliably recover TCR sequences from 3' barcoded single-cell RNA sequencing libraries, such as those produced by most popular single-cell platforms. The method would ideally be easy to carry out, requiring minimal customized reagents, and be applicable to previously processed scRNA-seq samples. The second aim of this thesis is to apply the technique to clinical and animal studies to advance the understanding of antigen-specific responses in T cells. Considering these aims, the thesis is organized into two parts:

Part 1: Technology development

Chapter 2: Size selection enrichment of TCR sequences from single-cell sequencing libraries. Reliable recovery of TCR variable sequences is hindered by the 3' bias of short-read sequencing platforms. In this chapter, we investigate the use of careful, precise size-selection of fragmented sequencing libraries to preferentially enrich for the CDR3 region of TCR sequences in a nonbiased way. We attempted several size selection and enrichment methods, including several forms of gel electrophoresis and magnetic selection. We discovered significant challenges associated with this approach, including repeatability and inherent noise from imprecise expression of TCR transcripts. The work in this chapter highlights the challenge in developing a reliable technique that is also compatible with the limitations of broadly available sequencing technologies.

Chapter 3: Recovery of paired TCR sequences from single-cell Seq-Well libraries reveals clonotypic T cell signatures. After investigating the plausibility of recovering TCR CDR3 sequences through size selection, we attempted using multiplex V primer pools to construct TCR sequencing libraries that maintain single-cell resolution of the libraries. We successfully achieved this goal by using a combination of biotin-streptavidin enrichment of TCR transcripts and single-step primer extension to avoid direct amplification with the multiplex primer sets. We validated the method by applying it to murine T cell libraries containing known proportions of T cells from OT-I transgenic mice. We then applied the method to study clonotypic response in tetramer sorted T cells from animals that were immunized with HPV-E7 antigen. We detected multiple groups of clonotypes with different transcriptomic profiles. Finally, we applied the method to CD154 enriched T cells from peanut allergy patients. We identified expanded T cells that were likely to be antigen-specific for peanut antigens. These expanded T cells also coincided with expression of Th2 cytokines, a known signature of food allergies.

Part 2: Application to the studies of immunological disorders and diseases

Chapter 4: Application of Seq-Well and TCR recovery to the study of peanut oral immunotherapy. Peanut food allergy is a type 1 hypersensitivity disorder that is mediated by IgE antibodies. The exact roles of T helper cells in the condition remain largely unknown. We applied the improved versions of Seq-Well and TCR recovery methods to study longitudinal samples from patients undergoing oral immunotherapy (OIT). We identified T cells with known T helper cell profiles, including Th1, Th17, and several subtypes of Th2 phenotypes. We found that each of the subsets was largely clonally distinct. Interestingly, while we did not find evidence of OIT-induced changes in the TCR repertoires of patients, we found that T cell clones were phenotypically suppressed over the course of treatment, except for clonotypes of the Tfh phenotype. We also found evidence of clonal sequence convergence in each of the T helper subtypes, suggesting that phenotypes in peanut reactive T cells could be in part driven by their TCR sequences. Our findings suggest that OIT is likely only effective in modulating a subset of T helper cells, leading to varied clinical outcomes.

Chapter 5: Application of TCR recovery to other biological systems. Due to the nature of the TCR recovery technique, it is applicable to processed samples *post hoc*, allowing us to retroactively study TCR profiles of samples that have already been processed using standard scRNA-seq techniques. In this chapter, I will highlight two examples of such efforts. In collaboration with the Alex Shalek lab, I applied our method to longitudinal HIV samples collected in South Africa. We were able to detect clonal CD8 T cells that have likely expanded in response to infection. Next, I applied the method to a study of tuberculosis (TB) in an animal model of cynomolgus macaques. We were successful in adapting the method for this species, which necessitated a new design of primers. We were able to detect clonal expansion in TB granulomas with high bacterial burden, suggesting that the cells were clonally expanded in response to infection.

2. Size selection enrichment of TCR sequences from single-cell sequencing libraries

This chapter describes our initial attempts at recovering TCR sequences, in particular the CDR3 sequences, from 3' barcoded single-cell libraries. 3' barcoding is a common strategy for various high-throughput single-cell sequencing platforms. We show that by carefully selecting for the size of the sequencing libraries, we were able to mitigate the effects of 3' bias and recover CDR3 sequences from a significant portion of T cells. Additional testing of the methodology, however, revealed several unexpected challenges. These included issues with repeatability and scalability. By carefully analyzing the bottlenecks of this approach, we were able to use the results to inform better technique design, detailed in Chapter 3.

2.1 Motivation

2.1.1 Challenges in recovery TCR CDR3 from massively parallel single-cell libraries

While several different platforms for high-throughput single-cell sequencing have emerged over the years, most follow similar design principles to achieve their higher throughputs. In general, tissues are dissociated into single-cell suspension, then the cells are individually segregated into separate compartments with a barcoded mRNA capturing reagent. Cells could be separated into single-cell compartments through various methods, such as FACS sorting into individual wells, laser capture, droplet emulsion, or random loading into nanowells[36,44]. Once the cells are separated into individual compartments, the cells are then lysed, and their mRNA transcripts are captured with a randomly barcoded reagent that maintains single-cell resolution. Once the barcoded mRNA transcripts are recovered and transcribed, the recovered material can then be pooled and processed in a single reaction. The barcodes can then be read out by sequencing, and single-cell resolution can be deconvoluted.

The most common mRNA barcoding method is 3' barcoding, or barcoding on the 3' side of the mRNA. In this strategy, single-cell barcodes are included in the poly-d(T) primers, which are then used to generate first-strand cDNAs through reverse transcription and thereby incorporating the barcodes into the transcribed molecules. The full-length cDNA products are then amplified to generate enough materials for sequencing library preparation. Due to the restriction of common short-read sequencing technologies, the amplified cDNAs need to be first fragmented, whether physically or enzymatically, into smaller pieces before the final library can be prepared. During the subsequent library preparation, the 3' side of the fragmented cDNAs is preferentially amplified to ensure that the single-cell barcodes could be sequenced. This process results in pairing of the single-cell barcodes with their corresponding 3' mRNA sequences, which can be computationally processed for downstream analysis.

Seq-Well, which was developed in the J. Christopher Love lab in collaboration with the Alex Shalek lab, is one such single-cell sequencing platform (**Figure** 2-1). A single-cell suspension is loaded onto a PDMS chip with nanowells, each loaded with a mRNA capture bead. Each bead is coated with poly(dT) primers that are barcoded with single-cell barcodes and UMIs that uniquely label each captured mRNA molecule. Once the cells are loaded, the nanowell array is then sealed with a polycarbonate semi-permeable membrane that allows for buffer exchange to promote mRNA capture. Once the cells have been lysed and the mRNAs hybridized to the beads, the beads can then be recovered from the array, and the captured mRNAs can be pooled and amplified, and processed for sequencing and analysis.

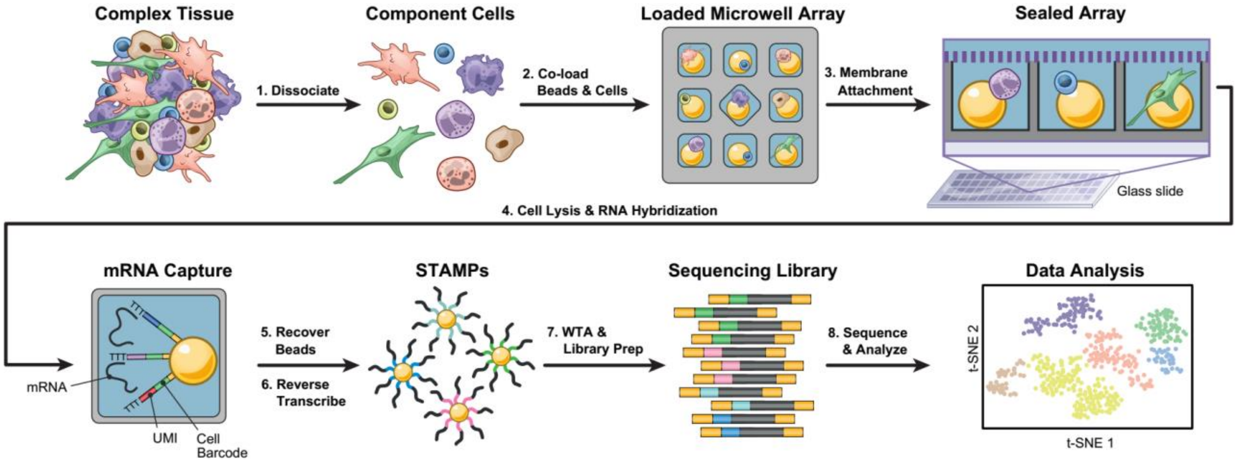


Figure 2-1. Single-cell RNA-seq processing by Seq-Well. Tissue of interest is dissociated into a single-cell suspension before being loaded into a PDMS nanowell array preloaded with barcoded mRNA capture beads. After cells are loaded into the array, a semi-permeable membrane is attached to seal the array. Cells are then lysed and the mRNAs are hybridized onto the beads. The beads are then removed from the array for bulk processing. The resulting data are demultiplexed by the original bead barcodes to achieve single-cell resolution. Figure adopted from Gierahn, T. M. *et al*[46].

3' barcoding presents several advantages. Firstly, it is straightforward. The strategy does not require significant changes to the reverse transcription reaction. Secondly, barcoded primers are also relatively easy to synthesize on solid substrates through split-and-pool synthesis. Though the process introduces significant 3’ bias, it is sufficient for quantifying frequencies of most genes (**Figure** A2-1). However, as mentioned in Chapter 1, 3’ bias presents significant challenges for TCR sequencing, as the CDR3 is located closer to the 5’ side of the TCR transcripts. Furthermore, in order to accurately determine the CDR3 sequences, TCR transcripts would have to be sequenced relatively deeply. In the whole transcriptome library data, TCR transcripts are often relatively lowly expressed, and as such are not adequately sequenced in most workflows.

Amplification with TCR specific primers presents other issues as well. As noted in Chapter 1, constant region primers are situated upstream of the single-cell barcodes, and therefore would eliminate the barcodes after amplification. Amplification with V region primers alone, without constant region primers, often produce non-specific products, and therefore the resulting library could be difficult to sequence successfully.

To overcome these challenges, we attempted to develop a method that does not require amplification with TCR specific primers. In this chapter, we present an approach using size selection to enrich for CDR3 sequences.

2.2 Methods

Single-cell transcriptome sequencing. Human or murine T cells were processed for single-cell RNA sequencing via Seq-Well. In brief, up to 30,000 cells per sample were loaded into the arrays, resulting in roughly single-cell occupancy in each nanowell with a single barcoded poly(dT) bead. The arrays were then washed, and sealed with a semi-permeable polycarbonate membrane. The sealed arrays were then submerged in lysis buffer, and subsequently in hybridization buffer to allow mRNAs to hybridize onto the beads. The beads were then recovered from the arrays through centrifugation, and the captured mRNAs were reverse-transcribed, amplified, and prepared for sequencing using the Nextera XT kit. Libraries were sequenced on either the Illumina NextSeq or Novaseq.

Sequencing data preprocessing. Single-cell transcriptomic data was processed using Drop-seq tools (http://mccarrolllab.com/dropseq) as previously described[44,46]. In brief, barcodes and UMIs were collapsed with a single-base error tolerance. Cells with less than 500 detected genes and 1,000 UMIs were filtered. The resulting data were then natural-log normalized for each cell to account for library size, and variance due to detected mitochondrial genes were regressed from the data.

Biotin-streptavidin enrichment for TCR. Enrichment of TCR-encoding transcripts from whole transcriptome amplified (WTA) starting materials was done with the XGen Lockdown reagents (IDT; Cat.No.1072281), with modifications. Biotinylated TCRα and TCRβ probes were purchased (IDT Ultramer services), mixed, and diluted to 1.5 µM each. Up to 3.5 µL of WTA was added to 8.5 µL of xGen 2x hybridization buffer, 2.7 µL of buffer enhancer, 0.8µL of UPS primer (50 µM), and 0.5 µL of human cot-1 DNA (Invitrogen; Cat.No.15279011). The mixture was incubated at 95^{o}C for 10 min, and 1 µL of diluted TCRC mix was added. The final mixture was then incubated at 65^{o}C for 1 h. Then the remainder of the xGen Lockdown protocol was followed. 50 µL of streptavidin Dynabeads (Invitrogen; Cat.No.65306) was used for each sample. Each sample was eluted into 20 µL of water.

To amplify the TCRα and TCRβ transcripts after enrichment, five PCR reactions were done for each enriched sample with the following composition: 2 µL of eluted sample, 2 µL of UPS primer (10 µM), 8.5 µL of water, and 12.5 µL of 2x Kapa Hifi Hotstart Readymix (Kapa Biosystems). The following PCR cycling condition was used: 1 cycle of 95^{o}C for 3 min; 25 cycles of 98^{o}C for 40 s, 67^{o}C for 20 s, 72^{o}C for 1 min; and 1 cycle of 72^{o}C for 5 min. The five reactions were then pooled to a final volume of 100 µL. Products of >1,000bp were purified using homemade purification reagents outlined by Rohland and Reich[52]. The purified product was eluted into 15 µL of water. Quality of the final product was assessed using fragment analyzer (Advanced Analytical/Agilent).

Assessment of TCR transcript enrichment. A qPCR assay was used to assess enrichment of *TCRA* or *TCRB* transcripts after affinity enrichment. Three rounds of TCR affinity enrichment were performed as described. TaqMan Fast Advanced Master Mix (Applied Biosystems; Cat.No.4444556) was used along with FAM TaqMan primer mixes (*TCRA*, HS00354482_m1; *TCRB*, HS01588269_g1; *GAPDH*, HS02758991_g1). Amplification was done according to manufacturer's instruction. qPCR was done before and after TCR enrichment, and the difference of crossing point (ΔCp) for *TCR* and *GAPDH* was calculated. ΔCp for each the enriched sample was compared to that of the unenriched sample, and the difference was calculated (ΔΔCp). ΔΔCp was used to calculate relative increase in concentration of *TCR* transcripts after enrichment compared to *GAPDH*.

Size selection enrichment of TCR variable sequences through SPRI. After barcoded single-cell cDNAs were made, or after the cDNAs have been enriched for TCR sequences, a portion of the resulting material was used for size selection enrichment of the TCR variable region. Full length material (pre- or post-enrichment) was used in tagmentation with low proportion of Nextera tagmentation enzyme (50% of amount recommended by manufacturer). After amplification, library fragments greater than 1kbp was enriched using either AmpureXP or homemade SPRI reagent. The resulting libraries were quantified on fragment analyzer (Advanced Analytical/Agilent).

Size selection enrichment of TCR variable sequences through Pippin-prep. Similar to size selection through SPRI, after sequencing libraries were tagmented (pre- or post-TCR enrichment), the products were used for size selection through the pippin prep. Pippin prep (BluePippin) was used to select for library fractions larger than 1kbp. The collected fractions were concentrated using SPRI and analyzed using fragment analyzer (Advanced Analytical/Agilent).

Sequencing of size-selected TCR libraries. After TCR-enriched libraries were constructed, they were quantified via qPCR and sequenced on the Illumina MiSeq. Read 1 was performed with the Seq-Well sequencing primer to sequence the single-cell barcodes. Standard Nextera sequencing primer was used in Read 2 to sequence the enriched variable region of the TCR. Read 1 was performed for 20 cycles, and Read 2 was performed for 150 cycles.

Analysis of TCR sequences. TCR sequencing data were filtered by mapping to the TCR-encoding loci (chromosome *TCRB* 7 and *TCRA* 14 for human, and *Tcrb* 6 and *Tcra* 14 for mouse, respectively). The filtered data were categorized by cell barcodes and unique molecular identifiers (UMIs). Cell barcodes and UMIs with at least 10 filtered reads were kept and the rest were discarded. Each set of reads was then mapped to *TCRV* and *TCRJ* IMGT (imgt.org) reference sequences via IgBlast, and mapping to each V and J region were tabulated. The reads were then filtered for “strong plurality,” wherein the ratios of the most frequent V and J calls to their respective second most frequent calls were calculated, resulting in possible values of 0.5 to 1. Cell barcodes with top V and J ratios of greater than 0.6 were kept, and the rest were filtered out. Within each cell barcode group, reads with the top V and J calls were then used for CDR3 calling, and a similar ratio was calculated based on the nucleotide sequence of the CDR3 region. For CDR3 calling, nucleotides corresponding to the 104-cysteine and 118-phenylalanine were identified according to IMGT references, and amino acid sequences in between the residues were translated. TCR sequences were then matched to single-cell data via the cell barcodes. If multiple TCRα and β chains were detected for a cell barcode, the TCRβ sequence with highest numbers of UMIs and raw reads were kept. Up to two TCRα sequences with the top two highest numbers of UMIs and raw reads were kept. We note that non-functional CDR3s (i.e. CDR3s with stop-codon or out-of-frame sequences) are often a result of initially unsuccessful V(D)J recombination, and are often shared in clonal cells[43]. As such, nonfunctional CDR3s were excluded from additional functional phenotype analysis, but used as unique markers for clonal tracking.

2.3 Results

2.3.1 Enrichment of TCR transcripts through biotin-streptavidin pull-down

To increase the proportion of TCR transcripts within the amplified cDNA pool, we decided to use biotinylated oligonucleotide probes designed to target the constant regions of the TCRα and TCRβ chains. The oligonucleotides were designed to be roughly 90 nucleotides long, with a number of degenerate nucleotides to accommodate different alleles of the constant regions. Oligonucleotide probes specific for both murine and human sequences were designed. While the probes were used in combination with the IDT xGEN enrichment kit in accordance to manufacturer's instruction, we believe that any other similar protocol for biotin-streptavidin pull-down enrichment would be equally effective.

To quantify the effects of the enrichment, we used qPCR to quantify the relative concentrations of the TCR transcripts to housekeeping genes, such as actin or GAPDH. We performed the enrichment on cDNA samples generated from a variety of different starting materials with varying proportions of T cells. We also performed the enrichment for multiple successive rounds to determine the optimal number of enrichments that we should perform.

The results are shown in (**Figure** 2-2). Across the samples, we were able to reach 10^4-10^5 fold enrichment with just one round of enrichment. Interestingly, depending on the relative proportions of T cells in the different samples, subsequent rounds of enrichment provided varying benefits. In cases where T cells were a small proportion of the total population, we observed a larger increase in fold enrichment between first and second rounds. However, in samples with relatively high proportions of T cells, we observed little to no increase in relative enrichment in subsequent rounds. This observed increase in TCR transcript concentration was consistent across TCRα and TCRβ transcripts. Due to the efficiency of the enrichment, we decided to use just one round of enrichment for a majority of following experiments. Additional enrichments were performed only when necessary.

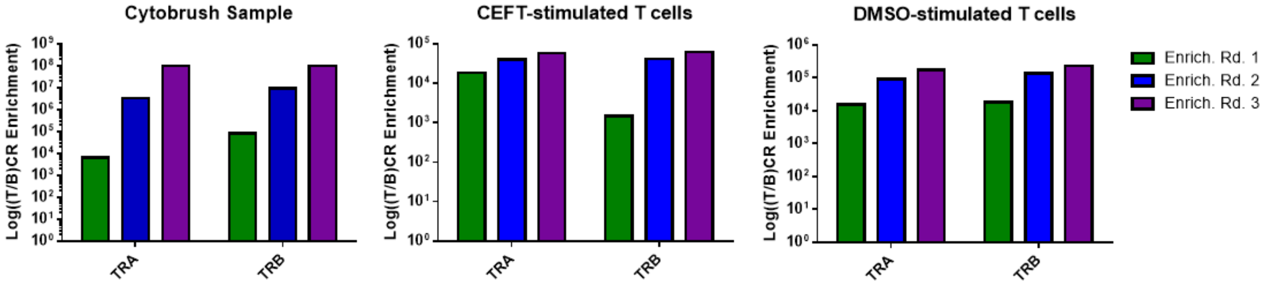


Figure 2-2.qPCR enrichment of TCRα and β chain by biotin-streptavidin pulldown. Enrichment is calculated as increase of ratio relative to GAPDH transcript before and after each enrichment. Cytobrush sample consisted minority T cells (~5%), while CEFT and DMSO-stimulated T cells consisted majority T cells (>99%).

After TCR enrichment, we used a low ratio of tagmentation enzyme to complete preparation of TCR enriched sequencing libraries, resulting in a distribution of products of different sizes (**Figure** 2-4). Before the final product can be sequenced, lower-sized products must be removed.

2.3.2 Enrichment of TCR variable region through size selection

We reasoned that, given that the sequencing library is enriched for TCR transcripts after biotin-streptavidin enrichment, we would be able sequence the variable region of the TCR efficiently if we could preferentially sequence the longer fragments of the library (**Figure** 2-3). Given the penchant for current short-read sequencing platforms to favor shorter library fragments, we would have to remove the shorter fraction of the library, so that the short reads could be targeted to the longer fraction.

For NGS library preparation, our preferred method for size selection is solid phase reversible immobilization (SPRI). SPRI is a common technique to preferentially purify DNA fragments of a desired size. In a SPRI protocol, a mixture of magnetic beads and DNA crowding reagents (frequently a combination of buffers and a soluble polymer such as polyethylene glycol, or PEG) is added to the amplified cDNA pool at a predetermined ratio. The crowding reagent preferentially drives DNA fragments of a certain size to bind to the beads. The magnetic beads can then be removed, and the bound nucleotides eluted. Based on the ratio of the crowding reagents in the reaction, DNA fragments of different sizes can be preferentially selected.

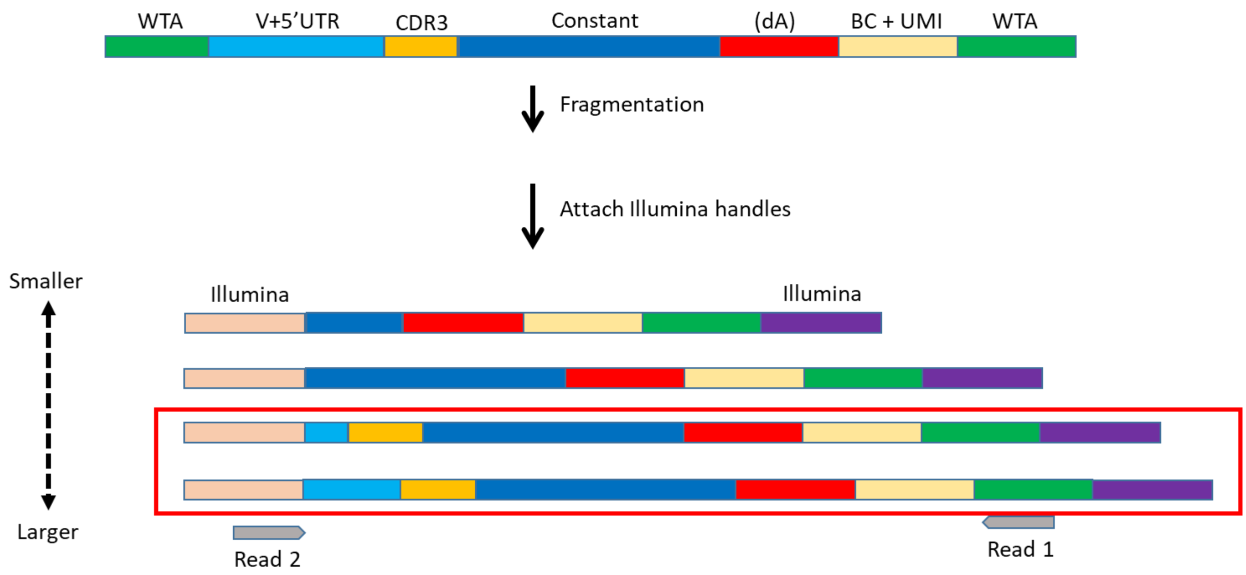


Figure 2-3. Size selection of larger fragments allows paired-sequencing of cell barcode and CDR3. After fragmentation of full-length cDNAs, larger fragments contain the CDR3 region (yellow), while the smaller fragments contain only the constant region (dark blue). Sequencing of larger fragments allow for CDR3 sequences (Read 2) to be linked to the cell barcode (beige; Read1).

One commonly used reagent for SPRI is the AmpureXP reagent from Beckman Coulter, though it is also possible to use homemade reagents. We tested several different SPRI conditions using AmpureXP as well as a homemade version of SPRI reagent. We found that, for enriching fragments of greater than 1kbp, the homemade reagent resulted in better, cleaner enrichment. This is likely due to the slight differences in concentrations of polymer (i.e. PEG) and buffer salts, resulting in differing size selection properties from that of AmpureXP.

By using an aggressive SPRI purification protocol, we were able to achieve relatively precise size enrichment of large nucleotide fragments. However, due to the nature of the protocol, trace amounts of the lower-sized fragments can still be detected by fragment analyzer.

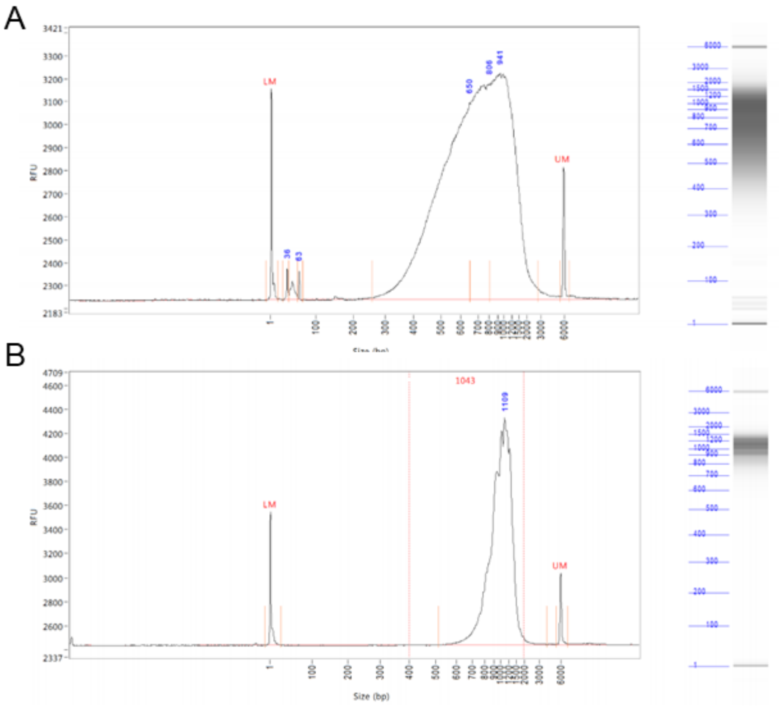


Figure 2-4. Size-selection using SPRI selects for fragments larger than 1kbp. **A**, Typical size distribution of fragmented cDNA using low tagmentase ratio to input cDNA. **B**, Typical size distribution after TCR pull-down enrichment and SPRI purification of fragmented cDNA pool shown in **A.** Combination of the two processes resulting in libraries with average sizes of greater than 1kbp.

2.3.3 Sequencing of size-selected libraries

We then attempted to sequence the size-selected libraries on the Illumina MiSeq. The MiSeq was chosen for several reasons. Firstly, since we were only interested in the TCR transcripts for each cell, we did not need the higher throughput of the NextSeq or the HiSeq. Secondly, we suspected that due to the lower complexity of the sequencing libraries, the sequencing results could contain low diversity regions (such as the TCR constant regions). MiSeq uses 4-color imaging, which is better suited for resolving low-diversity sequencing libraries.

To investigate the feasibility of the size selection approach, we sequenced libraries that were selected to have different average sizes, ranging from 500bp to roughly 900bp (**Figure** 2-5). As expected, larger libraries produced higher proportions of sequencing reads mapping to the

CDR3 region of the TCR, though large portions of the reads still mapped only to the constant regions of the TCR.

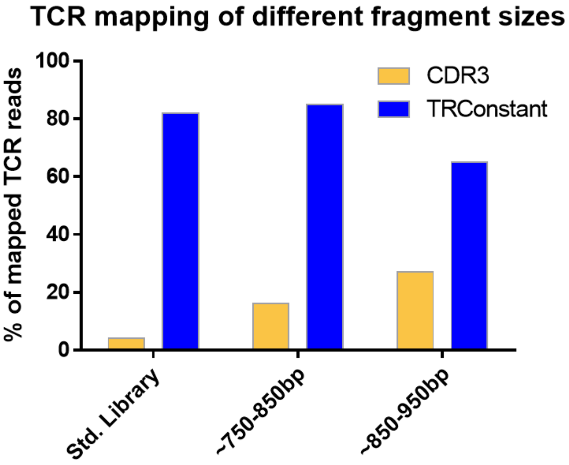
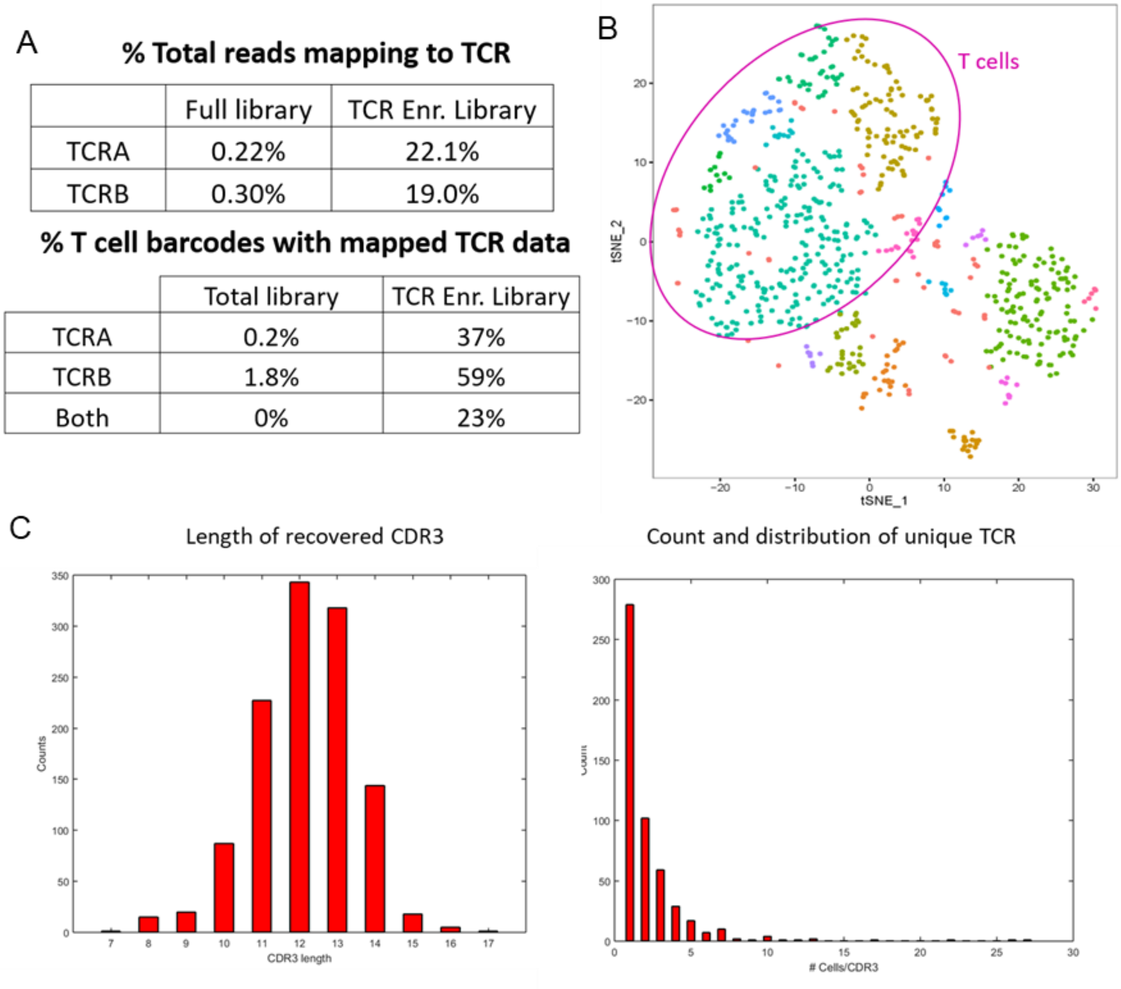


Figure 2-5. Size-dependent mapping of TCR regions. Standard library had average size of roughly 500bp. Percentage of reads mapping to the CDR3 region (yellow) is compared to reads mapping to constant regions (including reads that also span parts of the CDR3 region).

We then attempted to combine the TCR sequencing results with the corresponding single-cell transcriptomic data. We first attempted the experiment in a murine library of mostly T cells. From the TCR-enriched sequencing library, we were able to recover TCRα CDR3 sequences for 37% and TCRβ CDR3 sequences for 59% of T cells in the dataset, compared to less than 0.5% and 2% for TCRα and TCRβ, respectively, in the whole transcriptome data (**Figure** 2-6). In total, over 50% of T cells had recovered TCRα or TCRβ chain, and the proportion of T cells with mapping to both was consistent with the expected result of independent recovery of both TCR chains. The numbers of reads mapping to the TCR genes were also a magnitude higher in the TCR-enriched library compared to the whole transcriptome library. As expected, the lengths of the recovered TCR sequences had a gaussian distribution. The clonotypes were also each largely only attributed to one cell.

We then repeated the experiment with human as well as other murine samples, and we quickly noticed issues with repeatability. One issue was that the stringent size-selection required to adequately enrich for the CDR3 region of the TCR was often difficult to repeat consistently. As shown in **Figure** 2-7, even low amounts of carryover of lower-sized fragments often led to sequencing of predominantly just the constant regions of the TCR. Even though the lower-sized products were a minor fraction of the libraries, due to the sequencer’s preference for shorter fragments, they were still preferentially sequenced. This result was observed multiple times across separate samples, regardless of the species or the sample origin.



% Total reads mapping to TCR

	Full library	TCR Enr. Library
TCRA	0.22%	22.1%
TCRB	0.30%	19.0%

% T cell barcodes with mapped TCR data

	Total library	TCR Enr. Library
TCRA	0.2%	37%
TCRB	1.8%	59%
Both	0%	23%

Figure 2-6. Mapping of TCR sequences onto transcriptomic data in a murine sample. **A**, (top) Proportion of total reads mapping to TCR by the type of sequencing library. TCR-enriched libraries have roughly two magnitudes greater in frequency of TCR mapping reads. (bottom) Percentage of cell barcodes with mapped TCR reads in the total transcriptomic sequencing library compared to the TCR-enriched libraries. **B**, tSNE representation of the murine transcriptomic data. Clusters of T cells are circled in pink. **C,** (left) Amino acid length of recovered CDR3 sequences and (right) counts of cells that share the same CDR3.

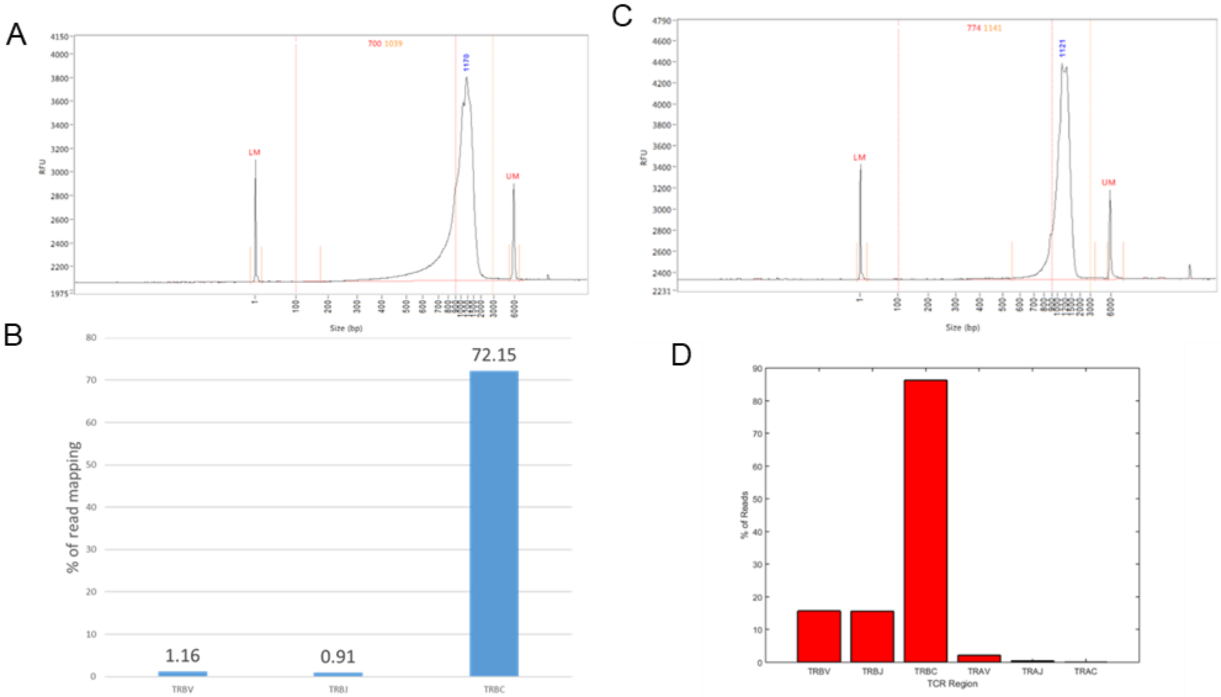


Figure 2-7. Examples of low mapping of TCR variable regions due to incomplete size selection. **A**, Example of incomplete size selection after SPRI, resulting in carryover of lower-sized products. **B**, TCRβ mapping results of library shown in **A**. **C**, Pull-down enrichment of sample shown in **A**. **D**, TCRβ mapping results of library shown in **C**.

2.3.4 Modification to the size-selection process

We attempted various solutions to ensure more consistent size selection, Including multiple steps of SPRI purification, additional rounds of TCR enrichment, and DNA gel extraction. We discovered several modifications to the protocol that improved the selection of CDR3 containing fragments. Firstly, we found that the design of the TCR-enrichment oligonucleotide probes can substantially influence the results. By doing an enrichment using probes targeting the most 5' ends of the constant regions before amplification, we were able to more effectively select for fragments that contained the CDR3 sequences (**Figure** 2-7). Enrichment with the re-designed probes must be done on already tagmented libraries to be effective. As a result, we modified the protocol to perform TCR enrichment after, instead of before, tagmentation.

We also noted that physical size selection methods (i.e. those that allow users to directly separate different fractions of the sample, such as DNA gel extraction), were often more effective than SPRI. Instead of traditional DNA gel extraction, which often requires additional steps of column-based purification, we decided to use the Pippin Prep system, which selectively elutes DNA fragments of a preselected size without column purification. Using the Pippin Prep, we were able to dramatically increase our enrichment efficiency (**Figure** 2-8). Sequencing of the resulting libraries confirmed dramatic increases in mappings of V and J genes, and consequently recovery of CDR3 sequences.

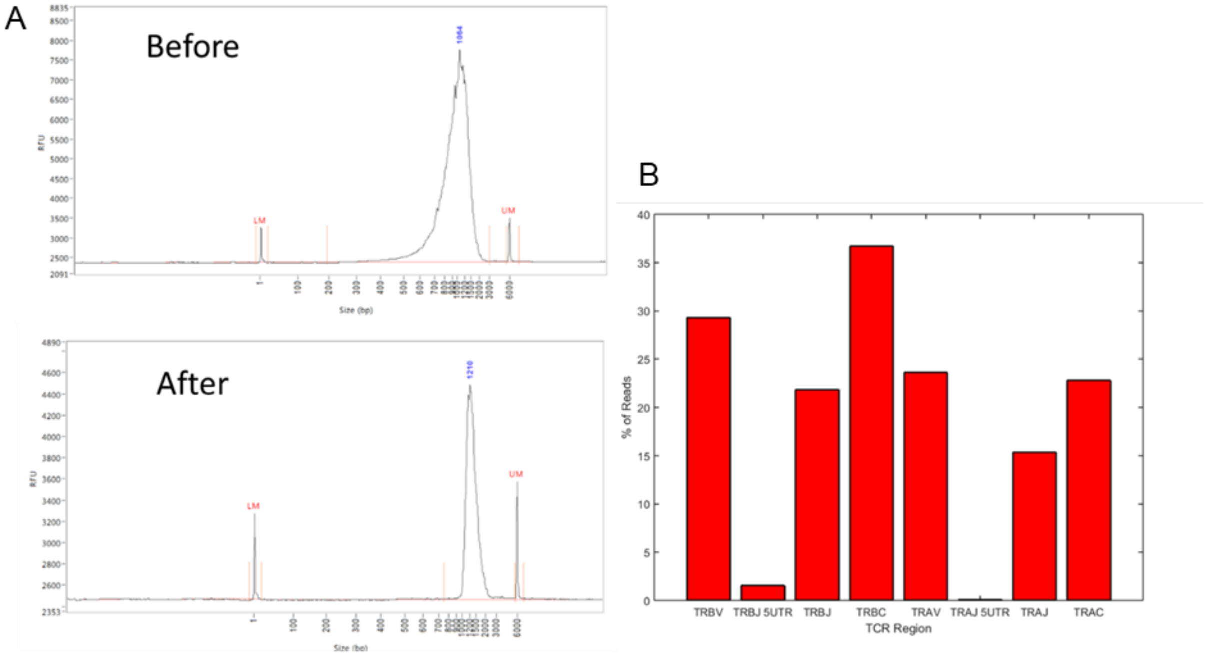


Figure 2-8. Improvement in V and J gene mappings after Pippin Prep size selection. **A**, An example of a TCR-enriched sequencing library (top) before Pippin Prep and (bottom) after Pippin Prep. **B**, TCRα and TCRβ mapping results of Pippin Prep library shown in **A**.

Despite these efforts, we noticed several issues that were still difficult to solve. Firstly, with gel selection and additional steps of pull-down enrichment, we often needed to amplify the samples with additional cycles of PCR amplification to generate enough materials for sequencing. Due to uneven amplification and PCR-jackpotting, additional amplification could reduce the diversity of the final product, leading to lower TCR recovery[53]. Secondly, we also often observed a significant amount of “sterile” TCR transcripts, wherein the constant region of the TCR transcript was not joined to the recombined VDJ segment. We observed that a large portion of these transcripts were mapped to the UTR regions of various J genes, or otherwise intronic regions flanking the V genes (**Figure** 2-9). This seems to be due to constitutive expression of un-recombined transcripts, and since the protocol, as described in this chapter, only directly targets the constant genes, these sterile transcripts could not be effectively removed.

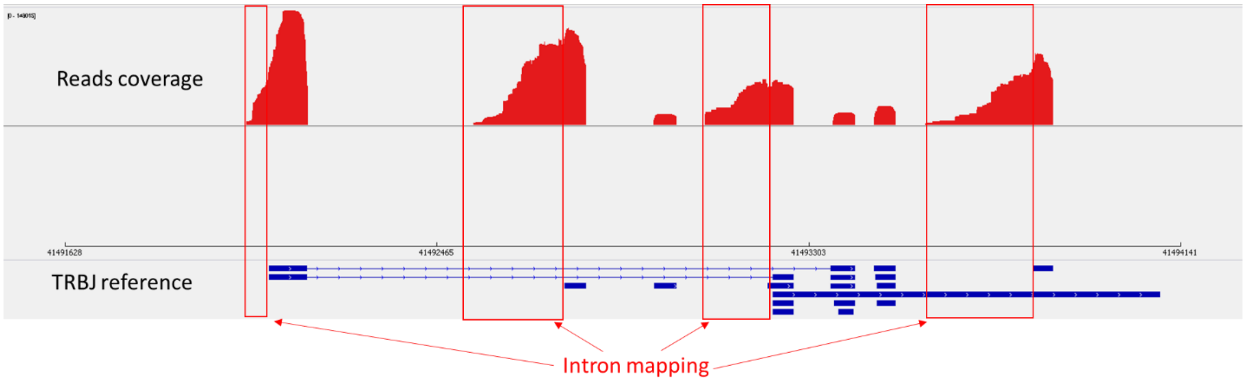


Figure 2-9. An example of sequencing reads mapping to introns of J genes. Top track shows a histogram of reads mapping to the TRBJ reference on bottom track. Reads mapping to the intron regions are boxed in red.

2.4 Discussion

In this chapter, we show that it is feasible to recover the variable region of TCR transcripts by simply selecting for larger fragments from single-cell barcoded sequencing libraries. This method presents several advantages, the chief of which is simplicity. This method does not rely on multiplex primer sets specific to each of the V genes, meaning it can be easily applied to different species without additional optimization. The constant region sequences are typically much better defined and annotated than the V genes, particularly for species with less established genome references. A method that does not require a prior knowledge of V region sequences could be particularly useful for these species.

Nonetheless, we found this approach difficult to repeat consistently. As mentioned before, current short-read sequencing technologies have extreme biases for shorter library segments. In the case of Illumina platforms, this is caused by preferentially clustering of shorter fragments on the flow cell. Meaning, given a sequencing library with a distribution of lengths, the shorter segments will be preferentially sequenced. Magnetic bead-based purification (e.g. SPRI) alone is insufficient in selecting for larger fragments. While physical size selection of the libraries (i.e. gel selection) is effective, it is labor-intensive, and often results in low yields of recovered libraries. Therefore, size selection by gel electrophoresis would be difficult to scale-up efficiently and could present significant risks for processing clinical samples where multiple attempts may not be possible.

Nevertheless, we established biotin-streptavidin pull-down as an efficient approach to enrich TCR transcripts from whole-transcriptome libraries. With just one enrichment step, we were able to achieve roughly 10^4-fold enrichment of TCR transcripts, which translated to higher proportion of TCR-mapping reads in the sequencing data. The enrichment process could also be repeated for multiple rounds in cases where TCR transcripts may be in low abundance. In the following work, we decided that pull-down enrichment would be an appropriate method to increase the frequencies of TCR transcripts before examining other methods of sequencing library preparation.

3. Recovery of paired TCR sequences from single-cell Seq-Well libraries reveals clonotypic T cell signatures

This chapter is in part adapted from: A.A. Tu, T.M. Gierahn, et al[54].

This chapter follows up on the size selection approach in Chapter 2. While we were able to achieve some success by size selection, we had difficulties with repeatability and scalability. As a result, we decided to re-evaluate our strategy.

In this chapter, we incorporate multiplex primer sets specific to V genes into our technique to more effectively enrich for the CDR3 region of the TCR transcripts. We modified our approach to minimized amplification bias from the primer sets. We further modified our sequencing approach to incorporate custom sequencing primers to more effectively target sequencing reads to the CDR3 region.

Here, we also demonstrate the utility of our approach in both a murine T cell model and human samples from peanut allergic patients. In both cases, we were able to detect clonotype-specific transcriptomic signatures that highlight the importance of TCR clonotypic data in better understanding scRNA-seq results.

3.1 Motivation

3.1.1 Learning from difficulties in previous TCR enrichment approaches

After we determined that size selection alone is insufficient in enriching for the CDR3 region of TCR transcripts from 3' barcoded single-cell libraries, we reconsidered the necessary criteria for a useful technique for TCR recovery. Firstly, the method must be repeatable. We anticipated that we would have to reliably process an upwards of 20-50 clinical samples for each project, based on our conversation with collaborators and the needs of several projects within our group. In the same vein, the method must be scalable. While we did not anticipate on utilizing high-throughput machineries, such as a liquid-handling machine, we wanted to avoid steps that could significantly bottleneck throughput, such as gel extraction.

Secondly, we wanted to avoid unnecessary amplification. While the transcripts in our libraries are single-cell barcoded with UMIs, and as a result are resistant to PCR bias, excessive amplification could still result in other artifacts, such as chimeric products and PCR errors. Lastly, we wanted to drastically increase our sequencing efficiency of the CDR3 region. While we were able to recover some CDR3 sequences in Chapter 2 and map them onto the transcriptomic data, it was often done with large amounts of total sequencing reads. This is because a large portion of the reads was mapping exclusively to the constant regions, contributing no useful information regarding the CDR3. A higher sequencing efficiency would allow us to utilize the single-cell barcodes more effectively and sequence more T cells using fewer reads, and thus lowering the cost of the method.

In the remainder of the chapter, we detail our attempts to meet all the above-mentioned criteria by modifying our approach with V gene primers and utilizing a custom sequencing scheme compatible with Illumina sequencing platforms. We then used the modified method to investigate clonotypic immune responses in both murine and human samples.

3.2 Methods

Mouse splenocyte processing for OT-I spiked-in experiments. Spleens were taken from C57BL/6NTac (Taconic) wild-type (WT) mice and B6 *Rag1*$^{+/+}$*Rag2*$^{+/+}$ OT-I (C57BL/6-Tg; Jackson) transgenic mice. Mice were male at age of 8-12 weeks. T cells from each spleen were isolated using magnetic bead-based enrichment (StemCell; Cat.No.19853). Cell concentration was estimated by counting on a hemocytometer. Four mixes of WT and OT-I cells were made with 10%, 1%, 0.1%, and 0.01% of OT-I T cells. Each cell mixture was processed via Seq-Well as previously described[46]. No OT-I *Tcr*a or *Tcrb* chain was observed in the 0.01% spiked-in sample. The resulting single-cell libraries were sequenced on the Illumina NextSeq 500 as previously described. A portion of each constructed library was used for TCR recovery as described below. All animal work was conducted under the approval of the Massachusetts Institute of Technology (MIT) Division of Comparative Medicine in accordance with federal, state, and local guidelines (CAC protocol #01717-076-20, #0917-092-20).

Mouse splenocyte processing for HPV-E7 experiment. C57BL/6NTac (Taconic) mice (female, 8 weeks of age) were primed with 100 µg of MSA-E7 and 25 µg of cyclic di-GMP subcutaneously in the tail base (Day 0). The mice were boosted with the same mixture at Day 14, and at Day 20 spleens from the mice were collected. Splenocytes were stimulated for 6 hours with 10 µg/mL of E7 peptide (RAHYNIVTF) in RPMI with 10% FBS. The cells were then stained with anti-CD8-APC (clone 53-6.7; BioLegend) and E7-tetramer-PE (MBL; Cat.No.TB-5008-2) and flow sorted with a FACSAria II instrument (BD Biosciences) for double-positive T cells. The sorted cells were processed via the updated version of Seq-Well (Seq-Well S^3)[29]. All animal work was conducted under the approval of the Massachusetts Institute of Technology (MIT) Division of Comparative Medicine in accordance with federal, state, and local guidelines (CAC protocol #01717-076-20, #0917-092-20).

Human subjects. The human subjects in this study were all screened for participation in a peanut oral immunotherapy trial (NCT01750879), and some were included in a high-threshold peanut challenge study (NCT02698033), at the Food Allergy Center at Massachusetts General Hospital. All subjects were recruited with informed consent, and the study was approved by the Institutional Review Board of Partners Healthcare (protocol no. 2012P002153) and MIT (protocol no. 1312006071). The participants all had a previous diagnosis of peanut allergy, a history of peanut-induced reactions consistent with immediate hypersensitivity, and confirmatory peanut- and Ara h 2 (a dominant peanut allergen)-specific serum IgE concentrations (> 0.35 kU/l; ImmunoCAP; Thermo Fisher). Blood samples were taken at the time of patient intake, before any treatment of peanut allergy.

Human PBMC processing for allergy samples. PBMCs were isolated from patient blood samples by density gradient centrifugation (Ficoll-Paque Plus; GE Healthcare). Fresh PBMCs were cultured in AIM V medium (Gibco) for 20 h with 100 µg/ml peanut protein extract. The peanut extract was prepared by agitating defatted peanut flour (Golden Peanut and Tree Nuts) with PBS, centrifugation, and sterile-filtering. Anti-CD154-PE (clone TRAP1; BD Biosciences) was added to the cultures for the last 3 h. After harvesting, the cells were labeled with anti-CD3-AF700 (clone UCHT1), anti-CD4-APC-Cy7 (RPA-T4), anti-CD45RA-FITC (HI100), anti-CD154-PE (all from BD Biosciences), anti-CD69-AF647 (FN50; BioLegend), and Live/Dead Fixable Violet stain (L34955; Thermo Fisher). Live $CD3^{+}CD4^{+}CD45RA^{-}$ activated $CD154^{+}$ were sorted with a FACSAria II instrument (BD Biosciences). The sorted cells were processed via Seq-Well[46].

Construction of TCR sequencing libraries. TCRV-UPS2 primers were purchased[30,43] (Eurofin). Two primers mixes were made (one for TCRα and one for TCRβ), and diluted to 10µM each. TCRα and TCRβ reaction mixes were made with 4 µL of purified enriched product, 6 µL of water, 2.5 µL of TCRV-UPS2 primer mix (TCRα or TCRβ), and 12.5 µL of 2x Kapa Readymix. Primer extension was done with the following conditions: 1 cycle of 98^{o}C for 5 min, 1 cycle of 55^{o}C for 30 s, and 1 cycle of 72^{o}C for 2 min. The final product was purified as previously described, and eluted into 11 µL of water.

Complete sequencing handles were added to the final product using the following PCR mix: 0.5 µL of UPS2-N70x primer (10 µM), 0.5 µL of UPS2-N50x primer (10 µM), 9 µL of water, and 12.5 µL of 2x Kapa readymix were added to 2.5 µL of previously eluted product. Four reactions were performed for each sample, using a total of 10 µL of the eluted product. Amplification was done using the following cycling conditions: 1 cycle of 95^{o}C for 2 min; 12-15 cycles of 95^{o}C for 30 s, 60^{o}C for 30 s, and 72^{0}C for 1.5 min; and 1 cycle of 72^{o}C for 5 min. All four reactions were pooled and purified for products >1000bp as previously described. Final product was assessed using fragment analyzer, and a major peak of around 1100 bp was observed for TCRβ product, and 1300 bp for TCRα products. Library concentrations were assessed using the KAPA Library qPCR quantification kit (Kapa Biosystems). A more detailed step-by-step protocol (including example size distributions of successfully amplified samples) is available on http://shaleklab.com/resources/ and on the Nature Protocol Exchange.

Conditions for TCR sequencing. TCRα and TCRβ libraries were pooled at equimolar concentration. For single-end sequencing, 1-2 nmol of the final library was used to sequencing on the Illumina MiSeq. 150 cycles was performed on read 1 using the TCR-specific sequencing primers, and 20 cycles was performed on index 1 using Seq-Well sequencing primer. Sequencing primers were used at a final concentration of 2.5 µM. We aimed for 8-12 * 10^{6} pass filter reads per lane (cluster density of roughly 450K/mm^{2}). Based on the whole-transcriptome data, we allotted ~6,000 T cells per lane.

Single-cell analysis of T cells from mice immunized with HPV-E7. Single-cell analysis was performed as previously described, with modifications[46]. The modifications are as follows: we identified 461 variable genes with log-mean expression values greater than 0.1 and dispersion (variance/mean) of greater than 1. Principal component analysis (PCA) was performed on the variable genes using the RunPCA function in Seurat. Principal components (PCs) were analyzed with the PCElbowPlot function in Seurat, and five significant components were identified. A two-dimensional tSNE visualization was then generated from the PC loadings for these first five PCs. Clusters were identified using the FindClusters function in Seurat with resolution = 0.4. Genes shown in **Figure 3-5** was chosen by using the FindAllMarkers function in Seurat using the previous defined clusters. The resulting list was filtered for genes that show average fold change of greater than 2 with an adjusted *P* value of less than 0.001. To calculate gene expression by clonotypes, normalized gene count of cells sharing the same clonotypes were averaged. Then each gene was scaled to produce a *z*-score with maximum and minimum of +/- 2, respectively. The clonotype gene expression was clustered using ward.D2. Module 2 and 3 were also separately queried against C5 reference set. Signatures from *Singer et al.*[55] was implemented on the dataset using the AddModuleScore function in Seurat.

Single-cell analysis of T cells in peanut allergy. Single-cell analysis was performed as previously described, with modifications[46]. The modifications are as follows: for dataset including all four patients, we identified 486 genes with log-mean expression values greater than 0.1 and dispersion (variance/mean) of greater than 1. PCA was performed on the variable genes using the RunPCA function in Seurat. A two-dimentional tSNE visualization was then generated from the PC loadings 20 most significant PCs. For patient 77, we identified 701 variable genes with log-mean expression values greater than 0.1 and dispersion (variance/mean) of greater than 1. PCA was performed on the variable genes using the RunPCA function in Seurat. PCs were analyzed with the PCElbowPlot function in Seurat, and 15 significant components were identified. A two-dimentional tSNE visualization was then generated from the PC loadings for these first 15 PCs. UMAP was also separately applied to the loadings of these identified principal components. The cluster of T cells with highest clonal expansion was used pseudotemporal analysis by Monocle 3[56]. UMAP dimension reduction was used for pseudotemporal analysis, as was suggested by the authors of Monocle 3. Trajectory was calculated using the implementation of DDRTree in the learnGraph function in Monocle. Signatures from *Wei, et al.*[57] were implemented using the AddModuleScore function in Seurat.

MsigDb signature enrichment analysis. Genes of interest (Modules from **Figure** 3-5, Cluster 3 and 4 from **Figure** 3-10) were queried against gene sets from MsigDb to calculate significant overlap of gene signatures[58,59]. For the murine dataset, Module 2, 3, and 4 were compared against H, C2, and C7 reference sets. The results were filtered for signatures relevant to T cells, and the top five most significant signatures for each Module were shown in **Figure** 3-6. Module 1 was compared against H reference set, and the top five most significant signatures were shown in **Figure** 3-6. For the human dataset, Cluster 3 and 4 were compared against C7 reference set. The results were filtered for signatures relevant to T cells, and the top 10 most significant signatures for each Cluster were shown in **Figure** 3-11**.** Each signature was manually assigned a short description to aid visualization.

Primer sequences, MsigDb signature enrichment results, and other supplementary information are available online at A.A. Tu, T.M. Gierahn, et al[54].

3.3 Results

3.3.1 V gene selection using multiplex primer sets

We reasoned that, in addition to targeting the constant regions of TCR transcripts by the biotin pull-down, we also needed to select for the 5' side, or the V genes, of the transcripts. For this purpose, we examined the use of V gene specific primers. Multiplex primer sets specific to the V genes are often used in conventional methods of TCR sequencing, as described in Chapter 1. By using the V gene primers in our protocol, we could potentially select against the sterile transcripts detected by the size selection method, since these transcripts are less likely to have V genes on the 5' side.

Multiplex primer sets specific to the V genes are often difficult to design. As there are an upwards of roughly 50-70 V genes for each chain, the primer sets would often contain 30-60, or even more depending on the exact protocol, separate primers. Off-target interactions between the primers are difficult to predict, and the resulting PCR artifacts could significantly degrade the quality of the sequencing results. This issue is compounded by the number of cycles needed in amplification using these primer sets, wherein the PCR artifacts could be exponentially amplified[60,61].

For our application, however, we reasoned that because we had already enriched and amplified the TCR transcripts from our samples, we would only need to use the V gene primer sets to attach the sequencing handles upstream of the CDR3 region, and not to further amplify our cDNA materials. We hypothesized that by avoiding amplification with the primer sets, we would also avoid much of the issues associated with multiplex primer sets.

As such, we decided to only use the multiplex primer sets in a one-step primer extension to add partial sequencing handles to the TCR cDNA, then following up with a final amplification using primers targeting shared priming sites to complete the extension of the sequencing handles (**Figure** 3-1). The resulting sequencing library can then be sequenced as normal to recover the CDR3 sequences.

We first tested the approach using an established primer set designed for human TCR sequences. TCR transcripts were first enriched using pull-down enrichment as before, then a one-step primer extension was performed using Kapa Hifi polymerase, followed by sequencing handle amplification. The products were sequenced on Illumina MiSeq using standard paired-end Illumina sequencing primers. First, we noted that, unlike the size selection method, the modified method produced sequencing libraries with much less size variation, with almost no detectable short-length products (**Figure** 3-2). Second, upon analysis and comparison to the size selection sequencing results, we recovered almost identical distribution of V genes for both alpha and beta chains (**Figure** 3-2). This indicates that despite the V region specific selection, we introduced very little selection bias into our protocol.

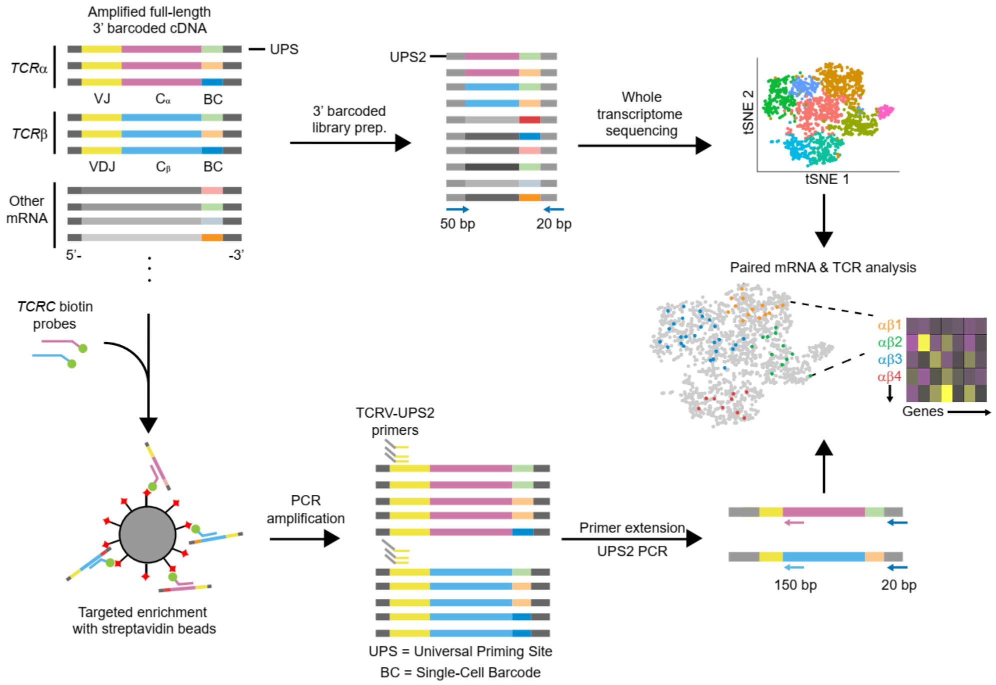


Figure 3-1. Strategy for TCR recovery from 3′ barcoded single-cell sequencing library. Barcoded cDNA libraries (WTA products) including TCRα and TCRβ transcripts in addition to other transcripts (top). Fragmentation and selective amplification of cDNA results in sequencing library used for transcriptomic sequencing, and analyzed via 3′ gene mapping as previously described[46]. *TCR* enrichment of same cDNA library through affinity capture with biotinylated oligonucleotides results in produces amplified products enriched in *TCR*α and β transcripts. Sequencing library is made by primer extension with V region primer sets followed by PCR amplification using the UPS2 handles (bottom). The CDR3 region is sequenced using Illumina MiSeq with custom sequencing primers, and merged with the transcriptomic data based on single-cell barcodes.

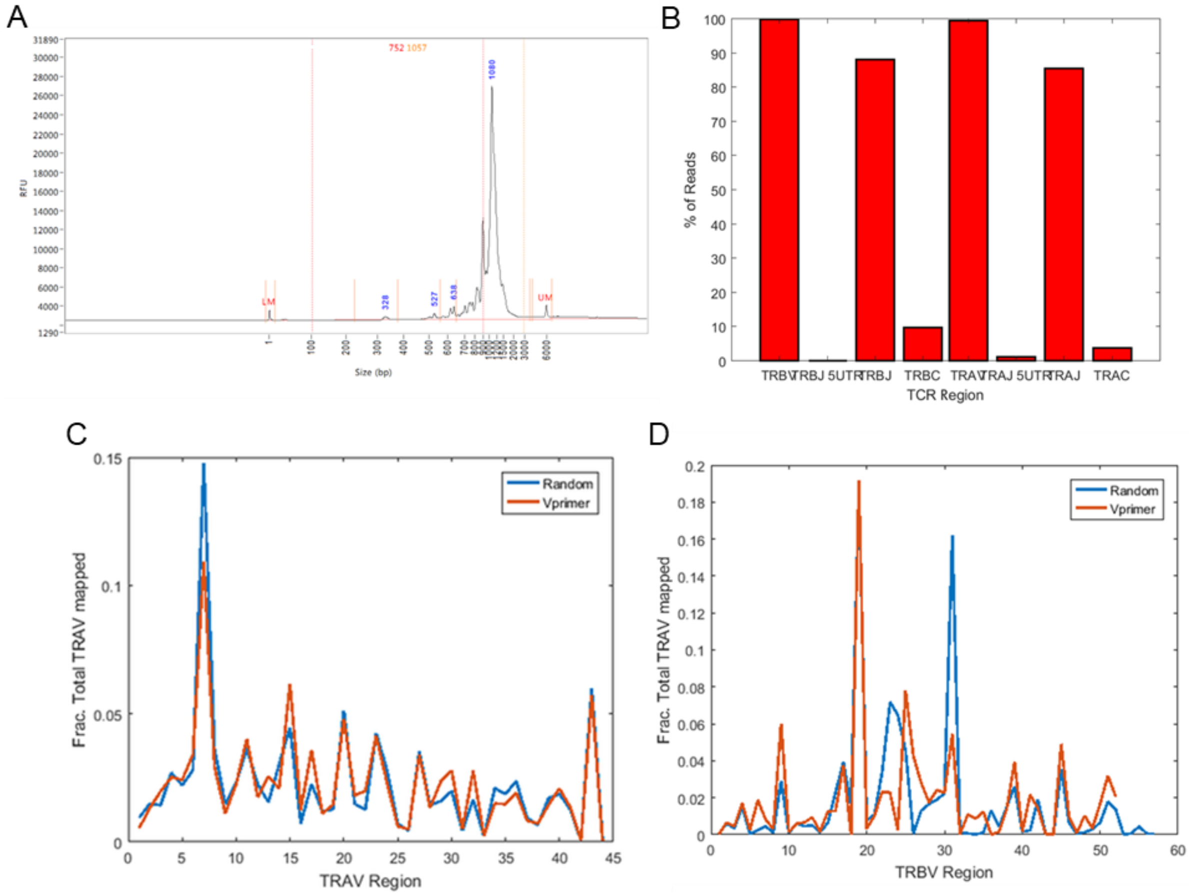


Figure 3-2. Incorporation of V gene primers improves mapping efficiency of TCR-enriched library. **A**, Typical size distribution of TCR enriched libraries with V gene selection. The major product is typically ~1300bp for TCRα, and ~1100bp for TCRβ. **B**, TCR mapping results of library shown in **A**. **C**, Comparison of *TRAV* gene mapping between V gene-selected library (orange) and size-selected library (blue). **D**, same as **C**, but for *TRBV* genes.

3.3.2 Sequencing using custom sequencing primers

Next, we looked for ways to further optimize the sequencing protocol to ensure optimal sequencing quality and cost efficiency. Up to this point, we had been relying on conventional paired-end sequencing (i.e. Read 1 and Read 2) to recover and pair single-cell barcodes to CDR3 sequences. In this fashion, Read 1 would be used to sequence the cell barcodes, and Read 2 (on the opposite strand) would target the CDR3 sequences. While this approach was acceptable in our initial testing, it did contain several inefficiencies. Firstly, despite several steps of enrichment, our final sequencing libraries often still contained a small fraction of non-TCR products, which may still be preferentially clustered on the Illumina flow cell depending on the lengths of the products. Secondly, sequencing quality on Read 2 of the paired-end reads is often worse than that of Read 1. This is due to the additional on-chip cluster amplification required to sequence Read 2 of the paired-end reads.

While neither issue was severe enough to hinder successful sequencing, we nevertheless thought there was significant room for improvement, and thus decided to optimize the process by using custom sequencing primers specific to the TCR transcripts. Illumina sequencing platforms allow for the use of custom-designed sequencing primers to target specific DNA sequences. An example of a custom primer is the Drop-Seq sequencing primer, which is also used to sequence Seq-Well libraries. We reasoned that it should be possible to leverage the constant region flanking the 3' side of the CDR3 to directly target the sequencing reads to the CDR3 region.

As a result, we designed custom sequencing primers specific to the constant sequences of both TCRα and TCRβ chains (**Figure** 3-3). We used the primers to perform Read 1 to directly sequence the CDR3 region. Then, leveraging the Seq-Well custom sequencing primer, we sequenced the single-cell barcodes using 20 cycles on the index read (Index 1), which would normally be used to demultiplex pooled samples. The entire process was done with a single-end sequencing scheme, removing the need for paired-end reads.

As shown in **Figure** 3-3, we were able to achieve high sequencing quality using this sequencing scheme, and because we used a TCR specific sequencing primer, virtually all resulting reads mapped to the variable regions of the TCR. We did note that there is an abrupt drop of quality at around 120th cycle in Read 1. This is likely due to some reads sequencing into the sequencing handle, due to the design of the primers. In the future, the primers should be adjusted to produce longer inserts. We also note that while we could no longer use Index 1 for sample barcodes, it is still possible to incorporate Index 2 on the 5' side of the sequencing libraries.

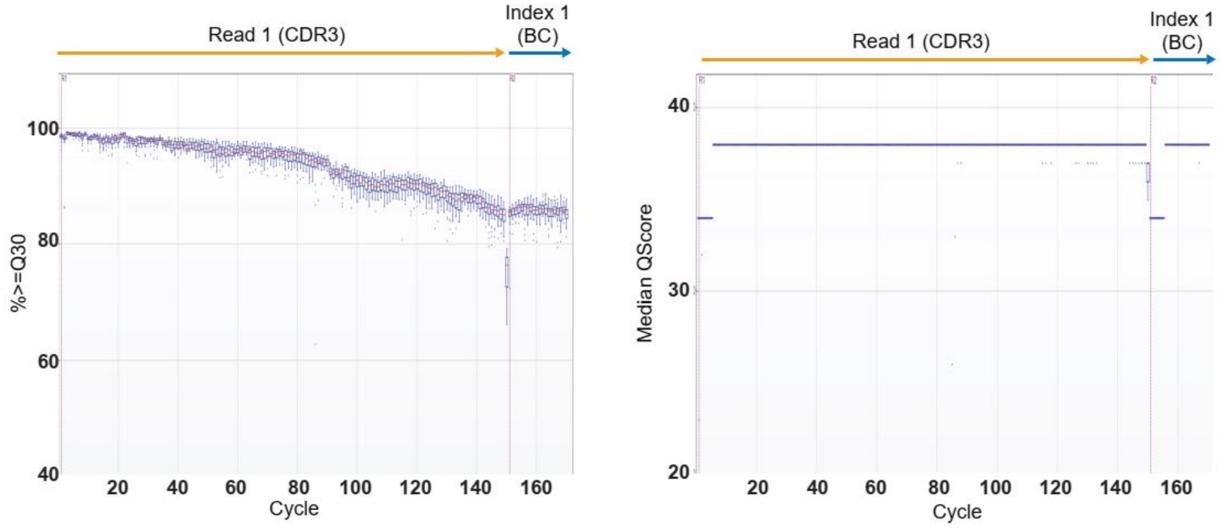


Figure 3-3. Sequencing quality of data produced with constant region sequencing primers. A total of 170 cycles were used to sequenced the CDR3 (1-150) and the cellular barcodes (151-170). (left) Cumulative percentages of reads with overall QScore >=30 by cycle number. (right) Median QScore of each cycle.

3.3.3 Assessing accuracy by OT1 spike-in

To test the sensitivity of our approach, we enriched TCR transcripts from Seq-Well WTA products derived from mouse splenic T cells spiked with T cells from an OT-I transgenic *Rag1*$^{+/+}$*Rag2*$^{+/+}$ mouse (0.01-10%). We recovered *Tcra* (25+/-3%) and *Tcrb* (65+/-4%) *CDR3* sequences from cells detected in our whole transcriptome data (*n* = 4 samples). Both chains were recovered for 20 +/- 3% of cells—similar to the predicted rate of recovery, assuming the capture of each transcript to be an independent event (16+/-7%; **Figure** A3-1). Mapped *Tcr* sequences coincided with expression of T cell markers, such as *Cd3e* (**Figure** A3-1). Sequences sharing the same UMI also had high degree of sequence consensus (**Figure** A3-1**; Methods**). The proportion of OT-I *CDR3* sequences recovered in each sample was consistent with expectations (**Figure 3-4**). Cells with an OT-I *Tcra* chain almost exclusively matched to the expected OT-I *Tcrb* chain (97.7%; 33/34 cells); cells with OT-I *Tcrb* chains, meanwhile, primarily matched with the expected OT-I *Tcra* chain (70.5%; 33/45), though not exclusively (**Figure 3-4**). These results are similar to previous studies, wherein *Rag1*$^{+/+}$*Rag2*$^{+/+}$ OT-I T cells were observed to produce functional TCRα chains in addition to the OT-I TCRα chain[62].

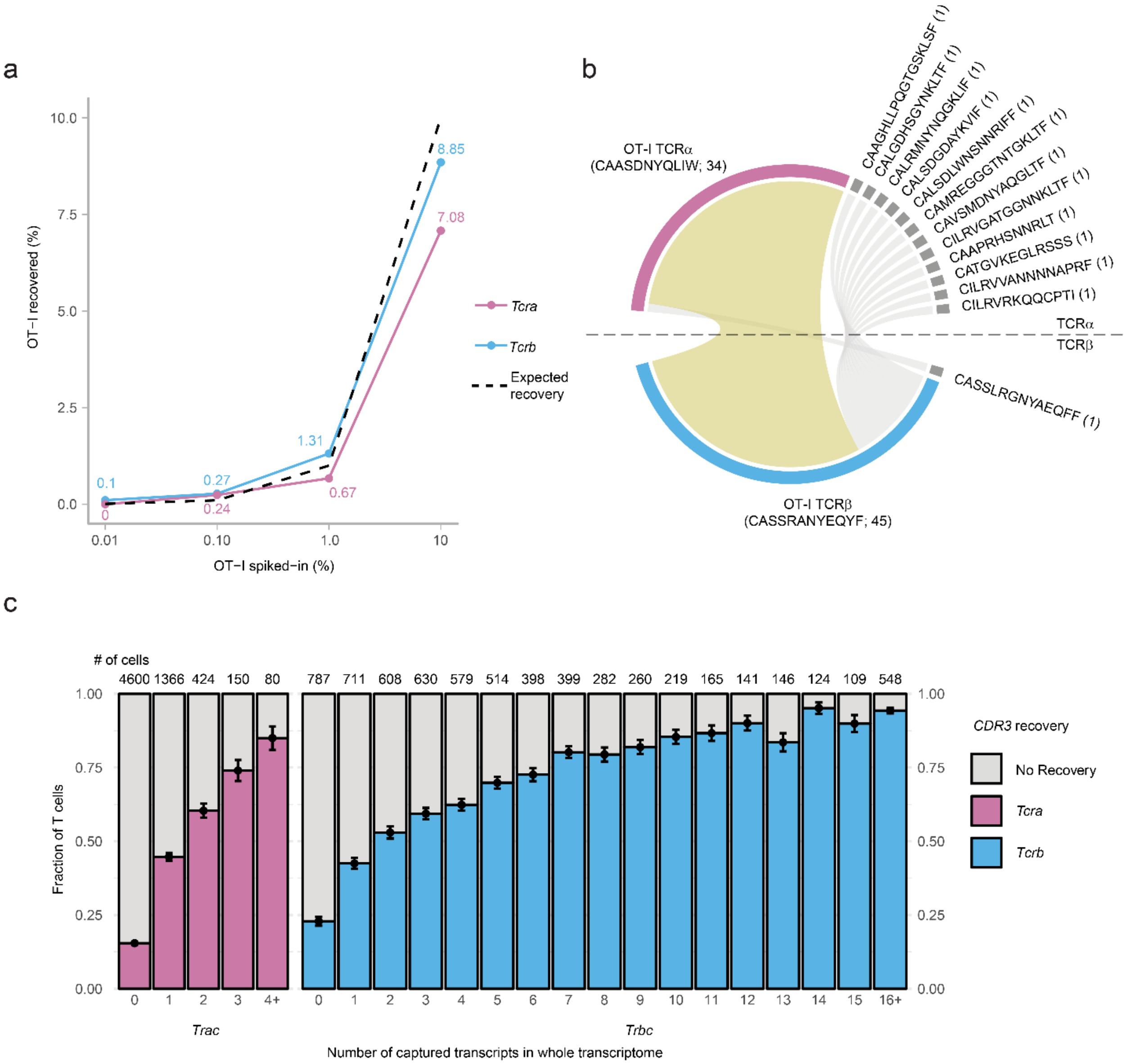


Figure 3-4. Recovery of OT-I *Tcra* and *Tcrb CDR3s***. a**, Proportions of recovered OT-I *Tcra* and *Tcrb* sequences from murine samples spiked-in with 10%, 1%, 0.1%, and 0.01% OT-I *Rag1*$^{+/+}$*Rag2*$^{+/+}$ T cells. Dash line indicates expected recovered proportions. **b**, Pairing of recovered OT-I *Tcra* and *Tcrb* chain from cells in all spiked-in libraries with either OT-1 *Tcra* or *Tcrb* chain sequences. Number of detected cells indicated in parenthesis. Yellow band indicates pairing of OT-I *Tcra* and *Tcrb* sequences from recovered cells. **c**, Proportion of T cells with successful *CDR3* recovery (y-axis) as a function of the constant region mapping via scRNA-Seq by Seq-Well (x-axis). Number of cells with the corresponding number *TCR* constant region

transcripts within their 3′ scRNA-Seq data are indicated above the respective column. Error bar indicates standard deviation of estimated binomial distribution.

We next assessed the relationship between the fraction of cells expressing TCR transcripts in their whole transcriptome data (based on TCR constant region mapping) and the percentage of CDR3 sequences recovered from the same cells (**Figure 3-4**). We observed a high correlation: cells with more copies of TCR transcripts in their whole transcriptome data yielded higher rates of CDR3 recovery from the TCR-targeted libraries (*Tcra*: $r_s = 1$, n = 5, *P* value =0.017 by Spearman; *Tcrb*: $r_s = 1$, n = 15, *P* value < 10^{-6} by Spearman). Overall, excluding classified T cells with no detected TCR genes in their transcriptomic data, we recovered CDR3β from an average of 70+/-4% of cells, and CDR3α from 52+/-3%, resulting in combined pairings of *Tcra* and *Tcrb* sequences for 40+/-4% of T cells (**Figure** A3-1). Finally, we investigated the reproducibility of our approach by comparing technical replicates of the TCR-targeted libraries produced from the same starting WTA material. Across replicates, 94% of detected cellular barcodes were the same, and 99.7% of detected shared transcripts (12,849 out of 12,883) resulted in identical assignments of clonotypes (**Table** A3-1). Taken together, our results show that the method allows consistent and reproducible recovery of CDR3 sequences with high yields.

3.3.4 TCR recovery reveals clonal expansion in immunized mice

Antigen-specific T cells are often enumerated by flow cytometry using tetrameric reagents comprising known antigenic peptides bound to recombinant MHC molecules. The same peptide-MHC complex, however, can select multiple T cell clonotypes[63]. This intrinsic multiplicity can obscure the underlying relationships between phenotypic states and associated clonotypes. We sought to resolve the clonotypic diversity of tetramer-sorted T cells by applying our approach to murine T cells specific to a canonical envelope antigen (E7) from human papilloma virus (HPV16). After immunization and challenge, splenocytes were harvested from mice. Half of the splenocytes were stimulated *ex vivo* with E7 antigen for six hours, and half of the cells were not (**Methods**).

E7 tetramer^{+} CD8^{+} T cells were then sorted from both groups of splenocytes (**Figure** A3-2). These cells were prepared for scRNA-seq using Seq-Well, and their TCR CDR3 sequences were recovered (**Figure** A3-2). In total, 14,424 cells from across four mice were included in the study. We found a diverse set of clonal, expanded T cells within the tetramer-sorted populations isolated from individual animals. For each animal, the 20 most expanded *Tcrb* clones accounted for 69% to 89% of recovered T cells (mean = 908+/-332 cells). Between 77% to 90% of the recovered T cells had clonal *Tcrb* chains. In total, over 900 unique *Tcra* and 1200 *Tcrb* clonotypes were detected.

We next analyzed the clonality of these cells with respect to their whole transcriptomes (**Figure** 3-5). The majority of stimulated cells were transcriptionally distinct from unstimulated cells isolated directly *ex vivo* (**Figure** 3-6). Computationally-determined clusters of cells (PCA followed by tSNE visualization; **Methods**) were preferentially enriched for either unstimulated or stimulated cells; only cluster 5 contained nearly equal portions of both (**Figure** 3-6). The degree of expansion observed among the clonotypes—that is, the number of cells sharing the same clonotype in the dataset—associated strongly with phenotypic clusters of T cells determined based on scRNA-Seq (**Figure** 3-5)**.** The most expanded clonotypes were observed in clusters 0 through 4. These clusters were enriched (compared to cluster 5) in genes associated with cytotoxic effector functions such as *Gzmb* and *Id2*. The least expanded clonotypes were concentrated in cluster 5, and were characterized by enrichments for genes encoding naïve or central memory markers, such as *Ccr7* and *Sell*[64] (compared to clusters 0-4) (**Figure** 3-5 and **Figure** A3-3). This association between the degree of clonal expansion and T cell activation affirms common principles of antigen-dependent activation among T cells[65].

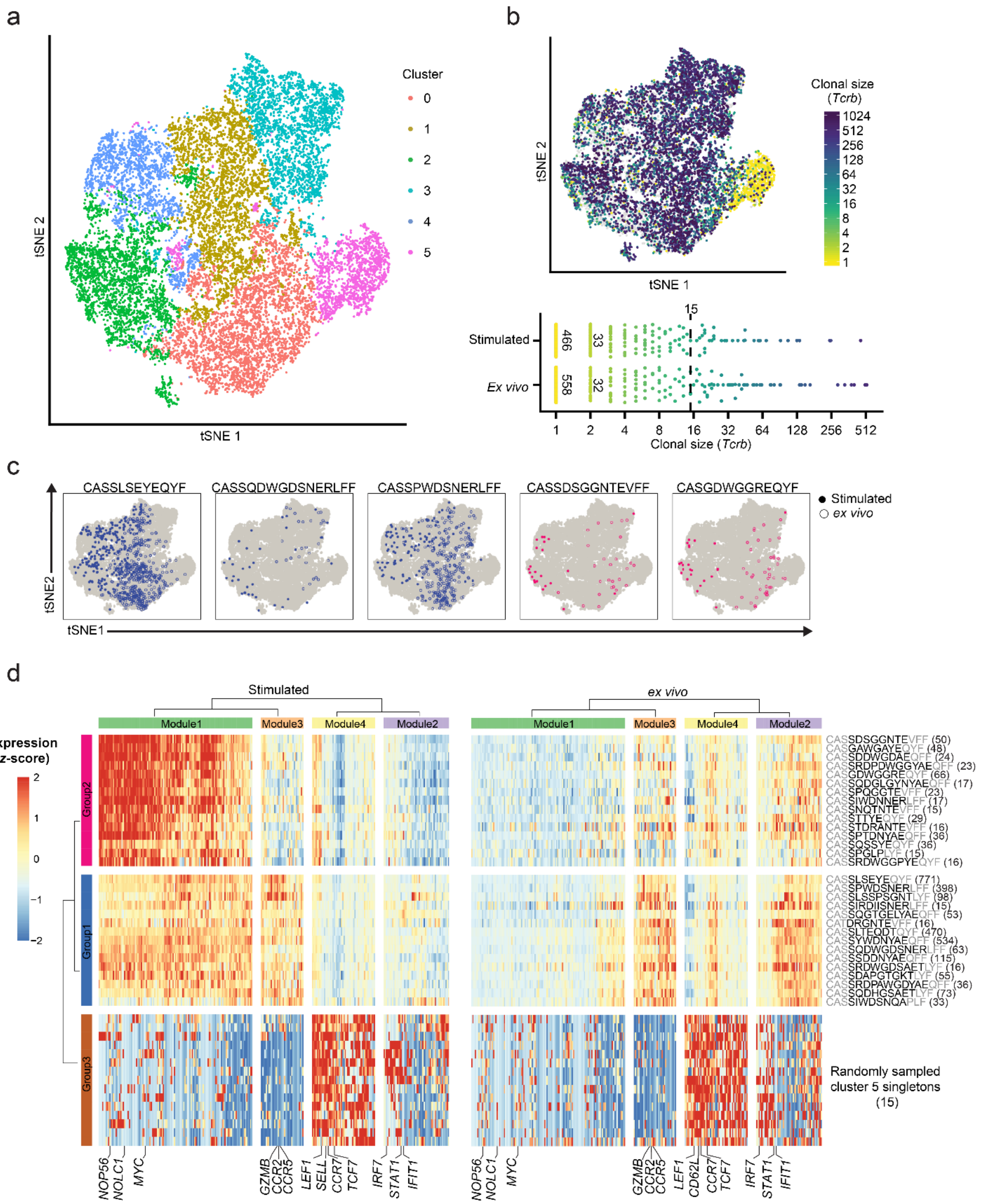


Figure 3-5. scRNA-Seq and TCR analysis of HPV-E7 immunized mice. **a**, tSNE visualization of all cells colored by computationally determined clusters based on transcriptomic data (n = 14,424 cells). **b,** (top) Clonal size of *Tcrb* chain mapped on tSNE visualization of scRNA-Seq results. Cells are colored by the clonal size of their detected *Tcrb* clonotype. Clonal size is defined as the number of cells that share the particular clonotype. (bottom) Distribution of

differentially expanded clonotypes between the stimulated and *ex vivo* conditions. Each colored circle indicates a unique clonotype. **c**, Example mappings of selected clonotypes from Group 1 (blue) and 2 (magenta) shown in **d** on tSNE visualization of all cells. **d**, Heatmaps of differentially expressed genes amongst the expanded clonotypes (>=15 cells) and 15 randomly sampled non-expanded cells (singletons) from Cluster 5 (see **a**) between the *ex vivo* and antigen-stimulated conditions. Gene expression represent scaled averages within cells of the same clonotype across the two conditions. Number of cells shown in parenthesis. Data represent combined data from four independent experiments of four mice total (**a**-**d**).

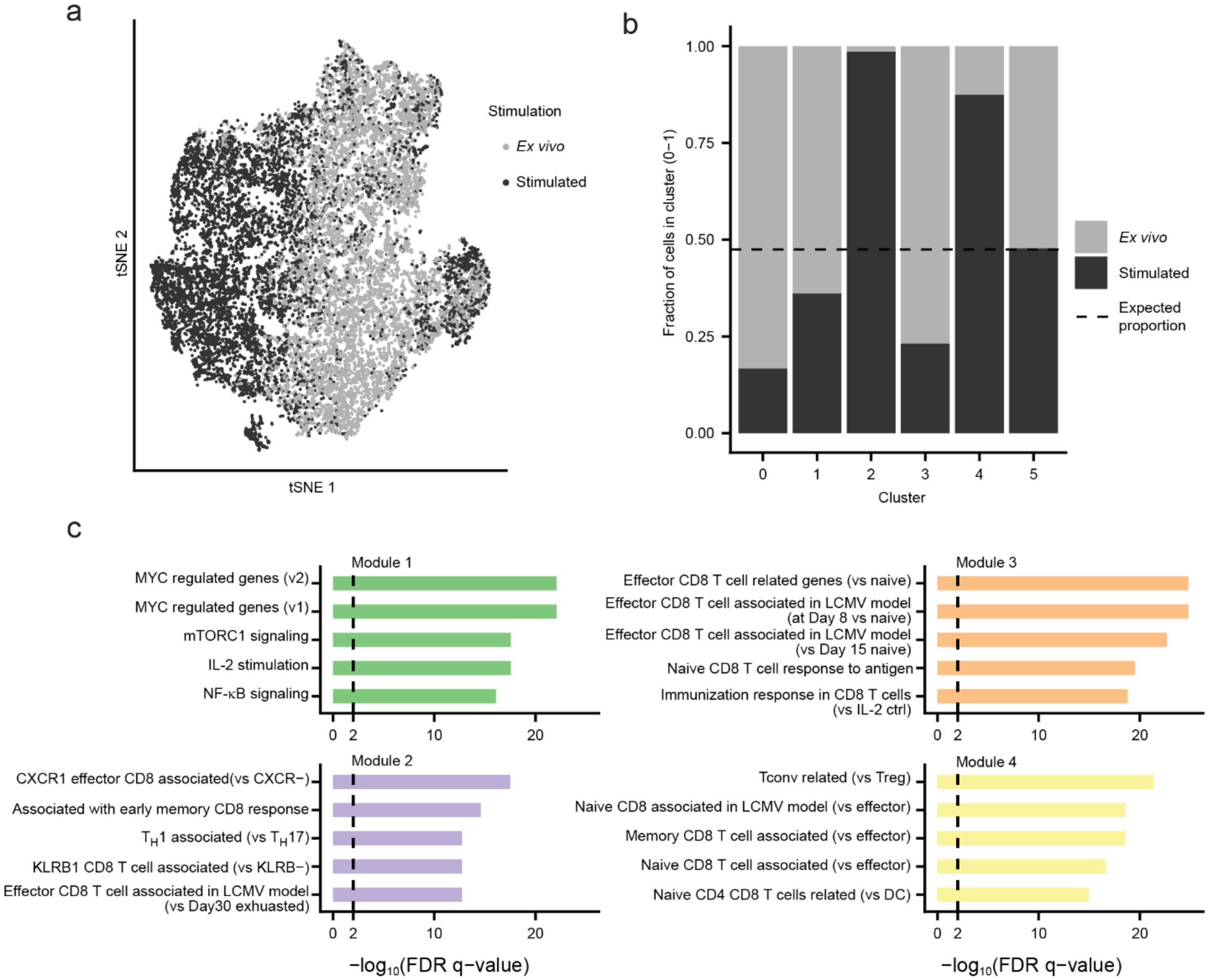


Figure 3-6. Stimulated and *ex vivo* cells are transcriptionally distinct. **a**, tSNE visualization of all cells colored based on stimulation condition (n = 6,912 stimulated cells, dark grey; 7,512 *ex vivo* cells, light grey). **b**, Proportions of stimulated and *ex vivo* cells in each of the computationally determined clusters shown in **Figure 3-5**. Dash line indicates expected proportions assuming even distributions of cells from both conditions. **c,** Enriched MsigDb signatures of the four modules of genes identified in **Figure 3-5**. FDR q-values represent Benjamini and Hochberg-corrected, one-tailed hypergeometric *P* values. 50, 49, 35, and 48 genes are included in Module 1, 2, 3, and 4, respectively for enrichment calculation. Data represent combined data from four independent experiments of four mice total (**a-c**).

3.3.5 Stimulated cells show clonotype-associated transcriptional profiles

To further investigate transcriptomic differences between the expanded clonotypes, we filtered the data to expanded clonotypes (detected in at least 15 cells) that were shared between the stimulated and the *ex vivo* groups (**Figure** 3-5). We also included 15 randomly sampled singletons (i.e. clonotypes that were only detected once in the dataset) from the naïve cluster (cluster 5) in each stimulation condition as a point of comparison. We examined the gene expression among these clonotypes *ex vivo* and after antigenic stimulation (**Figure** 3-5). We observed three groups of clonotypes associated with four modules of differentially expressed genes. Unsurprisingly, by comparing to annotated gene sets (MsigDb; **Methods**), we found that the sampled singletons (group 3) associated with a set of naïve and central memory T cell-related genes (e.g. *Sell, Ccr7*; Module 4) across both *ex vivo* and stimulated conditions[64]. We also observed another group (group 2) of clonotypes that strongly upregulated cell-cycle related genes, characterized by *Myc* and *Myc*-targeted genes (module 1). The last group of clonotypes (group 1) exhibited higher expression of canonical cytotoxic effector markers such as *Gzmb, Ccr2,* and *Ccr5* (module 3), but only moderately upregulated genes in module 1 upon stimulation[64] (**Figure** 3-5 and **Figure** 3-6). These observed signatures were also consistent with previously published signatures of effector CD8$^+$ T cells[55] (**Figure** A3-4).

While module 2 and module 3 were both associated with phenotypes of effector T cells, module 2 contained markers of cytokine signaling and interferon response such as *Irf7* and *Ifit1,* as opposed to the cytotoxic markers in module 3 (**Figure** 3-5). Module 2 was accordingly enriched in cytokine-mediated signaling signatures, while module 3 was enriched in cell motility signatures (**Figure** A3-5)**.** Module 3 was differentially expressed between group 1 and group 2 of clonotypes, but module 2 was upregulated directly *ex vivo* for both groups of clonotypes. Downregulation of module 2 upon stimulation may represent a transcriptional response to TCR-dependent activation[66]. We also note that group 1 clonotypes were also significantly more expanded than group 2 clonotypes, suggesting that the two groups of clonotypes may have experienced different levels of activation and expansion *in vivo* (*P* value < 0.001 by Mann-Whitney U test; **Figure** 3-7). Overall, the clonotypes within groups 1 and 2 responded similarly upon exposure to antigens. That is, similar genes were up- or down-regulated upon stimulation in both groups (**Figure** 3-7). The two groups of clonotypes differed, however, in the magnitude of transcriptional changes, particularly for genes in module 1 (*Myc*-related genes), and in the expression of genes in module 3 (cytotoxic-associated genes) (**Figure** 3-5; **Figure** 3-8). Together, these results highlight how our method can further reveal clonotype-specific transcriptional responses not delineated by scRNA-seq alone.

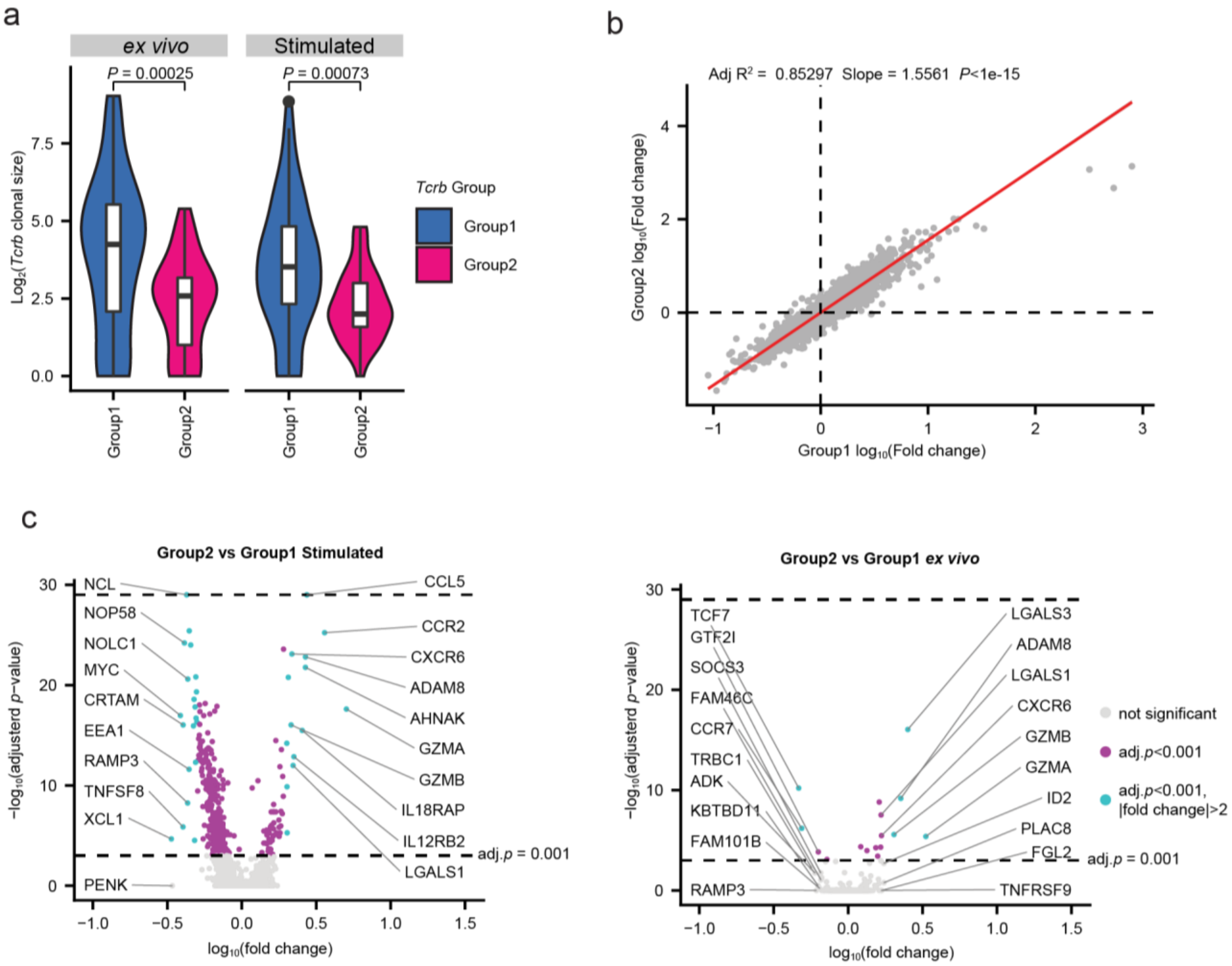


Figure 3-7. Group 1 and 2 clonotypes differ in expansion and gene expression upon stimulation. **a,** Clonal sizes of Group 1 and 2 clonotypes in the stimulated and *ex vivo* conditions shown in **Figure 3-5**. *P* value calculated by two-sample Mann-Whitney U test (Stimulated: n = 74 clonotype groups in Group 1; 37 Group 2 clonotypes. *ex vivo:* n = 82 Group 1 clonotypes; 40 Group 2 clonotypes). Box and whisker plots indicate the (box) 25th and 75th percentile along with (whisker) +/- 1.5*interquartile range. Violin plots represent estimated density of clonotypes. **b,** Gene expression fold changes between stimulated and *ex vivo* cells in (x-axis) Group 1 clonotypes and (y-axis) Group 2 clonotypes. Each point represents a shared gene across Group 1 and 2 clonotypes. Red line indicates fitted linear model. *P* value calculated by one-tailed *F* statistics (F(1,5908)) of the linear regression. n = 5908 genes. **c,** Volcano plots of differentially expressed genes between Group 1 and 2 clonotypes in the (left) stimulated and the (right) *ex vivo* conditions. *P* values were determined using a two-tailed likelihood ratio test, and adjusted by Bonferroni correction. Top 10 genes with positive or negative fold changes are labeled. Cells in Group 1 and 2 have been downsampled to 300 each (n = 300 cells for each of the groups).

3.3.6 Public clonotypes exhibit similar CDR3 sequences

We next investigated public clones that were shared among the four animals. We detected 76 unique *Tcrb* sequences shared in at least two of the four animals (**Figure** 3-8). We focused our analysis on the 21 clonotypes detected in at least three of the four mice. Amongst the public clones, we observed five sequences across mice that exhibited clear convergence, wherein only two amino acid residues (7th and 8th residues) varied across the CDR3 sequences (**Figure** 3-8). Among these, Leu-Gly account of 70% of cells, Ser/Ala/Gly-Gly for 20%, and a shortened Asp-only sequence the remaining 10%. Analysis of shared CDR3β sequences revealed variable pairing with CDR3α sequences both within and across mice, but several identical αβ pairings were observed in multiple animals (**Figure** 3-8). The most common CDR3β (CASSQDLGNYAEQFF) and its two distinct CDR3α partners (CAMREGLMATGGNNKLTF and CAVSNSGGSNYKLTF) were present in the same cells in three of the four animals (*n* = 225 cells; **Figure** 3-8 and **Table** A3-2). These data suggest that these cells may possess dual functional TCRα chains, a molecular feature that may play an important role in infection-induced autoimmunity[42,43,67].

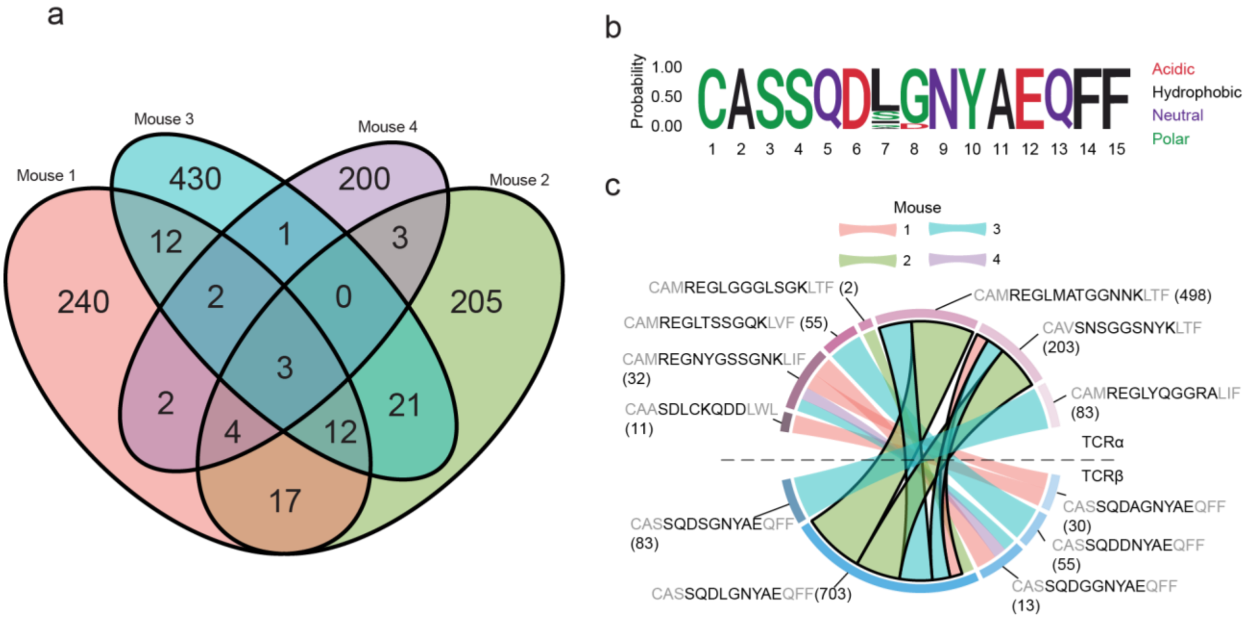


Figure 3-8. Analysis of shared clonotypes across four E7-HPV immunized mice. **a,** Venn diagram of shared unique *Tcrb* clones across the four mice. **b**, Amino acid logo plot of TCRβ sequences that show high similarity among public clones shared by at least three of four animals. Individual sequences are shown in **c. c**, TCRα and β matching of highly similar clonotypes found in public clones shared by at least three of four animals. Bold outline indicates dual TCRα chains found in the same cells (**Table A3-2**). Structural amino acids shown in grey. Number of cells shown in parenthesis. TCRα and β sequences that were detected in less than two cells were excluded for visualization.

3.3.7 Clonally expanded T cells detected in peanut-allergic patients

We next adapted the technique for use with human antigen-reactive CD4$^+$ T cells. Antigen-specific MHC-tetramers are often not available for human T cells, making identification of disease-relevant T cells difficult compared to standard inbred mouse models. Instead, it is common to use either proliferation or expression of proteins associated with antigen-dependent activation (e.g., CD154) as a proxy for response to disease-relevant antigens[68,69]. We applied our approach to profile T cells isolated from patients with peanut allergy—a type 1 hypersensitivity condition linked to dysregulation of CD4$^+$ T cells[70]. Peripheral blood mononuclear cells (PBMCs) from four patients were incubated overnight with peanut antigens and then sorted for CD154$^+$ expression to enrich antigen-activated cells[68,69] (**Figure** A3-6; **Methods**). Single-cell RNA-Seq was then performed on the sorted cells via Seq-Well, and the corresponding TCR sequences were recovered (**Figure** A3-6 and **Figure** A3-7). We note that there were differences observed in the gene expression of cells from different patients, even after controlling for technical sources of variation (i.e. sequencing depth, mitochondrial content; **Methods**). The differences observed were due in part to varied expression of a number of genes associated with basal cellular functions, including sex-linked genes (*XIST* and *RPS4Y1)* that were upregulated in cells from female and male subjects, respectively (**Figure** A3-7). Overall, 2,712 cells from four patients were included in the analysis. Contrary to what was observed for the tetramer-sorted mouse cells, the majority of the human CD154^{+} T cells were not clonal (mean = 75+/-12%). One of the patients (patient 77) exhibited a substantial expansion of T cells sharing common TCR sequences relative to the others (All patients: **Figure** A3-6 and **Figure** A3-7**;** patient 77: **Figure** 3-9). These clonally expanded T cells expressed genes associated with activation, such as *CD154*, *CD69* and *TNFRSF4*, as well as *GATA3*, a transcription factor associated with T$_{H}$2 cells[68,71]. In contrast, the non-expanded cells exhibited genes associated with central memory or naïve cells including *CCR7*, *SELL* and *LEF1*[72,73](**Figure** 3-9). Composite scores of known signatures of CD4^{+} T cell subtypes confirmed an enrichment of T$_{H}$2 signature in some of the expanded T cells[57] (**Figure** A3-7)**.**

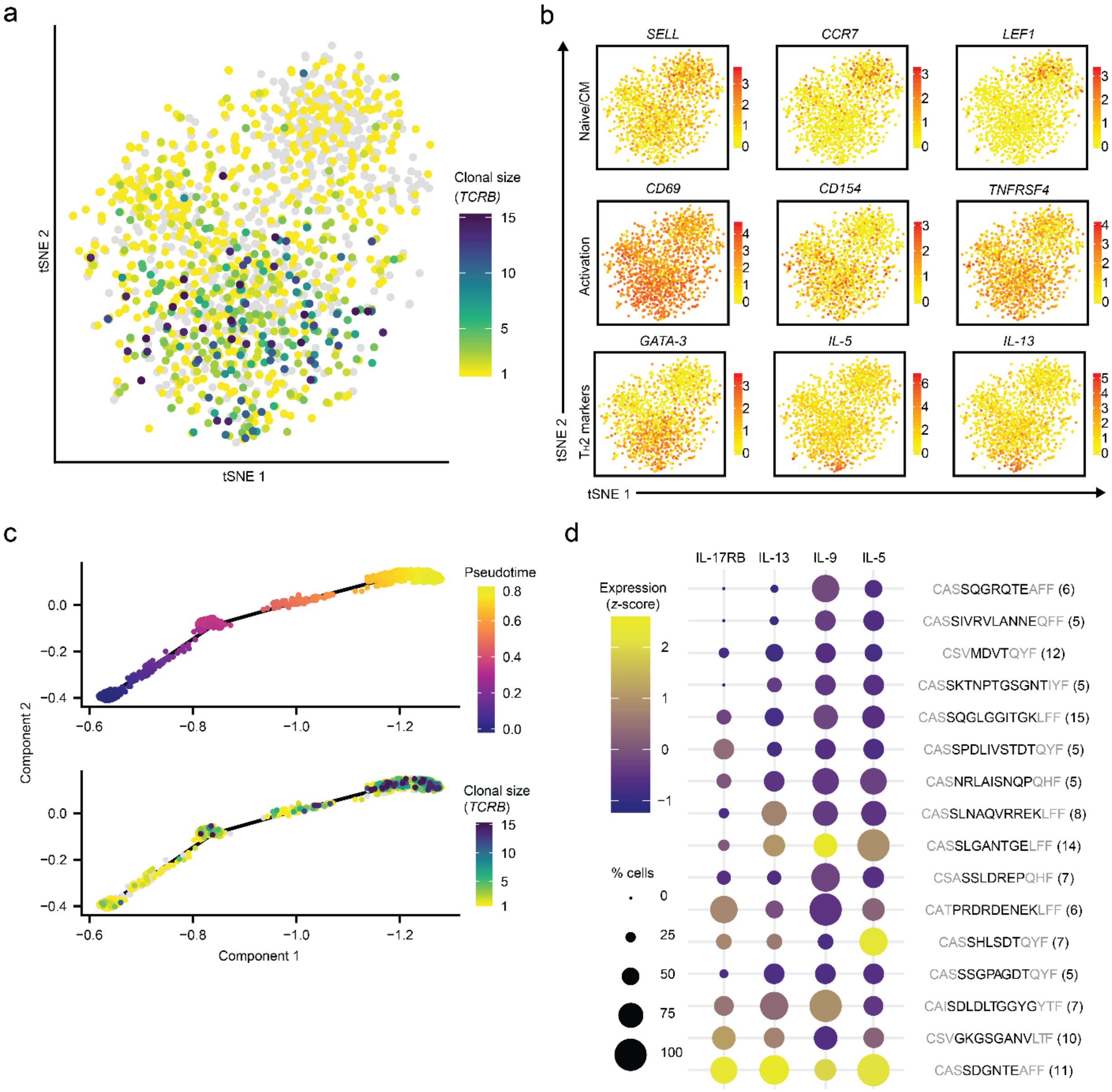


Figure 3-9. ScRNA-Seq and TCR analysis of peanut-dependent activated T cells from one of the peanut-allergic individuals (patient 77) combined with pseudotemporal analysis. **a**, Clonal size of *TCRB* clonotypes mapped onto tSNE visualization of transcriptomic data (n = 1,496 cells). **b**, Expression of canonical markers associated with naive/central memory, T cell activation, and T_H2 phenotypes. Color indicates log-normalized gene expression (yellow to red). **c**, (top) Pseudotime trajectory of scRNA-Seq results with (bottom) *TCRB* clonal size mapped. **d**, T_H2 pathogenic markers expression amongst expanded (n>=5 cells per clonotype, resulting in 16 total clonotypes included) *TCRB* clonotypes with high pseudotime value (mean > 0.4, see **Figure 3-11**). Gene expression represent averages within each clonotype group and scaled across all groups. Structural amino acids shown in grey. Number of cells shown in parenthesis. Color indicates scaled and log-normalized gene expression (purple to yellow).

3.3.8 Expanded clonotypes exhibit varied expression of Th2 genes

To examine whether these heterogeneous T cells from patient 77 might represent a spectrum of activation states, we next performed pseudotemporal analysis that showed a trajectory correlated with the degree of T cell stimulation, marked by increased expression of *JUN*, *FOS*, *NFKB* and *CD154*, among others[68,74] (**Figure** 3-10). Genes associated with early pseudotime were enriched in canonical markers for naïve T cells, while genes associated with late pseudotime were enriched in markers for effector T cells (**Figure** 3-10). The T cells most associated with activation on the trajectory were also the most clonally expanded ($r_s = 0.39$, $n = 851$, P value < 0.001 by Spearman; **Figure** 3-9). Further, our pseudotemporal trajectory correlated strongly with expression of *IL-5*, *IL-9*, *IL-13* and *IL-17RB*, known to encode markers of pathogenic T_H2 cells[70,71] (**Figure** 3-10, cluster 1). From these data, we posit that clonotypes that are both expanded and located towards the end of the trajectory may enumerate activated peanut-specific T cells. Among such cells, a subset of clonotypes exhibited T_H2 functional signatures (**Figure** 3-9). In particular, only one clonotype (CASSDGNTEAFF) had high expression for all four T_H2 markers, suggesting a robust polyfunctionality that may represent a highly differentiated T_H2-polarized clone involved in the allergic state of the individual[75,76]. Although the majority of expanded clonotypes were located at the end of the pseudotime trajectory, we noted some clonotypes showed higher variation in phenotypic states than others (**Figure** 3-11). It is possible that other factors may also contribute to the observed cell state and expansion of these T cells, such as transcriptional pulsing or bystander activation[77–79]. Taken together, our data suggest that our method can resolve differential degrees of antigen-dependent activation among clonotypes from human T cells and potentially highlight clonotypes among enriched pools of activated T cells that are most relevant to a disease state of interest.

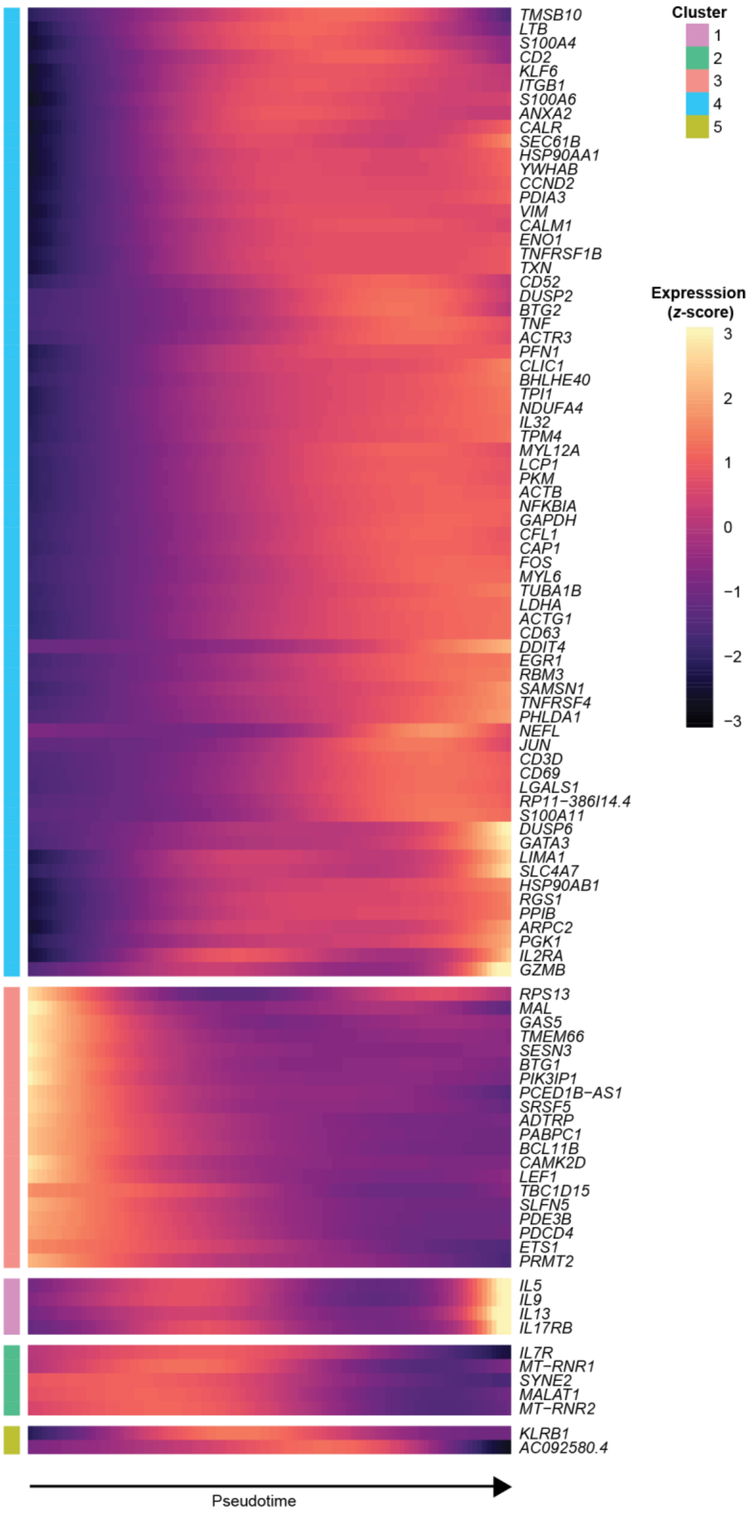


Figure 3-10. Distinct patterns of gene expression correlate with pseudotime. **a,** Expression of top 100 most significant genes visualized across pseudotime. Genes were clustered via Ward.D2 based on their patterns of expression. Data represent an individual experiment with 1847 single-cells from one patient (patient 77).

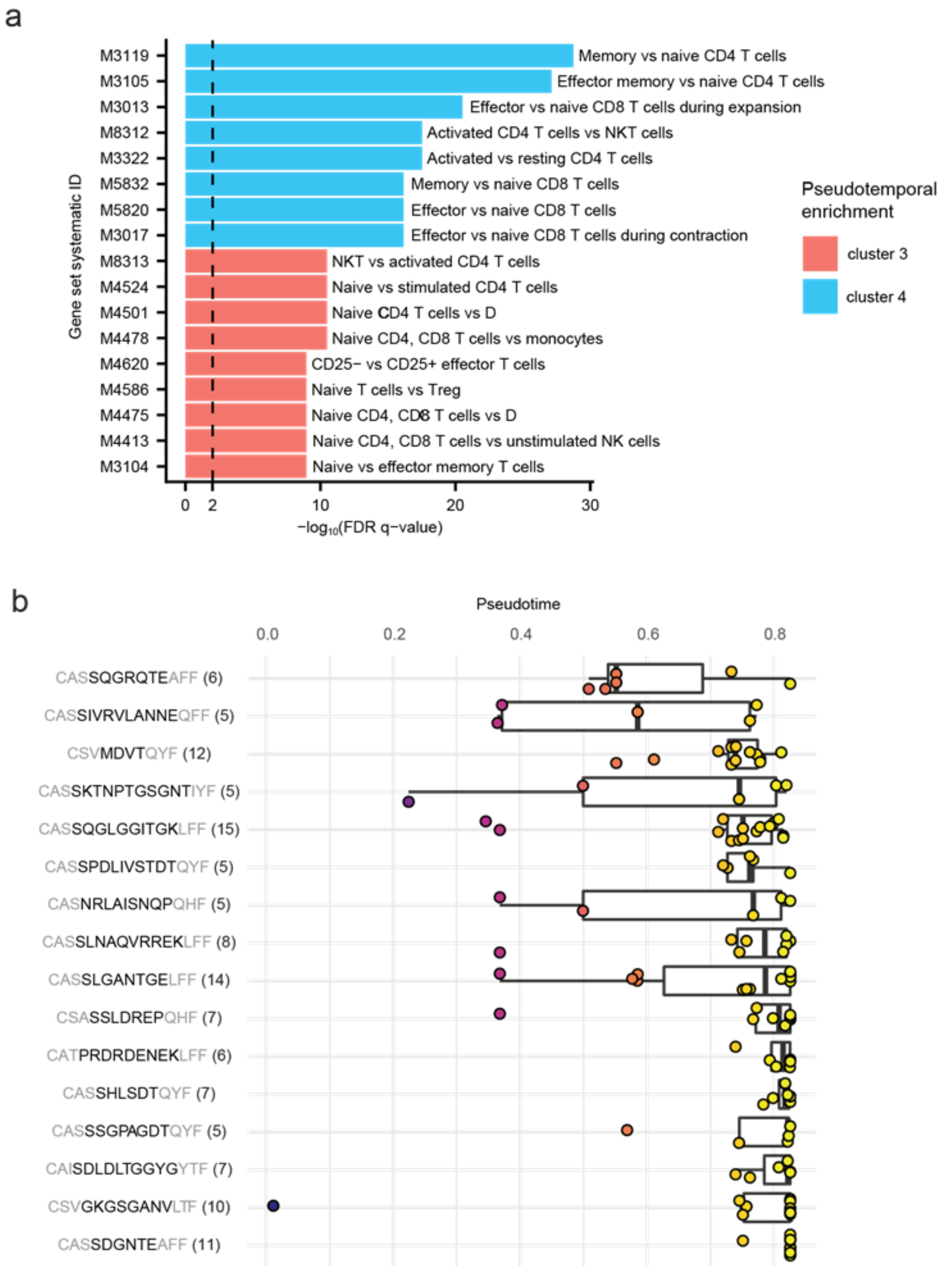


Figure 3-11. Psuedotime correlates with effector T cell signatures. **a**, MsigDB analysis of genes enriched early (cluster 3 in **Figure 3-10;** n = 38 genes) or late (cluster 4 in **Figure 3-10**; n = 123 genes) on the pseudotemporal trajectory. Description indicates cell state enriched with the corresponding gene set in comparison to another cell state. FDR q-values represent Benjamini and Hochberg-corrected, one-tailed hypergeometric *P* values. **b**, Pseudotime distribution of expanded clones shown in **Figure 3-9**. Number of cells for each clonotype group indicated in parenthesis. A total of 16 clonotype groups are shown. All box and whisker plots indicate the (box) 25th and 75th percentile along with (whisker) +/- 1.5*interquartile range (**b**). Data represent an individual experiment with 1847 single-cells from one patient (patient 77; **a**,**b**).

3.4 Discussion

In this chapter, we took the lessons we learned from the size selection method to develop a more reliable method for recovery of CDR3 sequences from 3’ single-cell barcoded libraries. We continued with the biotin enrichment of TCR sequences, and by incorporating a V region selection using multiplex primer sets for primer extension, but not amplification, we achieved consistent recovery of sequences without introducing noticeable biases.

The protocol also does not require overly precise pipetting, as the size selection methods often did. For the samples shown in this chapter, we routinely used multichannel pipettes, allowing us to process up to 96 samples at once. This feature will become important in later chapters, as studies become larger and necessitate more samples.

The modified method also presents significant computational advantages. Previously, because the sequencing libraries were randomly tagmented, each read would be randomly mapped onto the TCR genes. As such, many of the reads would not cover the junctional regions within the CDR3 (i.e. the junctions between the different gene segments), thus requiring a much higher total number of sequencing reads to recover the CDR3 sequences from each cell. While we attempted several different computational packages to reconstruct consensus sequences from the data, many of those were not successful. In the modified method, because each read is specifically targeted to cover the junctions between the V, D, and J genes, we can directly derive the junctional sequences without *de novo* assembly, resulting in a higher yield of CDR3 sequences.

Since this method relies on a universal primer element for PCR amplification (UPS2 and a modified UPS), the method simplifies adaptation to other model species, such as non-human primates, by minimizing the need for additional optimization with new multiplex pools of primers. We also foresee that our general approach can also be used for other targets where isoform information is paramount, such as identification of *CD45RA* or *CD45RO* in T cells[80].

Nonetheless, there are currently two practical limitations. First, the innate efficiency of the mRNA capturing reagents (i.e. 3′ barcoded beads for Seq-Well and Drop-Seq) limits the maximum potential representation (and thus recovery) of TCRs in libraries. Currently, for Drop-Seq, it is estimated that the barcoded beads capture 10-12% of available transcripts[44]. Improved quality of these reagents or the molecular biology used to generate the WTA products will likely address this limitation[29]. Second, due to the incorporation of V region primers, the method presently does not recover full-length TCR transcripts, and therefore our analysis is restricted to the CDR3 region. The CDR3 region contains the majority of variability in TCR transcripts, and therefore, is likely sufficient for assessing clonal expansion and clonal tracking analysis[60]. Further, the design of the V region primers depends on adequate annotation of V genes, which may not be available for less characterized model species.

In summary, we have developed a simple approach to recover TCR CDR3 sequences from whole transcriptome libraries produced by common high-throughput scRNA-seq techniques that rely on 3′ barcoding of transcripts. We found the technique reliable and also high yielding, limited predominantly by the efficiency of initial capture of mRNA. Our approach can map murine and human antigen-reactive T cells, and in principle, is extensible to other species (e.g., non-human primates) and target genes of interest (e.g., viral antigens, isoforms, germline and somatic variations of B cell receptors). Extension of this technique should be feasible so long as the targeted variable regions are flanked by suitable known sequences. Overall, our data demonstrate that enhancing the resolution of these populations of cells by the combined recovery of TCRs and scRNA-seq can further reveal phenotypic variations that emerge as a function of clonotype, and reveal convergent public clones with precision. We anticipate that our method will be especially useful for elucidating the intrinsic heterogeneity among antigen-specific T cells and their roles in immunological diseases, such as cancer, autoimmune disorders, and food allergy.

4. Application of Seq-Well and TCR recovery to the study of peanut oral immunotherapy

This chapter is in part adapted from: B. Monian, A.A. Tu, B. Ruiter, et al, *in prep.*

This chapter details the application of TCR recovery to further study peanut food allergy in a cohort of 12 patients undergoing treatment. Food allergy is a hypersensitivity condition with increasing prevalence globally, especially in developed countries. While some of the relevant cellular components have been identified, the mechanism of tolerance is still largely unknown. Due to the rarity of peanut-specific cells and the multitude of helper cells involved, peanut food allergy is apt for single-cell RNA and TCR sequencing.

We uncovered a wide variety of T helper cell functions. Some only represented an extremely minor fraction of reactive T cells, and were therefore unlikely to have been detected in bulk measurements. Fortuitously, this study coincided with not only the development of the final TCR recovery method, but also an improved version of Seq-Well that dramatically increased the numbers of genes and transcripts per cell. As a result, we believe this is one of the most comprehensive views of not just peanut-reactive, but all T helper cells in general to date.

4.1 Motivation

4.1.1 Immunotherapy is effective for some, but not all, peanut-allergic patients

Food allergy is an immune hypersensitivity condition that affects an estimated 8% of children in the US, with increasing severity and global prevalence[81]. Allergic reaction could be directed to a variety of allergens, including milk, shellfish, and tree nuts[82]. The condition is characterized by the presence of allergen-specific Th2 cells, which in turn mediate the production of allergen-specific IgE antibodies[83]. The antibodies prime effector cells, such as mast cells, basophils, and eosinophils, through FcεRI receptors that can be cross-linked in the presence of allergen. The resulting cellular degranulation leads to systemic release of histamine and other mediators[81]. Symptoms of the allergic reaction can range from mild to life-threatening (i.e. anaphylaxis).

At the time of writing, oral immunotherapy (OIT) is an emerging treatment for food allergy that was recently approved[84]. OIT involves daily ingestion of allergen wherein the dose gradually increases over time to promote clinical tolerance. The efficacy of OIT is variable: 80-85% of patients achieve desensitization (a loss in clinical reactivity with regular consumption of allergen), but most do not maintain unresponsiveness after treatment[85] (i.e. without continued allergen consumption).

Nevertheless, OIT has been proven effective in inducing at least temporary tolerance, and even sustained tolerance in some patients. Though the mechanism of the therapy is still unclear, studies have characterized some aspects of the immune response to OIT. The prevalence of circulating allergen-specific Th2 cells and their expression of Th2 cytokines, may decrease or be suppressed by anergic gene programs, and patients who achieve sustained tolerance may have higher frequencies of regulatory T cells (Tregs) post-treatment[86,87]. OIT may also result in suppression of other cell types, such as basophils and eosinophils[88]. Th17 and Th1 responses, whether measured in expression of the respective cytokines or in cell numbers, have also been linked to allergic status, suggesting roles of non-Th2 response in food allergy[89].

Beyond the canonical markers of Th2 cells, including GATA3, IL-4, IL-5, and IL-13, distinct subsets of Th2 cells have also been reported[90]. Wambre et al. reported a subset of Th2 cells, termed Th2A, that correlated with allergic conditions[91]. In addition to the canonical cytokines, Th2A cells also exhibited increased expression of *KLRB1* and *IL17RB*. Similarly, Mitson-Salazar et al. reported a pathogenic subset of TH2 cells, called peTh2 cells, that also similarly associated with expression of *IL17RB* as well as *PTGDS*[92]. Gawtham et al. proposed a model of different T follicular helper cells (Tfh cells) that helps explain class switching of B cells in response to foreign antigens. The group proposed that Tfh13, marked by expression of CXCR5 in addition to Th2 cytokines, may promote food allergen specific IgE antibodies, leading to allergic outcomes[93].

Despite the advances in the understanding of allergen-reactive T cells, most studies of the immunological responses induced by peanut OIT have relied predominantly on population-level measurements of T cells and other allergic effectors. As a result, past studies have relied on isolation of specific cell populations based on *a priori* knowledge, limiting the breadth of the studies. Similarly, study of TCR repertoire in OIT have been limited to measuring bulk changes, limiting further association of specific clonotypes to specific phenotypic response.

In the work presented in this chapter, we used scRNA-seq to study peanut-reactive T cells from patients undergoing OIT. We identified subsets of Th2 cells that shared markers with previously identified subsets in prior studies. By incorporating TCR recovery, we were also able to track common T cell lineages within patients across multiple timepoints.

4.2 Methods

Patients. Peanut-allergic individuals aged 7 and up were enrolled in a peanut OIT trial (NCT01750879) at the Food Allergy Center at Massachusetts General Hospital. All subjects were recruited with informed consent, and the study was approved by the Institutional Review Board of Partners Healthcare (protocol 2012P002153). Subjects were first screened for a diagnosis of peanut allergy by medical history, evidence of peanut-specific IgE per skin prick test (reaction wheal ≥5mm larger than saline) or serum peanut-specific IgE titer (≥5 kU/L), and Ara h 2-specific serum IgE > 0.35 kU/L. Subjects then underwent a double-blind, placebo-controlled food challenge (DBPCFC) up to a maximum dose of 443 mg of peanut protein. Patients who reacted during the challenge, and had passed the prior screening, were eligible for inclusion in the study.

Oral immunotherapy (OIT) study. The main objective of this phase I/II, double-blind placebo-controlled, interventional study was to provide additional safety and mechanistic data on OIT for people with IgE-mediated peanut allergy. Enrolled patients were randomized to receive either treatment (peanut flour) or placebo (roasted oat flour) at a ratio of 3:1. Treatment consisted of a modified-rush protocol, followed by a build-up phase lasting for 44 weeks or when the patient reached 4000mg, whichever came first. Treatment dose was administered daily, and dosing escalation was incremental (based on previous OIT studies), occurring every two weeks. After the buildup phase, patients entered a maintenance phase in which treatment was continued at the top tolerated dose for each patient for 12 weeks. Finally, patients underwent an avoidance phase, an additional 12 weeks off therapy while strictly avoiding dietary peanut protein, in order to assess the durability of any desensitization resulting from OIT. During each phase of the study, a blood sample was taken, for four samples total per patient: two weeks prior to the start of treatment at baseline, fourteen weeks into the buildup phase, eight weeks into the maintenance phase, and eight weeks into the avoidance phase.

Clinical assessments were made by double-blind placebo-controlled food challenge at baseline (DBPCFC1), at the end of 12 weeks of maintenance therapy (DBPCFC2), and at the end of 12 weeks of avoidance (DBPCFC3). Clinical outcomes were defined as: 1) treatment failure (failure to achieve the minimum maintenance dose (600 mg) of peanut protein by 12 months, or an eliciting dose less than 1443 mg at DBPCFC2, or less than 443mg at DBPCFC3, OR less than 10-fold more than at DBPCFC1), 2) partial tolerance (eliciting dose less than 4430mg at DBPCFC3 but at least 430 mg AND more than 10-fold more than at DBPCFC1), and 3) tolerance (ingestion of 4430 mg of peanut protein at DBPCFC3 without symptoms).

Cell purification and sorting. After a blood sample was collected, PBMCs were immediately isolated by density gradient centrifugation (Ficoll-Paque Plus; GE Healthcare) and cryopreserved in FBS with 10% DMSO. After the study was completed, PBMCs from a patient at all four time points (15-30 x 10^6 PBMCs per timepoint) were simultaneously thawed, washed with PBS, and cultured in AIM-V medium (Gibco) with 100 µg/ml peanut protein extract for 20h, at a density of 5 x 10^6 PBMCs in 1 mL medium per well in 24-well plates. (Peanut protein extract was prepared by agitation of defatted peanut flour with PBS, centrifugation, and sterile-filtering.) Anti-CD154-PE antibody (BD Biosciences; clone TRAP1) was added to the cultures at a 1:50 dilution for the last 3h. After harvesting, cells were labeled with anti-CD3-AF700 (BD Biosciences; UCHT1), anti-CD4-APC-Cy7 (BD Biosciences; RPA-T4), anti-CD45RA-PE-Cy7 (BD Biosciences; HI100), anti-CD154-PE (BD Biosciences; TRAP1), anti-CD137-APC (BD Biosciences; clone 4B4-1), and Live/Dead Fixable Blue stain (Thermo Fisher; cat. no. L23105). Cells were then sorted on a FACSAria Fusion instrument (BD Biosciences). Cells were gated as live CD3+CD4+CD45RA- and sorted as either CD154+CD137+/- (referred to as “CD154+”), CD154-CD137+ (“CD137+”), or CD154-CD137- (referred to as “DblNeg”).

Gene module discovery. Coexpressed gene modules were generated based on a sparse PCA approach described by Witten et al and implemented in the R package “PMA”[94]. This method employs an L1 norm penalty to reduce and eliminate gene loadings that do not materially contribute to each component. Prior to running sparse PCA, the gene expression matrix was randomly downsampled to have an equal number of cells from the top 70 (out of 109) samples, in order to limit bias caused by unequal cell numbers and to decrease computational time. Genes were filtered to the union of immune genes (as defined by the sets of gene lists available on ImmPort) and the variable genes in the dataset using the ‘var.genes’ command in the R package “Seurat”[95]. Finally, the gene expression data was scaled with respect to genes, and sparse PCA was run using the command “SPC” (with “orth” parameter set to TRUE and tuning parameter “sumabsv” set to 1.8). Gene module scores were calculated as the scaled gene expression input matrix multiplied by the outputted loadings matrix “v”.

Cells were deemed to “express” a module (that is, “positive” for a module) using a gating strategy similar to flow cytometry gating. Module scores of CD154-CD137- cells were used as a negative control, and a gate was set such that no more than 0.1% of CD154-CD137- cells were in the positive population.

Each gene module was analyzed for contributions from the individual patients. Proportions of cells that scored positive for a module from each of the patients were tabulated. Modules with over 65% of cells from a single patient were removed from subsequent analysis (**Figure** A4-1).

Distance analysis of TCR sequences. Pairwise similarity of TCRβ CDR3 sequences was evaluated using an adapted version of the TCRdist method published by Dash et al[63]. In brief, for two TCRβ CDR3 amino acid sequences of the same length, each residue position was compared and a penalty was assessed for every mismatch. The penalty for two different amino acid residues i and j was assessed using the BLOSUM62 matrix and was defined as min(4 – BLOSUM62[i, j], 4). Each substitution thus incurred a penalty between 1 and 4. The overall distance between two CDR3s was calculated as the sum of penalties at all positions. In the case of two CDR3s of unequal length, the sequences were aligned in all possible ways and the minimum overall penalty was taken, with each gap incurring a penalty of 8. In this way, a pairwise distance matrix for all CDR3 sequences was generated. To accrue sufficient numbers for comparison, close CDR3 pairs were binned according to the following distances: 1-4, 5-8, 9-12, 13-16, 17-20, and 20 or more.

Probability-based association between TCR and gene expression. Probability-based analysis was used to determine the degree of association between a categorical transcriptional feature (such as cluster or status of gene module expression) and a TCRβ CDR3 sequence. A likelihood ratio of association was defined as P/P_0, where P was the probability of two cells, drawn randomly without replacement from all cells sharing the same TCRβ CDR3 sequence (in the case of TCRdist = 0) or a defined pair of sequences (with a defined TCRdist > 0), both expressing a gene module or belonging to the same cluster of cells. The probability is normalized by P_0, the probability of two cells, drawn randomly from all cells, both expressing the module or belonging to the same cluster. P_0 represents the prior probability without the constraint of TCRβ information; thus, the ratio P/P_0 represents the gain in probability due to the knowledge of TCR sequence. A ratio of 1 represents random co-occurrence of TCR sequence and the transcriptional feature, while a ratio of 2 represents a two-fold increase in the likelihood of shared transcriptional features given the same TCRb sequence.

scRNA-seq and TCRαβ recovery. Sorted subsets of CD4 memory T cells were processed for scRNA-seq and TCR recovery as previously described in Chapter 3.

Visualization and clustering of single-cell RNA-Seq data. Visualization and clustering were done with the Python package “scanpy.” Prior to visualization, the normalized gene expression data was transformed using a standard “regress-out” approach to mitigate batch effects. A multiple linear regression was performed on all genes with two covariates that could be batch-associated: numbers of transcripts per cell, and percent of transcripts aligning to the mitochondrial chromosome. The residuals from this regression were taken as the transformed data.

Next, a principal components analysis was performed, and the top 10 components were used to generate a visualization with UMAP[96] (uniform manifold approximation and projection). Clustering was performed on the top 10 principal components using the Louvain graph-clustering method.

Clustering and clonotype assignment of Th1, Th2, and Th17 subsets. Cells that were positive for the Th1 (module 38), Th2 (module 7), and Th17 (module 9) modules were clustered via UMAP separately. In this scheme, cells that were positive for multiple modules were included in the analyses of all relevant modules. For **Figure** 4-4, clonotypes were each assigned to the subset with the highest proportion of cells of the clonotype.

Correlation of gene modules to clinical outcome. An average score was calculated for each gene module in the CD154+ compartment of each patient in the treatment group. The correlation of each module to clinical outcome was calculated by Spearman's correlation. P values were adjusted by Bonferroni correction.

4.3 Results

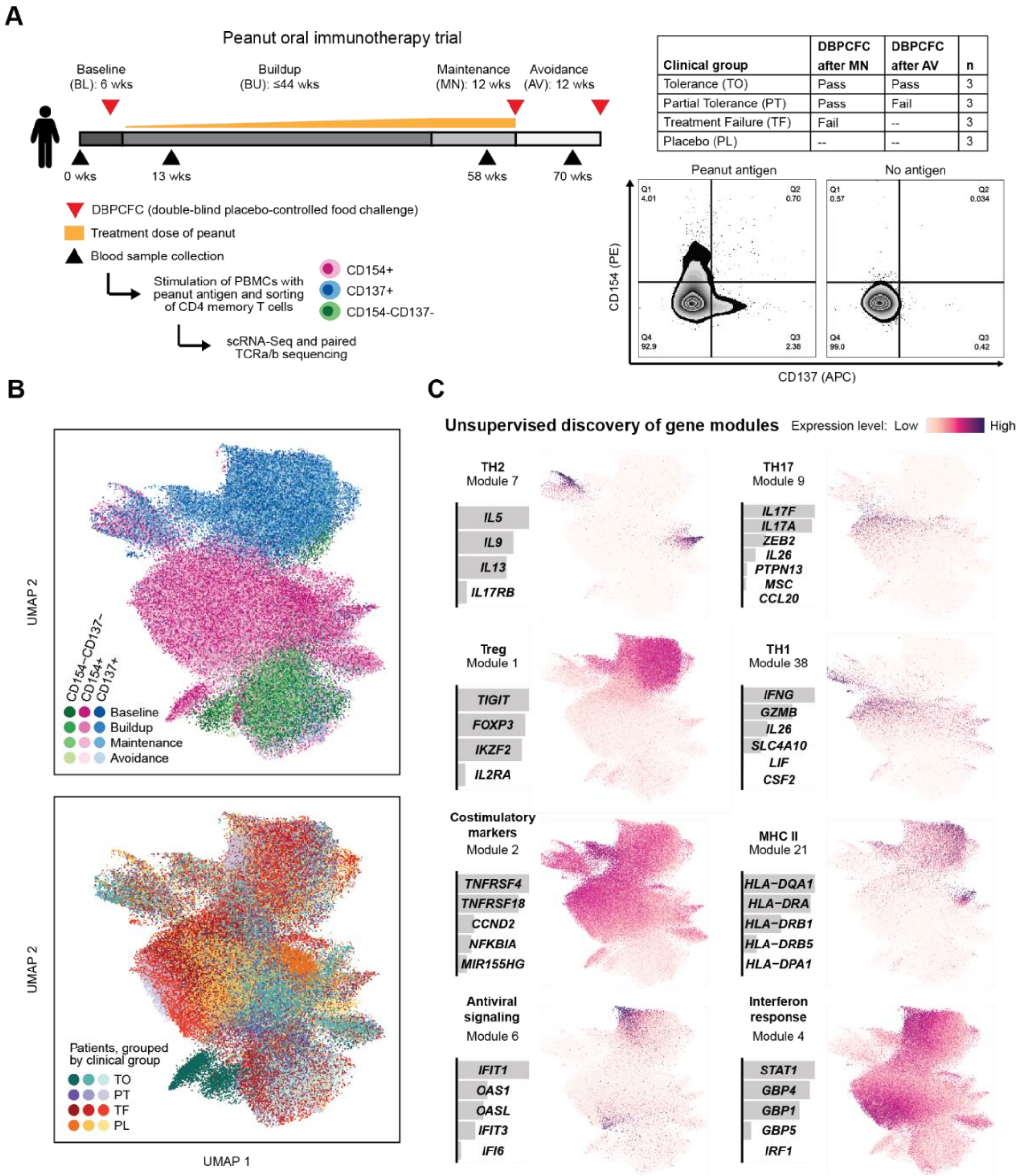
4.3.1 Single-cell transcriptomic landscape of the patients undergoing peanut OIT

Despite the higher throughputs of Seq-Well and other scRNA-seq platforms, studying peanut-specific T cells directly from patient still presents several challenges. Firstly, the number of peanut-specific T cells is expected to be low. This due to the relatively small sample size of T cells derived from PBMCs of any single blood sample, compared to the total number of T cells within an individual. The sampling issue is compounded by the fact that patients undergoing OIT have most often been under peanut-avoidance for extended periods of time. As such, peanut-reactive T cells are not expected to have gone through recent clonal expansion or activation.

Therefore, to enrich for peanut-reactive T cells in our study, we used a peanut activation assay to identify cells expression activation markers CD154 and CD137 (**Figure** 4-1). PBMCs collected from four time points during the trial were cultured with peanut extract, and peanut-activated CD4 memory T cells were enriched via FACS (**Figure** 4-1). The sorted cells were then processed for scRNA-seq and TCR recovery. We elected this strategy to not restrict our study only to cells reactive to known antigens. Peanut reactive cells were rare, consisting only less than 5% of memory CD4 cells (as defined by CD154 expression by FACS).

A UMAP representation of the transcriptomic data show that cells largely clustered with the sorted subsets (**Figure** 4-1). CD154+ and CD137+ cells showed distinct transcriptional states, with top differentially expressed genes including *CD40LG* in CD154+ cells and *TNFRSF9* as well as the regulatory markers *FOXP3* and *TIGIT* in CD137+ cells (**Figure** 4-2). Despite normalizing for technical factors such as library size and expression of mitochondrial genes, we also observed patient-dependent variation within each cluster (**Figure** 4-1). Many patient-related differences could be attributed to features such as sex-linked genes and baseline inflammation, suggesting that these features represent inherent biological differences rather than batch effects. An analysis of the patient contribution in each of the modules also confirmed that some of patient-specific clusters were driven by genes specific to a subset of patients (**Figure** A4-1). Overall, there was no association of time points with broad transcriptional states, indicating that OIT-induced effects were likely to be subtle.

To observe the finer-grain subsets among peanut-reactive T cells, we developed an unsupervised approach to discover immune-related gene expression programs. The dataset was filtered to 1,500 immune-related or variable genes. Then, co-expressed genes were aggregated into gene modules using an implementation of sparse principal components analysis (PCA) to derive a set of 50 gene modules, each containing 4-10 genes. Known relevant functional states of T cells were recovered as individual modules, such as Th1, Th2, and Th17. (For more details about generation of the modules, please see thesis work of Brinda Monian.)



Clinical group	DBPCFC after MN	DBPCFC after AV	n
Tolerance (TO)	Pass	Pass	3
Partial Tolerance (PT)	Pass	Fail	3
Treatment Failure (TF)	Fail	--	3
Placebo (PL)	--	--	3

Figure 4-1. Peanut-reactive T cells from individuals undergoing oral immunotherapy have diverse transcriptional signatures**.** **a**, OIT study design, definition of outcomes, and experimental workflow. CD3+CD4+CD45RA- memory T cells were further sorted as CD154+CD137+/-, CD154-CD137+, or CD154-CD137-. **b**, Two-dimensional UMAP visualization of all single-cell transcriptomes, colored by sorted subset and time point (top) or by patient and clinical group (bottom). Data represent 134,129 total cells (74,646 CD154+, 41,186 CD137+, and 18,297 CD154-CD137-) **c**, Selected gene modules discovered from the data using sparse principal components analysis overlaid on UMAP from **b**. For each module, a putative name, the weights of each contributing gene, and an overlay of module score on the UMAP coordinates are shown.

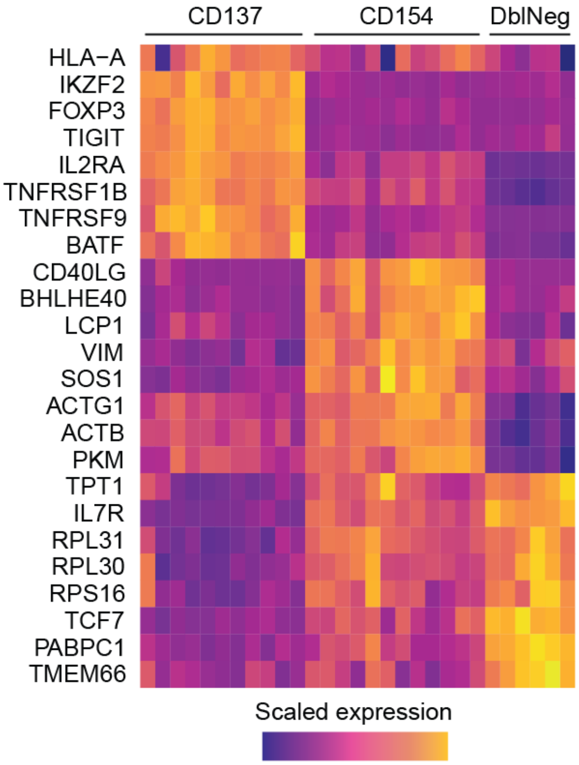


Figure 4-2. Top differentiating genes of the three sorted subsets of memory T cells. CD154 (*CD40LG*) and CD137 (*TNFRSF9*) are mainly differentiated by transcripts of their respective surface protein markers.

4.3.2 TCR recovery from peanut OIT samples

Next, to investigate clonal T cell responses to peanut antigens, we recovered paired TCR sequences for each cell. We recovered paired TCRβ sequences for 60% (sd = +/-17%) of cells, TCRα for 55% (+/-15%) of cells, and both chains for 36% (+/- 12%) of cells for each patient. Recovery was uniform across samples, and expanded clones largely localized within certain areas of the UMAP, suggesting an association between expansion and transcriptional state (**Figure** 4-3).

The vast majority of expanded TCRβ sequences were paired with a single TCRα (**Figure** A4-2). As a result, we used TCRβ for all downstream analyses involving clonotypes. The diversities of CD154+ and CD137+ repertoires were significantly less than that of the CD154-

CD137- (double negative) cells, suggesting that these markers enriched for a pool of expanded, antigen-specific clonotypes (**Figure** 4-3).

Across the three sorted subsets, we detected clear distinction of clonotypes between CD154 and CD137 population, suggesting that the two subsets represent distinct clonal lineages of peanut-reactive T cells. Further, we noted similar, though much less distinct separation of clonotypes between CD154 and double negative populations (**Figure** A4-3). We believe this result is expected, considering that CD154 has been noted as an early and transient activation marker. It is likely that cells of identical clonotypes would not expression the marker at uniformly the same time, or even at the same level. Therefore, our CD154 enrichment may only capture some fraction of cells of a particular clonotype. Taken together, we interpret this to mean that we were successful in enriching for peanut-reactive cells based on the two activation markers.

To determine how the gene modules associated with clonal T cell expansion, we first assigned cells as positive or negative for each module based on their respective module scores (**Methods**). We then calculated the average TCRβ clonal size for cells associated with each module as well as the average expression of that module in the CD154+ cells relative to the double negative cells (**Figure** 4-3; Methods). We found that modules representing Th1, Th2, and Th17 functions exhibited strong upregulation in the CD154+ compartment and were associated with expanded T cell clonotypes, suggesting that these modules were associated with expanded, peanut-responsive T cells. As expected, modules unrelated to T cell effector functions were not associated with upregulation in the CD154+ compartment. Of these, modules representing general T cell activation and survival were also associated with clonal expansion (such as module 6, which included interferon response-related genes, and module 26, which included *PDE4A,* a T cell activation marker).

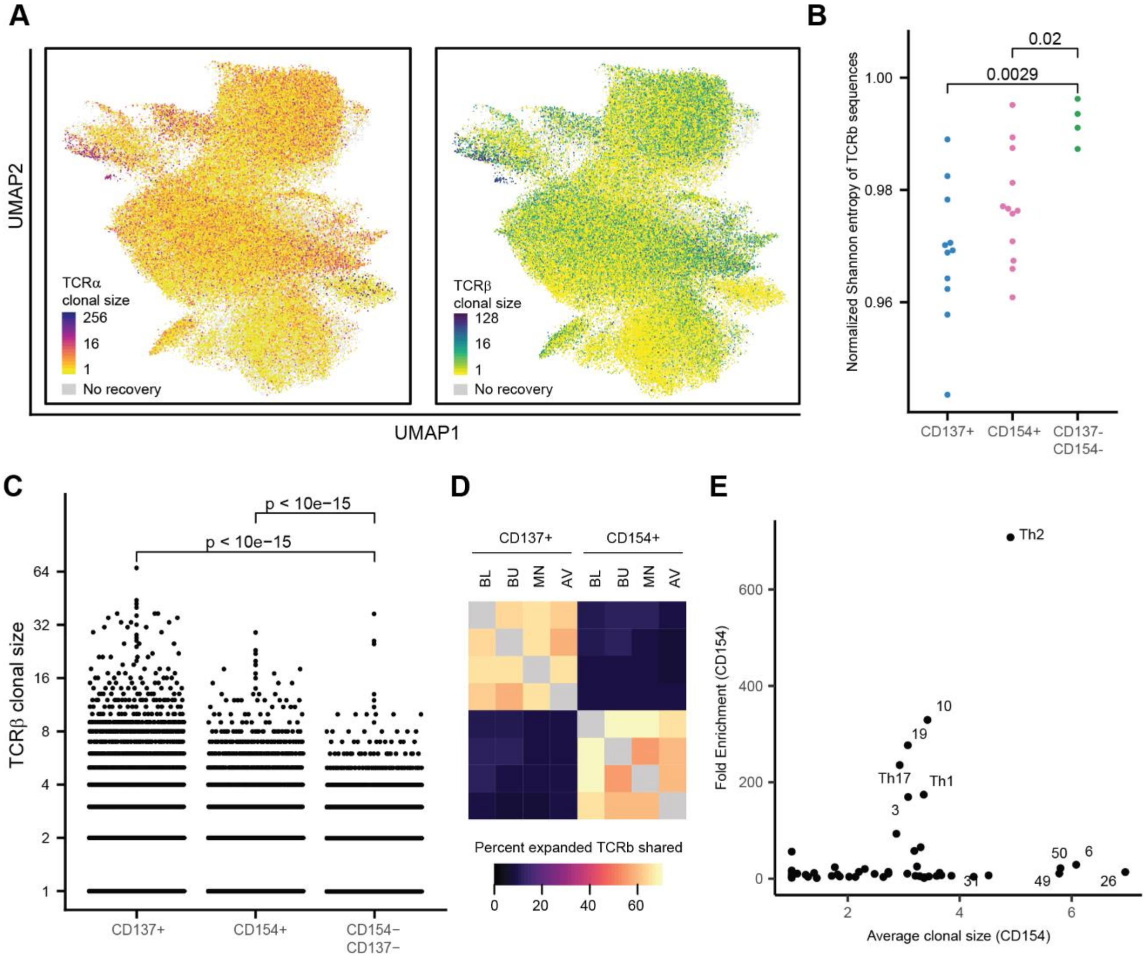


Figure 4-3. Functional gene modules are associated with clonal expansion and enriched expression in activated cells. **a**, Clonal size of TCRα sequence (left) or TCRβ sequence (right) overlaid onto the UMAP. Clonal size is defined as the number of cells sharing a TCR sequence. **b**, Diversity (Shannon index) of TCR repertoire by sorted subset. Each data point represents the repertoire from one patient at all time points. Comparisons between subsets are annotated with p-values from Wilcox rank-sum test. **c**, Distribution of TCRβ clonal sizes (defined as the number of cells sharing a TCR sequence), within each sorted subset. Comparisons between subsets are annotated with p-values from Wilcox rank-sum test. **d**, Heatmap of the percentage of TCRβ sequences shared between time points and activated subsets. 'Percent shared' is defined as the number of unique TCRβ sequences detected in both conditions, divided by the geometric mean of the number of unique TCRβ sequences in each of the two conditions. Sequences from all treatment-group patients were pooled. **e**, Average clonal size and CD154 enrichment in expression above CD154-CD137- cells for every gene module. Clonal size was calculated with respect to all cells and then averaged for those cells expressing the module. Expression enrichment is defined as the average expression of CD154+ cells, divided by the average expression in CD154-CD137- cells (**Methods**).

4.3.3 T helper cells comprise six clonally distinct subtypes

Due to their strong enrichment in the CD154 compartment and known contribution to food allergy, we further analyzed the heterogeneity among cells expressing the Th1, Th2, and Th17 modules. We found three distinct clusters of Th2 cells, two clusters of Th1 cells, and only one cluster of Th17 cells. Our results suggest heterogeneity within the Th1 and Th2, but not Th17, cells (**Figure** 4-4).

The three clusters of Th2 cells corresponded to a Tfh2-like population (high in *CXCR5* and *PDCD1*), a Treg-like state (*FOXP3* and *TNFRSF9*), and a Th2A-like population6 (*GATA3*, *IL17RB*, and *PTGDR2*) (**Figure** 4-4). The Tfh2-like population showed similarity to a previously-described pathogenic Tfh13 subset, while the Th2A-like-high population shares markers previously identified in Th2A and peTh2 populations[91,93].

Within the Th1 cells, the clusters corresponded to a Tfh1-like population and another larger population with canonical Th1 signatures (**Figure** 4-4). Both of these clusters expressed high levels of *IFNG* and *GZMB*, and the Tfh1-like cluster exhibited high overlap of upregulated genes with the Tfh2-like population, including *ICOS*, *PDCD1* and *TNFRSF9*. Unlike the Th1 and Th2 cells, we did not observe distinct subsets within the Th17 population, which is marked by expression of *IL17A* and *IL17F*.

Next, we examined how highly expanded TCR clonotypes were distributed between these six subsets. We found that most clones were primarily associated with a single subset (**Figure** 4-4), suggesting that these subsets represent distinct clonal lineages. We also observed a pattern of overlapping clones between the Th1 and Th17 states. This is in accordance to previous studies that have reported overlap of Th1 and Th17 phenotypes[97].

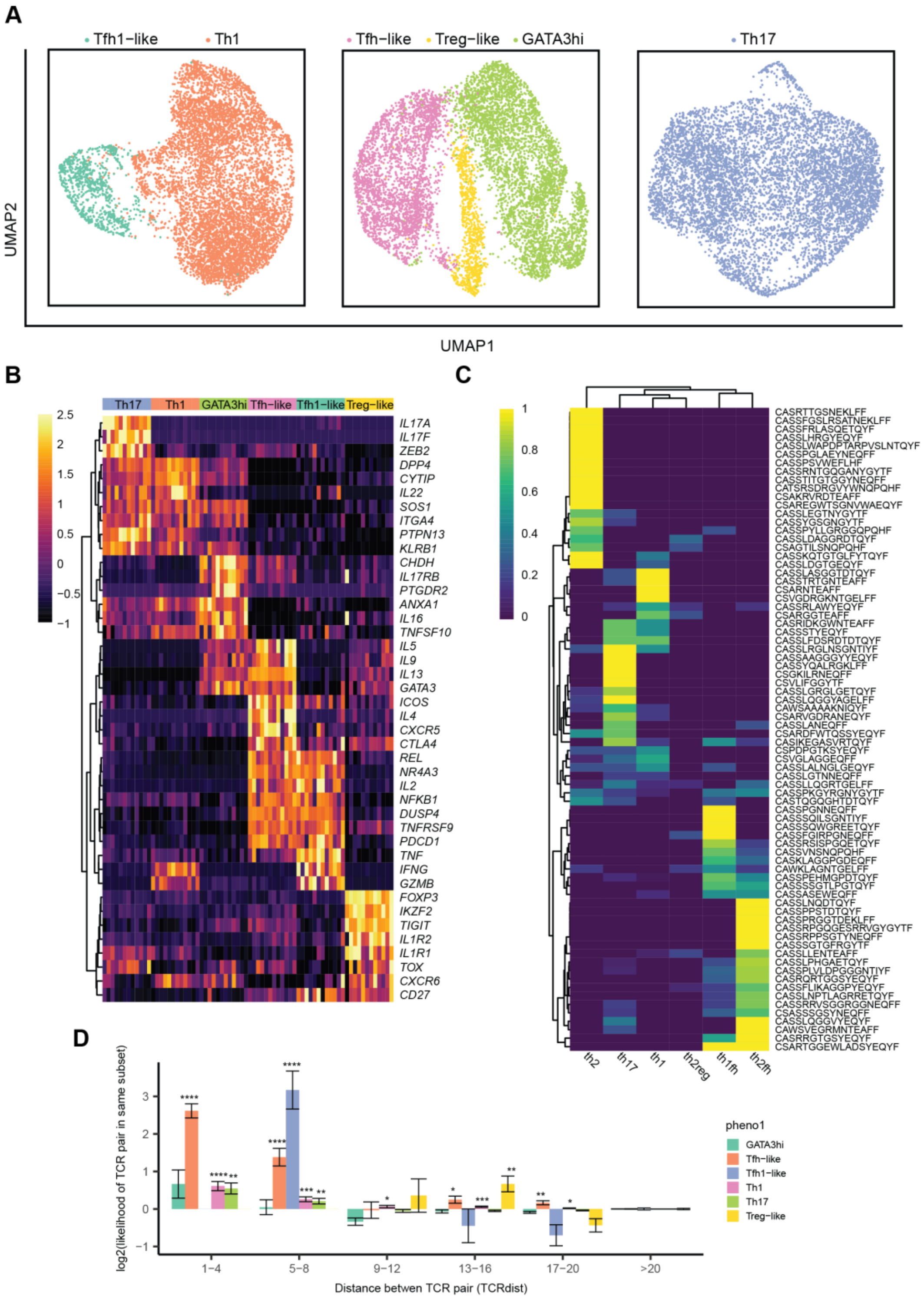


Figure 4-4. Subsets of T helper are clonotypically distinct. **a,** UMAP visualization of Th1, Th2, and Th17 scoring cells. Clusters are annotated by their putative identities. **b**, Differentially expressed genes in each subset, averaged by patient. Genes were selected using a ROC test and manual curation. Each column represents the average expression of all cells in a given patient, for a given gene (row). Average expression levels were row-normalized. **c**, Heatmap of the percentage of TCRβ sequences shared between the subsets. Sharing is defined as the

number of cells of particular TCRβ sequence (row) detected in each of the subset, divided by the total number of cells within the clonotype. Sequences from all patients were pooled. **d**, TCR distance analysis of TCRβ sequences in the subsets. The x-axis represents bins of successively greater pairwise TCR distance, calculated using TCRdist. The y-axis represents likelihood ratio of cells with highly similar TCRs (of similarity indicated on the x-axis) to be of the same subset,relative to the prior probability of any two cells belonging to that subset. '****' refers to a p-value of <0.0001 by a Chi-square proportion test, '***' refers to p-value of <0.001, and '**' refers to p-value of <0.01.

To determine whether the association between clonotypes and phenotypes was likely influenced by TCR structure, we looked for homology using TCRdist, which quantifies homology between TCR sequences using a modified FLOSUM matrix[63] (**Methods**). We found that pairs of cells with TCRβ sequences with a similarity distance of less than 9 had a significantly increased likelihood of both cells belonging to the same T helper cluster, with the exception of cells in the Th2A-like cluster (**Figure** 4-4d). This result indicates a convergence onto common TCR motifs within each of the subsets, suggesting that clonotype-specific factors such as TCR affinity or epitope selection may drive the induction of specific T helper phenotypes.

4.3.4 Th1 and Th2, but not Tfh, expression is suppressed by OIT

We next assessed the influence of OIT on the TCR repertoire and the identified T helper phenotypes. One possible mechanism of OIT is by either inducing expansion of T cell clones associated with regulatory function, or by promoting contraction of reactive clones. We quantified the frequencies of expanded clonotypes at different timepoints. The majority of expanded Th2-expressing clonotypes were present at all four timepoints, and the clonal frequencies at baseline was significantly correlated with clonal frequency at maintenance, suggesting that OIT introduced little change in the peanut-reactive TCR repertoire (**Figure** A4-4). Furthermore, we assessed whether any of the timepoints were associated with emergence of lowly-expanded clonotypes, which could represent emergence of newly activated clonal T cells. We did not find any of the four timepoints to be associated with higher amount of unique clonotypes than expected (**Figure** A4-4).

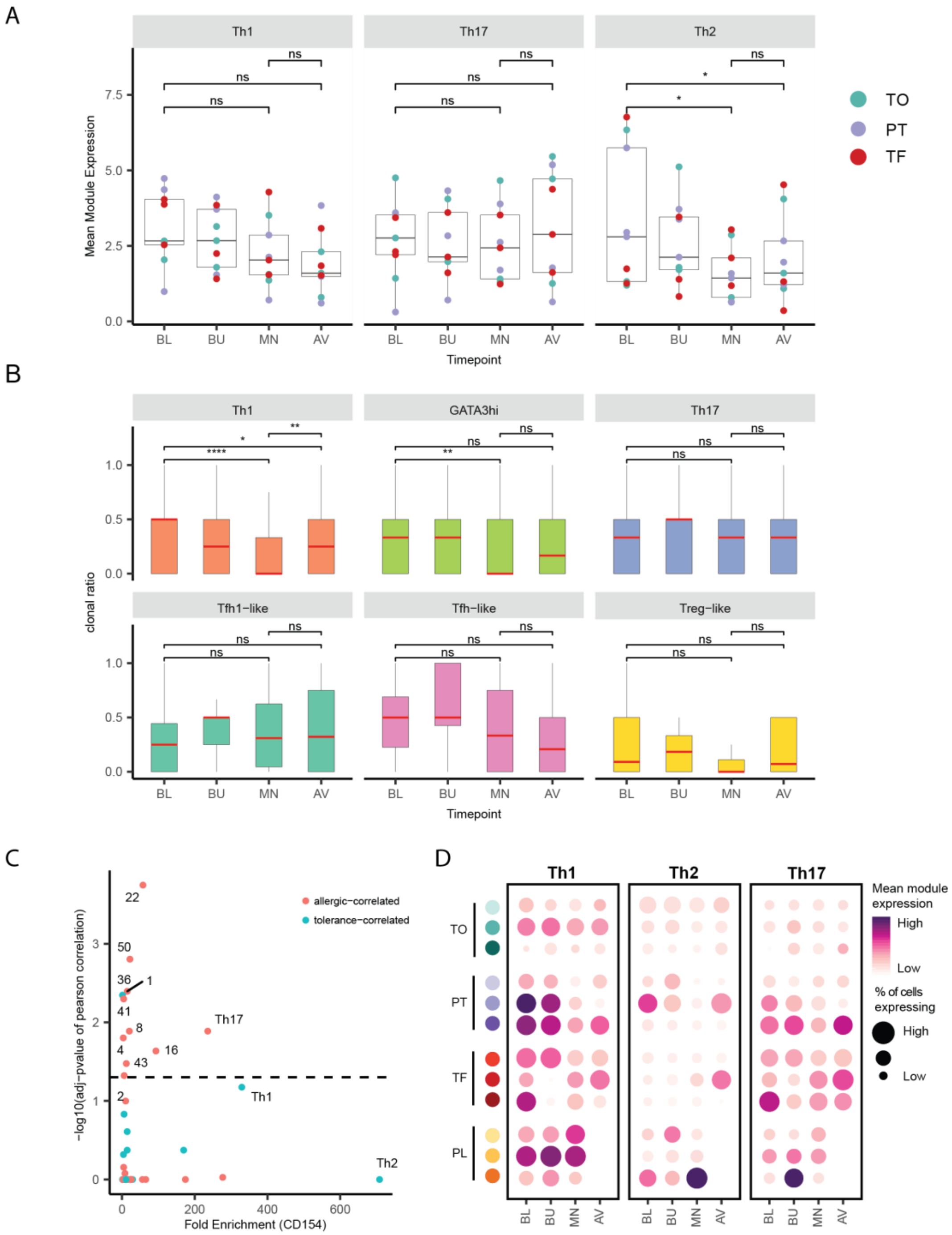


Figure 4-5. Th1 and Th2, but not Tfh, subsets are suppressed by OIT. **a,** Average Th1 Th2 and Th17 module expression by patient, over time, for treatment patients. Data points are colored by the clinical outcome for each patient. '*' refers to an adjusted p-value of <0.05 by a Wilcox rank-sum test. **b**, Proportion of Th1+, Th2+, and Th17+ cells within clonotypes of six identified T helper subsets. The fraction of cells within each clonotype that were positive-scoring for their respective module at a given time point was plotted for each clonotype with at least two cells detected at that time point. '*' refers to an adjusted p-value of <0.05 by a Wilcox rank-sum test, "**" refers to adjusted p-value of <0.005, and "****" refers to adjusted p-value of <0.0005. **c**,

Correlation of modules to clinical outcome and their CD154 enrichment. Dashed line indicates adjust p-value = 0.05. Each module colored by their correlation to allergic (TF and PT) or tolerance (TO). Data represents all patients at all timepoints with the exception of placebo (PL) patients. **d**, Expression of Th1, Th2, and Th17 modules by patient and timepoint. Size of the circle indicates scaled percent of CD154+ cells expressing the respective modules. Color indicates the level of expression, within cells that express the respective modules.

We then evaluated the mean expression of Th1, Th2, and Th17 modules in patients undergoing treatment. We found that only the Th2 module showed significant changes over the course of OIT (**Figure** 4-5). We did not detect significant changes in patients treated with placebo (**Figure** A4-5).

Next, we quantified the clonotype-specific impact of OIT. Due to the varying expression of Th1, 2, and 17 genes by different clonotypes, we decided to quantify the uniformity of clonotypes in their expression of Th1, 2, or 17 module, instead of their average expression directly. For each clonotype associated with each subset, we calculated the proportion of cells (of the same clonotype) that meet the respective module gating threshold at each timepoint. This resulted in a "clonal ratio" metric that demonstrates suppression and activation of clonal cells.

We found that clonotypes associated with Th1 and Th2A-like populations showed decreases in proportions of Th1+ and Th2+ cells, respectively, over the course of treatment, indicating that OIT induced a suppression of inflammatory activities in these populations, a trend which was not detected in placebo-treated patients (**Figure** A4-5). In contrast, we did not detect statistically significant changes in the Tfh1-like, Tfh2-like, or the Th17 subset (**Figure** 4-5), suggesting that these populations are more resistant to the effects of OIT.

Interestingly, we found that the expression of the Th17, but not Th1 or Th2, module significantly correlated with clinical outcome: across all timepoints, patients in the treatment failure or the partial tolerance group exhibited higher expression of Th17 than patients in the tolerance group (**Figure** 4-5). The Th17 module was also significantly enriched in the CD154+ population over the double negative compartment, suggesting that its expression was also specifically induced by peanut antigens. Other modules that significantly correlated with the allergic patients were mostly associated with signatures of general T cell activation, such as *OX40L* (Module 22), *LAG3* (Module 8), and *STAT1* (Module 4).

4.3.5 Clonotypes associated with increased expression are more resistant to OIT

To more precisely determine the effect of OIT on clonotypes, we decided to further quantify clonotypic changes overtime. We saw the largest changes in Th1 and Th2 expression levels between baseline (BL) and maintenance (MN) timepoints. Therefore, we decided to focus our analysis on just those two timepoints. We filtered the dataset to just clonotypes from treated patients that have been detected at both timepoints. Then, we separated clonotypes into ones that are Th2+ at just BL, Th2+ at just MN, and Th2+ at both timepoints and neither timepoints (**Figure** 4-6).

We compared the mean Th2 scores of the clonotypes. We found that clonotypes that were classified as 'both' had higher mean Th2 scores and clonal ratios at both timepoints than clonotypes that were only classified as Th2+ at one, but not both timepoints (**Figure** 4-6). Clonotypes in the four categories were not significantly different in clonal expansion. Our results would suggest that clonotypes with higher expression of Th2 genes were more likely to maintain their expression over the two timepoints. Interestingly, we did not find significant enrichment of the categorized clonotypes in any of the Th2 subsets (**Figure** 4-7).

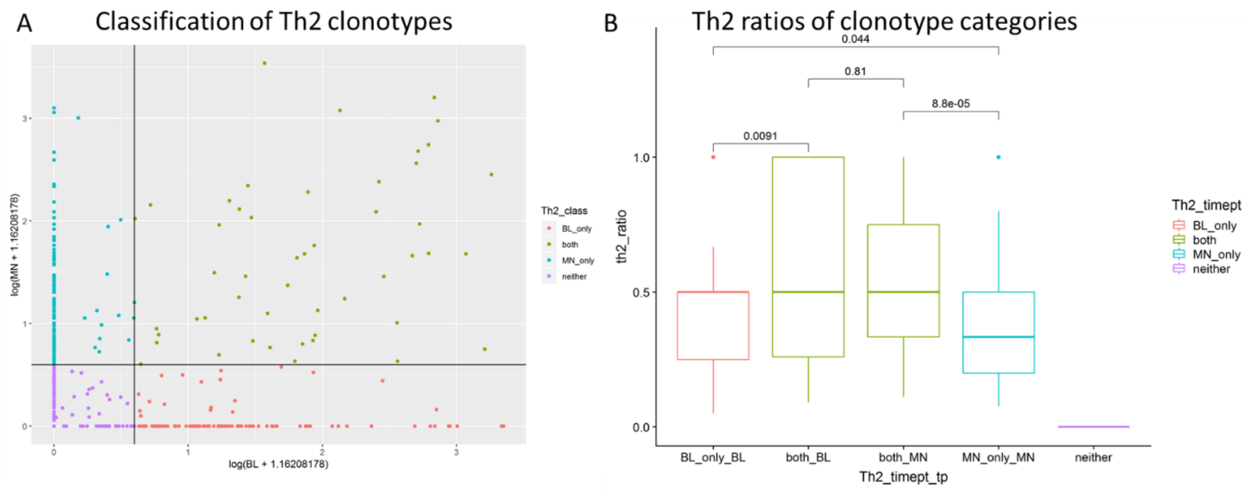


Figure 4-6. Highly Th2 clonotypes are more likely to remain highly activated at both BL and MN timepoints. **a,** Th2 module scores of clonotypes at BL (x-axis) and MN (y-axis). Each point represents average expression of a clonotype. Vertical and horizontal lines represent Th2 gating on the module score (**Methods**). Clonotypes are then classified as 'BL-only', or Th2 + at only BL timepoint; 'MN-only', or Th2+ at only the MN timepoint; 'Neither' for not Th2+ at either timepoints; or 'Both' for Th2 + at both timepoints. **b**, Proportion of Th2+ cells for each of the categories of clonotypes identified in **a**. p-value shown for wilcoxon rank-sum test.

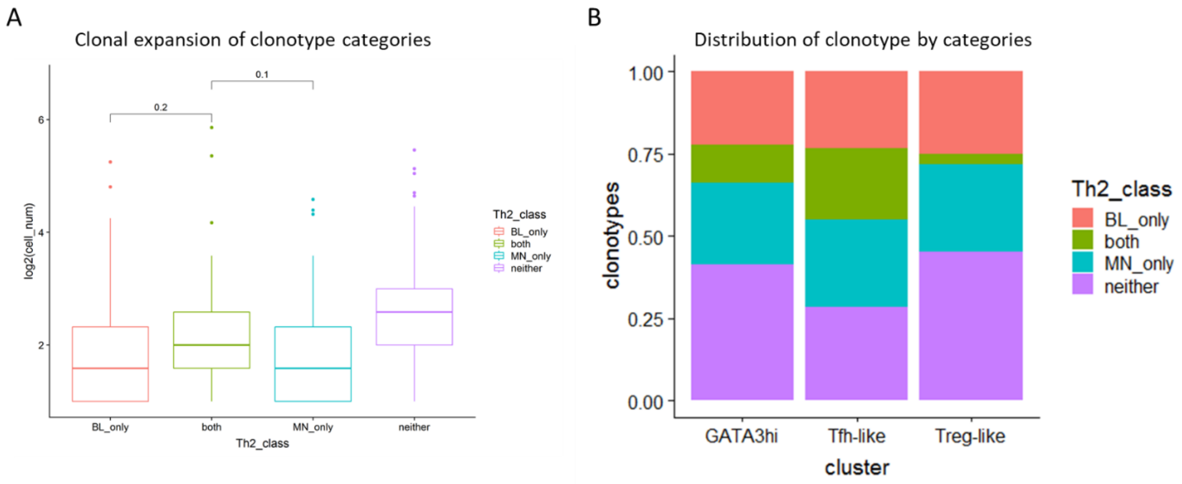


Figure 4-7. Clonotypes that are Th2+ at BL and MN are not more clonally expanded or more likely to be of a particular Th subset. **a,** Clonal expansion of categories of clonotypes shown in **Figure** 4-6. **b,** Distribution of the categories of clonotypes among the three subsets of Th2 cells.

4.4 Discussion

In this chapter, we applied scRNA-seq and TCR recovery to study peanut-reactive T cells. By using an antigen activation assay and leveraging the resolution of scRNA-seq, we were able to delineate multiple subsets of relevant T helper cells simultaneously, despite the rarity of these populations. We were also able to quantify clonotypic responses associated with each of these populations. To our knowledge, this work represents the most comprehensive characterization of T helper cells in the context of food allergy.

Our results suggest that OIT may only be effective in modulating certain subsets of T helper cells. In our data, Tfh cells, which may be most directly responsible for the initial production of allergen-specific IgE antibodies, were less likely to be suppressed by OIT. Nevertheless, we detected a significant decrease in the response of other Th2 cells. Our findings suggest a possible explanation for the temporary effects of peanut OIT. It is possible that only the non-Tfh Th2 cells are responsive to therapy, and as a result IgE-driven immune responses, which may be associated with Tfh cells, could return after the end of treatment. Furthermore, our analysis of clonotypes detected in both BL and MN would indicate that the most pathogenic Th2 cells may also be the most resistant to therapy, further indicating that OIT may only modulate a subset of peanut-reactive cells.

Surprisingly, we did not detect an emergence of Treg response in our dataset. Treg module did not change significantly with time, nor did it associate with clinical outcome. The Treg subset within the Th2 cells is the smallest subset, making the study of this subset over the four timepoints more difficult.

Our study had several limitations. First, it is likely that the peanut activation assay likely also induced some bystander activation of T cells. We believe that by incorporating TCR clonotype information, we were at least able to in part ameliorate this source of noise and identify the likely antigen-specific T cells. Secondly, due to the size of the study, we were unable to determine gene expression modules predictive of clinical outcome at baseline, simply due to p-value adjustment. We believe that if we were to increase the size of the study, our correlation could reach statistical significance. Nevertheless, by including datapoints from all timepoints (and not just those from baseline) we were able to detect statistically significant correlations between clinical outcome and expression of the Th17 module as well as other modules related to general T cell activation. Our analysis indicates that Th17 as well as other signatures of T cell activation may be upregulated in patients with worse outcomes.

Our results suggest that while OIT is effective in inducing changes in the Th2 cells, it is likely ineffective in modulating key populations of T cells, limiting the effectiveness of the treatment. In the future, treatment that can more directly target Tfh cells may lead to more lasting benefits.

5. Application of Seq-Well and TCR recovery to the studies of other diseases

This chapter is in part adapted from: S.W. Kazer, T.P. Aicher, … A.A. Tu, et al[98], and T.K. Hughes, D. Gideon, … A.A. Tu, et al, *in prep.*

In this chapter, we detail the application of our TCR recovery technique to previously processed samples. One of the main advantages of our method is that it is applicable to any 3' library in general, and as such is applicable to processed samples *post hoc*. We applied the technique to two clinical studies. First, we recovered TCR sequences from HIV samples originally processed in South Africa to find T cell clonality associated with known markers of activated T cells. Then, we expanded the application of the method to cynomolgus monkeys by designing new primers specific to the TCR V genes of the cynomolgus genome.

5.1 Motivation

5.1.1 Retroactive study of TCR from processed clinical samples

Single-cell sequencing is an increasingly common technique used to study immunological questions. There are an increasing number of published and commercially available platforms for research groups to choose from. The costs of these platforms are often high both in terms of the monetary costs and the learning curve associated with technology adoption. Therefore, it is uncommon for any individual group to utilize multiple platforms, and as a result, different research labs often utilize different platforms, depending on various factors such as the specific application and the ease of adoption.

As such, our outlined TCR recovery technique presents several advantages. Firstly, it is applicable to any single-cell 3' barcoded library, regardless of how the library is generated, whether via Seq-Well, or another commercial solution such as those of 10X genomics. Secondly, our method relies on short-read sequencing, which is more widely available than long-range sequencing to most research groups. Thirdly, because library amplification is only done with universal primers, it is relatively easy to adapt the technique for other species or other gene targets by changing the gene-specific primers, since these primers would need less optimization.

The first two points offer further advantages for sample processing. Namely, a T cell sample could be process and analyzed before deciding whether sequencing of the TCR repertoire would likely be fruitful. Furthermore, it also means that the technique could be applied to archived clinical samples that were previously processed, should the T cells in these samples warrant further analysis. We leveraged this characteristic to further study T cells in the context of human immunodeficiency virus (HIV) infection using samples that were collected and processed in South Africa.

Predicting primer interactions in a multiplex primer pool is often difficult, and therefore optimization often requires laborious trial and error. Much of the difficulty comes from the PCR amplification process, which could intensify imperfections in the primer design by amplifying off-target products. Therefore, we reasoned that it would be relatively simple to design new V primer sets in our protocol since the primers would only be used for a single-step extension instead of a multi-cycle amplification. We demonstrated this principle by extending our technique to study T cells in a cynomolgus monkey model of tuberculosis (TB). We designed the primers based on limited knowledge of V genes, and was able to recovered TCR sequences in a wide variety of T cells.

5.2 Methods

FRESH study subjects. The FRESH study recruits HIV negative women, age 18-24, and tests for HIV-1 RNA in the plasma twice weekly for one year. Each subject participates in peer-support groups and receive a stipend for each visit. If a plasma test resulted positive, the participant in asked to return to the clinic the same day to collect a blood sample. Thereafter, blood samples are collected weekly through first 6 weeks of infection, and regularly afterward as long as the participant continues to return to the study center. At the time of positive plasma test, subjects also initiate anti-retroviral therapy, as per standard treatment guidelines.

Cynomolgus macaque animals. Four Cynomolgus macaques (*Macaca fascicularis*), >4 years of age, (Valley Biosystems, Sacramento, CA) were housed within a Biosafety Level 3 (BSL-3) primate facility. Animals were infected with low dose *M tuberculosis* (Erdman strain) via bronchoscopic instillation of 7-12 colony-forming units (CFUs)/ monkey to the lower lung lobe. Animals were infected for a period of 10 weeks and Infection was confirmed by tuberculin skin test conversion. Serial clinical, microbiologic, immunologic, and radiographic examinations were also performed.

Necropsy of cynomolgus macaques. an ^{18}F-FDG PET-CT scan was performed on every animal 1-3 days prior to necropsy to measure disease progression and identify individual granulomas. At necropsy, monkeys were maximally bled and humanely sacrificed using pentobarbital and phenytoin (Beuthanasia; Schering-Plough, Kenilworth, NJ). Individual lesions previously identified by PET-CT and those that were not seen on imaging from lung and mediastinal lymph nodes were obtained for histological analysis, bacterial burden, and immunological studies.

Design of cynomolgus TCR primers. Cynomolgus TCR gene references were provided by the Sam Behar lab. Human primer sets were mapped to the reference via Blast, and primer sequences were adjusted to include all genes in the cynomolgus reference. Homologies between different genes were leveraged to decrease the total number of primers needed.

5.3 Results

5.3.1 Recovery of TCR sequences from HIV samples

Due to the complex immune response involved in the disease, scRNA-seq is an attractive tool for the study of HIV infection. The dynamics of immune response during the early stages of infection, particularly Fiebig Stage I and II, as well as before and at peak viral load, could be important for identification of better therapeutics targets[99,100]. To this effort, the Females Rising through Education, Support and Health (FRESH) study, which combined education, job training, and administration of treatment, was started to track high-risk individuals in South Africa before and after acute infection[101]. The study involved regular collection of blood samples from enrolled individuals. The individuals were tested for HIV infection, and samples before, during, and after acute infections were processed for scRNA-seq.

Due to the scarcity of the samples, Seq-Well was well suited for the application, as it can process low-input samples with minimal losses. Samples were acquired and processed in an ongoing basis, often directly onsite in South Africa. As a result, transcriptomic data from the samples were acquired and analyzed before the development of work described in this thesis (the data was also collected using an earlier version of Seq-Well). The data indicated upregulation of interferon stimulated genes as well as pro-inflammatory T cell differentiation. Modules of viral response genes were also identified in subsets of CD8 T cells.

We were interested to investigate whether we could also detect clonally expanded T cells in individuals over the course of the study. We recovered TCR sequences from four individuals in the study (**Figure** 5-1). While we were able to recover on average 40% of TCRβ from each of the patients, we were only able to recover roughly 25% of TCRα, leading to low pairing of alpha and beta chain. As a result, we focused our analysis on TCRβ sequences.

While we were able to detect clonal cells in the dataset, the overall clonal expansion was low, and there was no observable trend in each of the patients (**Figure** 5-2). This could be due to low numbers of T cells recovered for each of the patients, making consistent discovery of expanded clonotypes difficult (**Figure** A5-1). Furthermore, the samples were not enriched for antigen-specificity or reactivity, meaning the number of antigen-reactive T cells in the dataset may be inconsistent from timepoint to timepoint. Nevertheless, we detected some clonal T cells, particularly around two-four weeks after infection. As expected, clonal expansion was higher in the cytotoxic T lymphocyte (CTL) populations (**Figure** 5-3). The detected expanded clones could represent T cell expansion in response to acute infection.

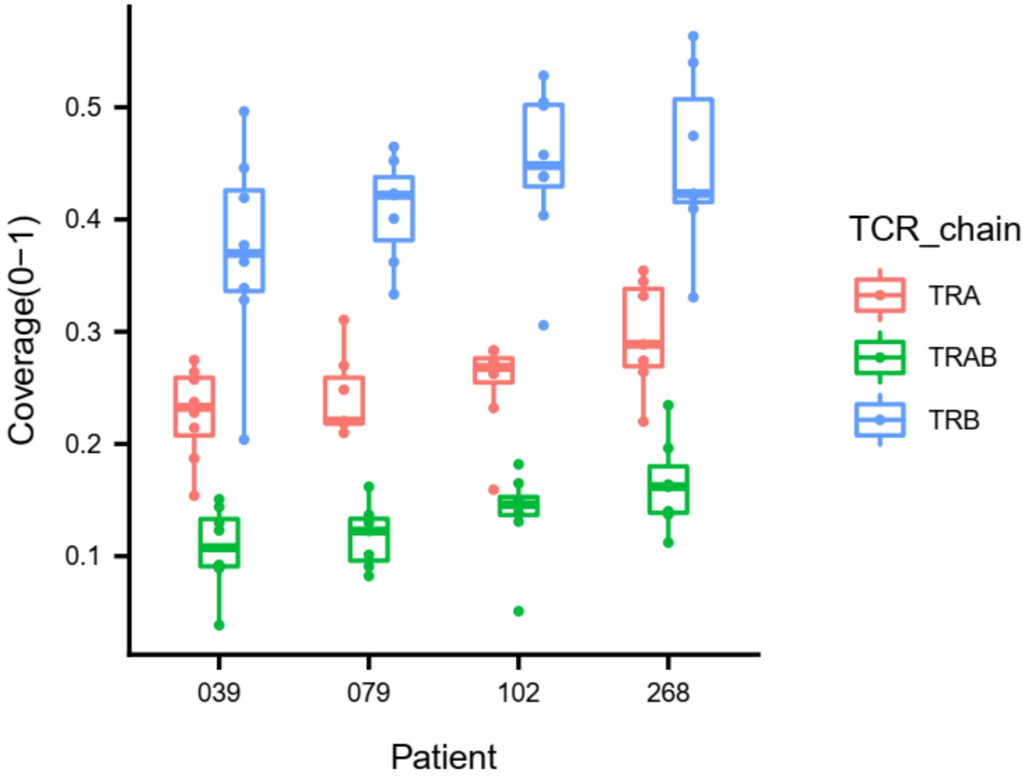


Figure 5-1. Recovery rate of TCRα and TCRβ sequences from human HIV samples. Each data point represents a sample from one patient at one timepoint.

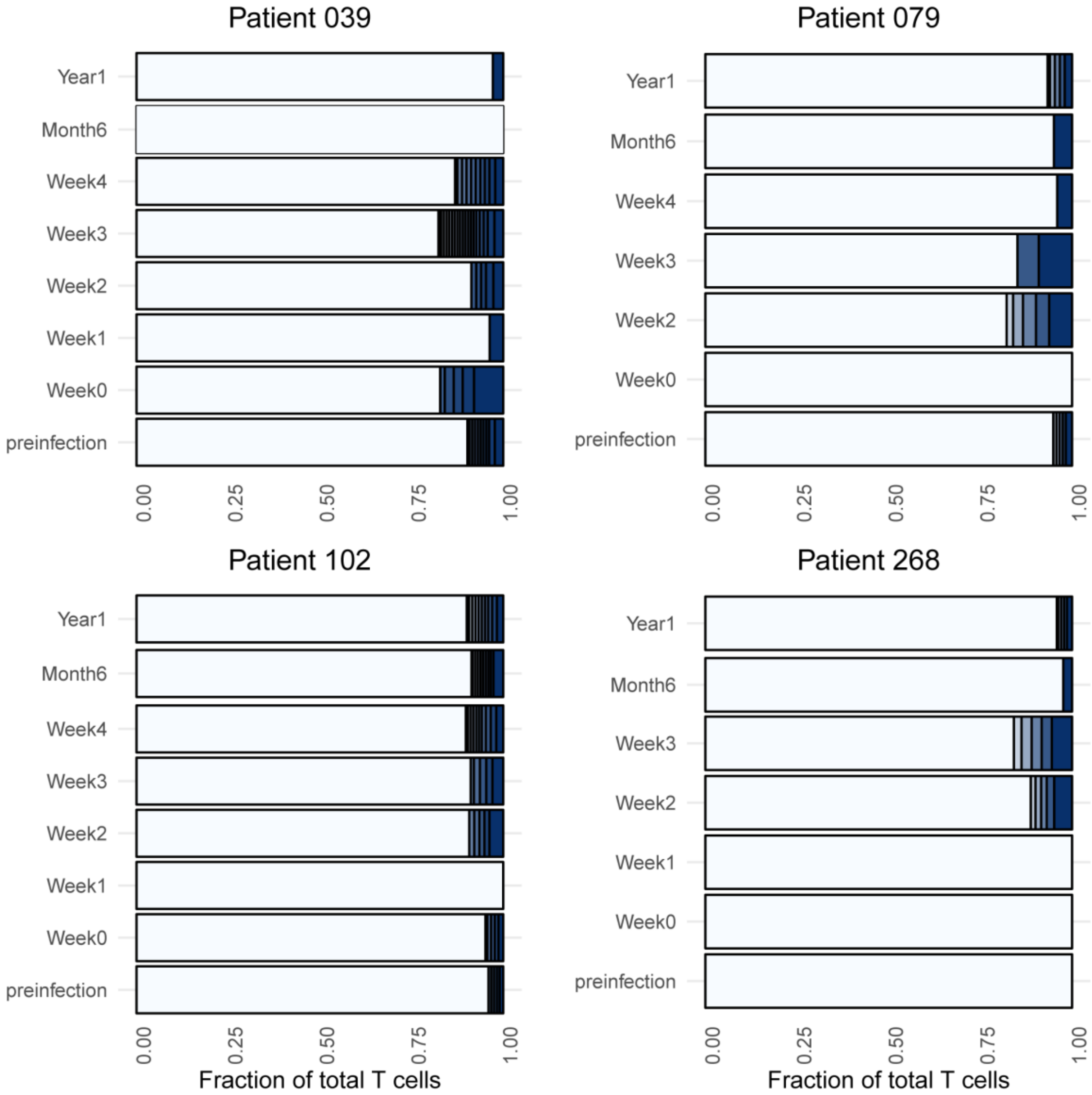


Figure 5-2. Clonal expansion of T cells from each of the HIV patient samples. Dark blue segments indicate proportion of cells with expanded clonotypes. Each segment indicates 2 unique clonotypes from the most expanded (i.e. rank 1 and 2) to least expanded. White portion indicates singletons, or clonotypes detected in only one cell.

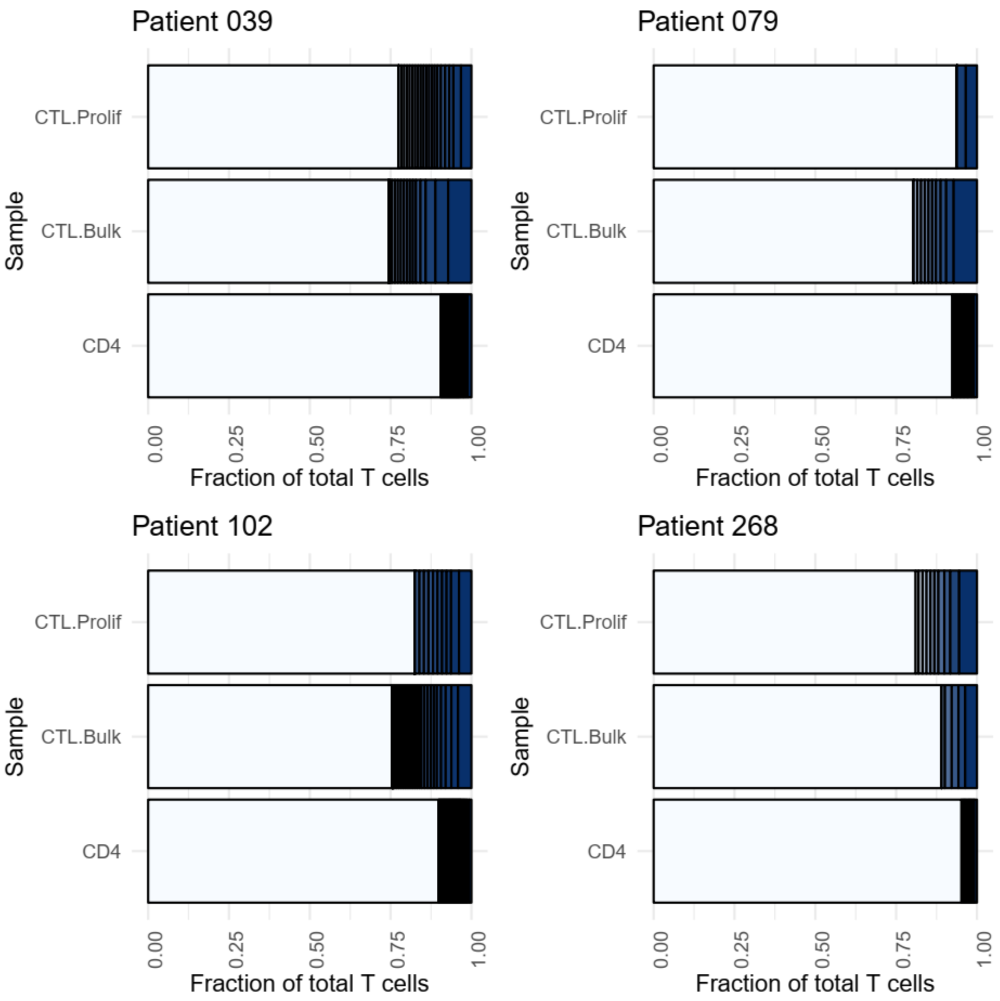


Figure 5-3. Clonal expansion by T cell cluster. Cytotoxic T lymphocyte (CTL) clusters are separated into one with proliferation markers (CTL.Prolif) and one without (CTL.Bulk). Blue segments indicate proportion of expanded cells.

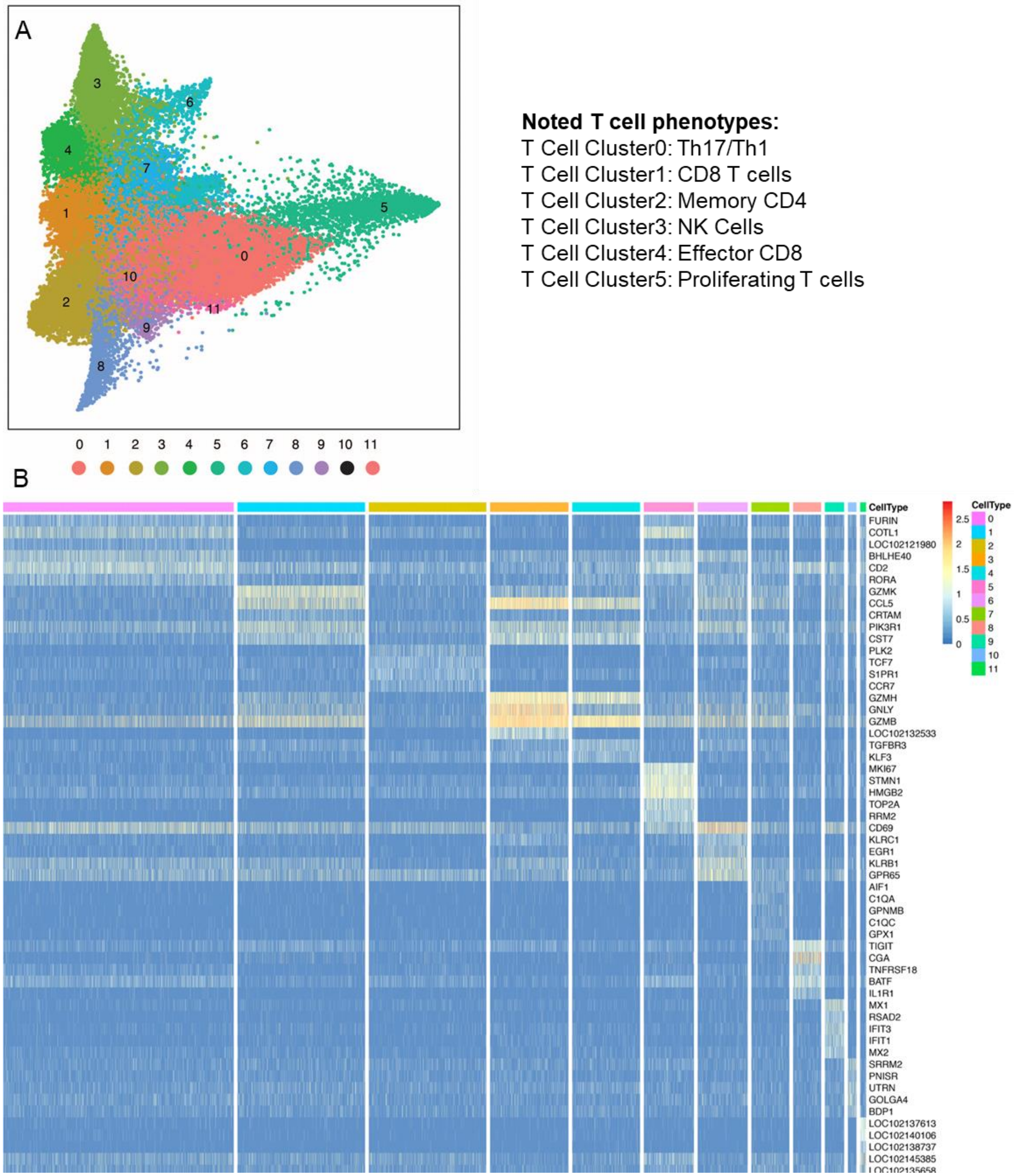
5.3.2 Recovery of cynomolgus monkey TCR sequences

So far, we have shown applications of our method on samples derived from murine and human tissues. However, many research groups also utilize non-human primates (NHPs), particularly to study infectious diseases[102]. Therefore, extension of our method to NHP species could prove valuable for those studies.

TCR sequencing of NHP models can be particularly challenging, since the relevant loci (particularly those of V and J genes) of the species may not be well-annotated, and well-validated primer sets may not be available. Working with the Samuel Behar lab, we adapted our human primer sets to cynomolgus references. While there are many homologies between the human and the cynomolgus references, many of the primers required modifications to avoid erroneous matching of primer and target sequences, and to ensure that all relevant V genes were targeted by at least one primer in the pool.

We then applied our modified method to study *Mycobacterium tuberculosis* (Mtb) in an NHP model. Mtb is the major causative agent of TB, which is estimated to cause 1.5 million deaths per year[103]. Mtb infection is characterized by the formation of granuloma, a structure of immune and stromal cells formed in response to Mtb. Immune cells within the granuloma are critical to the control of persistent infections[104,105]. Due to the nature of the disease, human samples of granulomas are typically only available in cases when surgery is needed (often for unrelated reasons) or posthumously. NHP model provides an opportunity to study Mtb granulomas in different stages of disease.

We recovered TCR sequences from single-cell libraries of 28 granulomas from four cynomolgus macaques. The recovered T cells of the samples exhibited a variety of phenotypes, including a Th1/Th17 (cluster 0), a cytotoxic CD8 (cluster 1 and 4), and a proliferating (cluster 5) phenotype (**Figure** 5-4). Recovered clonotypes from the T cells indicated high amount of clonality among the T cells, suggesting that the cells have gone through extensive proliferation (**Figure** 5-5). Clonal expansion was highest among the Th1/17, proliferating, and cytotoxic clusters. Further, sharing of the clonotypes was also the highest among these three clusters of T cells, suggesting that these clusters of cells had shared antigen-recognition (**Figure** 5-6). Besides these phenotypes, most TCR sequences were restricted to just one T cell cluster. We also found that TCR clonotypes were mostly specific to each individual animal, with very few public (or shared) clones (**Figure** A5-2).



Noted T cell phenotypes:

T Cell Cluster0: Th17/Th1
T Cell Cluster1: CD8 T cells
T Cell Cluster2: Memory CD4
T Cell Cluster3: NK Cells
T Cell Cluster4: Effector CD8
T Cell Cluster5: Proliferating T cells

Figure 5-4. T cells of granulomas form distinct clusters. **a**, tSNE representation of T cell transcriptomic data from granulomas. Cells are colored by putative clusters. Clusters with known canonical phenotypes are noted on the right. **b**, Heatmap of gene expression of T cells shown in **a.**

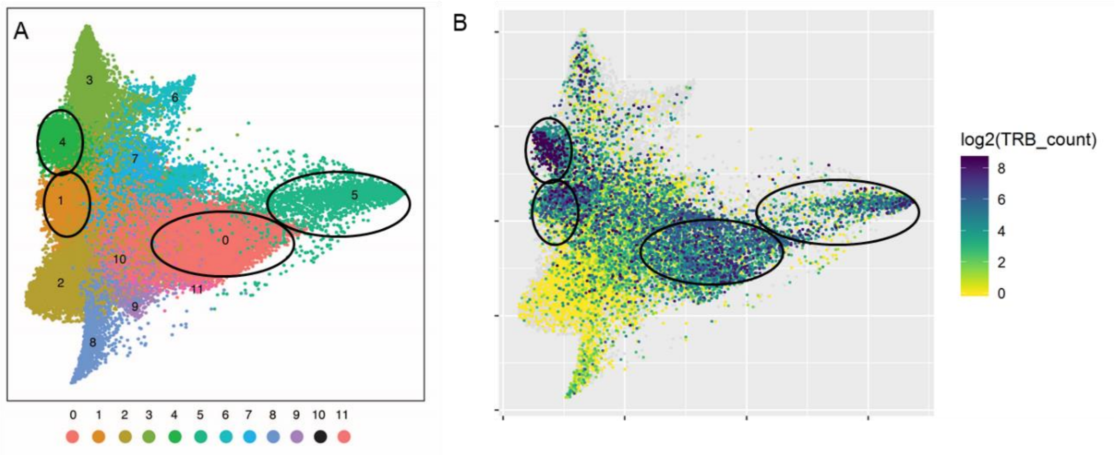
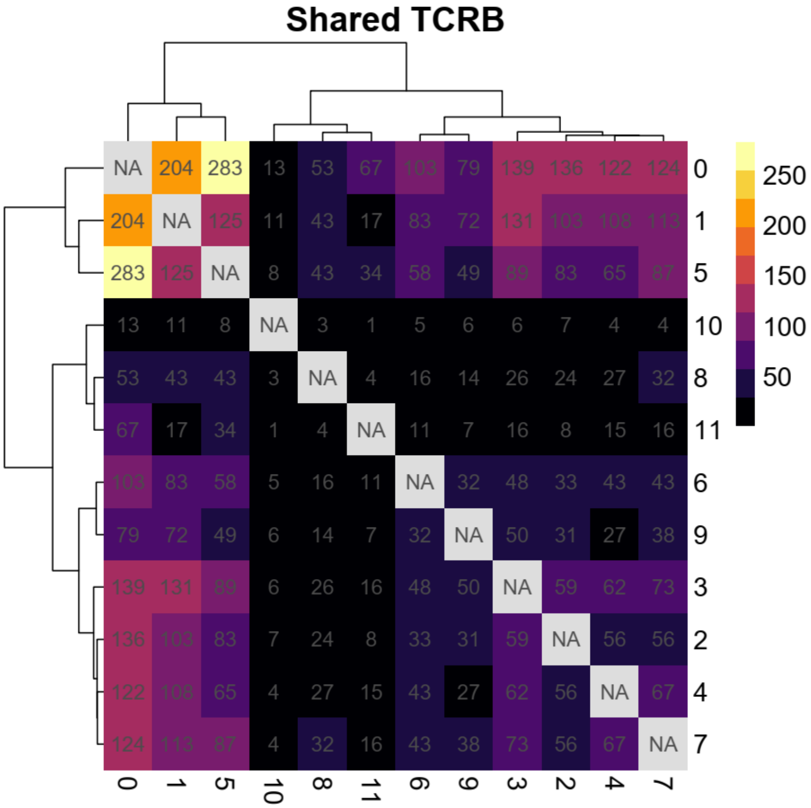


Figure 5-5. Clonal expansion detected in T cells from granulomas. **a,** tSNE representation of T cells, with expanded T cell clusters circled. **b,** tSNE representation of T cells colored by TCRβ clonal size, defined as the number of cells in the dataset sharing the same clonotypes. Darker blue represents more clonal T cells (on log2 scale).



	0	1	5	10	8	11	6	9	3	2	4	7
0	NA	204	283	13	53	67	103	79	139	136	122	124
1	204	NA	125	11	43	17	83	72	131	103	108	113
5	283	125	NA	8	43	34	58	49	89	83	65	87
10	13	11	8	NA	3	1	5	6	6	7	4	4
8	53	43	43	3	NA	4	16	14	26	24	27	32
11	67	17	34	1	4	NA	11	7	16	8	15	16
6	103	83	58	5	16	11	NA	32	48	33	43	43
9	79	72	49	6	14	7	32	NA	50	31	27	38
3	139	131	89	6	26	16	48	50	NA	59	62	73
2	136	103	83	7	24	8	33	31	59	NA	56	56
4	122	108	65	4	27	15	43	27	62	56	NA	67
7	124	113	87	4	32	16	43	38	73	56	67	NA

Figure 5-6. Number of TCRβ sequences shared between clusters of T cells shown in Figure 5-4. The number in the heatmap indicates the count of unique clonotypes shared between the two respective clusters.

5.4 Discussion

In this chapter, we demonstrate the utility of our TCR recovery method to previously processed samples. In both the study of HIV and the study of TB, samples were processed without the expressed intent of characterizing the TCR repertoires. It was only after the transcriptomic data were analyzed, indicating strong T cell-associated immune responses, was the decision made to study the TCR repertoires of the samples.

Due to the simplicity of our method, we were able to adapt the technique to recover TCR sequences from the archived samples (that is, frozen WTA samples), and incorporate them into the processed data. We believe that this workflow offers notable cost-saving potential: since NGS sequencing is still often the most expensive part of a single-cell workflow, it is often more cost-efficient to process and analyze the whole transcriptomic data first, before deciding whether the additional costs of TCR sequencing would likely yield useful information. We believe this workflow may be attractive to groups with limited resources.

Similarly, as 3' barcoding is one of the most common schemes employed in various scRNA-seq platforms, we believe our method could be especially useful for research groups with archived libraries of various platforms. The potential to augment existing samples presents an interesting avenue to further study T cell immunology.

6. Conclusion and outlook

The maturation of NGS platforms followed by the development of high-throughput single-cell methodologies represents leaps in technical ability similar to that of flow cytometry for immunologists. The immune system is necessarily heterogenous, making it a suitable application for scRNA-seq to investigate biological questions without *a priori* knowledge. While data produced by these techniques seem comprehensive, there are still nevertheless biologically critical characteristics missed by the currently established technologies. One example is the TCR sequence, which has largely eluded available scRNA-seq platforms due to technical reasons. While new, emerging sequencing platforms and technologies hold promise to better capture TCR sequences at single-cell resolution, a cost-effective method that is easily adaptable to existing platforms would still provide significant advantages to many research groups that are already analyzing and processing valuable clinical samples.

In Part I of this thesis, I detail our attempts at developing such a method through the perhaps naïve approach of size selection. We reasoned that by preparing the sequencing library in such a way as to preferentially sequence larger fragments, we could more effectively capture the CDR3 sequences of the TCRs. While we were correct in principle, we encountered several difficulties that would prove ultimately too difficult to overcome. Firstly, precise size selection was simply difficult to repeat consistently from sample to sample without a gel electrophoresis-based approach, which significantly limited throughput, making the approach impractical. Secondly, the expression of TCR mRNA was less strictly regulated than expected: we captured a significant number of transcripts with incompletely spliced TCR genes. These segments would not be efficiently filtered by size selection alone. Interestingly, we also seemed to see similar results in bulk TCR sequencing techniques that do not rely in V region primers (such as the 5' RACE-based approached outlined in **Figure** 1-2). These methods, in general, seem to capture a higher proportion of non-functional sequences.

We took the lessons learned from the size selection approach, and decided to incorporate a V gene selection in the form of primer extension by primers specific for V genes. By using the multiplex primers for just a single-step extension instead of amplification, we by and large avoided issues with PCR artifacts. The modified technique resulted in sequencing libraries characterized by monodispersed sizes that can be sequenced with TCR-specific sequencing primers, resulting in high yields of CDR3 sequences that can be combined with the transcriptomic data. To demonstrate the utility of our method, we applied the technique to characterize MHC-tetramer sorted CD8 T cells in mice immunized with HPV-E7, and in patients with peanut food allergy. In both cases, we detected clonal T cells associated with specific T effector functions, such as Th2 cells and cytotoxic CD8 T cells.

In Part II of this work, we further characterized T helper cell responses in food allergy patients undergoing treatment. We profiled putatively peanut-reactive T cells enriched by an T cell activation assay. By using scRNA-seq, we identified subsets of T helper cells with Th1, Th2, Th17, Tfh, and Treg effector phenotypes. By further studying their clonotypes, we observed that these subsets were clonotypically distinct. We also found that though the TCR repertoires of each patient was stable through OIT, regardless of clinical outcome, Th1 and Th2 clonotypes were suppressed in their effector functions, while Tfh clonotypes were unchanged by treatment. Interestingly, the Th17 gene module, instead of Th2, correlated with clinical outcome along with other signatures of T cell activation. Furthermore, these signatures did not change significantly throughout the course of OIT. By tracking clonotypes over the course of treatment, we also found that more highly activated clonotypes were less likely to be modulated by therapy.

Next, we applied the technique to previously processed clinical samples, including samples from studies of HIV. We were able to identify clonally expanded T cells correlated to activated phenotypes. We also extended the technique to characterize the TCR repertoires in TB granulomas of cynomolgus macaques. By changing the biotin pull-down probe to target the constant regions, which are well-annotated, of cynomolgus monkey, we were able to enrich the corresponding TCR transcripts. Then by adapting the V gene primers to the V genes of the monkeys, we were able to successfully modify the method to the species. In the study of TB, we detected clonal expansion correlated to activated subsets of T cells. We also identified clusters with shared TCR sequences, indicating shared T cell lineages among those T cells.

The work described here demonstrates an example of combining and utilizing conventional techniques with cutting-edge methods to enhance our ability to study complex biological systems. By identifying the importance of including T cell clonal information in scRNA-seq dataset, and by analyzing why this information was not commonly captured by available technology, we were able to design a methodology to complement existing workflows. Much of this work was also guided by practical and secondhand experiences of adopting and working with scRNA-seq technologies. We understood that adopting new technologies, and all the associated downstream data analyses, is often a labor- and cost-intensive process for many research groups, and as such methods that are compatible with a wide array existing platform would be especially useful. We believe the methods described in this work would be especially helpful for groups that have already adopted one of the common scRNA-seq platforms, and are looking to also complement their existing datasets with T cell clonotypic information.

Nevertheless, we recognize that as the NGS technological landscape continues to develop, there will be other technologies that will eventually supersede current workflows and, therefore, the work presented in this thesis. The advent of long-range sequencing, such as those of Pacific Biosciences and Oxford Nanopore, holds promise to improve our ability to not just sequence relatively short pieces of mRNA at a time, but rather full-length sequences of the transcripts. Though currently, these technologies are still limited by their sample efficiency (i.e. number of reads per nanogram of cDNA) and sequencing quality (which presents difficulties in accurately determining single-cell barcodes), we expect that these technologies will continue to improve and eventually allowing accurate sequencing of not just TCR and BCR clonotypes, but also transcript isoforms and single-nucleotide mutations.

As for the future of high-resolution characterization of adaptive immune system, we believe that the work presented here, along with works presented by others elsewhere, indicates that inclusion of TCR and BCR sequences is crucial to better understanding of antigenic immune responses. It has been our experience that T cells subsets are often difficult to segregate into useful subsets. Effector phenotypes that are clearly defined by previously defined panel of surface proteins and receptors are often much less useful when analyzing large sets of single T cell transcriptomics, as the differences among the T cells are often much subtler. Incorporating clonotypic information allows us to group cells by shared antigen-recognition and lineage, often giving us a better understanding of the recent history of antigen exposure in the populations of cells. Despite the relatively elementary analyses of TCR and transcriptome in this work, we have already shown the utility of such approach by tracing T cell lineages over time and identifying cells that are more likely to be truly antigen-specific. We expect further incorporation of sophisticated analytical techniques such as binding-affinity prediction and similarity classification of TCR sequences will further advance our understanding of antigen-specific T cells in different disease contexts.

7. Acknowledgement of contributions

The TCR recovery method presented in this thesis were co-developed and co-designed by Todd M. Gierahn and me. Alex Shalek and J. Christopher Love provided indispensable guidance and input throughout the technology development process. I would also like to acknowledge the many helpful discussions with the Paul Blainey lab and the Michael Birnbaum lab that all contributed to the completion of this work.

Study of T cells from peanut allergic patients was conducted in collaboration with the Wayne Shreffler lab at Massachusetts General Hospital. The clinical trial (NCT01750879) and the PBMC isolation were conducted by Wayne Shreffler's clinic. PBMC stimulation and FACS was done by Bert Ruiter and Neal Smith from the Shreffler lab. Seq-Well experiments of the patient samples were done with the help of Brinda Monian, Patrick Petrossian, and Rachel Barry. Transcriptomic and TCR analyses were done in concert with Brinda Monian and Duncan M. Morgan. Adaptation of the sparse PCA method was done by Brinda Monian.

HPV-E7 immunization of mice, the subsequent stimulation, and FACS of isolated T cells were done by Naveen K. Mehta from the Dane Wittrup lab. TCR analysis of FRESH samples was done in concert with Sam W. Kazer of the Shalek lab, who also led the collaborative effort for the study. TCR recovery and analysis of the cynomolgus granulomas were done in collaboration with Travis K. Hughes from the Shalek lab, who also led the collaborative effort for the study. Sequencing was done at the MIT BioMicro Center with the help of Noelani Kamelamela and Stuart Levine.

This work was supported in part by the Koch Institute Support (core) NIH Grant P30-CA14051 from the National Cancer Institute, as well as the Koch Institute - Dana-Farber/Harvard Cancer Center Bridge Project. This work was also supported by the Food Allergy Science Initiative at the Broad Institute and the NIH (5P01AI039671, 5U19AI089992, U19AI095261).

8. References

1. Boylston, A. The origins of inoculation. *J. R. Soc. Med.* **105**, 309–313 (2012).
2. Kaufmann, S. H. E. Remembering emil von behring: From tetanus treatment to antibody cooperation with phagocytes. *mBio* **8**, (2017).
3. Lucrative Biosimilars Space to Erode Biologics Market From 2019. Available at: https://drug-dev.com/lucrative-biosimilars-space-to-erode-biologics-market-from-2019-2/. (Accessed: 23rd February 2020)
4. Ribatti, D., Crivellato, E. & Vacca, A. The contribution of Bruce Glick to the definition of the role played by the bursa of Fabricius in the development of the B cell lineage. *Clinical and Experimental Immunology* **145**, 1–4 (2006).
5. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: Genetic events and selective pressures. *Nature Reviews Genetics* **11**, 47–59 (2010).
6. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 3628–3632 (1976).
7. Janeway, C. A., Travers, P., Walport, M., Shlomchik, M. J. & others. The generation of diversity in immunoglobulins. in (Garland Science, 2001).
8. Arstila, T. P. *et al.* A direct estimate of the human alphabeta T cell receptor diversity. *Science* **286**, 958–961 (1999).
9. Gorski, J. *et al.* Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. *J. Immunol.* **152**, 5109–19 (1994).
10. Kirsch, I. R. *et al.* TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Sci. Transl. Med.* **7**, 1–13 (2015).
11. Schrama, D., Ritter, C. & Becker, J. C. T cell receptor repertoire usage in cancer as a surrogate marker for immune responses. *Semin. Immunopathol.* **39**, 255–268 (2017).
12. Lossius, A. *et al.* High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8(+) T cells. *Eur. J. Immunol.* **44**, 1–41 (2014).
13. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352**, 189–196 (2016).
14. Khodadoust, M. S. *et al.* Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature* **543**, 723–727 (2017).
15. 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. *Sci. Transl. Med.* **4**, 128ra42 (2012).
16. Joshi, K. *et al.* Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nat. Med.* **25**, 1549–1559 (2019).
17. Ruiter, B. *et al.* Expansion of the CD4+ effector T-cell repertoire characterizes peanut-allergic patients with heightened clinical sensitivity. *J. Allergy Clin. Immunol.* **145**, 270–282 (2020).
18. Crosby, E. J. *et al.* Complimentary mechanisms of dual checkpoint blockade expand unique T-cell repertoires and activate adaptive anti-tumor immunity in triple-negative breast tumors.

Oncoimmunology **7**, e1421891 (2018).

19. Thommen, D. S. & Schumacher, T. N. T Cell Dysfunction in Cancer. *Cancer Cell* **33**, 547–562 (2018).

20. Stadtmauer, E. A. *et al.* CRISPR-engineered T cells in patients with refractory cancer. *Science* (2020). doi:10.1126/science.aba7365

21. Castro, C. D., Luoma, A. M. & Adams, E. J. Coevolution of T-cell receptors with MHC and non-MHC ligands. *Immunol. Rev.* **267**, 30–55 (2015).

22. Wanger, A. *et al.* Overview of Molecular Diagnostics Principles. in *Microbiology and Molecular Diagnosis in Pathology* 233–257 (Elsevier, 2017). doi:10.1016/b978-0-12-805351-5.00012-0

23. Oakes, T. *et al.* Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Front. Immunol.* **8**, (2017).

24. De Simone, M., Rossetti, G. & Pagani, M. Single cell T cell receptor sequencing: Techniques and future challenges. *Frontiers in Immunology* **9**, 1638 (2018).

25. Brandariz-Fontes, C. *et al.* Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Sci. Rep.* **5**, 8056 (2015).

26. Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J. & Chain, B. *Sequence and primer independent stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding*. (2014). doi:10.1101/011411

27. Bolotin, D. a *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

28. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).

29. Hughes, T. K. *et al.* Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. *bioRxiv* 689273 (2019). doi:10.1101/689273

30. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).

31. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).

32. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).

33. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nat. Publ. Gr.* (2017). doi:10.1038/nature22976

34. Scheper, W. *et al.* Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nat. Med.* **25**, 89–94 (2019).

35. Davis, M. M. & Boyd, S. D. Recent progress in the analysis of αβ T cell and B cell receptor repertoires. *Curr. Opin. Immunol.* **59**, 109–114 (2019).

36. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268–272 (2018).

37. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).

38. Avraham, R. *et al.* Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. *Cell* **162**, 1309–1321 (2015).

39. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2017).

40. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

41. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).

42. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).

43. Dash, P. *et al.* Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J. Clin. Invest.* **121**, 288–295 (2011).

44. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

45. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).

46. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).

47. Saikia, M. *et al.* Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nat. Methods* **16**, 59–62 (2019).

48. Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.* **10**, 3120 (2019).

49. Zemmour, D. *et al.* Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.* **19**, 291–301 (2018).

50. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

51. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).

52. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–46 (2012).

53. Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J. & Chain, B. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci. Rep.* **5**, 14629 (2015).

54. Tu, A. A. *et al.* TCR sequencing paired with massively parallel 3’ RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol.* **20**, 1692–1699 (2019).

55. Singer, M. *et al.* A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell* **166**, 1500-1511.e9 (2016).

56. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

57. Wei, G. *et al.* Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30**, 155–67 (2009).

58. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

59. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

60. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).

61. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).

62. Blüthmann, H. *et al.* T-cell-specific deletion of T-cell receptor transgenes allows functional rearrangement of endogenous α- and β-genes. *Nature* **334**, 156–159 (1988).

63. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

64. Mousset, C. M. *et al.* Comprehensive Phenotyping of T Cells Using Flow Cytometry. *Cytometry Part A* **95**, 647–654 (2019).

65. Farber, D. L., Yudanin, N. A. & Restifo, N. P. Human memory T cells: generation, compartmentalization and homeostasis. *Nat. Rev. Immunol.* **14**, 24–35 (2014).

66. Huang, W. & August, A. The signaling symphony: T cell receptor tunes cytokine-mediated T cell differentiation. *J. Leukoc. Biol.* **97**, 477–85 (2015).

67. Padovan, E. *et al.* Expression of two T cell receptor alpha chains: dual receptor T cells. *Science (80-.).* **262**, 422–424 (1993).

68. Bacher, P. & Scheffold, A. Flow-cytometric analysis of rare antigen-specific T cells. *Cytom. Part A* **83A**, 692–701 (2013).

69. Chattopadhyay, P. K., Yu, J. & Roederer, M. Live-cell assay to detect antigen-specific CD4+ T-cell responses by CD154 expression. *Nat. Protoc.* **1**, 1–6 (2006).

70. Syed, A., Kohli, A. & Nadeau, K. C. Food allergy diagnosis and therapy: where are we now? *Immunotherapy* **5**, 931–944 (2013).

71. Seumois, G. *et al.* Transcriptional Profiling of Th2 Cells Identifies Pathogenic Features Associated with Asthma. *J. Immunol.* **197**, 655–664 (2016).

72. Mueller, S. N., Gebhardt, T., Carbone, F. R. & Heath, W. R. Memory T Cell Subsets, Migration Patterns, and Tissue Residence. *Annu. Rev. Immunol.* **31**, 137–161 (2013).

73. Nish, S. A. *et al.* CD4+ T cell effector commitment coupled to self-renewal by asymmetric cell divisions. *J. Exp. Med.* **214**, 39–47 (2017).

74. Foletta, V. C., Segal, D. H. & Cohen, D. R. Transcriptional regulation in the immune system: all roads lead to AP-1. *J. Leukoc. Biol.* **63**, 139–152 (1998).

75. Müller, U. *et al.* Lack of IL-4 receptor expression on T helper cells reduces T helper 2 cell polyfunctionality and confers resistance in allergic bronchopulmonary mycosis. *Mucosal Immunol.* **5**, 299–310 (2012).

76. Upadhyaya, B., Yin, Y., Hill, B. J., Douek, D. C. & Prussin, C. Hierarchical IL-5 expression defines a subpopulation of highly differentiated human Th2 cells. *J. Immunol.* **187**, 3111–20 (2011).

77. Ritvo, P.-G. *et al.* High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9604–9609 (2018).

78. Raj, A. & van Oudenaarden, A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* **135**, 216–226 (2008).

79. Han, Q. *et al.* Polyfunctional responses by human T cells result from sequential release of cytokines. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1607–12 (2012).

80. Hermiston, M. L., Xu, Z. & Weiss, A. CD45: A Critical Regulator of Signaling Thresholds in Immune Cells. *Annu. Rev. Immunol.* **21**, 107–137 (2003).

81. Sicherer, S. H. & Sampson, H. A. Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *J. Allergy Clin. Immunol.* **141**, 41–58 (2018).

82. Sampath, V., Sindher, S. B., Alvarez Pinzon, A. M. & Nadeau, K. C. Can food allergy be cured? What are the future prospects? *Allergy* all.14116 (2019). doi:10.1111/all.14116

83. Patil, S. U. *et al.* Peanut oral immunotherapy transiently expands circulating Ara h 2–specific B cells with a homologous repertoire in unrelated subjects. *J. Allergy Clin. Immunol.* **136**, 125-134.e12 (2015).

84. FDA approves first drug for treatment of peanut allergy for children | FDA. Available at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-drug-treatment-peanut-allergy-children. (Accessed: 19th February 2020)

85. Chinthrajah, R. S., Hernandez, J. D., Boyd, S. D., Galli, S. J. & Nadeau, K. C. Molecular and cellular mechanisms of food allergy and food tolerance. *J. Allergy Clin. Immunol.* **137**, 984–997 (2016).

86. Syed, A. *et al.* Peanut oral immunotherapy results in increased antigen-induced regulatory T-cell function and hypomethylation of forkhead box protein 3 (FOXP3). *J. Allergy Clin. Immunol.* **133**, 500-510.e11 (2014).

87. Ryan, J. F. *et al.* Successful immunotherapy induces previously unidentified allergen-specific CD4+ T-cell subsets. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1286–E1295 (2016).

88. Patil, S. U. & Shreffler, W. G. BATting above average: Basophil activation testing for peanut allergy. *J. Allergy Clin. Immunol.* **134**, 653–654 (2014).

89. Sampath, V. & Nadeau, K. C. Newly identified T cell subsets in mechanistic studies of food immunotherapy. *Journal of Clinical Investigation* **129**, 1431–1440 (2019).

90. Prussin, C., Yin, Y. & Upadhyaya, B. TH2 heterogeneity: Does function follow form? *J. Allergy Clin. Immunol.* **126**, 1094–1098 (2010).

91. Wambre, E. *et al.* A phenotypically and functionally distinct human T H 2 cell subpopulation is associated with allergic disorders. *Sci. Transl. Med.* **9**, eaam9171 (2017).

92. Mitson-Salazar, A. *et al.* Hematopoietic prostaglandin D synthase defines a proeosinophilic pathogenic effector human T H 2 cell subpopulation with enhanced function. *J. Allergy Clin. Immunol.* **137**, 907-918.e9 (2016).

93. Gowthaman, U. *et al.* Identification of a T follicular helper cell subset that drives anaphylactic IgE. *Science (80-.).* **365**, eaaw6433 (2019).

94. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).

95. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, (2018).

96. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

97. Kotake, S., Yago, T., Kobashigawa, T. & Nanke, Y. The Plasticity of Th17 Cells in the Pathogenesis of Rheumatoid Arthritis. *J. Clin. Med.* **6**, 67 (2017).

98. Kazer, S. W. *et al.* Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nat. Med.* 1–8 (2020). doi:10.1038/s41591-020-0799-2

99. Stekler, J. D. *et al.* No Time to Delay! Fiebig Stages and Referral in Acute HIV infection: Seattle Primary Infection Program Experience. *AIDS Res. Hum. Retroviruses* **34**, 657–666 (2018).

100. Robb, M. L. & Ananworanich, J. Lessons from acute HIV infection. *Current Opinion in HIV and AIDS* **11**, 555–560 (2016).

101. Ndung'u, T., Dong, K. L., Kwon, D. S. & Walker, B. D. A FRESH approach: Combining basic science and social good. *Sci. Immunol.* **3**, (2018).

102. Friedman, H. *et al.* The Critical Role of Nonhuman Primates in Medical Research - White Paper. *Pathog. Immun.* **2**, 352 (2017).

103. WHO | Global tuberculosis report 2019. *WHO* (2020).

104. Russell, D. G., Barry, C. E. & Flynn, J. L. Tuberculosis: What we don't know can, and does, hurt us. *Science* **328**, 852–856 (2010).

105. Lin, P. L. *et al.* Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat. Med.* **20**, 75–79 (2014).

A. Appendix

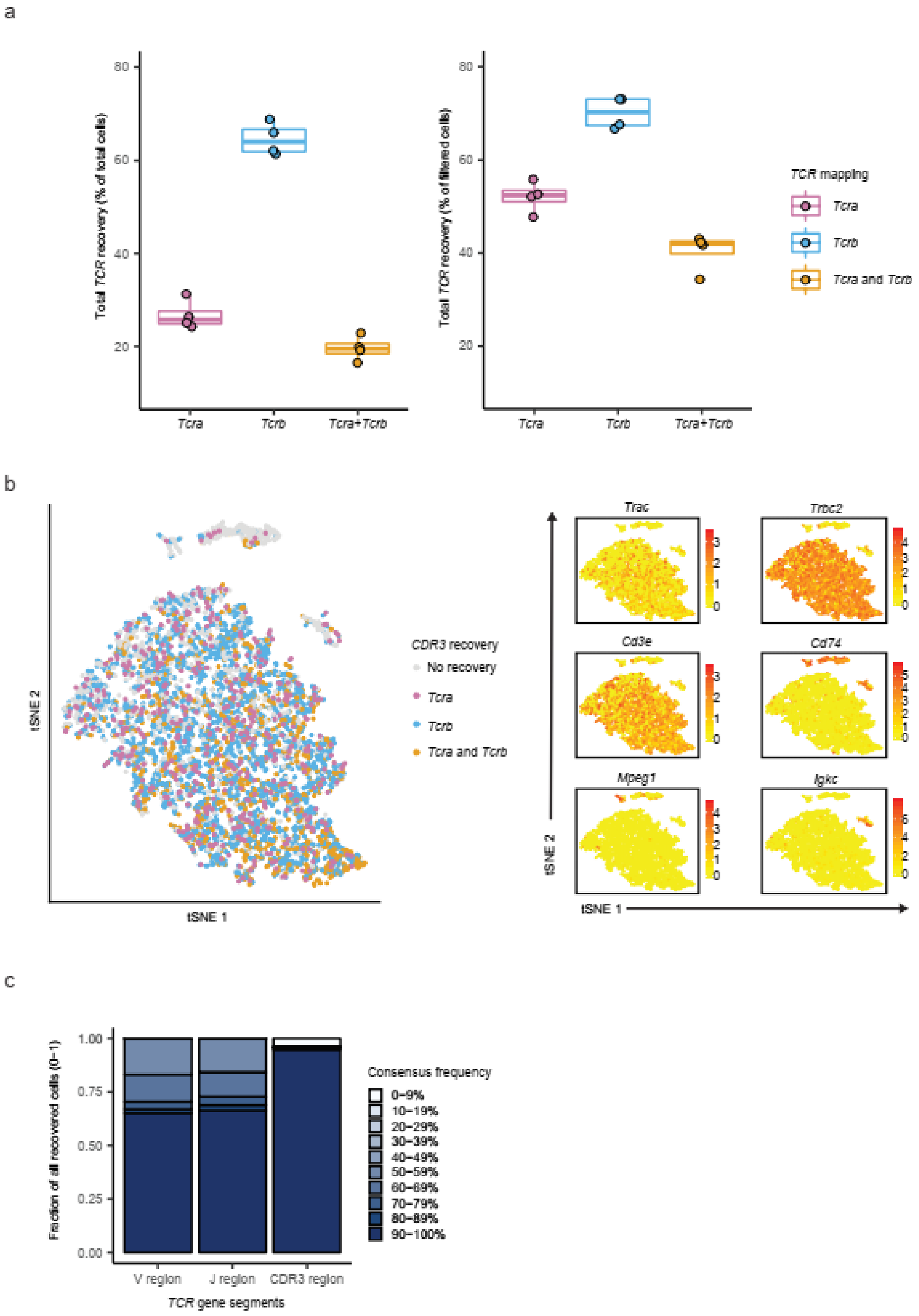
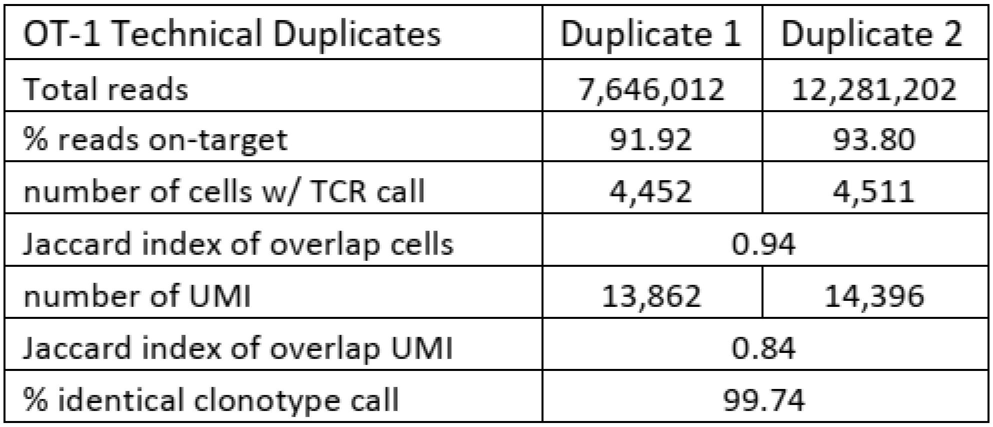


Figure A3-1. **a**, Total TCR recovery across four OT-I Spiked-in libraries (n = 4 samples). (left) Overall CDR3 recovery rates for all cells. (right) CDR3 recovery rates after removing cells without mapped TCR transcripts in whole transcriptome data. **b**, (left) *CDR3* recovery mapped on tSNE visualization of whole transcriptomes, and (right) key cell type surface markers (n = 6,620 cells). (right) Color indicates log-normalized gene expression (yellow to red). T cells were marked by expression of *Cd3e*, *Trac* and *Trbc*. Small numbers of other cell types were present due to incomplete magnetic enrichment. These included B cells (*Igkc*), macrophages (*Mpeg1*), and myeloid cells (*Cd74*). A small number of *TCR* sequences was also recovered from these clusters, correlating with trace amount of *Cd3e* expression in these clusters. These clusters were removed in subsequent analysis. **c**, Ratios of most frequent V,J, and CDR3 call for each UMI relative to either the second most frequent call (for V and J segments, resulting in "consensus frequency" between 0.5-1) or to the total number of reads (for CDR3, resulting in consensus frequency between 0-1).



OT-1 Technical Duplicates	Duplicate 1	Duplicate 2
Total reads	7,646,012	12,281,202
% reads on-target	91.92	93.80
number of cells w/ TCR call	4,452	4,511
Jaccard index of overlap cells	0.94	
number of UMI	13,862	14,396
Jaccard index of overlap UMI	0.84	
% identical clonotype call	99.74	

Table A3-1. Repeatability statistics from OT-1 spiked-in samples technical duplicates

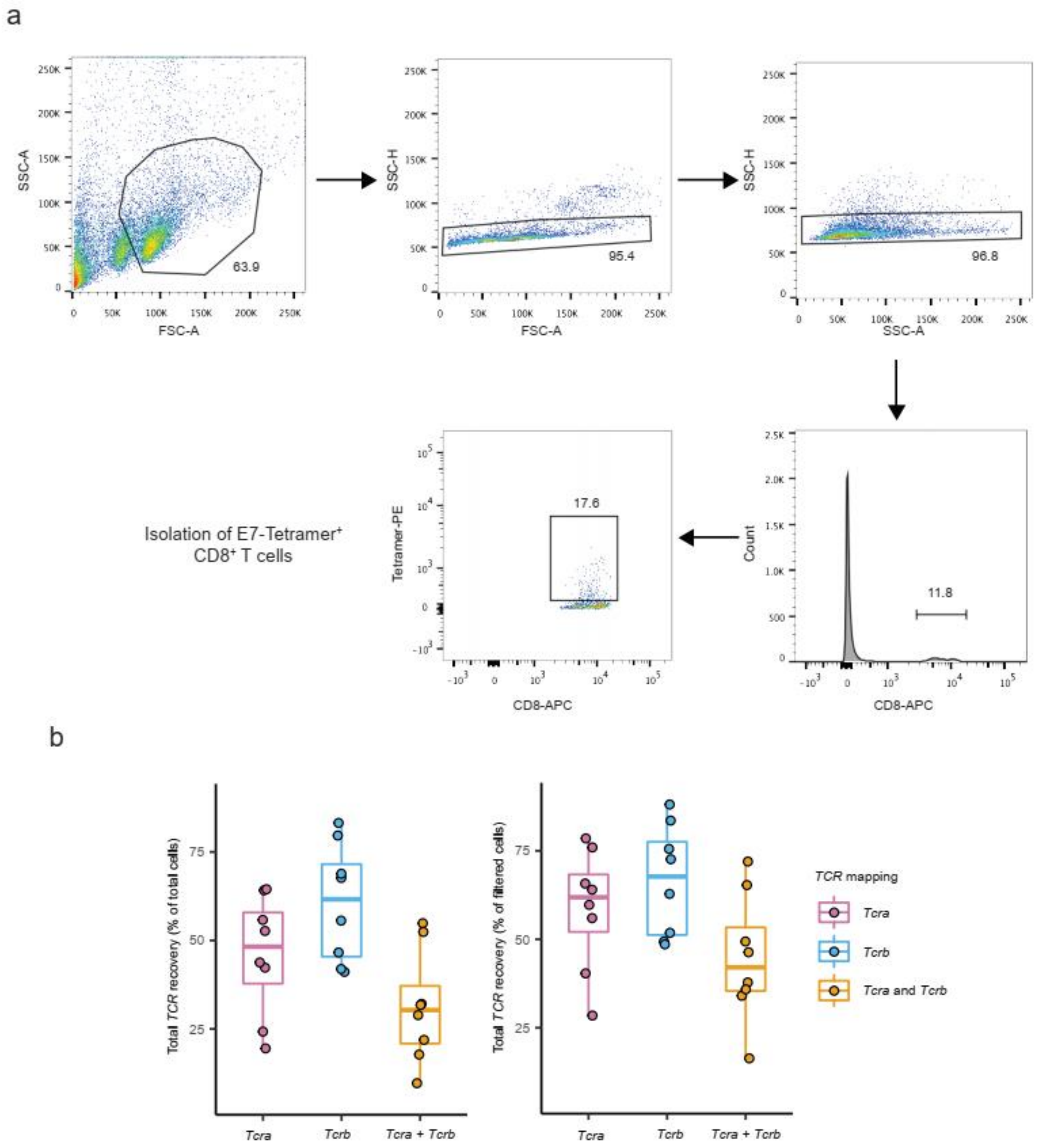


Figure A3-2. **a**, Gating strategy for flow cytometry sorting of E7-tetramer+ CD8+ T cells. **b**, Total *TCR* recovery across four HPV-E7 immunized mice (n = 4 animals). (left) Overall *CDR3* recovery rates for all cells. (right) *CDR3* recovery rates after removing cells without mapped *TCR* transcripts in whole transcriptome data.

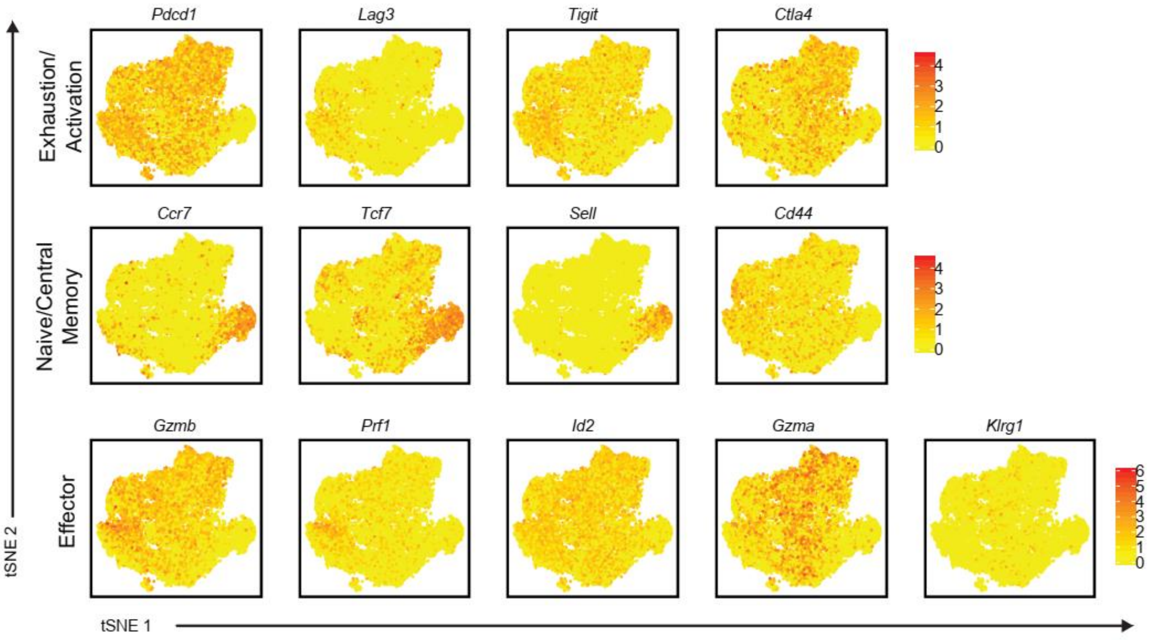


Figure A3-3. Expression of canonical markers associated with naïve/CM, effector, and T cell activation/exhaustion phenotypes. Color indicates log-normalized gene expression (yellow to red). Color scales apply to each respective row separately.

TRB_CDR3	TRA_CDR3	TRA.2_CDR3	AnimalID	TRAV	TRAJ	TRBV	TRBJ	TRAV.2	TRAJ.2	#cells
CASSQDLGNYAEQFF	CAMREGLMATGGNNKLTF	CAMREGLMATGGNNKLTF	m2	TRAV16D	TRAJ56	TRBV2	TRBJ2-1	TRAV16N	TRAJ56	1
CASSQDLGNYAEQFF	CAMREGLMATGGNNKLTF	CAMREGLMATGGNNKLTF	m2	TRAV16N	TRAJ56	TRBV2	TRBJ2-1	TRAV16D	TRAJ56	2
CASSQDLGNYAEQFF	CAMREGLMATGGNNKLTF	CAVSNSGGSNYKLTF	m1	TRAV16D	TRAJ56	TRBV2	TRBJ2-1	TRAV7D-5	TRAJ53	2
CASSQDLGNYAEQFF	CAMREGLMATGGNNKLTF	CAVSNSGGSNYKLTF	m2	TRAV16D	TRAJ56	TRBV2	TRBJ2-1	TRAV7D-5	TRAJ53	143
CASSQDLGNYAEQFF	CAMREGLMATGGNNKLTF	CAVSNSGGSNYKLTF	m3	TRAV16D	TRAJ56	TRBV2	TRBJ2-1	TRAV7D-5	TRAJ53	11
CASSQDLGNYAEQFF	CAVSNSGGSNYKLTF	CAMREGLMATGGNNKLTF	m2	TRAV7D-5	TRAJ53	TRBV2	TRBJ2-1	TRAV16D	TRAJ56	60
CASSQDLGNYAEQFF	CAVSNSGGSNYKLTF	CAMREGLMATGGNNKLTF	m3	TRAV7D-5	TRAJ53	TRBV2	TRBJ2-1	TRAV16D	TRAJ56	6

Table A3-2. E7-tetramer sorted CD8+ T cells with dual functional TCRα transcripts

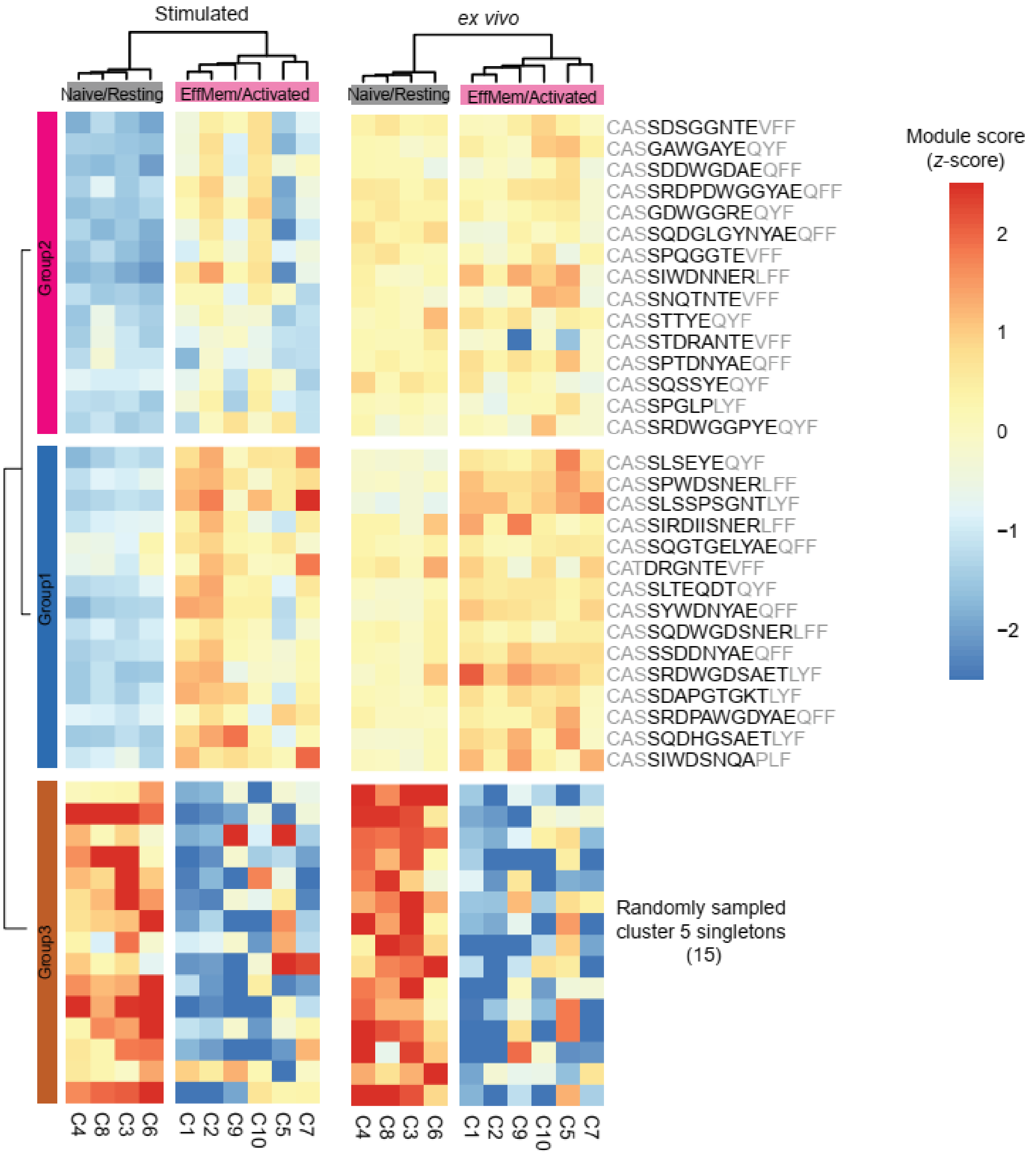


Figure A3-4. Average scores of CD8+ T cell modules identified by *Singer, M. et al.29* (C1-10, labeled on bottom of heatmap) in Group 1, 2, and 3 clonotypes shown in **Figure 3-5d**. In *Singer, M. et al.,* C3, C4, C5, C6, C8 were upregulated in sorted naïve or *TIM3- PD1- CD8+* T cells (naïve/resting). C1, C2, C7, C9, C10 were upregulated in sorted effector/effector memory, *TIM3+ PD1-*, or *TIM3+ PD1+ CD8+* T cells (Effmem/Activated).

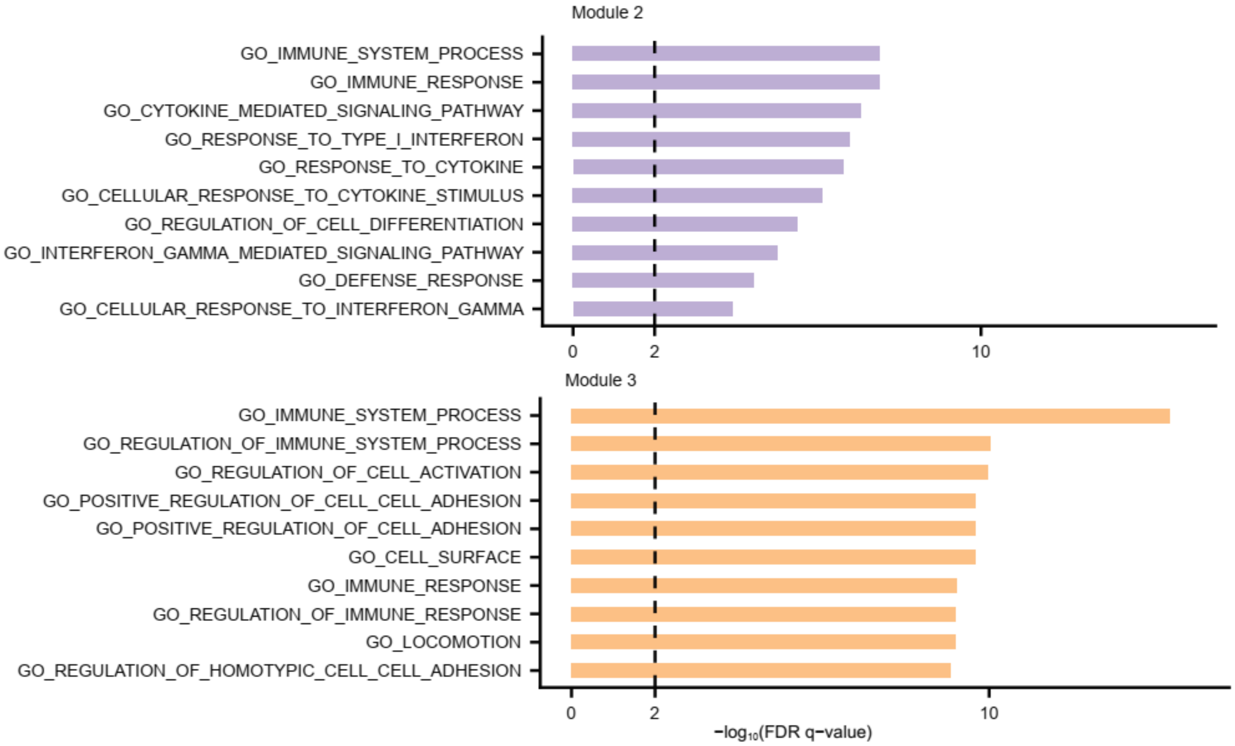


Figure A3-5. Gene Ontology (GO; C5) term enrichment of genes in Module 2 and 3. FDR q-values represent Benjamini and Hochberg corrected hypergeometric *P* values. 49 and 35 genes are used from Module 2 and 3, respectively.

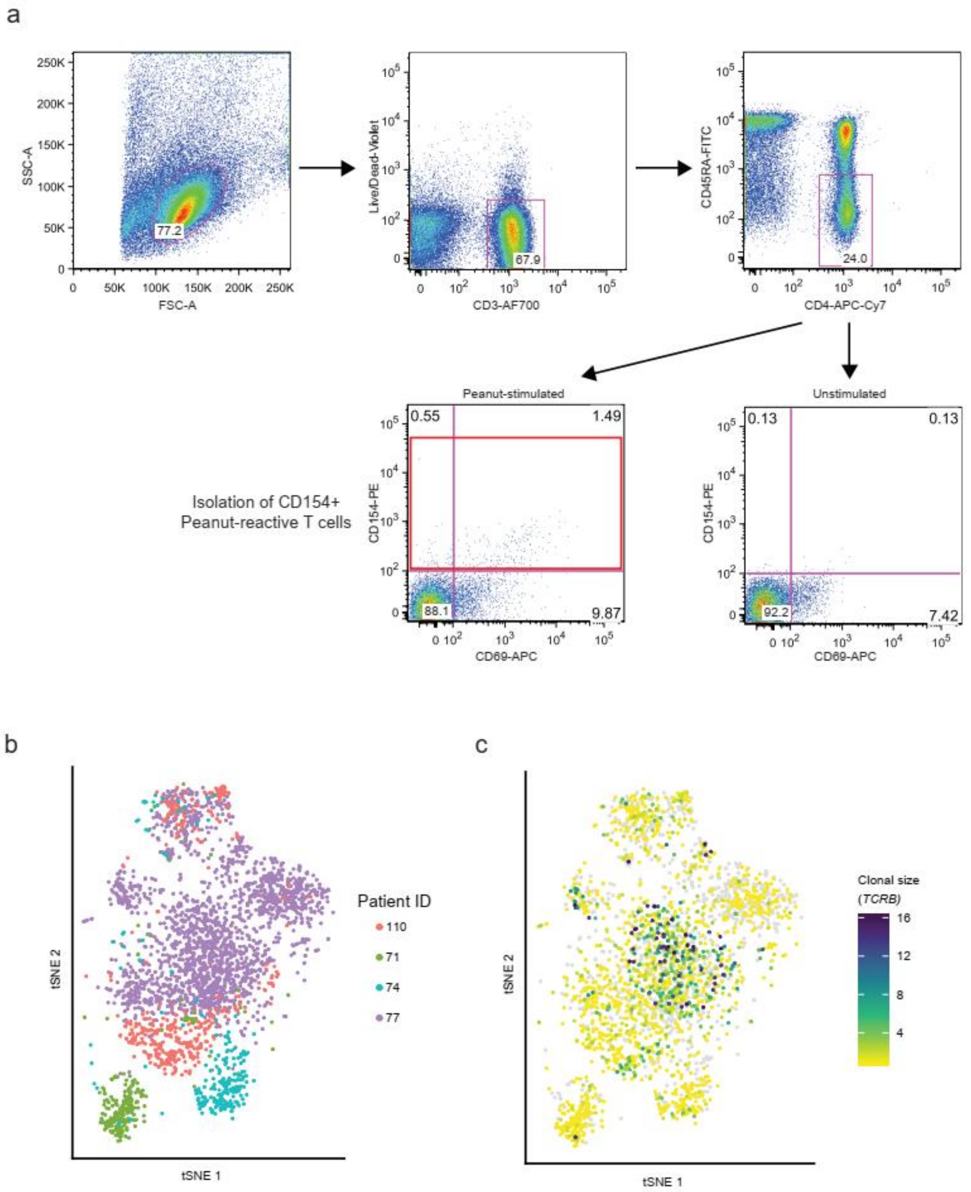


Figure A3-6. Gating strategy for flow cytometry sorting of CD154^{+} CD4+ T cells after *ex vivo* stimulation with peanut antigens. b, tSNE visualization of CD154^{+} T cells from four peanut-allergic patients, colored by patient identity (n = 2,712 cells). c, Clonal size of TCRβ mapped on the tSNE visualization (n = 2,712 cells).

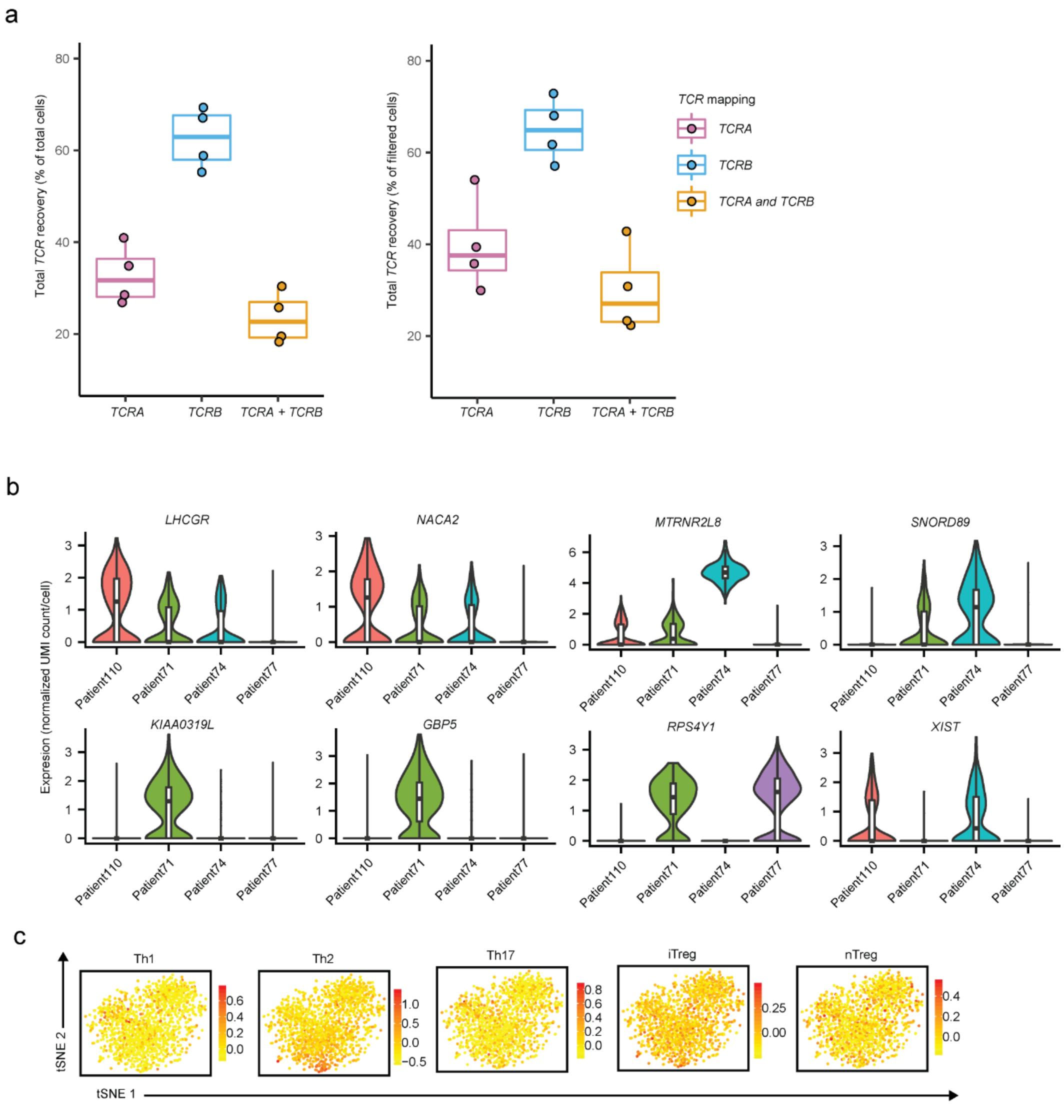


Figure A3-7. **a**, (left) *CDR3* recovery rates of all cells, and (right) *CDR3* recovery rates after removing cells without mapped TCR transcript in the whole transcriptome data. **b**, Expression of selected genes that most differentiated the four patients. Violin plots represent estimated density of cells (n = 398 cells for Patient110; 246 cells for Patient71; 221 cells for Patient74; and 1847 cells for Patient 77). **c**, Module scores (yellow to red) of CD4 effector T cell signatures outlined by *Wei, G., et al.38* mapped on the tSNE visualization of cells from Patient 77 (n = 1847 cells).

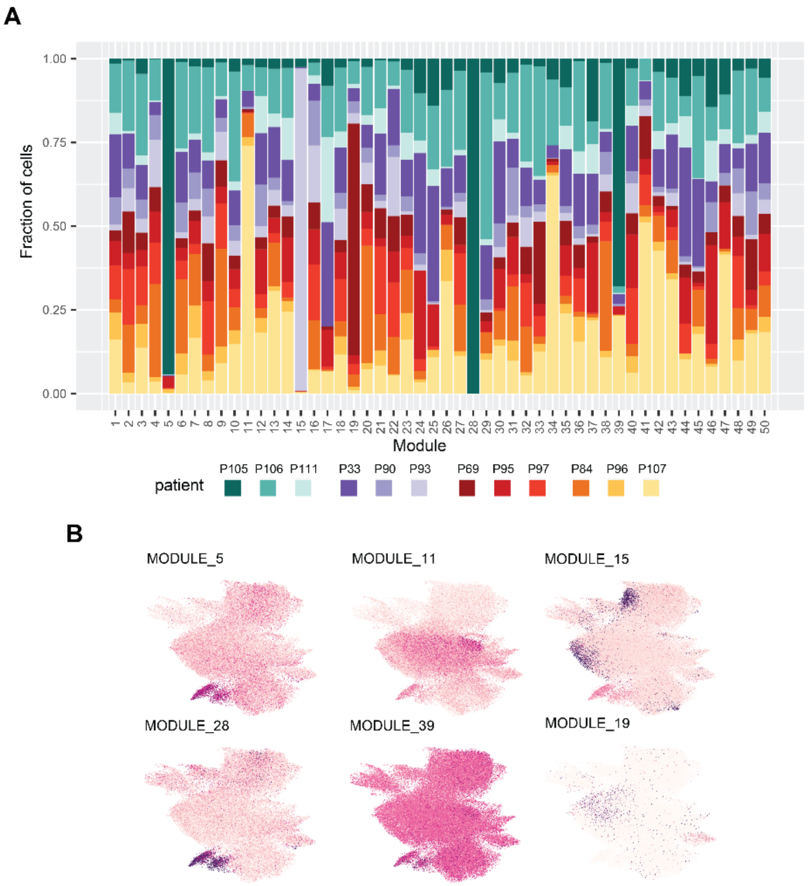


Figure A4-1. **a**, Distribution of module-expressing cells by patient, for the top 50 gene modules. “Module-expressing” cells were determined using the CD154-CD137- cells as described in **Methods**. **b**, UMAP overlay of module expression, and module loadings, for key patient-associated modules.

.

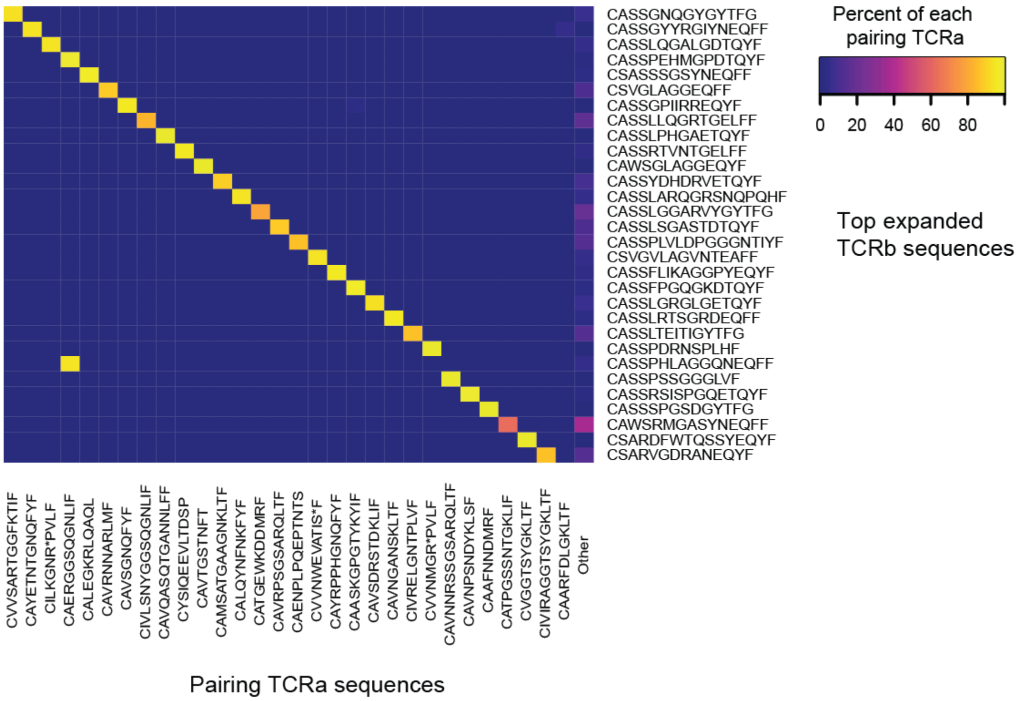


Figure A4-2. TCRα pairing for top expanded TCRβ sequences. Heatmap of TCRα pairing sequences (columns) found in cells with the top expanded TCRβ sequences (rows). Within each TCRβ clonotype, the percent of cells mapping to each TCRα is plotted. Rows are annotated with the majority patient in which the TCRβ clonotype was detected.

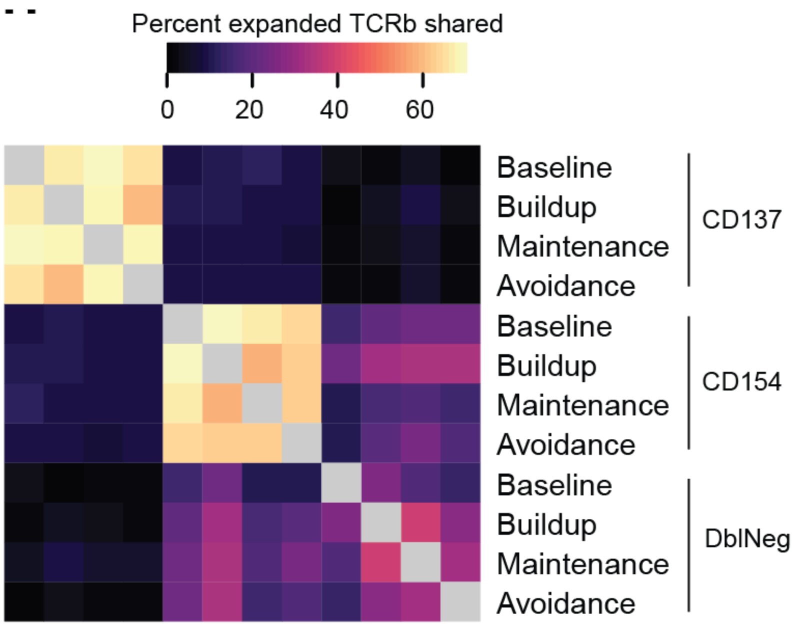


Figure A4-3. Sharing of TCRβ clonotypes between the three sorted across timepoints. Sharing of clonotypes is calculated as a geometric mean between the two respective sorted subsets and timepoints.

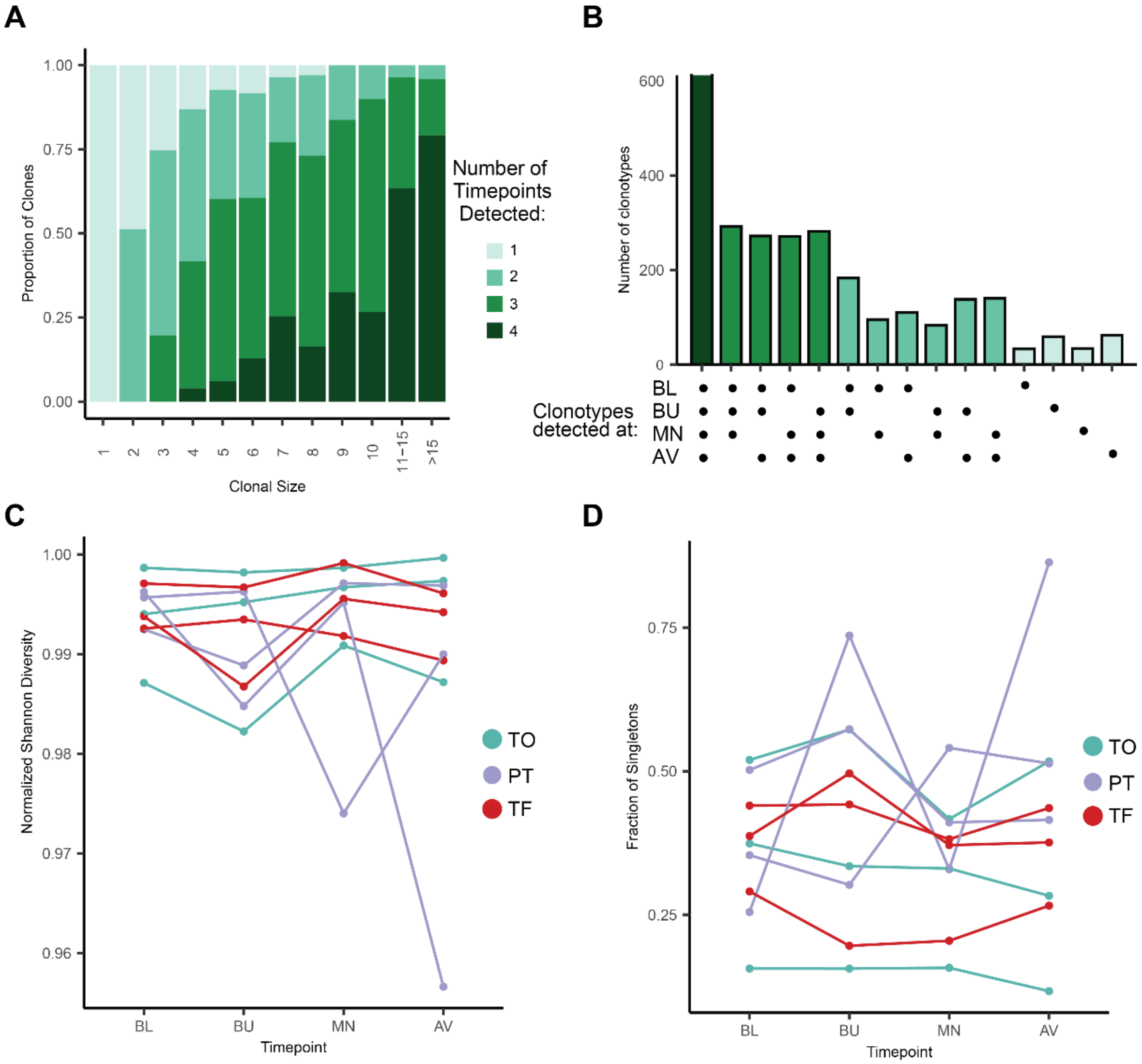


Figure A4-4. **a**, Proportions of clonotypes detected at one, two, three, or all four timepoints as a function of clonal size. **b**, Number of clonotypes detected at each combination of different timepoints. **c**, Normalized Shannon diversity of TCR repertoire by patient and timepoint. **d**, Fraction of singletons (clonotypes with clonal size of one) detected within each patient and at each timepoint.

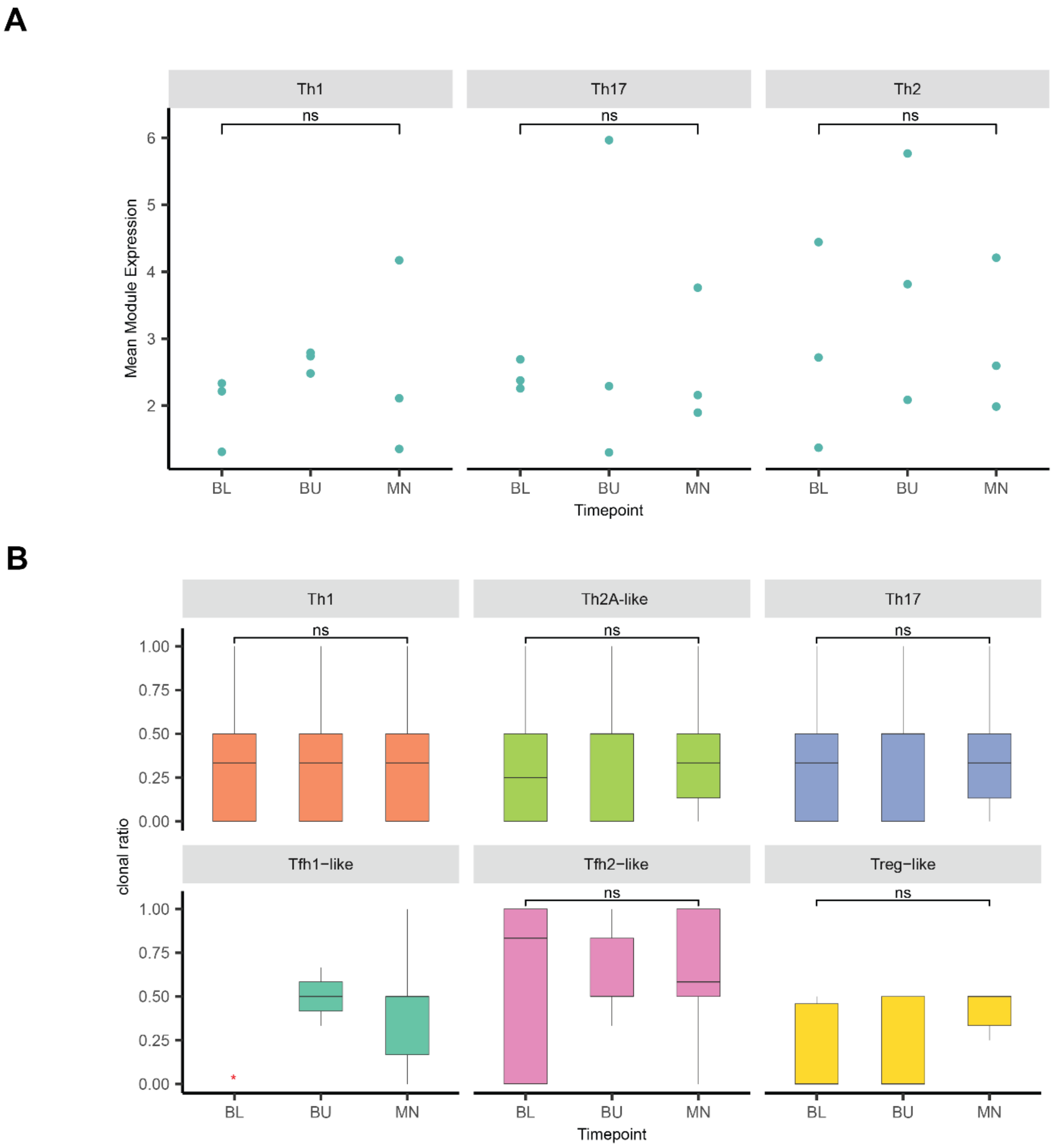


Figure A4-5. **a**, Mean expression of the Th1, Th2, and Th17 gene modules at baseline (BL), buildup (BU), and maintenance (MN) in CD154+ cells from each of the three placebo patients. **b**, Clonal expression ratio over time of clones from placebo patients in each Th subset. Clonal expression ratio was defined as the fraction of cells within each clonotype that scored as module-expressing for the relevant module (Th2, Th1, or Th17) at a given time point. Clonotypes were only included in the analysis at time points for which they had at least two cells recovered. P values were calculated using Wilcoxon rank-sum test.

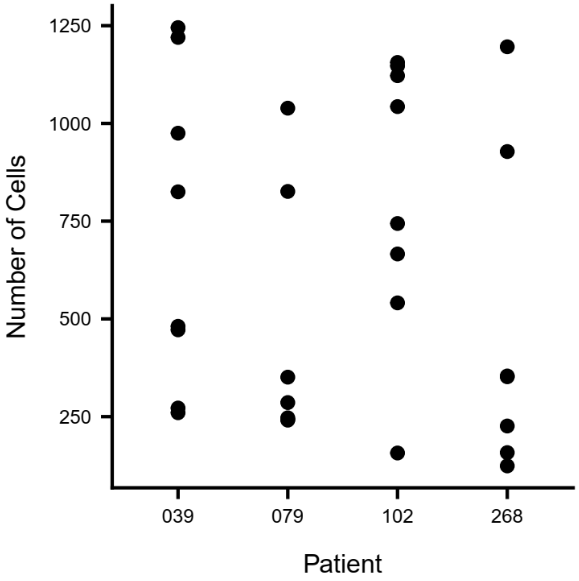


Figure A5-1. Number of T cells in each timepoint for each of the four patients. In many samples, less than 500 T cells were detected in the dataset.

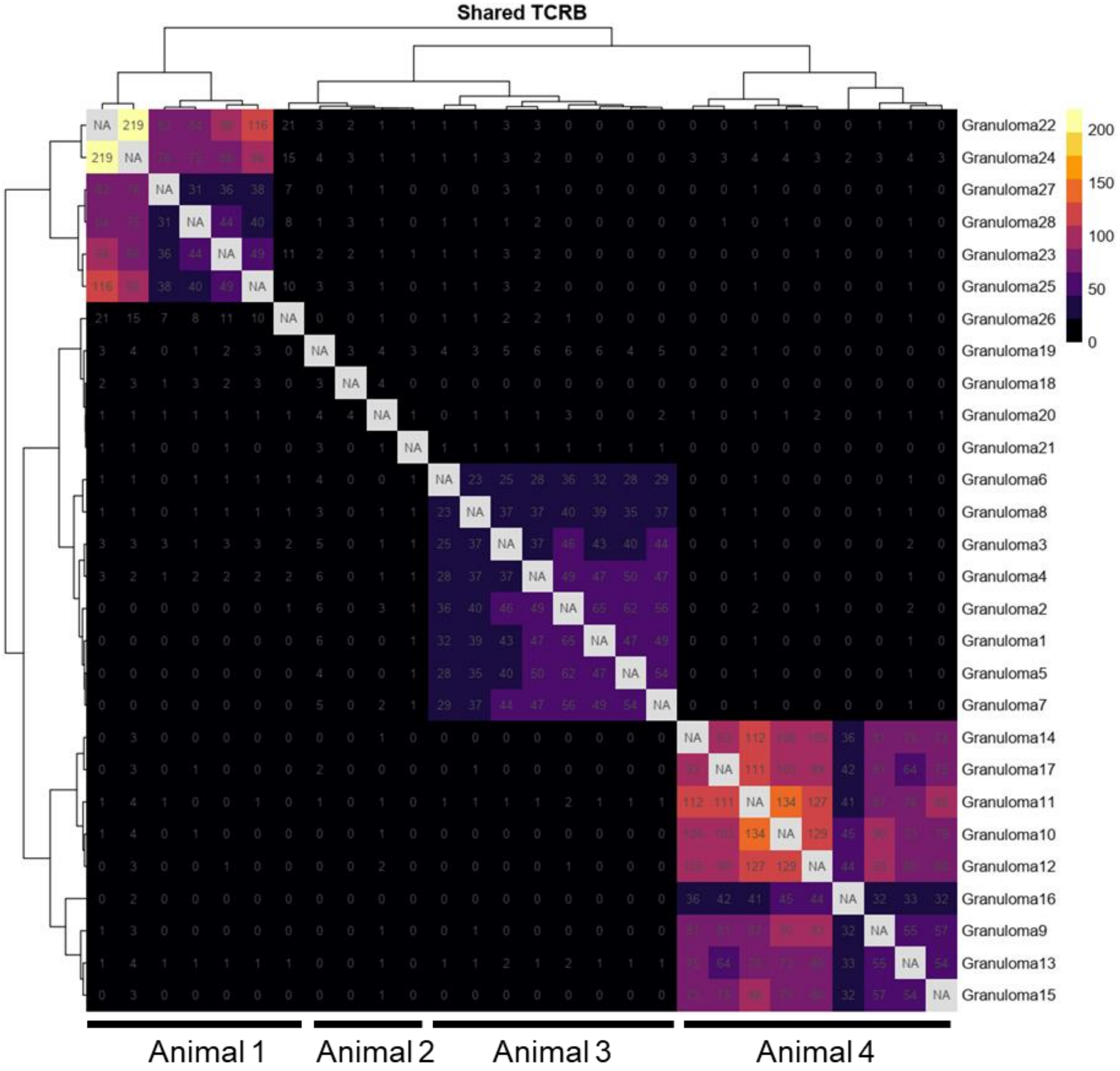


Figure A5-2. Unique TCRβ shared between granulomas from the four animals. Sharing of clonotypes between animals are mostly restricted within the individual animals.